THE UNIVERSITY OF ADELAIDE
SCHOOL OF COMPUTER SCIENCE

Real-Time Flight Delay Analysis and Prediction Based on the Internet of Things Data

2016 Master Project Thesis

Student: Abdulwahab Aljubairy (1659128)
Supervisors: Prof. Michael Sheng and Ali Shemshadi

04 November, 2016

Abstract

Flight delay is a significant problem resulting in the wasting of billions of dollars each year. Although this problem has been investigated in previous studies, all these previous studies rely on the historical records of flights provided by other agencies. Our work utilizes the emerging Internet of things (IoT) paradigm. It is now possible to collect and analyze sensors data in real-time. Our goal is to improve our understanding of the roots and signs of flight delays in order to be able to classify a given flight based on the features from flights and other data sources. We extend the existing works by adding new data sources and considering new factors in the analysis of flight delay. Through the use of real-time data, our goal is to establish a novel service to predict delays in real-time. In this project, we made a novel approach to collect the real time data from distributed sensors to study the flight delay. We create regression models to classify flights whether these flights are on-time or delayed as well as predicting how many minutes the delay would be. There are three main steps we conduct: first, we build a crawler to crawl the data from the pre-specified IoT data sources. Second, we implement an integration algorithm to integrate the data of all data sources using temporal and spatial criteria. Third, we conduct the analysis on the data with the aim to build a prediction model that could classify the flights and predict the delay time. This conducted analytical study provides three cases studies: Australia, China, and Europe. In addition, this project shows high correlation among the collected data. In addition, it shows that the prediction models in all case studies achieves very high accuracy. Comparing our models to others in previous studies, our model brings new factors that have impact on the flight delay as well as accomplish higher precision and recall.

Acknowledgment

I would like to thank my adviser Prof. Michael Sheng on his valuable supervision. His office's door is always open whenever I ran into a trouble or had a question about my research or writing. He consistently allowed this project to be my own work, but steered me in the right direction whenever he thought I needed it. He always give me valuable feedback that widen my vision and enhance my work. Prof. Michael has unique supervision style. He granted me the ability to approach any problem and find a solution for it. I am really thankful to him.

Also, I would like to thank Mr. Ali Shemshadi for the endless support and help. He provides me valuable feedback all the time. I really enjoyed with him the long meetings we have had. He is very punctual and approachable even if he has a packed schedule.

Finally, I should express my very profound gratitude to my parents and to my precious wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you very much.

Contents

Chapter 1

Introduction

With the rapid advances in the economy, air traffic has become one of the main means in the transportation industry, but the air traffic is suffering from the flight delay problem. Flight delay is a longstanding problem with the aviation industry, which massively affects the productivity of airlines and airports around the world. Thus, this problem cannot be ignored due to its impacts on the economy worldwide. Direct and indirect losses of flight delays are mind-blowing in terms of cost and span. A study by the National Center of Excellence for Aviation Operations Research (NEXTOR) estimates that the annual cost of air transportation delays only in the US surpass $32.9 billion in the year 2007 [23]. This number includes $8.3 billion airline component (consisting of increased expenses for crew, fuel, and maintenance, among others), $16.7 billion passenger component (based on the passenger time lost due to schedule buffer, delayed flights, flight cancellations, and missed connections) and $3.9 billion cost from lost demand. The indirect costs of flight delays can also be much higher in terms of the number and the span.

Many recent studies have investigated the flight delay problem in order to discover the issues that cause delays. Flight delays are often subjected to be caused by a number of sources of irregularity. However, many existing features and potentials are dismissed because the flight delay problem is very complex. Statistical studies suggest that nearly 20% of the flights are delayed for various reasons [2][4]. In particular, weather is responsible for nearly 75% of delays [24]. Moreover, due to the recent changes in weather patterns as an effect of global warming, we expect to see it a rise in those numbers as a result of increased harsh conditions.

It is a fact that each airport and airline operate with limited resources including airport capacity, number of aircrafts, number of flight crews and etc. Thus, many bottlenecks lead to delays in the scheduled flights. Based on the Federal Aviation Administration (FAA) policy, any flight departs or arrives after 15 minutes from the scheduled time is considered delayed [1]. Therefore, many researchers have studied this phenomenon in order to identify the major factors that cause the flight delay.

Previous studies of the flight delays assess this problem using the historical records that have been obtained from the bureau of transportation or the FAA [6][9]. In this project we use real-time IoT data to investigate this problem. We identify some important factors, which can suggest that a scheduled flight is going to be deferred. We provide a statistical analysis of the delays and their possible origins or signs. Furthermore, in this study we create a predictive model to predict the flight status in the future. We classify the flight status based on real-time sensors' data.

Prediction and analysis of flight delays are useful to reduce the direct and/or indirect associated costs. However, due to the highly dynamic environment, relying on a single historical dataset of flight delays in previous works [25] [29] may not be sufficient. For instance, the users of a flight delay prediction system would be interested to find out the chance of the delay for a scheduled flight rather than a flight in the past.

The emerging paradigm of the Internet of Things (IoT) aims at establishing a worldwide pool of sensors to interconnect physical devices [26] [3]. Thus, sensors will become the main generator of data on the Internet and enable a ubiquitous sensing of the environment. Based on the IoT data, Context-Aware Computing [27] can increase the effectiveness of the flight delay analysis. We use the following scenario in Figure 1.1 to illustrate this idea.

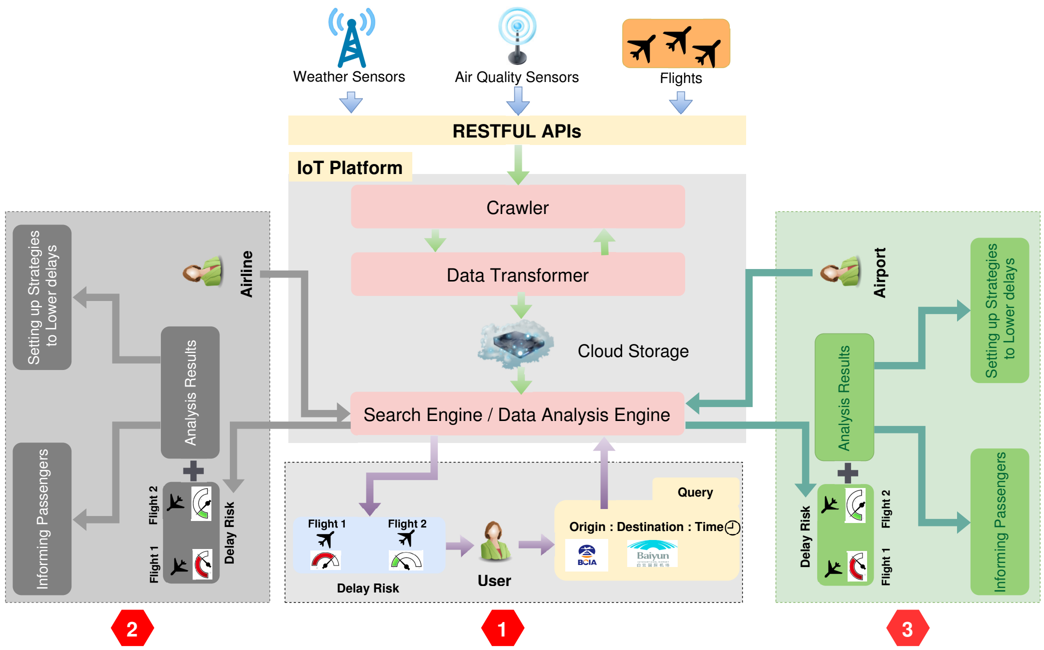


Figure 1.1: Online Service Scenario

In this study, we tackle a number of technical challenges to enable real-time flight delay analysis based on the IoT data. To the best of our knowledge, due to privacy issues, the access to the real-world IoT data remains very limited. In addition, none of the previous works has investigated the connection between contextual IoT data and flight schedules. In this project, we crawl and use real- world datasets to identify the correlation of the different data sources which consist of flight, weather and air quality data sources. We summarize our contributions as follows:

1. We create an IoT search engine to crawl the data from publicly available websites. In our crawler, we identify and standardize a set of steps to facilitate the Extract, Transform and Load processes in acquiring IoT data. In the context of IoT, users would normally be less interested in finding the pages of things (unlike finding Web pages in the Internet). Thus, we add the analysis of the flight delays to enhance the interests of the users in the result.

2. We crawl IoT data from different data sources. We examine the correlation between different datasets and the projected flight delays

dataset. We use two machine learning models. First model is multiple logistic regression to classify flights whether they are on-time or delayed. Second is multiple linear regression to investigate the effectiveness of each feature base on the crawled datasets and predict the delay time.

This study is a significant step as we obtain the data from IoT data sources in real-time. We also consider novel features and new data sources in our study. There are many applications to the results of our study. This project would be beneficial for helping all stakeholders. One of the applications is to enable airlines and airports to identify the sources of delays and resolve the issues in a short time. Moreover, customers can select the best flight for their journeys and get recommendations for flights with lower risks of delay, where it is applicable.

As mentioned above, we use various IoT data sources such as flights and flights schedules data, air quality data and publicly available weather stations records, all in real-time. The novelty of our work lies on the idea of correlating different datasets together rather than focusing on trends on a single dataset. This enables us to more effectively analyze the environmental and organizational features, which suggest that a given flight is going to be delayed. We identify a set of features from our dataset and use multiple linear regressions and categorized factor analysis to study their correlations. In the first step of our work, we focus on the domestic flights in two countries including Australia and China and international flights in some selected capitals. We plan to extend our study by including more regions and international flights. Furthermore, we anticipate that in the second part of our project, we develop a novel service to classify scheduled flights based on the trained model.

1.1 Motivating Scenario

Assume Alice has an important meeting in another city, and the best way to attend that meeting is to travel by airplane. She arranges everything, and she plans to arrive just in time before the meeting starts. She does not want to go a night before and waste her time. So, on the day of her flight, Alice arrives at the airport and discovers that her flight is delayed. She also notices that there is another flight, which is operated by different airline which is going to the same destination on the same time and on-time. Then she asks herself this question "why did I not select that flight?". In our study, we will enable such customer to determine the status of the desired flight in the future. We will create a service that can classify the flights based on new and real-time data. Figure 1.2 shows the abstract idea.

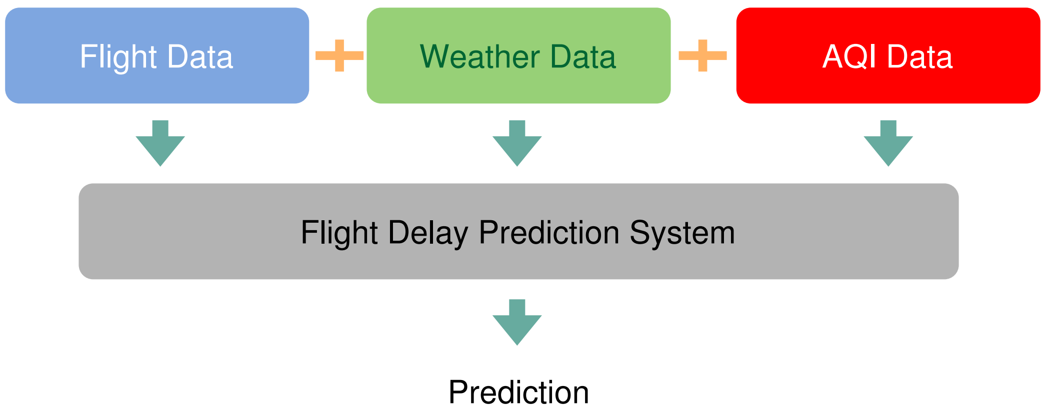


Figure 1.2: The Flight Delay - Abstract Idea

1.2 Project Benefits

The outcome of our project is a service that can predict the flight status in the future.

1. Travelers can utilize our service to reduce the chance of encountering unwanted flight delay. The can be able to plan their journeys effectively with the best options of flights. They can know in advance the status of the desired flight.

2. For airlines, this service will benefit them to improve their performance and engagement with their customer. They can be capable to manage their operations when they anticipate their flights future.

3. Flight delay has a large negative impact on the airports. When there is a delayed flight, that can increase congestion [3]. Also, it will impact the airports' management operations [7][8]. So, having this service, airports can alleviate the negative impact in advance impacts

by predicting the delays in advance. Furthermore, many airports and airlines can use this service to reduce their dependency on relatively expensive environmental data. This is achieved by using publicly available data, which is free.

1.3 Project Goals

The main objective of this project is to develop a better understanding toward the flight delay issue and develop a novel system that can identify the significant factors that contributes to the flight delay. We use new and real-time data to study this problem. Considering such data requires developing a tool that could help to collect the data we are targeting. Therefore, we are building a novel approach for that purpose. We aim to develop machine learning algorithms to classify flights and predicts the delay time for every individual flight.

1.4 Document Purpose

The main purpose of this document is to summarize the work I have done in my Master project. Firstly, this document highlights on the related work in this research area. Secondly, it presents the key background knowledge and the approaches applied in this project. Thirdly, it summarizes the comprehensive statistical analysis conducted for this project. Fourthly, it describes the implementation in details and the experimental results for evaluating the machine learning models.

Chapter 2

Related Work

This problem is not a new problem, and it has been considered by many researchers. Here it is the most relevant work to our work.

In [1] the authors analyzed the time factor influence of the flight delay in twenty airports in the US. They observed the changes of the delay rate using historical data. Their investigation aim was to predict the delay of each period based on their mode. They used ANOVA and k-means clustering model in order to demonstrate the periodic of the delay rate. After that they applied the Fast Furious Transform to find the period of the delay. Although their model was able to predict accurately for the first airport they were studying, they found out that their model should be improved in order to be applied to the other 19 airports. However, they did not consider the airline influence.

Liu and Yang studied in [2] the flight delay propagation in the flight chain. So they proposed a new algorithm that could estimate the delay from the beginning in order to to determine how much time the flights in chain could be delayed. Authors of [2] did not focus on the potential causes of the delay. They only modelled the problem utilizing the Bayesian Network.

Liu and Ma (2009) the authors of [3] analyzed how flight delay is influenced by delay propagation using Bayesian Network. First, they investigated the correlation between the departure delay and the arrival delay at a particular airport. They found that the majority of delays happens in the period between 8 am and 9 pm. They measured the delays as light, medium, or heavy. They proposed that canceling flights when there is a heavy delay in the chin will relief the problem. Even though canceling the flights will definitely help other subsequent flights in the chain to be on-time, other factors that may cause the flight delay should be taken in account.

In the study in [4], authors studied the major factors that contribute to flight delay. They developed a model to predict the flight delay using historical records of Denver International Airport. Basically, their model considers two types of delays. First is daily propagation patterns that might be caused by crew connection problems, propagated delay from previous flights, or other factors. Second is seasonal trend where weather or seasonal demand have impact on it. However, as in [1] predicting the status of the flight in the future would require additional dynamic resources that could enrich the model.

In [7] the authors looked at how the arrival delay could be propagated and impact the other subsequent flights in the stream. They believe all these types of delay only happen in busy hub-airports. They created three models. First, they had a propagation model after they investigated the relationships among flights. After that, they came up with an arrival delay model using Bayesian Network. Then they discussed the propagation delay in the hub-airport. They claim that the arrival delay is the source that mainly cause the departure delay.

Geng in his paper [11] provided statistical analysis of the flight delay. He listed all potential factors that may cause the flight delay. Some of these factors are airports, airlines, passengers, public safety, weather, fuel, departure control system, and air force. All these factors are actually play a role on the flight delay. Then he discussed some countermeasures in order to deal with the flight delay.

In [12] the authors focused on study the flight delay problem based on the random flight point delays. They used series analysis on airline data and presented an influence factor model of the random flight points.

The basic idea of this model is to combine the Bayesian Network with the Gaussian Matrix Model? expectation maximization algorithm. This model can predict the delay of the downstream.

As the best of our knowledge, there is no study has considered the real-time data to investigate the flight delay. In [1] the authors recommend for the future work to combine the analysis of historical data with real time data. That would predict the on-time performance of any airport. Our work will consider the real time data to predict the performance of individual flights.

Rebollo an Balakrishnan in [9] presented a new model to predict the flight delay. They consider the temporal and the spatial delay states as explanatory variables. Their approach is to predict the delay sometime in the future between 2 to 24 hours. They use the Random Forest algorithm to do so. Although this model predicts the flight status in the future, the aforementioned interval seems too short because people require time more than that when they book their flights.

Cheng 2014 [18] developed a prediction model for flight departure delay. First, he studied historical data for finding the main factors that cause flight delay. These data are weather, holiday influences and hourly pattern. After that, he used these factors as variables of mixed function to combine the weight function to a smoothing spline model with ARIMA models. His model can estimate delays for each flight on a specific day and hour and show a high accuracy result. It achieves actual probability is 91.8% with a delayed more than 60 minutes and the model achieves 2.78% with delayed more than 120 minutes.

In addition, some researches used machine learning method based on graph theory. Qianya .et al 2015 [19] developed a new analysis method which analyze and predict the delay during the flight based on Bayesian Network They tested the series experiments on actual airline data and the results show high accuracy 81.95%. Liu .et al 2008 [20] also used Bayesian Network to estimate the arrival delay and the propagation delay. They focused on one busy hub-airports and discuss the influence of propagation between the flights belonging to same air company.

Alonso and Loureiro 2015 [21] studied the flight departure delay. They focused on Porto Airport, and they treated the problem of predicting flight departure delay as an ordinal classification task and a suitable approach, based on the so-called unimodal model, is used to predict the delay. The unimodal model is implemented using neural networks. For comparison purposes, they also implemented the binomial model using trees(Hastie et al. 2009 [22]) . The neural networks outperform binomial model. It also obtained a better result using only half of the predictor variables used by the tree to predict the departure delay.

Chapter 3

Background and The Internet of Things (IoT) Data Sources

3.1 Background

In this section, several key conceptions that are referred to throughout this thesis are explained. It is important to understand these conceptions since they provide the foundation on which the project is built.

As stated above, this study will be based on the data produced from the distributed sensors. Doing that will definitely require an approach that help us to obtain the real-time data from their sources. Actually, in the beginning we will use several sources such as 24FlightRadar, Xively, waze, and others. Each source provides different type of data. Then we need to build a crawler discussed in [10] to collect the required real-time data from the aforementioned sources. After collecting the data, we will need to clean the data. Additionally, we will need to explore it by visualizing it and get familiar with it. That will help us to investigate the correlation among all collected variables in order to observe the potential impact of them on the flight delay. Furthermore, when you look at the flight delay problem, you would realize that there are several elements involved in it such as airports, airlines, weather, and others.

3.1.1 Internet of Things IoT

The IoT is the network of physical objects, and these objects are embedded with electronics, sensors, software [16][7]. These objects are also provided network connectivity. So they are able to collect and exchange data.

The Internet of Things paradigm increases the ability of objects communication among each other. A thing can be anything such people, items, animals, etc. This paradigm will help to reduce the amount of human intervention with physical objects [16][17] because these objects will be smart enough to determine the surround situation and act based on that.

3.1.2 Multiple Linear Regression

It is a statistical method, and it is used to study and measure the relationships among variables in form of a function. The variable that we want to predict is called the dependent variable. The variables that are used to predict the dependent variable are called predictors. These predictors determine the value of the independent variable. In our study we will use this technique in order to identify the major factors that cause the flight delay. As well, we will use this approach to determine the delay at departure for each flight.

3.1.3 Principal Component analysis

It is a mathematical approach that is used to convert the number of correlated variables into less number of uncorrelated variables called principal components. The purpose of this method is to ease the interpretation of these complex variables.

%chapterThe Internet of Things (IoT) Data Sources

3.2 The Internet of Things Data Sources

In order to be able to investigate the flight delay and implement a prediction system, we must have an adequate real-time datasets. Since all the previous studies only considered the historical data of flights and weather, our data model will be based on new and real-time data. As a result, we use real-time data obtained from sensors publically distributed in order to explore and analyze the flight delay phenomenon. We develop a novel approach to collect the real-time datasets we need in our study. In the subsequent sections, we describe the crawler engine we use. In addition, we provide a full description for each one of the real-time data sources along the features that we can extract.

This chapter describes the real-time data sources used in this study. The real-time investigation and prediction study has been conducted based on the Internet of Things data. We use three various real-time data sources:

1. Real-Time Flight Data
2. Real-Time Weather Underground Data
3. Real-Time Air Quality Index

Our main goals for this research are to identify the most important factors that contribute to the flight delay, to create a model that predict the possibility if an individual flight would be on-time or delayed, and finally to estimate the magnitude of this phenomenon. In the next section, we describe a subset of real-time data sources types with some examples that we use in this research. Then, we discuss all features from the data sources in order to provide some understanding of each one of them.

3.3 Data Sets

3.3.1 Real-time Flight Data

There are several data sources which provide live data of flights every day. They leverage data from several resources such as air traffic control systems. More importantly, they utilize the network ADS-B ground stations. As in Figure 3.1, Flightradar24 is an example of real-time data source that offer substantial data of large number of flights around the world. We can realize the flight number, the origin, the destination, the scheduled and actual departure time, the scheduled and actual arrival time, the aircraft type, and many other flight details. Figure 3.2 shows the feature list we extract from the FlightRadar.

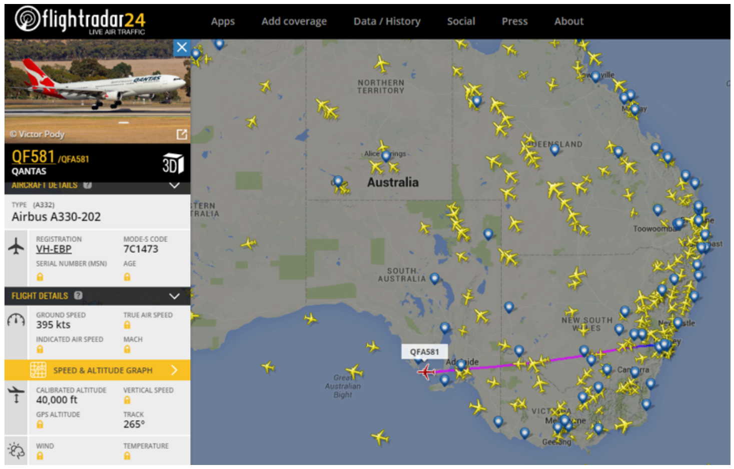


Figure 3.1: Information of a flight from FlightRadar24 [13]

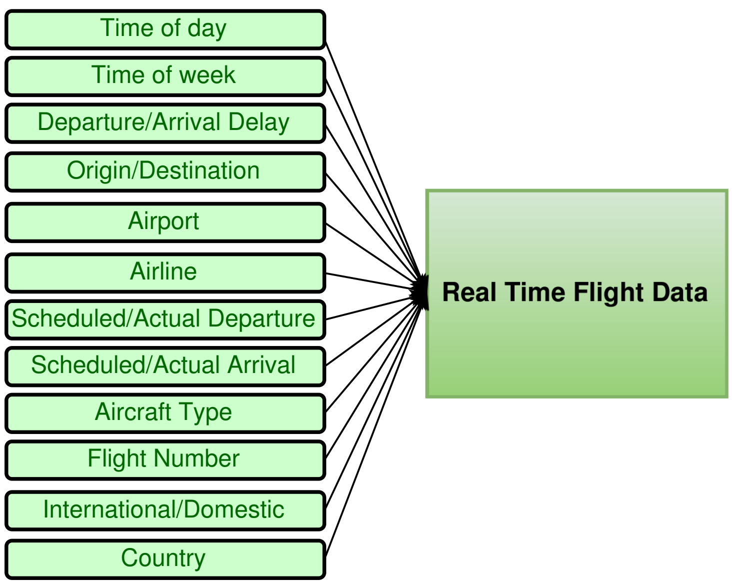


Figure 3.2: Flight Radar data source features list [13]

3.3.2 Real-time Weather Data

Many weather websites incorporate real-time weather data obtained from various weather and climate agencies. These sources offer wide range of relevant weather information. They deliver their data in various format such as XML or map format. Weather Underground is one of the well-known data sources that provide live weather data. Its large network includes more than 180,000 weather stations see Figure 3.3. In Figure 3.4, we can see the feature list of the weather underground data source.

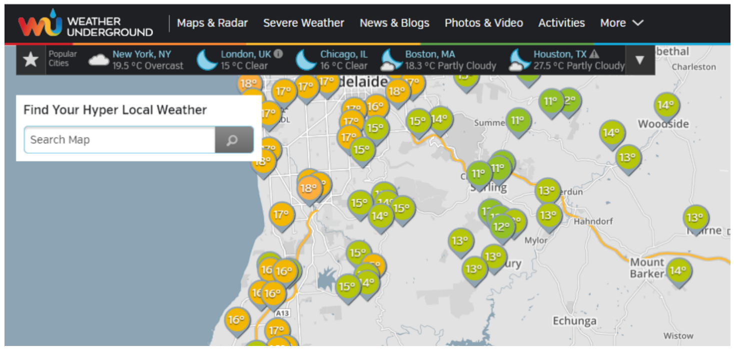


Figure 3.3: Weather information from WeatherUnderground [14]

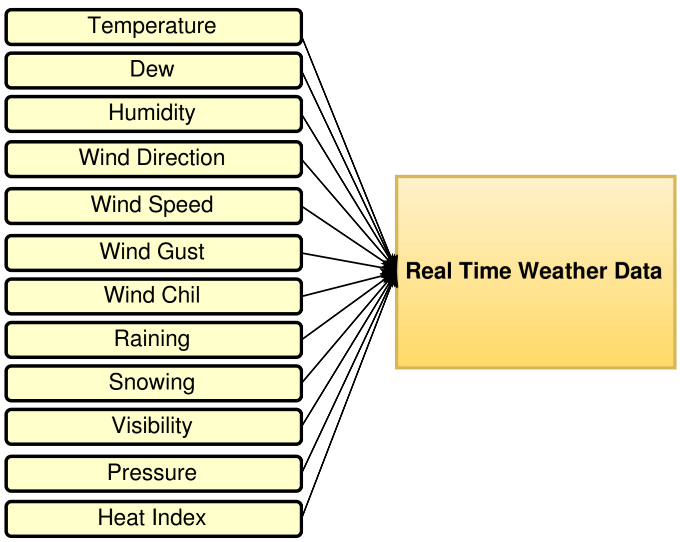


Figure 3.4: Weatherunderground data source features list [14]

3.3.3 Real Time Air Quality Index Data

An air quality index (AQI) (Figure 3.5) is an index that indicates the quality of air in a place. This number is measured by monitoring the air data, and this index reflects the air quality standards. It tells how the good or bad the air quality is. Figure 3.6 shows the feature list of this data source.

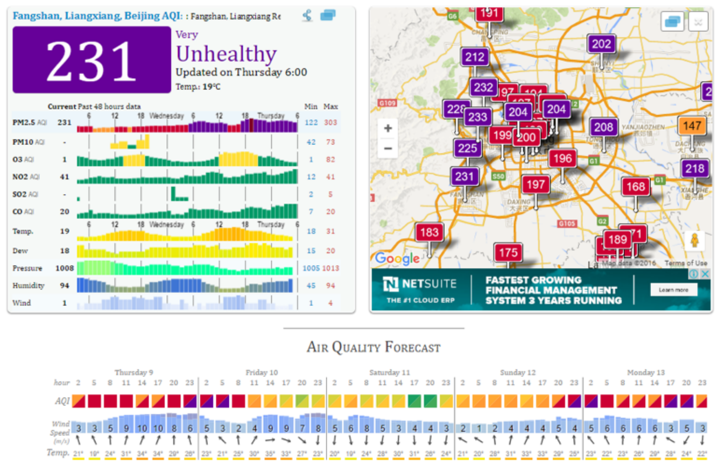


Figure 3.5:AQI of Beijing City, China from Air Quality Index [15]

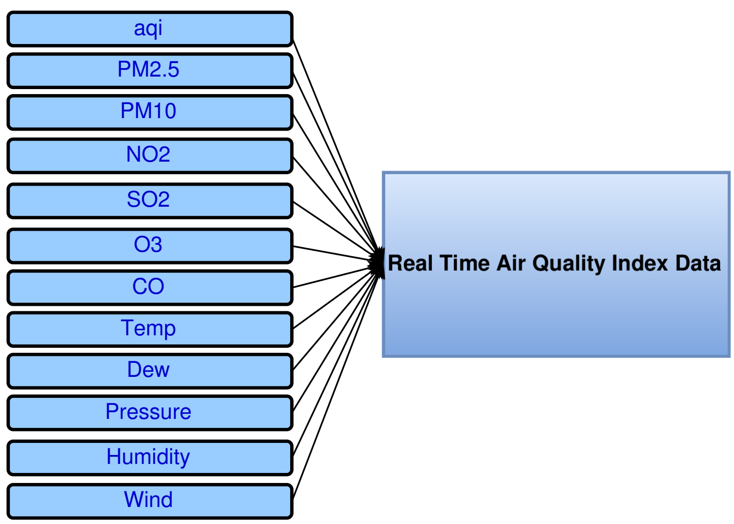


Figure 3.6: Air Quality Index data source features list [15]

We spend sometime on each one of these sources in order to understand the features and data they provide. All of these sources offer their data through APIs. We use these APIs in order to retrieve the data in our machine.

3.3.4 Features Description

This section will describe features we get from IoT sources. These features will be used later in our analysis while studying the flight delay problem and creating the predective model.

Real Time Flight Data

Table 1: Features of flight data

Feature	**Description**
Time of day	This feature represents the time of the flight during the day.
Day of week	This feature represents the day of the flight during the week.
Departure/Arrival Delay	This feature represents the departure delay and the arrival delay of the flight in minutes.
Origin/Destination	This feature represents the origin airport, city, country of the flight.
Airport	This feature represents the airport where the flight departs or arrives.
Airline	This feature represents the airline that operates the flight.
Scheduled/Actual Departure	This feature represents the scheduled/actual departure time of the flight.
Scheduled/Actual Arrival	This feature represents the scheduled/actual arrival time of the flight.
Aircraft Type	This feature represents the airplane type of the flight.
Flight Number	This represents the flight number.
International/Domestic	This feature represents if the flight domestic or international.

Real Time Weather Data and Air Quality Data

Table 2: The features of weather and air quality data

Feature	Description
Temperature	This feature represents the current temperature of the weather at the airport where the flight departs or arrive.
Dew	This feature represents the dew at the airport.
Humidity	This feature represents the Humidity at the airport.
Wind Direction	This feature represents the Wind Direction at the airport.
Wind Speed	This represents the Wind Speed at the airport.
Wind Gust	This feature represents the Wind Gust at the airport.
Wind Chill	This feature represents the Wind Chill at the airport.
Raining	This feature represents the Raining at the airport.
Snowing	This feature represents the Snowing at the airport.
Visibility	This feature represents the Visibility at the airport.
Pressure	This feature represents the Pressure at the airport.
Heat Index	This feature represents the Heat Index at the airport.
aqi	This feature represents the air quality index at the airport.

Chapter 4

The Crawler and Data Collection

4.1 The Crawler

Since there is no a ready tool for collecting data from the targeted IoT data sources, we develop a novel tool based on ThingSeek search engine [10] [30]. To minimize the required amount of work when collecting data from a new source, we have broken down the crawling procedure into a certain set of steps in a unified framework.

In the first step of crawling, a URL generator initializes the queue of queries. Each entry in the queue is supplied with certain parameters to construct a query to a page or a specific location. The parameters can be the time window, the boundaries of the querying region and/or other parameters. Then for each entity in the queue, a reader function reads the selected part of the page, and the contents are converted to a set of vectors and refined using a refiner. The refiner basically bind all read data from the previous step into subsets. The data for each subset is separately held until all subsets are refined where we merge all of the subsets of the resource's data. In this step, a specific enricher can be possibly used to collect the missing information, if any, from other sources. This can, for example, fill the incomplete fields such as IP address by acquiring them from Shodan. Finally, the collected data from different sources are integrated and stored on a distributed back-end.

Due to the size and dynamics of the sensor-generated data, IoT data sources often provide a subset of their data with a call to their API. Thus, pagination techniques such as location-based queries are deployed to present the data. We use the same mechanism through implementing the URL generator. The URL generator plays a key role in adjusting the workload on the data source. It converts a set of spatial segments to a sequence of queries which can be submitted via the API of the data source. Thus, a highly populated area can be placed multiple times in the processing queue while an empty area may appear only once (or not appear) in the queue. For example, through a URL generator, URL b will be repeated three times for others during a scan as it contains more dynamic objects than others.

To accomplish the data collection process, we need a tool that achieve this task. It was a tricky task to perform since there is no a ready tool that enables us to collect the required data. Therefore, we built a crawler in order to obtain the data from the aforementioned data sources using the proper API for each data source. The idea of building this tool was explained in [10].

We have developed our crawler using a set of tools to collect, process and visualize the dataset. Some of the tools we used are as follows: R programming language, SparkR, Apache Spark 1.4.1 and Rails framework. We initialized the crawler with around 3 data sources for air quality, weather watch and aircraft tracking.

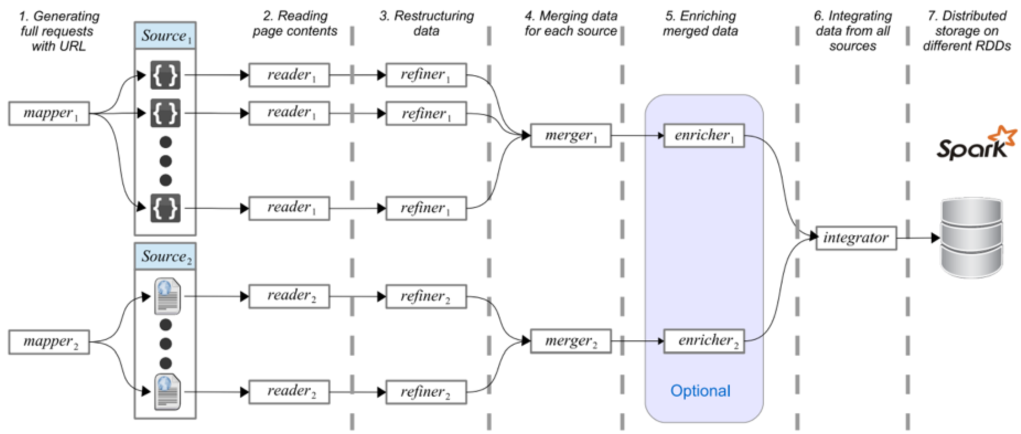


Figure 4.1: ThingSeek Search Engine [10]

As mentioned above, we will target the IoT data sources. For that reason, we borrow idea of implanting a crawler based on ThingSeek search engine in [10] [30]. We use the structure in order to be able to gather the required data for the targeted data sources. The following section will discuss the data collection phase including all steps as shown in Figure 4.1.

Input: IoT Data Source
Output: Data Matrix from The IoT Data Source
URLs = Generate list of URLs for the targeted data sources to start reading the data from;
for *i = 0 to length(URLs)* **do**
    Read the from each URL individually using the API of the IoT data source;
    Refining the response from URL by transforming the read data into data matrix;
end
After reading the data from the IoT data source, Store the data matrix
.

Algorithm 1: Crawling algorithm

4.2 Data Collection

4.2.1 The Collection Process

The data collection process depends on the steps mentioned in the crawler procedure. For any IoT data source we need to create a separate crawling procedure. We build a function for each step: generating URLs, reading data, refining data, and finally merging data. In the first step, a URL generator function generate a list of queries for the IoT data source using the API of the IoT data source. Each entry in the list is supplied with particular parameters to build a query to a page or a specific location. In the second step, a reader function will start reading each entry in that list. Then the refine function will refine the read content to be ready for the merge step. Finally, the merge function combine all read data for each entry in just one container. After having the data of a particular data source, we store it in our machine to be ready for the cleaning phase.

We crawl the IoT data sources and collect the data from the distributes including flights, weather and air quality sensors. In the beginning, our goal was to improve our understanding of the roots and signs of flight delays in order to be able to classify a given flight based on the features from flights and other data sources. We extend the existing works by adding new data sources and considering new factors in the analysis of flight delay. Through the use of real-time data, our goal is to establish a novel service to predict flight delays in real-time.

In our project we have constructed three different case studies. We consider specific criteria such as the weather factor, air quality factor, spatial size for cities, and distance between cities to select the areas. We study the flight delays issue in three case studies: China, Australia, and Europe. In those regions, we select the cities based on the previous criteria as we explained above.

Each data source has several variables. We referred to those variables in the chapter 3.

The temporal scope of the collected data for this study was from April 1st to September 30th 2016. The spatial scope included three various places: China, Australia, and Europe. The time of collection for the data was consistent since we use three blocks for each day: 14-16 pm to collect the data for China case study, 16-18 pm to collect the data for Australia case study, and 20-22 pm to collect the data for Europe case study. The raw data we get from the IoT data source is in json format.

Our data collection process is fully automatic. We utilize the crontab jobs in Mac systems. We automate the collection for each case study based on the specified time. Hint: We discovered later on that RStudio framework has built this feature in the newer version. We downloaded it, it does the same thing like crontab but in a graphical user interface.

The flight data consists 1,134,000 , 2,394,000 , and 810,000 records for Australia, China, and Europe respectively. The weather data contains 234,000 , 360,000, and 4,500,000 records for Australia, China, and Europe respectively. The air quality index data consists 14,580 , 558,000 , and 396,000 records for Australia, China, and Europe respectively.

4.3 China Case Study

4.3.1 Flight Data

We select 7 airports in 7 big cities in China. The cities are: Beijing (PEK), Shanghai (SHA), Guangzhou (CAN), Wuhan (WUH), Chengdu (CTU), Harbin (HRB), and Dalian (DLC). We collect all flights among these cities. The number of flights is 800 flights. The distribution of the number of flights varies according to the city size and the population number. The airlines operate the flight between these cities are: China Air (CA), Shanghai Airlines (FM), China Eastern Airlines (MU), China Southern Airlines (CZ), Juneyao Airlines (HO), Hainan Airlines (HU), Xiamen Airlines (MF), Sichuan Airlines (3U), Shandong Airlines (SC), Chongqing Airlines(OQ), Grand China Air (CN), Shenzhen Airlines (ZH), Spring Airlines (9C), Tibet Airlines (TV), Beijing Capital Airlines (JD), Chengdu Airlines (EU).

identification.row	identification.number.default	identification.number.alternative	status.live	status.text
2938392811	CA1855	CA1855	FALSE	Scheduled
2934625405	CA1855	CA1855	FALSE	Scheduled
2930890219	CA1855	CA1855	FALSE	Scheduled
2930890220	CA1855	CA1855	FALSE	Scheduled
2920607116	CA1855	CA1855	FALSE	Scheduled
2916225693	CA1855	CA1855	FALSE	Scheduled
2912409039	CA1855	CA1855	FALSE	Scheduled
2908630624	CA1855	CA1855	FALSE	Scheduled
2904833455	CA1855	CA1855	FALSE	Scheduled

Figure 4.1: Sample of the flight records we get when we run the crawler - China Case Study

4.3.2 Weather Data

The process of collecting the weather data for a country differs from the process we do when we collect flights data. To collect the flight data, we need to prepare the flights data set in order to allow the crawler to check them online and bring all the required records. However, to collect the weather data for a particular country such as China, we need to write a function to instruct the crawler to fetch the data from the stations that are in the range of the search. Otherwise, the crawler ignores reading data from the other stations. We do that for two reasons. First we only need weather data of China. Second, we build this method in order to generalized it for any other country.

winddir	windspeedmph	humidity	tempf	rainin	baromin	dewptf	weather
-999	0.0	72	87	-999	29.85	79	Partly Cloudy
290	2.3	75	88	-999	29.82	81	Partly Cloudy
230	5.8	72	80	-999	29.80	72	Partly Cloudy
270	9.2	73	86	-999	29.79	78	Scattered Clouds
270	5.8	42	93	-999	29.71	72	Scattered Clouds
230	2.3	43	93	-999	29.71	73	Scattered Clouds
230	17.3	77	89	-999	29.72	83	Mostly Cloudy
-999	0.0	71	83	-999	29.77	75	Mist

Figure 4.2 Sample of the weather records we get when we run the crawler - China Case Study

4.3.3 Air Quality Index Data

To collect the air quality index data, we use the same idea of collecting weather data. So we write a function that set the required location parameters of our case. In this case, we set the location parameters of China.

lat	lon	aqi	utime	sutime	stamp
14.349366683822	100.56853549578	21	Friday 28th August 08:00	2015-08-28 06:00:00	1440716400
14.683085094818	100.87513981078	49	Friday 28th August 08:00	2015-08-28 06:00:00	1440716400
14.523536277394	100.92917127159	17	Friday 28th August 09:00	2015-08-28 07:00:00	1440720000
14.040299305494	100.60873959621	-	Sunday 16th August 21:00	2015-08-16 19:00:00	1439726400
14.976785802969	102.10219652705	-	Thursday 27th August 10:00	2015-08-27 08:00:00	1440637200
14.5995124	120.9842195	108	Friday 28th August 08:00	2015-08-28 07:00:00	1440716400
14.6714904	120.93984669999998	166	Friday 28th August 08:00	2015-08-28 07:00:00	1440716400
14.65	120.96666700000003	98	Friday 28th August 09:00	2015-08-28 08:00:00	1440720000
14.554729	121.02444519999995	-	Tuesday 25th August 14:00	2015-08-25 13:00:00	1440478800

Figure 4.3 Sample of the AQI records we get when we run the crawler - China Case Study

4.4 Australia Case Study

4.4.1 Flight Data

For Australia case study, we select 6 airports in 6 big cities. We are interested in the main airport in each state in Australia. The cities are: Canberra (CBR), Melbourne (MEL), Sydney (SYD), Adelaide (ADL), Brisbane (BNE), and Perth (PER). We collect all flights among these cities. The number of flights is 526 flights. The distribution of the number of flights varies according to the city size and the population number.

In Australia case study, there are four airlines.These airlines operate the flight between these cities, and they are: Qantas Airways (QF), Virgin Australia International Airlines (VA), Jetstar Airways (JQ), and Tiger Airways Australia (TT).

Flight.ID	Status.with.Time	Status	Status.Type	Airline	Airline.IATA.Code	Orig.Airport	Orig.Airport.IATA
QF656	Scheduled	scheduled	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Scheduled	scheduled	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Scheduled	scheduled	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Scheduled	scheduled	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Scheduled	scheduled	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Scheduled	scheduled	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Estimated dep 18:00	estimated	departure	Qantas	QF	Adelaide Airport	ADL
QF656	Landed 20:44	landed	arrival	Qantas	QF	Adelaide Airport	ADL
QF656	Landed 20:54	landed	arrival	Qantas	QF	Adelaide Airport	ADL
QF656	Landed 20:39	landed	arrival	Qantas	QF	Adelaide Airport	ADL

Figure 4.4 Sample of the flight records we get when we run the crawler - Australia Case Study

4.4.2 Weather Data

To collect the weather data for Australia, we need to write a function to instruct the crawler to fetch the data from the stations that are in the range of the search. Otherwise, the crawler ignores reading data from the other stations.

epoch	ageh	agem	ages	type	id	lat	lon	adm1	adm2	country
1473065597	0	37	44	SYNOP	WMO94850	-39.88010025	143.88290405	King Island Airport		Australia
1473065598	2	32	29	SYNOP	BUOYC6FS9	-39.1	144.1	C6FS9	C6FS9	
1473065600	0	20	39	SYNOP	WMO94893	-39.12969971	146.42439270	Wilsons Promontory Light		Australia
1473065600	0	23	8	SYNOP	WMO94949	-39.22249985	146.98410034	Hogan Island Aws		Australia
1473065600	0	5	3	PWS	ITASLEEK2	-39.90038681	147.86463928	Leeka	TAS	AUSTRALIA
1473065621	0	23	48	SYNOP	WMO95826	-38.34389877	141.61360168	Portland Ntc Aws		AU
1473065621	0	21	0	SYNOP	WMO94826	-38.43059921	141.54370117	Cape Nelson		Australia
1473065621	0	21	0	SYNOP	WMO94828	-38.31480026	141.47050476	Portland Airport		Australia
1473065622	0	0	3	PWS	IVICTORI398	-38.38027954	142.51370239	Warrnambool	VICTORIA	AU
1473065622	0	1	24	PWS	IVICTORI1055	-38.32110214	142.32579041	Crossley	VICTORIA	AU

Figure 4.5 Sample of the weather records we get when we run the crawler - Australia Case Study

4.4.3 Air Quality Index Data

To collect the air quality index data, we use the same idea of collecting weather data. So we write a function that set the required location parameters of our case. In this case, we set the location parameters of Australia.

lat	lon	city	idx	stamp	pol	x	aqi	tz	utime
-38.23562	144.3030	GeelongSth., Australia	3967	1473058800	pm25	3247	18	+0900	2016-09-05 16:00:00
-38.19688	146.5056	Traralgon, Australia	3970	1473058800	pm25	3248	22	+0900	2016-09-05 16:00:00
-38.24000	146.3900	Morwell Sth., Australia	3968	1473058800	pm25	4749	20	+0900	2016-09-05 16:00:00
-37.79974	144.8997	Footscray, Australia	3963	1473058800	pm25	3242	21	+0900	2016-09-05 16:00:00
-37.67031	144.9331	Moe, Australia	3972	1473058800	pm25	8009	16	+0900	2016-09-05 16:00:00
-37.83469	144.8474	AltonaNorth, Australia	3960	1473058800	pm25	3239	15	+0900	2016-09-05 16:00:00
-37.82106	144.8350	Brooklyn, Australia	3961	1473058800	pm25	3240	15	+0900	2016-09-05 16:00:00
-37.90950	144.7519	Pt. Cook, Australia	3965	1473058800	pm25	3244	15	+0900	2016-09-05 16:00:00
-37.68298	144.5805	Melton, Australia	3964	1473058800	pm25	3243	5	+0900	2016-09-05 16:00:00
-37.77009	144.7727	Deer Park, Australia	3962	1442883600	pm25	3241	-	+0900	2015-09-22 10:00:00

Figure 4.6 Sample of the AQI records we get when we run the crawler - Australia Case Study

4.5 Europe Case Study

4.5.1 Flight Data

For Europe case study, we select 8 airports in 8 capitals. We are interested in the main airport in these capitals in Europe. The cities are: Madrid (MAD), Paris (CDG), Rome (FCO), Brussels (BRU), Berlin (TXL), London (LHR), Vienna (VIE), and Moscow (DME) . We collect all flights among these cities. The number of flights is 620 flights. The distribution of the number of flights varies according to the city size and the population number.

In Europe case study, there are several airlines.These airlines operate the flight between these cities, and they are: British Airways (BA) , Air Europa (UX), Austrian Airlines (OS), Air France (AF), Iberia (IB), Air Berlin (AB), Brussels Airlines (SN), Germanwings (4U), S7 Airlines (S7), Ryanair (FR), Alitalia (AZ), Kuwait Airways (KU), Vueling (VY), Niki (HG), EasyJet (U2), Ethiopian Airlines (ET), Korean Air (KE).

identification.row	identification.number.default	status.live	status.text	status.ambiguous	status.generic.status.text	status.generic.status.type
3545200269	VY8202	FALSE	Scheduled	FALSE	scheduled	departure
354081467 3545200269	202	FALSE	Scheduled	FALSE	scheduled	departure
3538594574	VY8202	FALSE	Scheduled	FALSE	scheduled	departure
3534085485	VY8202	FALSE	Scheduled	FALSE	scheduled	departure
3530166915	VY8202	FALSE	Scheduled	FALSE	scheduled	departure
3526224318	VY8202	FALSE	Scheduled	FALSE	scheduled	departure
3514745597	VY8202	FALSE	Landed 08:53	FALSE	landed	arrival
3510825677	VY8202	FALSE	Landed 08:49	FALSE	landed	arrival
3506866738	VY8202	FALSE	Landed 08:46	FALSE	landed	arrival
3547366528	VY8204	FALSE	Scheduled	FALSE	scheduled	departure

Figure 4.7 Sample of the flight records we get when we run the crawler - Europe Case Study

4.5.2 Weather Data

To collect the weather data for Europe, we need to write a function to instruct the crawler to fetch the data from the stations that are in the range of the search. Otherwise, the crawler ignores reading data from the other stations.

epoch	ageh	agem	ages	type	id	lat	lon	adm1
1476720928	5	45	12	SYNOP	BUOYDFPY2	30.0	-11.8	DFPY2
1476720930	0	1	3	ICAO	GMAD	30.32888985	-9.39944363	Agadir Al Massira
1476720930	0	3	57	PWS	IAGADIR3	30.41285133	-9.55869770	Agadir
1476720932	0	1	5	ICAO	GMMZ	30.93905258	-6.90943098	Ouarzazate
1476720935	0	59	47	SYNOP	WMO60602	30.13333321	-2.16666698	Beni Abbes
1476720939	0	0	35	ICAO	DAUE	30.57129478	2.85958600	El Golea
1476720948	6	56	30	SYNOP	WMO62120	30.38333321	13.58333302	Gariat El-Sharghia
1476720960	0	31	25	ICAO	HEBA	30.91769981	29.69639969	Alexandria Borg El Arab
1476720960	0	30	40	SYNOP	WMO62357	30.40250015	30.36333275	Wadi El Natroon

Figure 4.8 Sample of the weather records we get when we run the crawler - Europe Case Study

4.5.3 Air Quality Index Data

To collect the air quality index data, we use the same idea of collecting weather data. So we write a function that set the required location parameters of our case. In this case, we set the location parameters of Europe.

lat	lon	city	idx	stamp	pol	x	aqi
31.64571	34.674020	Sde Yoav, Southern Coastal Plain, Israel (ישראל,שדה יו...	1281	1476720000	pm25	2976	26
31.80438	34.655314	Ashdod (ישראל,dod)	2313	1476720000	pm25	5785	53
31.65374	34.550750	South Ashkelon, southern coastal plain, Israel (,ישראל...	1361	1476720000	pm25	7726	53
31.68610	34.636180	Nir Israel, Southern Coastal Plain, Israel (ישראל,ניר ישר...	1228	1476720000	pm25	2973	50
31.72793	34.740350	Jack. Malachi, southern coastal plain, Israel (ישראל,ק.מ...	1280	1476720000	pm25	2975	50
31.77196	34.627150	Tu District, southern coastal plain, Israel (ישראל,רובע ט...	1234	1476720000	pm25	5748	50
31.52796	34.601610	Sderot, the southern coastal plain, Israel (ישראל,שדרות...	1352	1476720000	pm25	2977	60
31.81355	34.778140	Pen, Southern Coastal Plain, Israel (ישראל,גדרה, מישור...	1278	1476720000	pm25	2969	45
31.58932	34.609690	Jack. Gvaram, Southern Coastal Plain, Israel (ישראל,ק.ג...	1351	1476720000	pm25	2974	92
31.65950	34.569750	Ashkelon, southern coastal plain, Israel (ישראל,אשקלון...	1279	1476720000	pm25	2972	78

Figure 4.9 Sample of the AQI records we get when we run the crawler - Europe Case Study

Chapter 5

Data Processing

5.1 Data Exploration - Uni-variate Data Analysis

After we collect the data from the IoT data sources, it is important for us to investigate the nature collected data in order to determine its quality, because the quality of the data determines how the modeling will be. For that reason, we conduct a comprehensive uni-variant analysis. At this stage, we examine variables individually in the datasets. In order to perform this analysis, first we need to identify the data type of each variable because choosing the proper method for conducting the uni-variate analysis depends on the data type of the variables. Thus, variables can be continuous or categorical. Whenever we know that, we can perform the statistical measures on these variables. However, there is one point we must be aware about. Sometimes, the data type of a variable can be categorical, but the values of this variable are continuous. In this case, we need to convert the type. Performing this requires a better understanding of the domain we target.

In case if the variable is continuous, we must understand the central tendency and the how the variable is spread. There are several statistical metrics and visualization methods in R language can help us to do this. If the variable is categorical, we need to understand the distribution of each category. We can do that by using frequency table or percentage of values under each category. Frequency table can measured by the metric Count, and the percentage can be measured by the Count%.

5.1.1 China Case Study

In China case study, we perform uni-variant analysis on the three datasets: flight dataset, weather dataset, and air quality dataset. On each dataset, we perform statistical analysis. First, we investigate the structure of the data. Second, we summarize the content of the data to see the distribution of values. Finally, we present a sample of some variables' density distribution.

Flight Data Uni-variate Analysis

Structure of the Data

```
$ identification.row                         : Factor w/ 22338 levels "3543134128","3540892663",..: 1 2 3 4 5 6 7 8 9 10 ...
$ identification.number.default              : Factor w/ 764 levels "CA1855","CA1857",..: 1 1 1 1 1 1 1 1 1 1 ...
$ status.live                                : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ status.text                                : Factor w/ 1920 levels "Scheduled","Landed 19:31",..: 1 1 1 1 1 1 1 1 2 3 ...
$ status.ambiguous                           : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ status.generic.status.text                 : Factor w/ 7 levels "scheduled","landed",..: 1 1 1 1 1 1 1 1 2 2 ...
$ status.generic.status.type                 : Factor w/ 2 levels "departure","arrival": 1 1 1 1 1 1 1 1 2 2 ...
$ status.generic.status.color                : Factor w/ 4 levels "gray","green",..: 1 1 1 1 1 1 1 1 2 2 ...
$ airline.name                               : Factor w/ 21 levels "Air China","Shanghai Airlines",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airline.code.iata                          : Factor w/ 21 levels "CA","FM","MU",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airline.code.icao                          : Factor w/ 21 levels "CCA","CSH","CES",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.name                        : Factor w/ 74 levels "Beijing Capital International Airport",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.code.iata                   : Factor w/ 73 levels "PEK","SHA","SFO",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.code.icao                   : Factor w/ 73 levels "ZBAA","ZSSS",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.latitude           : Factor w/ 73 levels "40.080109","31.19787",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.longitude          : Factor w/ 73 levels "116.584503","121.336304",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.country.name       : Factor w/ 6 levels "China","United States",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.country.code       : Factor w/ 6 levels "CN","US","IT",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.region.city        : Factor w/ 72 levels "Beijing","Shanghai",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.name               : Factor w/ 9 levels "Asia/Shanghai",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.offset             : Factor w/ 7 levels "28800","-25200",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.abbr               : Factor w/ 7 levels "CST","PDT","XJT",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.abbrName           : Factor w/ 7 levels "China Standard Time",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.isDst              : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.visible                     : Factor w/ 1 level "TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.name                   : Factor w/ 68 levels "Shanghai Hongqiao International Airport",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.code.iata              : Factor w/ 68 levels "SHA","CKG","PEK",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.code.icao              : Factor w/ 68 levels "ZSSS","ZUCK",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.latitude      : Factor w/ 68 levels "31.19787","29.71921",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.longitude     : Factor w/ 68 levels "121.336304","106.641602",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.country.name: Factor w/ 3 levels "China","Japan",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.country.code: Factor w/ 3 levels "CN","JP","US": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.region.city : Factor w/ 67 levels "Shanghai","Chongqing",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.name          : Factor w/ 6 levels "Asia/Shanghai",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.offset        : Factor w/ 4 levels "28800","32400",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.abbr          : Factor w/ 4 levels "CST","JST","XJT",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.abbrName      : Factor w/ 4 levels "China Standard Time",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.isDst         : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.visible                : Factor w/ 1 level "TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ time.scheduled.departure                   : Factor w/ 5391 levels "1477215000","1477128600",..: 1 2 3 4 5 6 7 8 9 10 ...
$ time.scheduled.arrival                     : Factor w/ 5478 levels "1477222800","1477136400",..: 1 2 3 4 5 6 7 8 9 10 ...
$ time.other.updated                         : Factor w/ 17993 levels "1476501687","1476412389",..: 1 1 1 1 1 1 1 2 3 4 ...
$ identification.id                          : Factor w/ 14466 levels "b50d50a","b4e27a8",..: NA NA NA NA NA NA NA NA 1 2 ...
$ identification.callsign                    : Factor w/ 865 levels "CCA1855","CCA1857",..: NA NA NA NA NA NA NA NA 1 1 ...
$ status.icon                                : Factor w/ 3 levels "green","yellow",..: NA NA NA NA NA NA NA NA 1 1 ...
$ status.generic.eventTime.utc               : Factor w/ 16600 levels "1476531093","1476443635",..: NA NA NA NA NA NA NA NA 1 2 ...
$ status.generic.eventTime.local             : Factor w/ 16566 levels "1476559893","1476472435",..: NA NA NA NA NA NA NA NA 1 2 ...
$ aircraft.model.code                        : Factor w/ 17 levels "A321","B789",..: NA NA NA NA NA NA NA NA 1 2 ...
$ aircraft.model.text                        : Factor w/ 60 levels "Airbus A321-232",..: NA NA NA NA NA NA NA NA 1 2 ...
$ aircraft.hex                               : Factor w/ 1512 levels "7807CC","78100A",..: NA NA NA NA NA NA NA NA 1 2 ...
$ aircraft.registration                      : Factor w/ 1511 levels "B-6823","B-7832",..: NA NA NA NA NA NA NA NA 1 2 ...
$ aircraft.serialNo                          : Factor w/ 1506 levels "4873","34309",..: NA NA NA NA NA NA NA NA 1 2 ...
$ aircraft.owner                             : Factor w/ 72 levels "Air China","Air China (Red Phoenix Liner Livery)",..: NA NA NA NA NA NA NA NA 1 1 ...
$ time.real.departure                        : Factor w/ 14704 levels "1476524981","1476437883",..: NA NA NA NA NA NA NA NA 1 2 ...
$ time.real.arrival                          : Factor w/ 10059 levels "1476530940","1476443640",..: NA NA NA NA NA NA NA NA 1 2 ...
```

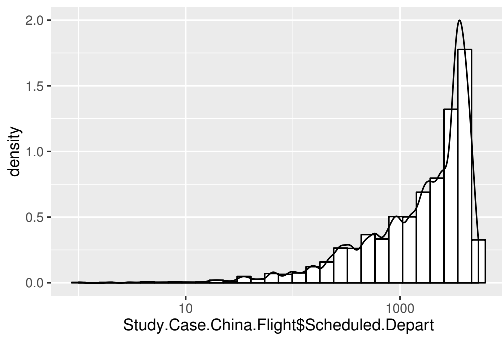
Figure 5.1 Sample of the flight dataset structure - China Case Study

Data Summary

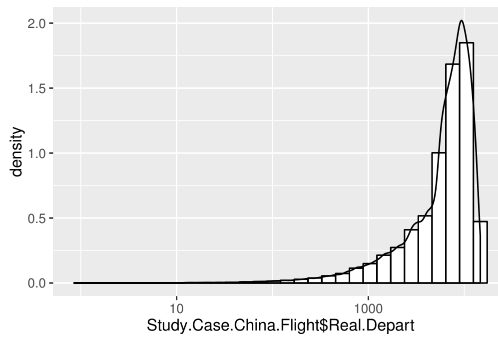
```
Flight.ID                  Status.with.Time        Status          Status.Type
3U8840 :    900   Scheduled           :79350   scheduled:80005   departure:89338
3U8839 :    600   Unknown             : 4380   landed    :63687   arrival  :67550
ZH9658 :    600   Canceled            :  889   canceled :  889
MU2296 :    600   Scheduled*          :  655   unknown  : 4380
CZ6433 :    600   Landed 10:35        :  150   estimated: 7490
CA8925 :    600   Estimated dep 08:00:  136   diverted :   84
(Other):152988   (Other)             :71328   delayed  :  353
Airline       Airline.IATA.Code
China Eastern Airlines :37581   MU      :37581
China Southern Airlines:37581   CZ      :37581
Air China              :29729   CA      :29729
Hainan Airlines        :11049   HU      :11049
Shenzhen Airlines      :10649   ZH      :10649
Sichuan Airlines       : 9626   3U      : 9626
(Other)                :20673   (Other):20673
Orig.Airport   Orig.Airport.IATA  Orig.Airport.Lat
Beijing Capital International Airport  :28487   PEK     :28487     40.080109:28487
Guangzhou Baiyun International Airport :22509   CAN     :22509     23.392429:22509
Xi'an Xianyang International Airport   :17765   XIY     :17765     34.447109:17765
Shanghai Hongqiao International Airport:14741   SHA     :14741     31.19787 :14741
Nanjing Lukou International Airport    :12789   NKG     :12789     31.742041:12789
(Other)                                :60467   (Other):60467     (Other)  :60467
NA's                                   :  130   NA's   :  130     NA's     :  130
Orig.Airport.Lon              Orig.Airort.Country Orig.Airport.City
116.584503:28487   China                   :156007   Beijing  :28487
113.298698:22509   United States           :   318   Guangzhou:22509
108.751503:17765   Italy                   :    82   Xi'an    :17765
121.336304:14741   Northern Mariana Islands:    51   Shanghai :15033
118.862   :12789   Japan                   :   222   Nanjing  :12789
(Other)   :60467   Australia               :    78   (Other)  :60175
NA's      :  130   NA's                    :   130   NA's     :  130
Dest.Airport   Dest.IATA.Code   Dest.Airport.Lat
Beijing Capital International Airport  :27137   PEK     :27137   40.080109:27137
Guangzhou Baiyun International Airport :19649   CAN     :19649   23.392429:19649
Shanghai Hongqiao International Airport:17515   SHA     :17515   31.19787 :17515
Xian Xianyang International Airport   :17028   XIY     :17028   34.447109:17028
Chengdu Shuangliu International Airport:14547   CTU     :14547   30.57852 :14547
(Other)                                :60212   (Other):60212   (Other)  :60212
NA's                                   :  800   NA's   :  800   NA's     :  800
Dest.Airport.Lon    Dest.Airport.Country Dest.Airport.City Scheduled.Depart  Real.Depart
116.584503:27137   China        :155397   Beijing  :27137   Min.   :   1     Min.   :    1
113.298698:19649   Japan        :   454   Guangzhou:19649   1st Qu.: 821     1st Qu.: 4084
121.336304:17515   United States:   237   Shanghai :17644   Median :2175     Median : 7217
108.751503:17028   NA's         :   800   Xi'an    :17028   Mean   :2242     Mean   : 7004
103.946999:14547                          Chengdu  :14547   3rd Qu.:3551     3rd Qu.: 9815
(Other)   :60212                          (Other)  :60083   Max.   :5391     Max.   :14704
NA's      :  800                          NA's     :  800   NA's   :957      NA's   :89637
Scheduled.Arrival  Real.Arrival    Aircraft.Model.Code
Min.   :   1       Min.   :    1   B738    :20581
1st Qu.: 922       1st Qu.: 2657   A320    :19357
Median :2181       Median : 4957   A321    :13770
Mean   :2285       Mean   : 4767   A333    : 5224
3rd Qu.:3596       3rd Qu.: 6740   A319    : 3635
Max.   :5478       Max.   :10059   (Other):10294
```

Figure 5.2 Sample of the weather dataset summary - China Case Study

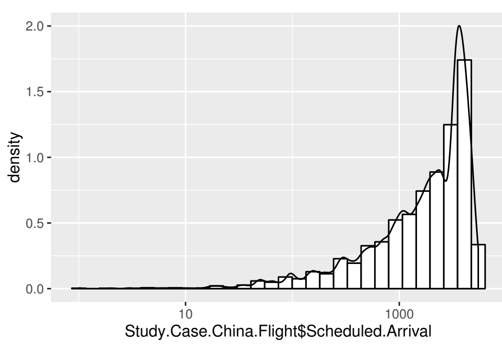
Sample of Density Analysis



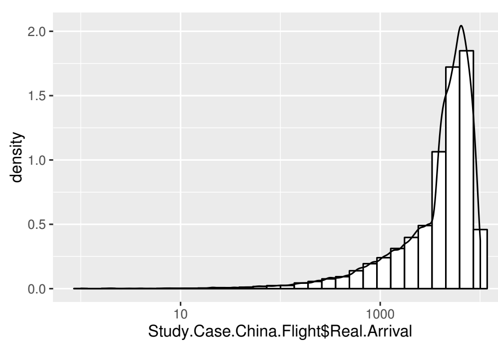
(a) Scheduled Departure Time for Flights



(b) Real Departure Time for Flights



(a) Scheduled Arrival Time for Flights



(b) Real Arrival Time for Flights

Figure 5.3: Sample of Continuous Variables Distribution of Flight Data - China Case Study

5.1.2 Weather Data Uni-variate Analysis

The Structure of the Data

```
$ epoch           : num  1.48e+09 1.48e+09 1.48e+09 1.48e+09 1.48e+09 ...
$ ageh            : num  0 3 2 10 10 0 0 0 0 0 ...
$ agem            : num  4 31 49 16 24 4 4 3 3 24 ...
$ ages            : num  35 48 35 52 59 36 37 27 28 30 ...
$ type            : Factor w/ 3 levels "SYNOP","PWS",..: 1 1 1 1 1 1 1 1 1 1 ...
$ id              : Factor w/ 2659 levels "WMO43226","WMO43225",..: 1 2 3 4 5 6 7 8 9 10 ...
$ lat             : Factor w/ 2527 levels "14.28333282",..: 1 2 3 2 4 5 6 7 8 9 ...
$ lon             : Factor w/ 2568 levels "74.44999695",..: 1 2 3 4 5 6 7 8 9 10 ...
$ adm1            : Factor w/ 1996 levels "Honavar","Karwar",..: 1 2 3 4 5 6 7 8 9 10 ...
$ adm2            : Factor w/ 259 levels "","VTJR","TANINTHARYI REGION",..: 1 1 1 1 1 1 1 1 1 2 ...
$ country         : Factor w/ 63 levels "India","IN","",..: 1 1 1 2 2 1 1 1 1 3 ...
$ dateutc         : Factor w/ 6872 levels "2016-10-15 15:00:00",..: 1 2 2 3 3 1 1 1 1 4 ...
$ winddir         : Factor w/ 363 levels "90","-999","140",..: 1 2 2 3 1 2 4 2 2 5 ...
$ windspeedmph    : Factor w/ 162 levels "1.2","0.0","4.6",..: 1 2 1 3 1 2 4 2 2 5 ...
$ humidity        : Factor w/ 101 levels "81","70","46",..: 1 2 2 3 4 5 6 7 8 9 ...
$ tempf           : Factor w/ 685 levels "78","85","84",..: 1 2 3 4 1 5 6 7 2 2 ...
$ rainin          : Factor w/ 38 levels "-999","-9999.00",..: 1 1 1 1 1 1 1 1 1 1 ...
$ baromin         : Factor w/ 287 levels "29.86","29.81",..: 1 2 3 4 4 1 5 5 6 6 ...
$ dewptf          : Factor w/ 807 levels "73","77","76",..: 1 2 3 4 5 6 7 5 8 9 ...
$ weather         : Factor w/ 51 levels "-999","Clear",..: 1 1 2 3 4 1 1 2 2 1 ...
$ icon            : Factor w/ 12 levels "-999","clear",..: 1 1 2 3 3 1 1 2 2 1 ...
$ clouds          : Factor w/ 9 levels "unknown","","-999",..: 1 1 1 1 1 1 1 1 1 1 ...
$ flightrule      : Factor w/ 5 levels "IFR","VFR","N/A",..: 1 2 1 2 2 1 1 1 1 3 ...
$ visibilitysm    : Factor w/ 59 levels "2","6","-999",..: 1 2 1 2 2 1 1 1 1 3 ...
$ windgustmph     : Factor w/ 137 levels "-999","2.2","4.9",..: 1 1 1 1 1 1 1 1 1 1 ...
$ snowin          : Factor w/ 2 levels "-999","-9999.00": 1 1 1 1 1 1 1 1 1 1 ...
$ name            : Factor w/ 1508 levels "Honavar","Karwar",..: 1 2 3 4 5 6 7 8 9 10 ...
$ elev            : Factor w/ 877 levels "194","13","148",..: 1 2 3 4 5 6 7 8 8 9 ...
$ windchillf      : Factor w/ 140 levels "-999","-9999",..: 1 1 1 1 1 1 1 1 1 1 ...
$ heatindexf      : Factor w/ 206 levels "-9999","78","83",..: 1 1 1 1 1 1 1 1 1 1 ...
$ updated         : Factor w/ 10558 levels "1476547232","1476534799",..: 1 2 3 4 5 1 1 6 6 7 ...
$ neighborhood    : Factor w/ 800 levels "WX ZOC","69 Moo 3",..: NA NA NA NA NA NA NA NA NA NA ...
$ partner_id      : Factor w/ 1 level "": NA NA NA NA NA NA NA NA NA NA ...
$ dailyrainin     : Factor w/ 187 levels "0.21","0.12",..: NA NA NA NA NA NA NA NA NA NA ...
$ softwaretype    : Factor w/ 61 levels "EasyWeather V8.8.0",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxtemp         : Factor w/ 634 levels "94.5","80.1",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxtemp_time    : Factor w/ 1085 levels "11:13AM","11:05PM",..: NA NA NA NA NA NA NA NA NA NA ...
$ mintemp         : Factor w/ 532 levels "75.9","80.1",..: NA NA NA NA NA NA NA NA NA NA ...
$ mintemp_time    : Factor w/ 1214 levels "5:52AM","11:05PM",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxdewpoint     : Factor w/ 685 levels "79.2","70.7",..: NA NA NA NA NA NA NA NA NA NA ...
$ mindewpoint     : Factor w/ 704 levels "72.9","70.7",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxpressure     : Factor w/ 276 levels "29.79","30.03",..: NA NA NA NA NA NA NA NA NA NA ...
$ minpressure     : Factor w/ 272 levels "29.63","30.03",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxwindspeed    : Factor w/ 103 levels "16","3","0","8.1",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxwindgust     : Factor w/ 122 levels "19","5","-999",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxrain         : Factor w/ 147 levels "-9999.00","0.71",..: NA NA NA NA NA NA NA NA NA NA ...
$ maxheatindex    : Factor w/ 255 levels "109","83","75",..: NA NA NA NA NA NA NA NA NA NA ...
$ minwindchill    : Factor w/ 179 levels "-999","-9999",..: NA NA NA NA NA NA NA NA NA NA ...
$ rtfreq          : Factor w/ 8 levels "5.0","36.0","2.5",..: NA NA NA NA NA NA NA NA NA NA ...
$ indoortempf     : Factor w/ 288 levels "74.5","93.0",..: NA NA NA NA NA NA NA NA NA NA ...
$ indoorhumidity  : Factor w/ 79 levels "77","56","66",..: NA NA NA NA NA NA NA NA NA NA ...
$ RawP            : Factor w/ 988 levels "29.77","30.03",..: NA NA NA NA NA NA NA NA NA NA ...
$ tzname          : Factor w/ 30 levels "Asia/Rangoon",..: NA NA NA NA NA NA NA NA NA NA ...
```

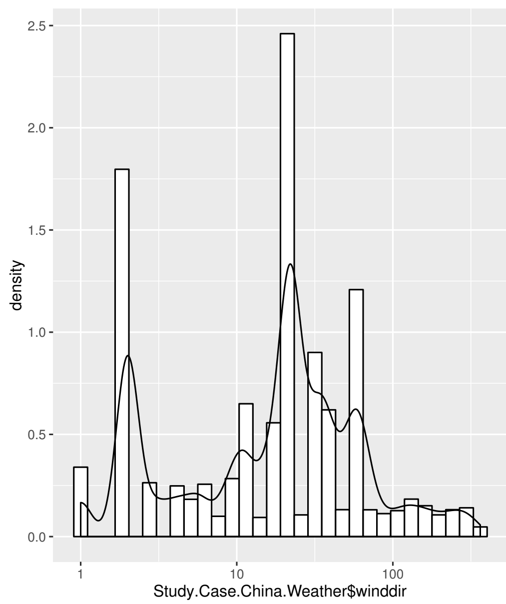
Figure 5.4 Sample of the Weather Dataset Structure - China Case Study

Data Summary

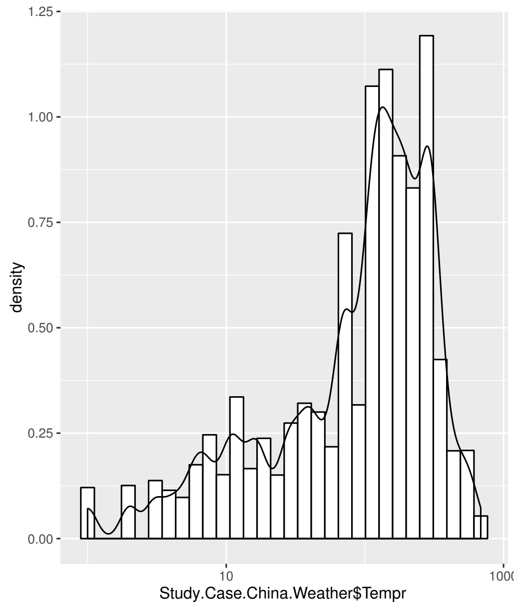
```
 TimeStamp                    Lat                       Lon             Country
 Min.    :1.477e+09   27.33333206:   55   74.53333282 :   54   China  :4079
 1st Qu.:1.477e+09   43.59999847:   36   118.15000153:   36   JP     :3716
 Median :1.477e+09   44.56666565:   36   91.73332977 :   35   India  :2266
 Mean   :1.477e+09   53.96666718:   36   80.36666870 :   33   Russia :1996
 3rd Qu.:1.477e+09   53.46666718:   36   77.16666412 :   32   CI     :1590
 Max.   :1.477e+09   54.36666870:   36   106.59999847:   24   (Other):9831
 (Other)     :23406   (Other)     :23427   NA's    : 163
 DateUTC               Wind.Dir      Wind.Speed.MPH     Humidity          Tempr
 2016-10-19 12:00:00: 1062   -9999   : 4787   Min.   :  1.00   Min.   :  1.00   Min.   :  1.0
 2016-10-23 12:00:00: 1002   -999    : 3724   1st Qu.:  3.00   1st Qu.: 17.00   1st Qu.: 39.0
 2016-10-25 00:00:00:  950   0       : 1178   Median : 19.00   Median : 31.00   Median :120.0
 2016-10-21 09:00:00:  948   270     :  805   Mean   : 17.09   Mean   : 36.31   Mean   :146.5
 2016-10-17 12:00:00:  944   320     :  738   3rd Qu.: 21.00   3rd Qu.: 54.00   3rd Qu.:213.8
 (Other)               :18572   (Other):12246   Max.   :162.00   Max.   :101.00   Max.   :685.0
 NA's                  :  163   NA's    :  163   NA's   :163     NA's   :163     NA's   :163
 Raining           Baromin          Dew.Point        Visibility     Wind.Gust.MPH
 Min.   : 1.000   Min.   :  1.00   Min.   :  1.0   Min.   : 1.00   Min.   :  1.000
 1st Qu.: 1.000   1st Qu.: 29.00   1st Qu.: 34.0   1st Qu.: 2.00   1st Qu.:  1.000
 Median : 1.000   Median : 53.00   Median :135.0   Median : 9.00   Median :  1.000
 Mean   : 2.877   Mean   : 56.21   Mean   :171.4   Mean   :12.79   Mean   :  4.044
 3rd Qu.: 4.000   3rd Qu.: 81.00   3rd Qu.:271.0   3rd Qu.:19.00   3rd Qu.:  4.000
 Max.   :38.000   Max.   :287.00   Max.   :807.0   Max.   :59.00   Max.   :137.000
 Terminal    63   NA's   :163      NA's   :163     NA's   :7502    NA's   :163
                  Heat.Index       Daily.Raining     Max.Presure      winddir
 Min.   :  1.0    Min.   :  1.00   Min.   :  1.000   Min.   :  1.00   Min.   :  1.00
 1st Qu.: 32.0    1st Qu.:  1.00   1st Qu.:  3.000   1st Qu.: 29.00   1st Qu.:  6.00
 Median :190.0    Median :  1.00   Median :  7.000   Median : 50.00   Median : 22.00
 Mean   :260.8    Mean   : 14.32   Mean   :  9.075   Mean   : 56.65   Mean   : 35.87
 3rd Qu.:444.0    3rd Qu.: 16.00   3rd Qu.:  7.000   3rd Qu.: 78.00   3rd Qu.: 38.00
 Max.   :877.0    Max.   :206.00   Max.   :187.000   Max.   :276.00   Max.   :363.00
 NA's   :163      NA's   :163      NA's   :14472     NA's   :14472    NA's   :163
```

Figure 5.5 Sample of the Weather Dataset Summary - China Case Study

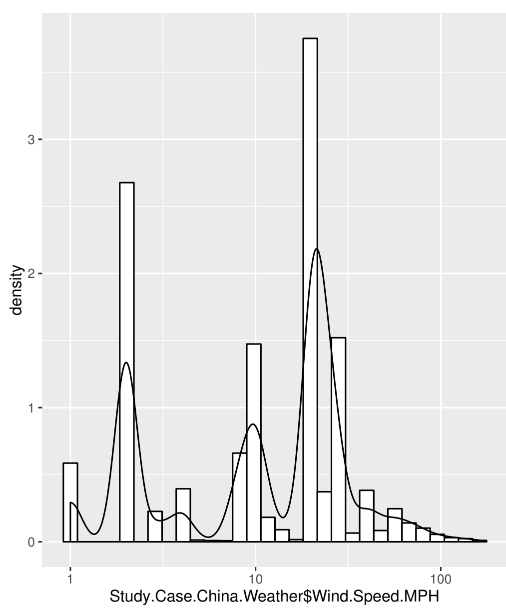
Sample of Density Analysis



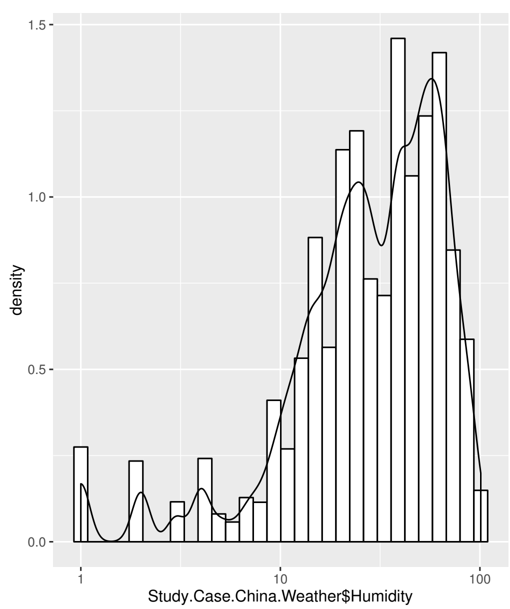
(a) Wind Direction



(b) Temperature



(a) Wind Speed



(b) Humidity

Figure 5.6: Sample of Continuous Variables Distribution of Weather Data - China Case Study

5.1.3 Air Quality Index Data Uni-variate Analysis

The Structure of the Data

```
$ lat  : num  14.3 14.7 14.5 14 15 ...
$ lon  : num  101 101 101 101 102 ...
$ idx  : num  967 969 996 1029 1006 ...
$ stamp: num  1.47e+09 1.47e+09 1.47e+09 1.47e+09 1.47e+09 ...
$ pol  : Factor w/ 3 levels "pm10","pm25",..: 1 1 1 1 1 2 2 2 2 3 ...
$ x    : Factor w/ 3330 levels "1815","1817",..: 1 2 3 4 5 6 7 8 9 10 ...
$ aqi  : Factor w/ 389 levels "-","109","59",..: 1 1 1 1 1 2 3 4 1 1 ...
$ tz   : Factor w/ 5 levels "+0700","+0800",..: 1 1 1 1 1 2 2 2 2 2 ...
$ utime: Factor w/ 372 levels "2016-09-03 12:00:00",..: 1 1 1 2 3 4 4 4 5 6 ...
```

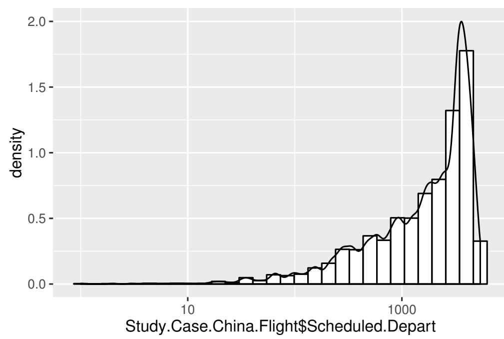
Figure 5.7 Sample of the AQI Dataset Structure - China Case Study

Data Summary

```
 lat              lon
 Min.   :14.04   Min.   : 75.24
 1st Qu.:29.43   1st Qu.:112.57
 Median :34.25   Median :118.18
 Mean   :33.05   Mean   :118.34
 3rd Qu.:37.06   3rd Qu.:127.04
 Max.   :50.43   Max.   :136.99

 city             idx
 Shaoguan University, Shaoguan (韶关韶关学院)                    :   24   Min.   : 216.0
 No.8 middle school, Shaoguan (韶关市八中)                       :   24   1st Qu.: 560.0
 Forest area, Shaoguan (韶关园林处)                              :   24   Median : 764.0
 Bìhú shānzhuāng, Shaoguan (韶关碧湖山庄)                         :   24   Mean   : 928.1
 City Environmental Monitoring Station, Xining (西宁市环境监测站):   24   3rd Qu.:1090.0
 Silu hospital, Xining (西宁四陆医院)                             :   24   Max.   :6069.0
 (Other)                                                          :36947
 stamp                  pol              x              aqi             tz
 Min.   :0.000e+00   pm10:  362   1815   :   12   Min.   :  1.00   +0700:  320
 1st Qu.:1.477e+09   pm25:36669   1817   :   12   1st Qu.: 16.00   +0800:26003
 Median :1.477e+09       :   60   1844   :   12   Median : 44.00   +0530:  446
 Mean   :1.474e+09                4693   :   12   Mean   : 55.38   +0600:   12
 3rd Qu.:1.477e+09                1854   :   12   3rd Qu.: 75.00   +0900:10310
 Max.   :1.477e+09                8665   :   12   Max.   :389.00
 (Other):37019
 utime
 2016-10-18 18:00:00: 2009
 2016-10-22 23:00:00: 2001
 2016-09-04 15:00:00: 1965
 2016-10-21 18:00:00: 1955
 2016-10-26 23:00:00: 1938
 2016-10-26 06:00:00: 1841
 (Other)            :25382
```

Figure 5.8 Sample of the Weather Dataset Summary - China Case Study



(a) Air Quality Index

Sample of Density Analysis

5.2 Australia Case Study

In Australia case study, we perform uni-variant analysis on the three datasets: flight dataset, weather dataset, and air quality dataset. On each dataset, we perform statistical analysis. First, we investigate the structure of the data. Second, we summarize the content of the data to see the distribution of values. Finally, we present a sample of some variables' density distribution.

5.2.1 Flight Data Uni-variate Analysis

The Structure of the Data

```
summary(Study.Case.Australia.Flight)

$ identification.row                     : Factor w/ 20204 levels "3382276551","3378612352",..: 1 2 3 4 5 6 7 8 9 10 ...
$ identification.number.default          : Factor w/ 514 levels "QF656","QF660",..: 1 1 1 1 1 1 1 1 1 1 ...
$ status.live                            : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ status.text                            : Factor w/ 1723 levels "Scheduled","Estimated dep 18:00",..: 1 1 1 1 1 1 2 3 4 5 ...
$ status.ambiguous                       : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ status.generic.status.text             : Factor w/ 7 levels "scheduled","estimated",..: 1 1 1 1 1 1 2 3 3 3 ...
$ status.generic.status.type             : Factor w/ 2 levels "departure","arrival": 1 1 1 1 1 1 1 2 2 2 ...
$ status.generic.status.color            : Factor w/ 4 levels "gray","green",..: 1 1 1 1 1 1 2 2 2 2 ...
$ airline.name                           : Factor w/ 4 levels "Qantas","Virgin Australia",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airline.code.iata                      : Factor w/ 4 levels "QF","VA","JQ",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airline.code.icao                      : Factor w/ 4 levels "QFA","VOZ","JST",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.name                    : Factor w/ 10 levels "Adelaide Airport",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.code.iata               : Factor w/ 10 levels "ADL","SYD","CBR",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.code.icao               : Factor w/ 10 levels "YPAD","YSSY",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.latitude       : Factor w/ 10 levels "-34.945","-33.946098",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.longitude      : Factor w/ 10 levels "138.530502","151.1772",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.country.name   : Factor w/ 1 level "Australia": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.country.code   : Factor w/ 1 level "AU": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.position.region.city    : Factor w/ 10 levels "Adelaide","Sydney",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.name           : Factor w/ 5 levels "Australia/Adelaide",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.offset         : Factor w/ 5 levels "34200","36000",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.abbr           : Factor w/ 5 levels "ACST","AEST",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.abbrName       : Factor w/ 5 levels "Australian Central Standard Time",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.timezone.isDst          : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.origin.visible                 : Factor w/ 1 level "TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.name               : Factor w/ 9 levels "Brisbane Airport",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.code.iata          : Factor w/ 9 levels "BNE","MEL","SYD",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.code.icao          : Factor w/ 9 levels "YBBN","YMML",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.latitude  : Factor w/ 9 levels "-27.3841","-37.673302",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.longitude : Factor w/ 9 levels "153.117493","144.843307",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.country.name: Factor w/ 1 level "Australia": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.country.code: Factor w/ 1 level "AU": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.position.region.city : Factor w/ 9 levels "Brisbane","Melbourne",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.name      : Factor w/ 5 levels "Australia/Brisbane",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.offset    : Factor w/ 5 levels "36000","34200",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.abbr      : Factor w/ 5 levels "AEST","ACST",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.abbrName  : Factor w/ 5 levels "Australian Eastern Standard Time",..: 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.timezone.isDst     : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ airport.destination.visible            : Factor w/ 1 level "TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ time.scheduled.departure               : Factor w/ 7378 levels "1473669000","1473582600",..: 1 2 3 4 5 6 7 8 9 10 ...
$ time.scheduled.arrival                 : Factor w/ 8303 levels "1473677400","1473591000",..: 1 2 3 4 5 6 7 8 9 10 ...
$ time.other.updated                     : Factor w/ 16796 levels "1473039656","1473039653",..: 1 1 1 2 2 3 4 5 6 7 ...
$ status.icon                            : Factor w/ 3 levels "green","red",..: NA NA NA NA NA NA 1 1 1 1 ...
$ status.generic.eventTime.utc           : Factor w/ 15216 levels "1473064200","1472985898",..: NA NA NA NA NA NA 1 2 3 4 ...
$ status.generic.eventTime.local         : Factor w/ 15253 levels "1473098400","1473021898",..: NA NA NA NA NA NA 1 2 3 4 ...
$ aircraft.model.code                    : Factor w/ 13 levels "B738","E190",..: NA NA NA NA NA NA 1 1 1 1 ...
$ aircraft.model.text                    : Factor w/ 21 levels "Boeing 737-838",..: NA NA NA NA NA NA 1 1 1 1 ...
$ aircraft.hex                           : Factor w/ 292 levels "7C6D98","7C6D8F",..: NA NA NA NA NA NA 1 2 3 4 ...
$ aircraft.registration                  : Factor w/ 292 levels "VH-VXM","VH-VXD",..: NA NA NA NA NA NA 1 2 3 4 ...
$ aircraft.serialNo                      : Factor w/ 288 levels "33483","29552",..: NA NA NA NA NA NA 1 2 3 4 ...
$ aircraft.owner                         : Factor w/ 18 levels "Qantas","Qantas (Retro Livery)",..: NA NA NA NA NA NA 1 1 1 1 ...
$ time.estimated.departure               : Factor w/ 2264 levels "1473064200","1473108000",..: NA NA NA NA NA NA 1 NA NA NA ...
```

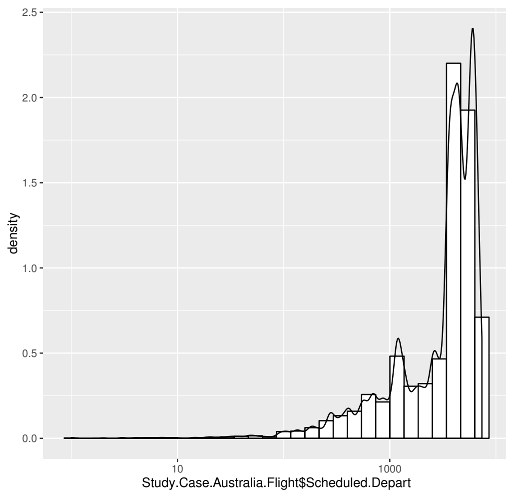
Figure 5.9 Sample of the flight dataset structure - Australia Case Study

Data Summary

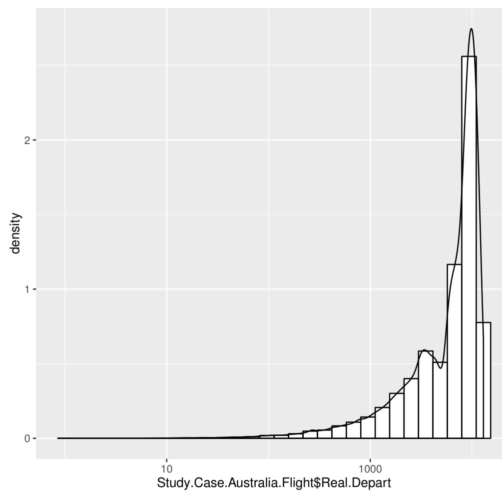
```
Orig.Airport   Orig.Airport.IATA   Orig.Airport.Lat
Melbourne Airport               :33698   MEL     :33698     -37.673302:33698
Sydney Kingsford Smith Airport:32562   SYD     :32562     -33.946098:32562
Brisbane Airport                :20270   BNE     :20270     -27.3841  :20270
Adelaide Airport                :15551   ADL     :15551     -34.945   :15551
Canberra International Airport: 9072   CBR     : 9072     -35.3069  : 9072
(Other)                         : 7112   (Other): 7112     (Other)   : 7112
NA's                            :    3   NA's   :    3     NA's      :    3
Orig.Airport.Lon Orig.Airort.Country Orig.Airport.City
144.843307:33698   Australia:118265    Melbourne:33698
151.1772  :32562   NA's     :      3    Sydney   :32562
153.117493:20270                        Brisbane :20270
138.530502:15551                        Adelaide :15551
149.195007: 9072                        Canberra : 9072
(Other)   : 7112                        (Other)  : 7112
NA's      :    3                        NA's     :    3
Dest.Airport   Dest.IATA.Code    Dest.Airport.Lat
Sydney Kingsford Smith Airport:30478   SYD     :30478   -33.946098:30478
Melbourne Airport               :29268   MEL     :29268   -37.673302:29268
Brisbane Airport                :22419   BNE     :22419   -27.3841  :22419
Adelaide Airport                :15958   ADL     :15958   -34.945   :15958
Canberra International Airport:10652   CBR     :10652   -35.3069  :10652
(Other)                         : 9232   (Other): 9232   (Other)   : 9232
NA's                            :  261   NA's   :  261   NA's      :  261
Dest.Airport.Lon Dest.Airport.Country Dest.Airport.City Scheduled.Depart
151.1772  :30478   Australia:118007     Sydney   :30478    Min.   :   1
144.843307:29268   NA's     :    261     Melbourne:29268    1st Qu.:1860
153.117493:22419                         Brisbane :22419    Median :4004
138.530502:15958                         Adelaide :15958    Mean   :3742
149.195007:10652                         Canberra :10652    3rd Qu.:5609
(Other)   : 9232                         (Other)  : 9232    Max.   :7378
NA's      :  261                         NA's     :  261    NA's   :291
Real.Depart     Scheduled.Arrival   Real.Arrival     Aircraft.Model.Code
Min.   :    1   Min.   :   1        Min.   :    1    B738    :35819
1st Qu.: 3518   1st Qu.:2121        1st Qu.: 2884    A320    : 7372
Median : 7572   Median :4473        Median : 6144    A332    : 5310
Mean   : 6824   Mean   :4214        Mean   : 5567    E190    : 4441
3rd Qu.: 9842   3rd Qu.:6254        3rd Qu.: 8039    A321    : 1400
Max.   :12948   Max.   :8303        Max.   :10606    (Other): 3957
NA's   :63868   NA's   :291         NA's   :65704    NA's    :59969
```

Figure 5.10 Sample of the weather dataset summary - Australia Case Study

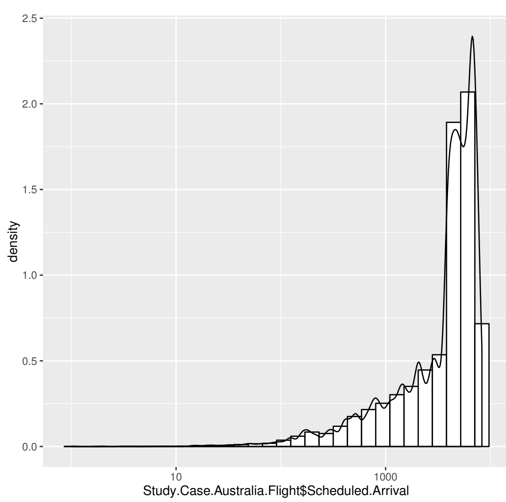
Sample of Density Analysis



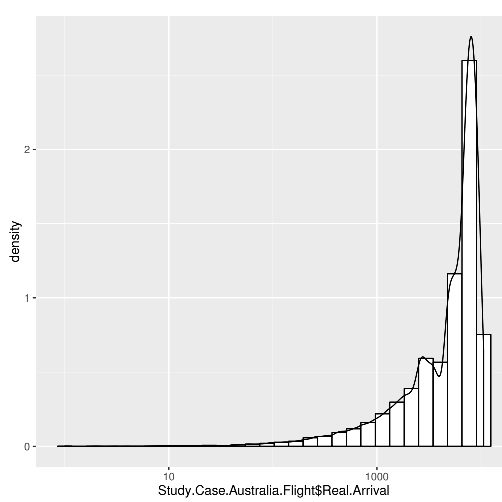
(a) Scheduled Departure Time for Flights



(b) Real Departure Time for Flights



(a) Scheduled Arrival Time for Flights



(b) Real Arrival Time for Flights

Figure 5.11: Sample of Continuous Variables Distribution of Flight Data - Australia Case Study

5.2.2 Weather Data Uni-variate Analysis

The Structure of the Data

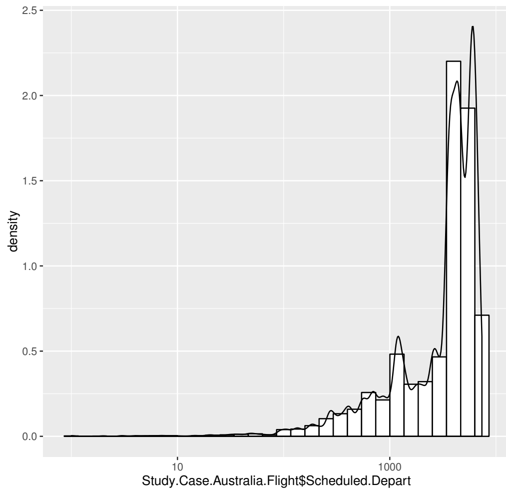
```
$ epoch          : num  1.47e+09 1.47e+09 1.47e+09 1.47e+09 1.47e+09 ...
$ ageh           : num  0 2 0 0 0 0 0 0 0 0 ...
$ agem           : num  37 32 20 23 5 23 21 21 0 1 ...
$ ages           : num  44 29 39 8 3 48 0 0 3 24 ...
$ type           : Factor w/ 3 levels "SYNOP","PWS",..: 1 1 1 1 2 1 1 1 2 2 ...
$ id             : Factor w/ 2446 levels "WMO94850","BUOYC6FS9",..: 1 2 3 4 5 6 7 8 9 10 ...
$ lat            : Factor w/ 2469 levels "-39.88010025",..: 1 2 3 4 5 6 7 8 9 10 ...
$ lon            : Factor w/ 2486 levels "143.88290405",..: 1 2 3 4 5 6 7 8 9 10 ...
$ adm1           : Factor w/ 1970 levels "King Island Airport",..: 1 2 3 4 5 6 7 8 9 10 ...
$ adm2           : Factor w/ 92 levels "","C6FS9","TAS",..: 1 2 1 1 3 1 1 1 4 4 ...
$ country        : Factor w/ 5 levels "Australia","",..: 1 2 1 1 3 4 1 1 4 4 ...
$ dateutc        : Factor w/ 9958 levels "2016-09-05 08:00:00",..: 1 2 1 1 3 1 1 1 4 5 ...
$ winddir        : Factor w/ 363 levels "270","280","275",..: 1 1 1 2 3 2 4 2 5 6 ...
$ windspeedmph   : Factor w/ 238 levels "15.0","20.7",..: 1 2 3 4 5 6 7 7 8 9 ...
$ humidity       : Factor w/ 103 levels "87","93","71",..: 1 2 3 4 5 6 1 7 8 9 ...
$ tempf          : Factor w/ 612 levels "54","55","53.4",..: 1 1 2 1 3 2 2 1 4 5 ...
$ rainin         : Factor w/ 48 levels "-999","0.00",..: 1 1 1 1 2 1 1 1 2 2 ...
$ baromin        : Factor w/ 407 levels "30.32","30.35",..: 1 2 1 1 3 4 4 2 5 6 ...
$ dewptf         : Factor w/ 690 levels "51","53","48",..: 1 2 3 4 5 6 7 7 8 9 ...
$ weather        : Factor w/ 32 levels "Clear","-999",..: 1 2 2 2 3 2 2 4 3 3 ...
$ icon           : Factor w/ 10 levels "clear","-999",..: 1 2 2 2 NA 2 2 3 NA NA ...
$ clouds         : Factor w/ 11 levels "unknown","","SCT",..: 1 1 1 1 2 1 1 1 2 2 ...
$ flightrule     : Factor w/ 5 levels "N/A","MVFR","VFR",..: 1 1 1 1 NA 1 1 2 NA NA ...
$ visibilitysm   : Factor w/ 35 levels "-999","5","6.2",..: 1 1 1 1 NA 1 1 2 NA NA ...
$ windgustmph    : Factor w/ 236 levels "-999","21.0",..: 1 1 1 1 2 1 1 1 3 4 ...
$ snowin         : Factor w/ 2 levels "-999","-9999.00": 1 1 1 1 NA 1 1 1 NA NA ...
$ name           : Factor w/ 473 levels "King Island Airport",..: 1 2 3 4 NA 5 6 7 NA NA ...
$ elev           : Factor w/ 790 levels "125","-999","318",..: 1 2 3 4 5 6 7 8 9 10 ...
$ windchillf     : Factor w/ 64 levels "-999","48","49",..: 1 1 1 1 1 1 1 1 1 1 ...
$ heatindexf     : Factor w/ 70 levels "-9999","53","56",..: 1 1 1 1 2 1 1 1 3 4 ...
$ updated        : Factor w/ 11171 levels "1473063333","1473056449",..: 1 2 3 4 5 6 3 3 7 8 ...
$ sstf           : Factor w/ 30 levels "56","55","67",..: NA 1 NA NA NA NA NA NA NA NA ...
$ neighborhood   : Factor w/ 1828 levels "Tanners Bay, Flinders Island",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ partner_id     : Factor w/ 1 level "": NA NA NA NA 1 NA NA NA 1 1 ...
$ dailyrainin    : Factor w/ 159 levels "0.04","0.00",..: NA NA NA NA 1 NA NA NA 2 2 ...
$ softwaretype   : Factor w/ 154 levels "weatherlink.com 1.10",..: NA NA NA NA 1 NA NA NA 2 1 ...
$ maxtemp        : Factor w/ 573 levels "59.2","61.7",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ maxtemp_time   : Factor w/ 1354 levels "1:32PM","3:05PM",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ mintemp        : Factor w/ 446 levels "53.2","51.4",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ mintemp_time   : Factor w/ 1373 levels "4:31AM","7:18AM",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ maxdewpoint    : Factor w/ 604 levels "54.0","55.0",..: NA NA NA NA 1 NA NA NA 1 2 ...
$ mindewpoint    : Factor w/ 672 levels "51.0","48.6",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ maxpressure    : Factor w/ 405 levels "30.31","30.42",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ minpressure    : Factor w/ 419 levels "30.21","30.33",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ maxwindspeed   : Factor w/ 105 levels "24","12","5",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ maxwindgust    : Factor w/ 110 levels "31","17","10",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ maxrain        : Factor w/ 149 levels "0.02","0.00",..: NA NA NA NA 1 NA NA NA 2 2 ...
$ maxheatindex   : Factor w/ 95 levels "59","62","61",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ minwindchill   : Factor w/ 96 levels "-999","45","47",..: NA NA NA NA 1 NA NA NA 1 1 ...
$ rtfreq         : Factor w/ 22 levels "2.5","5.0","600.0",..: NA NA NA NA 1 NA NA NA 2 1 ...
$ indoortempf    : Factor w/ 721 levels "55.4","61.0",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ indoorhumidity : Factor w/ 89 levels "64","78","54",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ RawP           : Factor w/ 2292 levels "30.31","30.37",..: NA NA NA NA 1 NA NA NA 2 3 ...
$ tzname         : Factor w/ 9 levels "Australia/Sydney",..: NA NA NA NA 1 NA NA NA 2 2 ...
```

Figure 5.12 Sample of the Weather Dataset Structure - Australia Case Study

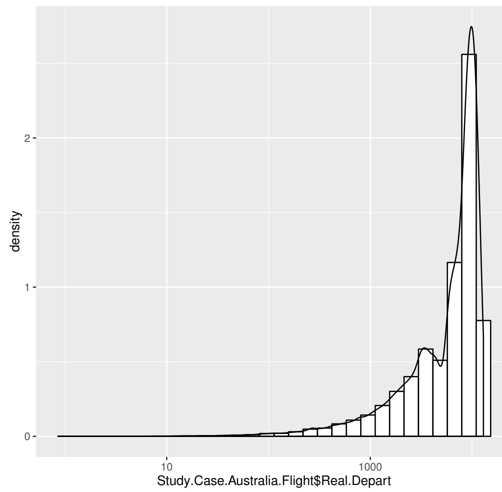
Data Summary

```
TimeStamp                     Lat                    Lon
Min.    :1.473e+09   -34.94689941:   36    146.00000000:   35
1st Qu.:1.473e+09   -34.25080109:   34    145.32501221:   23
Median :1.477e+09   -37.47268295:   23    143.88290405:   18
Mean   :1.476e+09   -36.67219925:   21    146.42439270:   18
3rd Qu.:1.477e+09   -39.88010025:   18    146.98410034:   18
Max.   :1.478e+09   -39.12969971:   18    147.86463928:   18
(Other)     :22317   (Other)     :22337
Country                    DateUTC             Wind.Dir      Wind.Speed.MPH
Australia : 2399   2016-09-05 23:00:00:  193   Min.   :  1.0   Min.    :  1.00
:  202   2016-10-25 23:00:00:  180   1st Qu.: 39.0   1st Qu.:  9.00
AUSTRALIA : 2562   2016-09-09 07:00:00:  179   Median :102.0   Median : 23.00
AU         :17295   2016-10-22 17:00:00:  179   Mean   :129.4   Mean   : 32.51
AUSTRALIEN:     2   2016-10-26 17:00:00:  176   3rd Qu.:204.0   3rd Qu.: 40.00
NA's      :     7   (Other)           :21553   Max.   :363.0   Max.   :238.00
NA's                :     7   NA's   :7        NA's   :7
Humidity          Tempr          Raining          Baromin
Min.   :  1.00   Min.   :  1.0   Min.   : 1.000   Min.   :  1.00
1st Qu.: 16.00   1st Qu.: 55.0   1st Qu.: 2.000   1st Qu.: 27.00
Median : 31.00   Median :128.0   Median : 2.000   Median : 64.00
Mean   : 36.06   Mean   :161.3   Mean   : 2.364   Mean   : 70.79
3rd Qu.: 54.00   3rd Qu.:240.0   3rd Qu.: 2.000   3rd Qu.: 94.00
Max.   :103.00   Max.   :612.0   Max.   :48.000   Max.   :407.00
NA's   :7        NA's   :7       NA's   :7        NA's   :7
Dew.Point        Visibility      Wind.Gust.MPH     Elevation      Heat.Index
Min.   :  1.0   Min.   : 1.000   Min.   :  1.0   Min.   :  1   Min.   : 1.00
1st Qu.: 49.0   1st Qu.: 1.000   1st Qu.:  3.0   1st Qu.: 67   1st Qu.: 2.00
Median :121.0   Median : 1.000   Median :  9.0   Median :203   Median :11.00
Mean   :159.2   Mean   : 2.787   Mean   : 18.5   Mean   :237   Mean   :14.52
3rd Qu.:228.0   3rd Qu.: 3.000   3rd Qu.: 20.0   3rd Qu.:367   3rd Qu.:22.00
Max.   :690.0   Max.   :35.000   Max.   :236.0   Max.   :790   Max.   :70.00
NA's   :7       NA's   :17166    NA's   :7       NA's   :7     NA's   :7
Daily.Raining     Max.Presure
Min.   :  1.00   Min.   :  1.00
1st Qu.:  2.00   1st Qu.: 26.00
Median :  2.00   Median : 56.00
Mean   :  7.85   Mean   : 73.44
3rd Qu.:  3.00   3rd Qu.:106.00
Max.   :159.00   Max.   :405.00
NA's   :4667     NA's   :4667
```

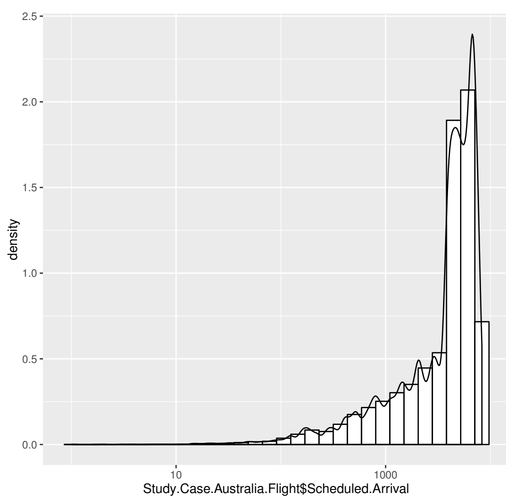
Figure 5.13 Sample of the Weather Dataset Summary - Aust Case Study



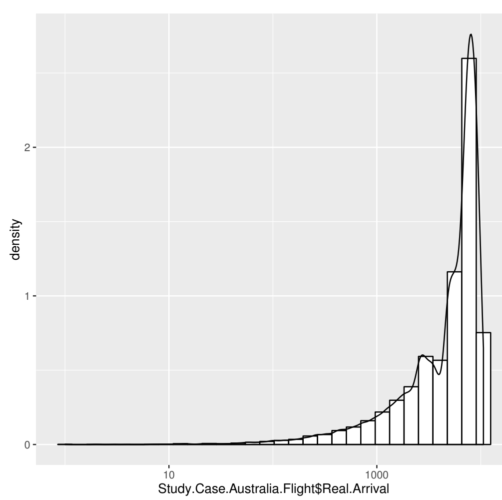
(a) Wind Direction



(b) Temperature



(a) Wind Speed



(b) Humidity

Figure 5.14: Sample of Continuous Variables Distribution of Weather Data - Australia Case Study

Sample of Density Analysis

5.3 Data Cleaning

5.3.1 The Cleaning Process

Data cleaning is an essential step in data mining and knowledge discovery. Therefore, after we gathered the data from all targeted IoT data sources and analyze all variables, we should start cleaning the data. First of all, based on the uni-variate analysis stated previously, we can realize that our data suffers from two main things missing values and outliers. Another point we should highlight, there are some variables that we do not need, so we summarize the data by selecting the most important variables based on our interest. In addition, since the columns' names are confusing, we change their names and the order of the columns to make them meaningful as you can see in the Figure 8.1. We do the same process for the weather data and the air quality data.

As you can see from Figure 8.1, the flight number is duplicated. Each record of that flight is in a different date, so we will use them in our study.

Flight ID	Status with Time	Status	Status Type	Airline	Airline IATA Code	Origin Airport	Origin Airport IATA	Orig Airport Latitude	Orig Airport Longitude
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503
CA1855	Scheduled	scheduled	departure	Air China	CA	Beijing Capital International Airport	PEK	40.080109	116.584503

Figure 5.15: Flight data after cleaning and changing the columns names

5.3.2 Dealing With The Missing Values and Outliers

After conducting the comperhensive uni-variant analysis, we realized that our datasets are suffering from the missing values and outliers. When we investigated the datasets, we found that the missing values are at random. For that reason, we set up 5% as a threshold for the missing values. Any variable fails to achieve this threshold will be ignored.

To resolve the missing values issue, we have two strategies. First strategy is to ignore the records that contain the NA values. Second is to is to impute the missing values by a particular means. Based on best practices, selecting the strategy depends on the situation that we face. We decided to use the imputation strategy. We could impute the missing values using median, mean or mode, but that is not advisable. Although doing this will keep the median, mean, or the mode unchanged, it would decrease the variance, and that is not desirable.

R language provides a good package called MICE. This package offers several functions to deal with this issue. We can get a better understanding of our data using md.pattern()function. This function will enable us to know the pattern of the missing data. Another function is mice(). This function takes care of the imputing process.

Also, our data contains outliers. We are able to capture the outliers values in our dataset by using boxplots for each variable. Figure 2 and 3 show some outliers in the variables. So far, we have not treated the outlier issue because we want to investigate the reason. Again, this case also depends on the situation. Sometimes outliers is the interesting part of the problem since it gives a hint of discovering something such as fraud cases. However, in our case, we are dealing flight and weather data. That means, it is impossible to find for example the Temperature variable holds a value equal to (- 999). As a result, whenever we detect an outlier value, we treated as we did with the missing value.

5.4 Data Integration

5.4.1 The Integration Process

After collecting the data from each source separately by running the crawler, we need to integrate all data from the three data sources in one place to prepare it for the next steps. Since we study the flight delay, we consider the flight data is main container that we integrate the other data sources with. The idea of the integration is to bring data from the weather and the aqicg containers based on some criteria. The integration is complicated because there is no common columns among the data sources except the latitude and longitude columns as names. We know that airports and sensors possess unique latitude and longitude. Therefore, our idea is to utilize these two columns for the integration. We use the haversine formula as shown in [oo] to calculate the distance among the airports and the sensors. After that we determine the proximity between an airport and the other sensors. Then we bring the records from these data sources that match these criteria and attach them with the flight data based on the date and the proximity of the stations. We implement an integration algorithm as shown below.

Performing exploration and analysis of the data to predict the flight delay requires us to combine all data sources in one container. Therefore, integratng data from different data sources requires sharing a common column among the datasets. Thus, the integration process we do in this project was tricky because there are no common data among our datasets.

However, after a deep investigation we made on the three data sources, we find that all data sources share the longitude and latitude column as a names, and the values in these columns are different. The longitude and the latitude represent the location of a particular object such as an airport and other sensors for weather and air quality index. That means the integration process should be done in a particular way. We can utilize these two columns in order to complete the integration process. Mathematically, it is possible to combine the flight data, weather data, and the air quality data based on the location of objects. As a result, we successfully integrate all data sources together by using haversine formula. We implement the integration algorithm as shown below.

Figure 5.16 shows our basic idea to integrate the data sources, and after that is the algorithm that we implement for the integration.

1. Get the location of the airport from the flight data.
2. For each weather record, check the distance between the airport and the weather station.
3. If the distance is 5 km or less then take the weather information from that station

- We set the distance to be 5 km because we tried to use smaller distances, but we did not find enough information. So 5 km seems a realistic distance to get the data from the nearest station.
- The calculation of the distance is based on the Haversine formula.

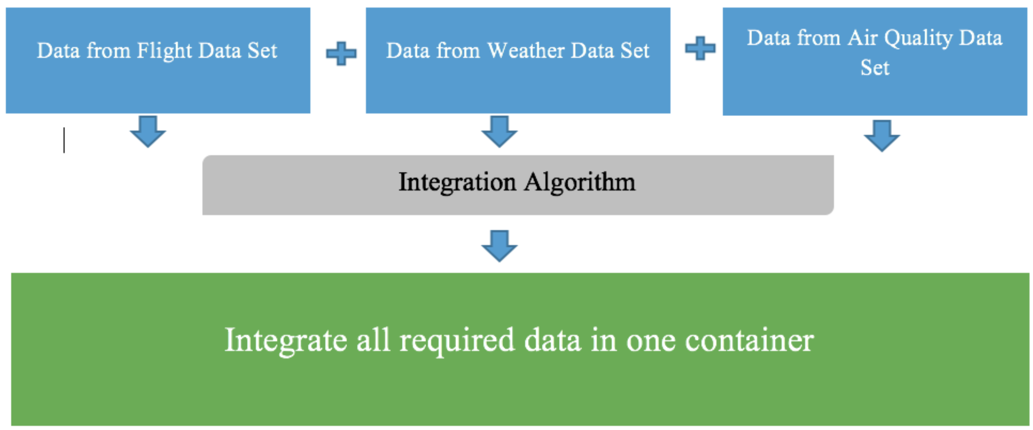


Figure 5.16: Integration Structure

Input: Flight Data and other sensors
Output: All Matched Data in one Container
for *i = 0 to length(sensor data records)* **do**
    Long1 = longitude of the sensor
    Lat1 = latitude of the sensor
    Point 1 = (Long1, Lat1)
    for *j = 0 to length(sensor data records)* **do**
        Long2 = longitude of the airport
        Lat2 = latitude of the airport
        Point2 = (Long2, Lat2)
        Distance = calculate the distance (Point1, Point2) using haversin formula
        if *Distance <5 km* **then**
            Take the matched data from the sensor data and attached it to the matched record in flight data
        end
    end
end

Algorithm 1: Integration algorithm

5.4.2 Feature Engineering

Furthermore, we create some extra columns by transforming the content of some columns and by doing some calculation such as the delay at departure and the delay at arrival in minutes.

The purpose of this study is study the flight delay issue in particular the delay at departure. Our data does not have the delay at departure in order to study it. However, using feature engineering technique, we can create several variables from the data we have. Thus, the delay at departure can be created by subtracting the real departure from the scheduled departure. In addition, we can construct six additional variables (data, day of the week, time, day of the month, month ,year) from the Timestamp variable.

Chapter 6

Data Exploration and Visualization: Bi-variate Analysis

6.1 The Purpose of Bi-variate Analysis

After conducting the data integration successfully, our data becomes ready for performing comprehensive bi-variate analysis. Our primary aim is to investigate the correlation among variables in the three data sources.

6.2 Variable Correlation

After having a deep insight of the data, we move to analyze all features of the data sources we have. We want to see the correlation among all variables from all data sets. We need to know how they are correlated to the delay at departure DAD because that will enable us to identify the potential factors of our predictive model. Here it is some observation as Figures 10.1, 10.2, and 10.3 show: what we are interested in is the correlation between the delay at departure (DAD) and the remaining variables. We can see there is a very strong correlation between DAD and the delay at arrival. Also, there is good correlation between the DAD and the weather elements (Temperature, Heat index, Dew, visibility, and elevation). When we look at the correlation among the weather data, we can observe that some of them have almost perfect correlation. So, we will continue to analyze the data more with having various study cases where the weather plays a significant role. We will also see the correlation again when we add more extra data sets.

6.2.1 Variable Correlation in Australia Case Study

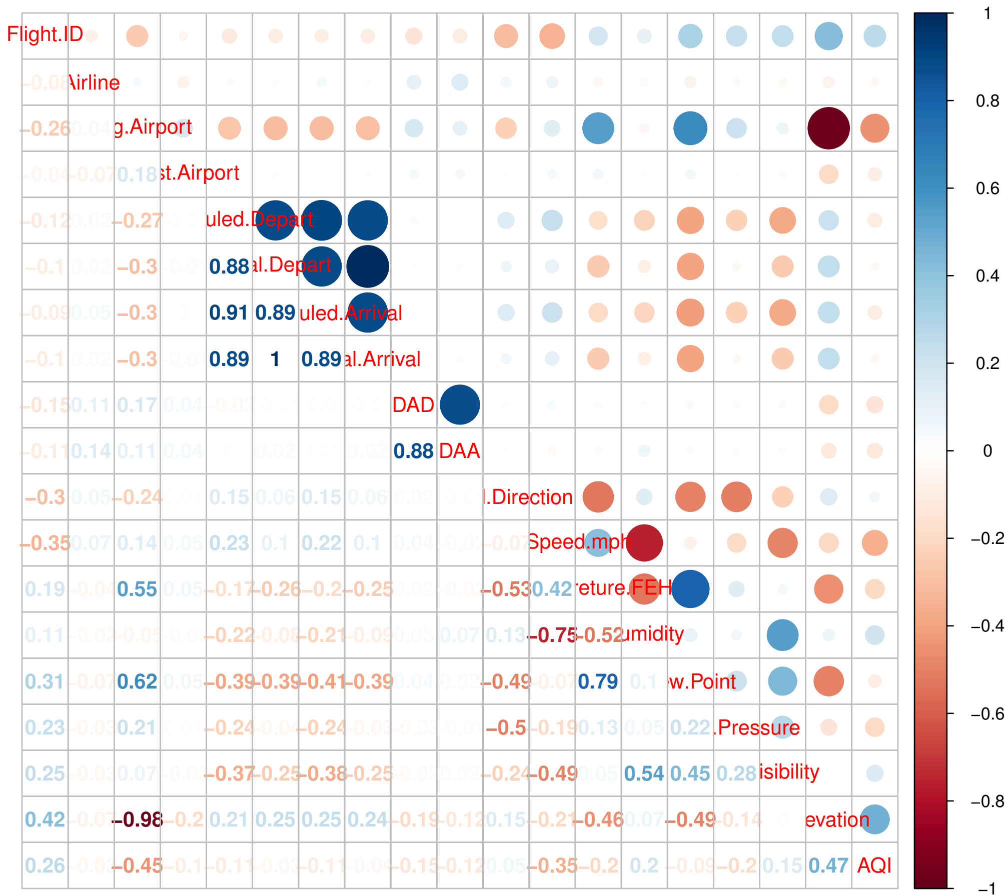


Figure 6.1: Correlation Analysis - Australia Case Study

6.2.2 Variable Correlation in China Case Study

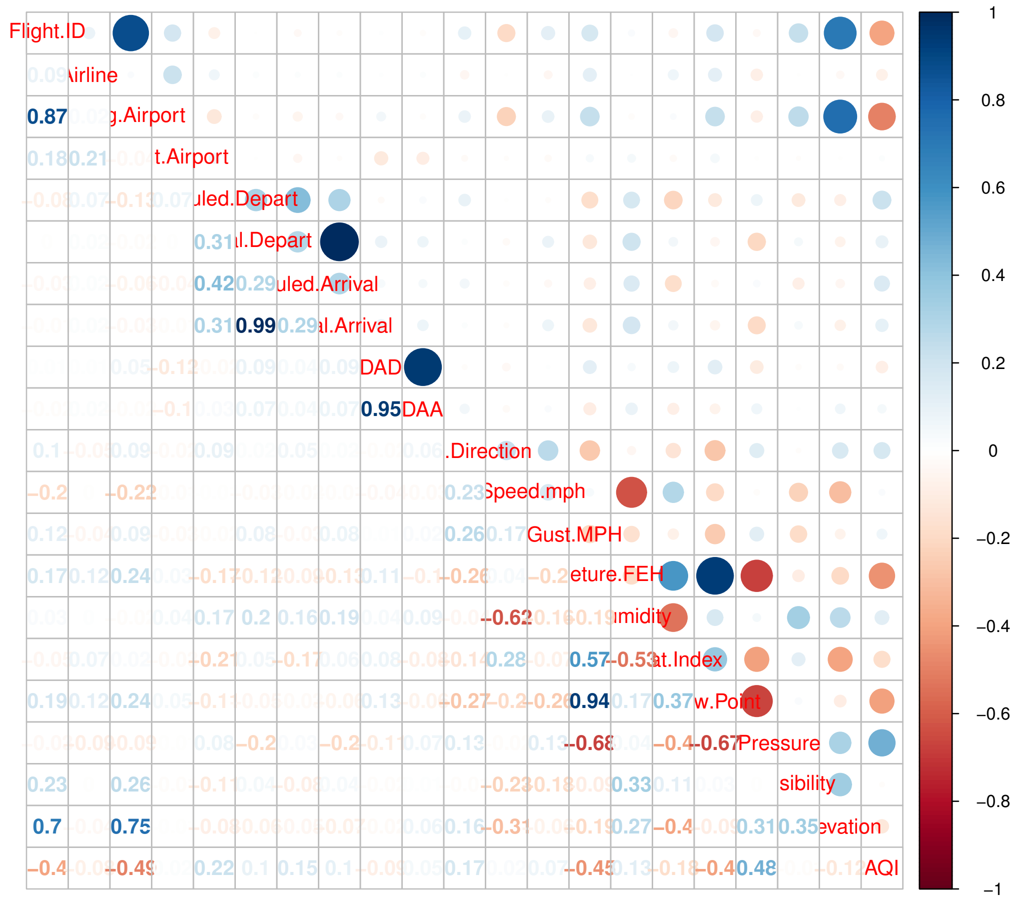


Figure 6.2: Correlation Analysis - China Case Study

6.2.3 Variable Correlation in Europe Case Study

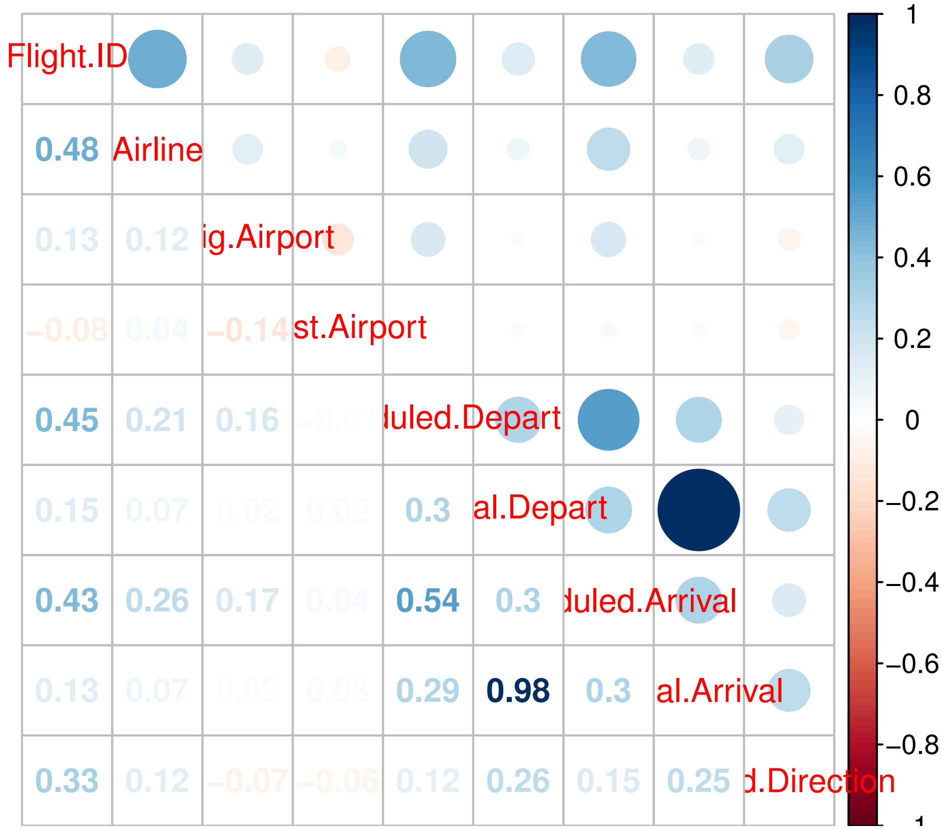


Figure 6.3: Correlation Analysis - Europe Case Study

6.3 Airlines and Airport Performance

6.3.1 Airlines Performance - China Case Study

We analyze the collected airline performance data in China case study by exploring them using one of the statistical methods. We use boxplot in order to find out the overall performance of each individual airline. We get interesting plot that gives us an indication about the airline impact on the flight delay problem. We believe that airline factor plays a significant role on the flight delay. As we can see from the Figure 10.4, some airlines do not operate the majority of their flight on-time. For example, in figure 10.4 the majority of the Xiamen Airlines (MF) flights are delayed likewise the Tibet Airlines (TV).

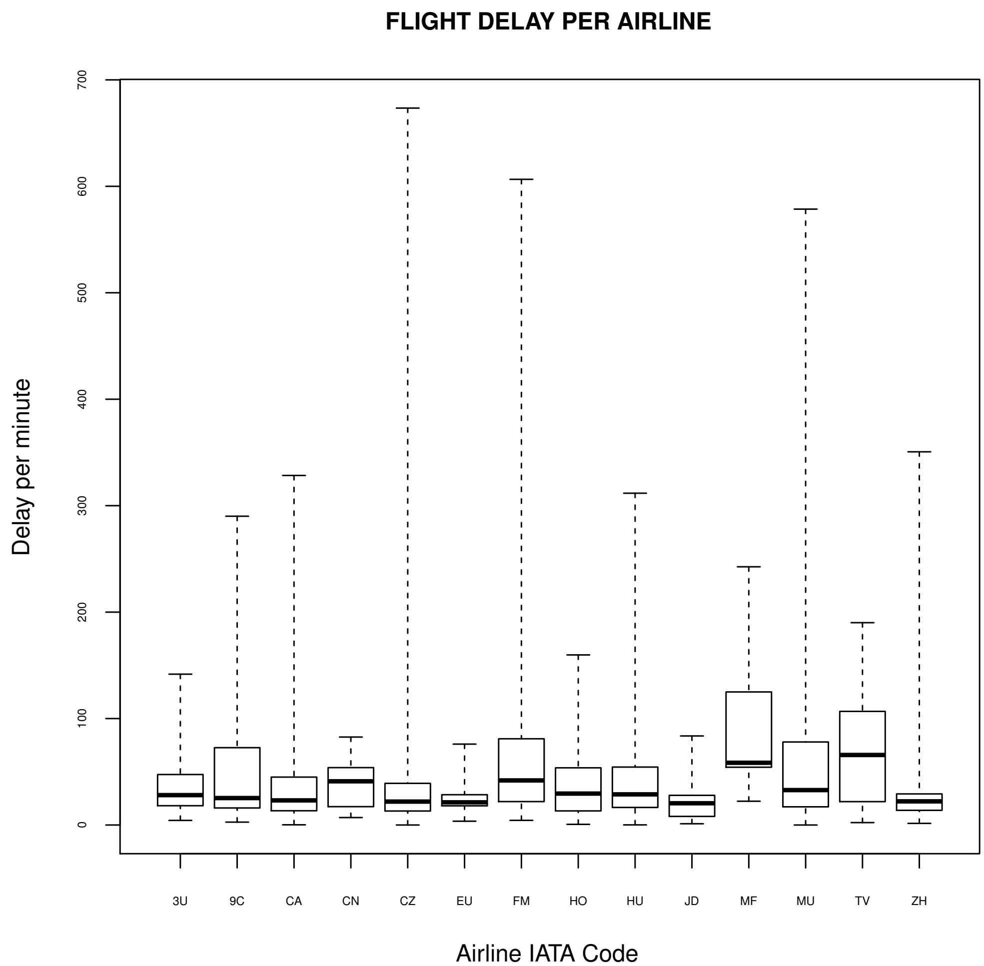


Figure 6.4: Delay at departure performance for airlines - China Case Study

6.3.2 Airports Performance - China Case Study

When looking at the individual airports, we also use boxplot to figure out the performance of each one of them. At this stage, we do not consider the capacity of the airport and how busy it is. We just want to see how much delay each airport have. Interestingly, we find that Shanghai airport (SHA) does not perform well since the majority of flights are delayed.

We know that Beijing airport (PEK) is as big as Shanghai (SHA), but the performance of (PEK) airport seems normal. As a result, this indicates that the airport factor should be taken in account in the future. Figure 10.5

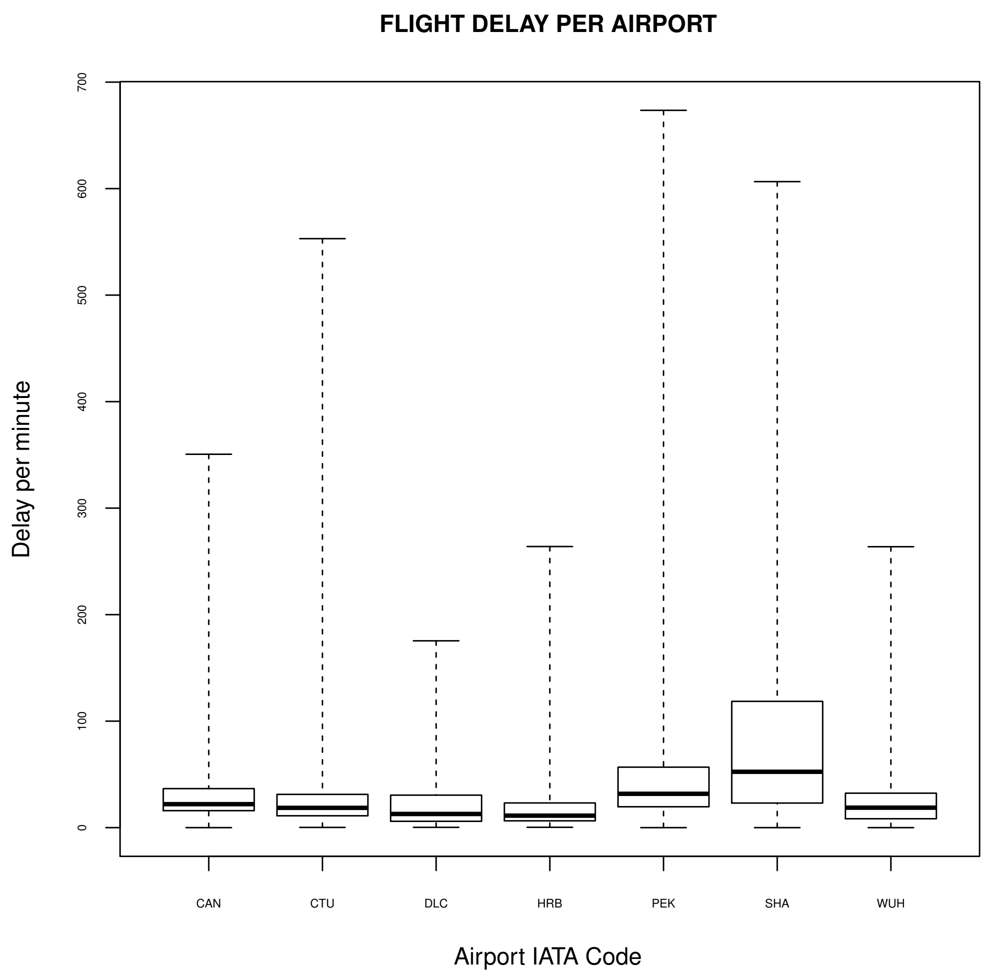


Figure 6.5: Delay at departure performance for airports - China Case Study

6.3.3 Airlines Performance - Australia Case Study

We analyze the collected airline performance data in Australia case study by exploring them using one of the statistical methods. We use boxplot in order to find out the overall performance of each individual airline. We get interesting plot that gives us an indication about the airline impact on the flight delay problem. We believe that airline factor plays a significant role on the flight delay in Australia case study. As we can see from the Figure 10.6, the majority of the Jetstar Airlines (JQ) flights are delayed.

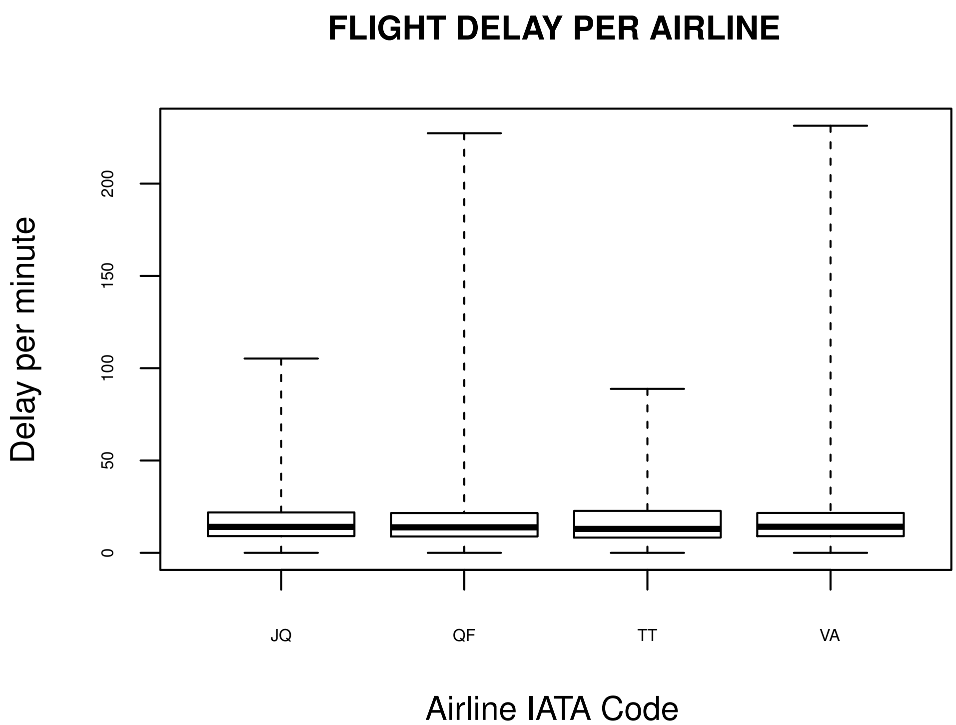


Figure 6.6: Delay at departure performance for airlines - Australia Case Study

6.3.4 Airports Performance - Australia Case Study

When looking at the individual airports, we also use boxplot to figure out the performance of each one of them. At this stage, we do not consider the capacity of the airport and how busy it is. We just want to see how much delay each airport have. Interestingly, we find that Sydney airport (SYD) does not perform well since the majority of flights are delayed. We know that Sydney airport (SYD) is as big as Brisbane (BNE), but the performance of (BNE) airport seems normal. As a result, this indicates that the airport factor should be taken in account in the future. Also, the majority of flights at Melbourne (MEL) airport are delayed. Figure 10.7

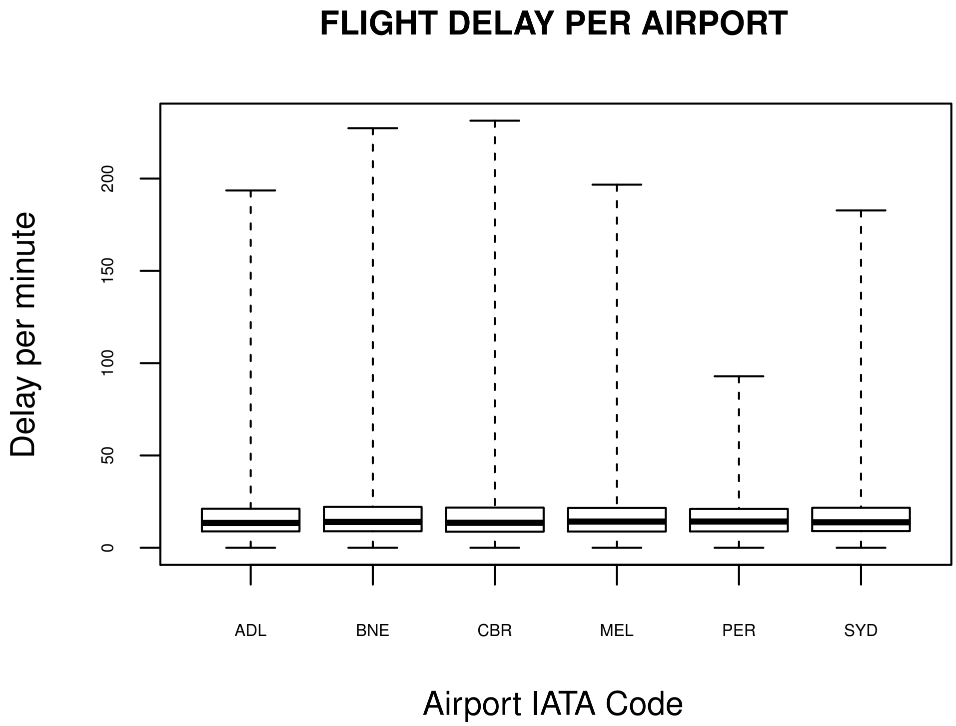


Figure 6.7: Delay at departure performance for airports - Australia Case Study

6.3.5 Airlines Performance - Europe Case Study

We analyze the collected airline performance data by exploring them using one of the statistical methods. We use boxplot in order to find out the overall performance of each individual airline. We get interesting plot that gives us an indication about the airline impact on the flight delay problem. We believe that airline factor plays a significant role on the flight delay. As we can see from the Figure 10.8, some airlines do not operate the majority of their flight on-time. For example, in figure 10.8 the majority of the Vueling airlines (VY) flights are delayed likewise the Ethiopian Airlines (ET).

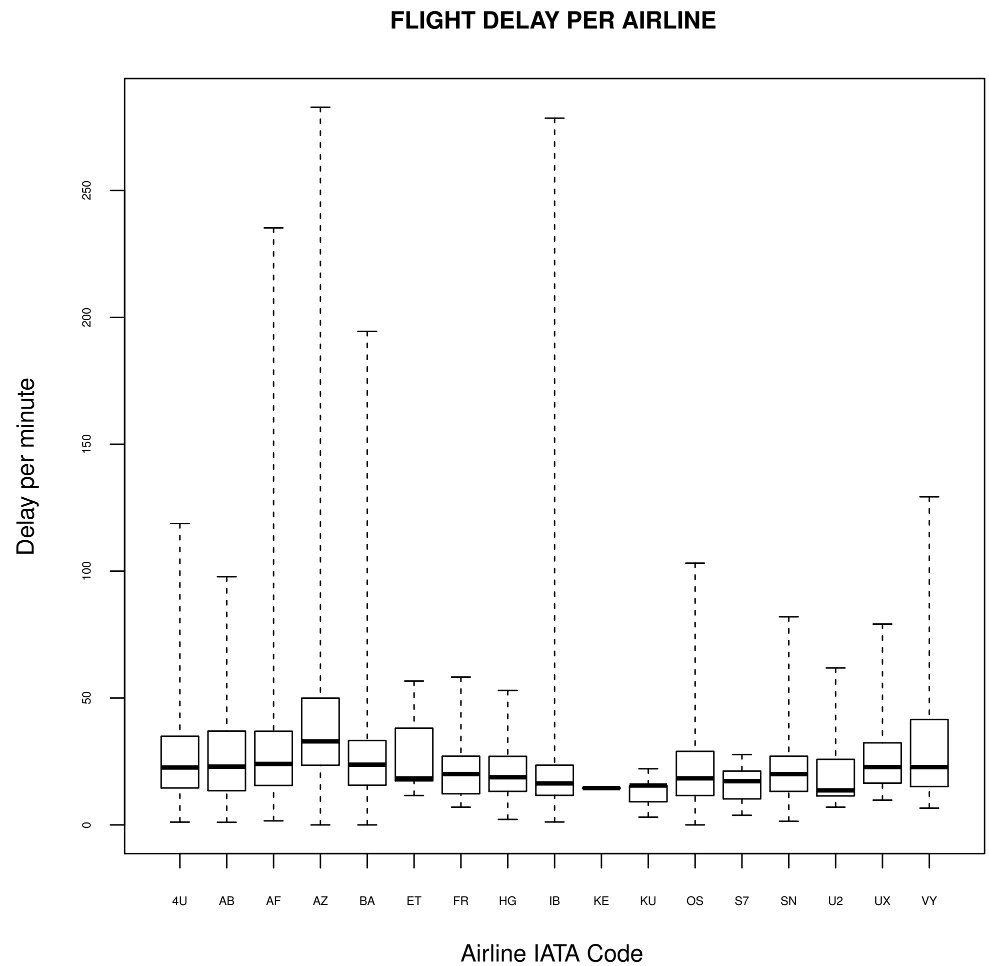


Figure 6.8: Delay at departure performance for airlines - Europe Case Study

6.3.6 Airports Performance - Europe Case Study

When looking at the individual airports, we also use boxplot to figure out the performance of each one of them. At this stage, we do not consider the capacity of the airport and how busy it is. We just want to see how much delay each airport have. Interestingly, we find that Rome (FCO) does not perform well since the majority of flights are delayed. We know that London airport (LHR) is larger than Rome (FCO), but the performance of

(LHR) airport seems normal. As a result, this indicates that the airport factor should be taken in account in the future. Figure 10.9

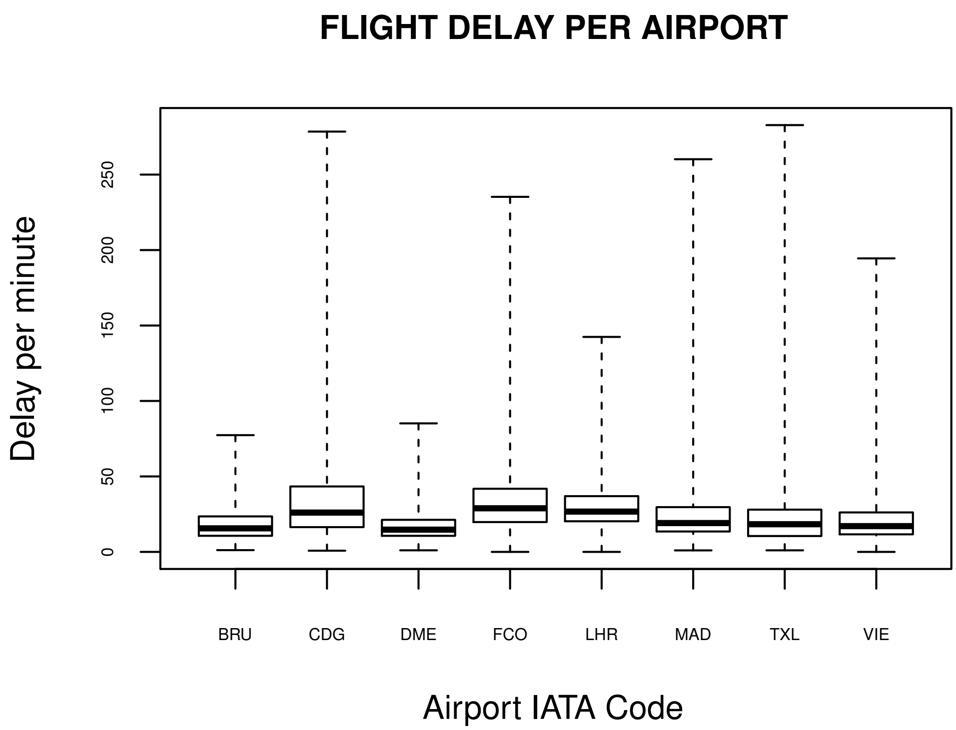


Figure 6.9: Delay at departure performance for airports - Europe Case Study

6.4 Heat Maps

6.4.1 Heat Map For China Case Study

We create a heat map in order to visualize the delay size of each airport. As the Figure 10.10 shows, most of the flight delays happen in large Shanghai, Beijing, and Guangzhou. This is due to the large size of these cities and the large number of flights in their airports. As a result, this is a good indication that the airports factor plays a significant role on the flight delay problem.

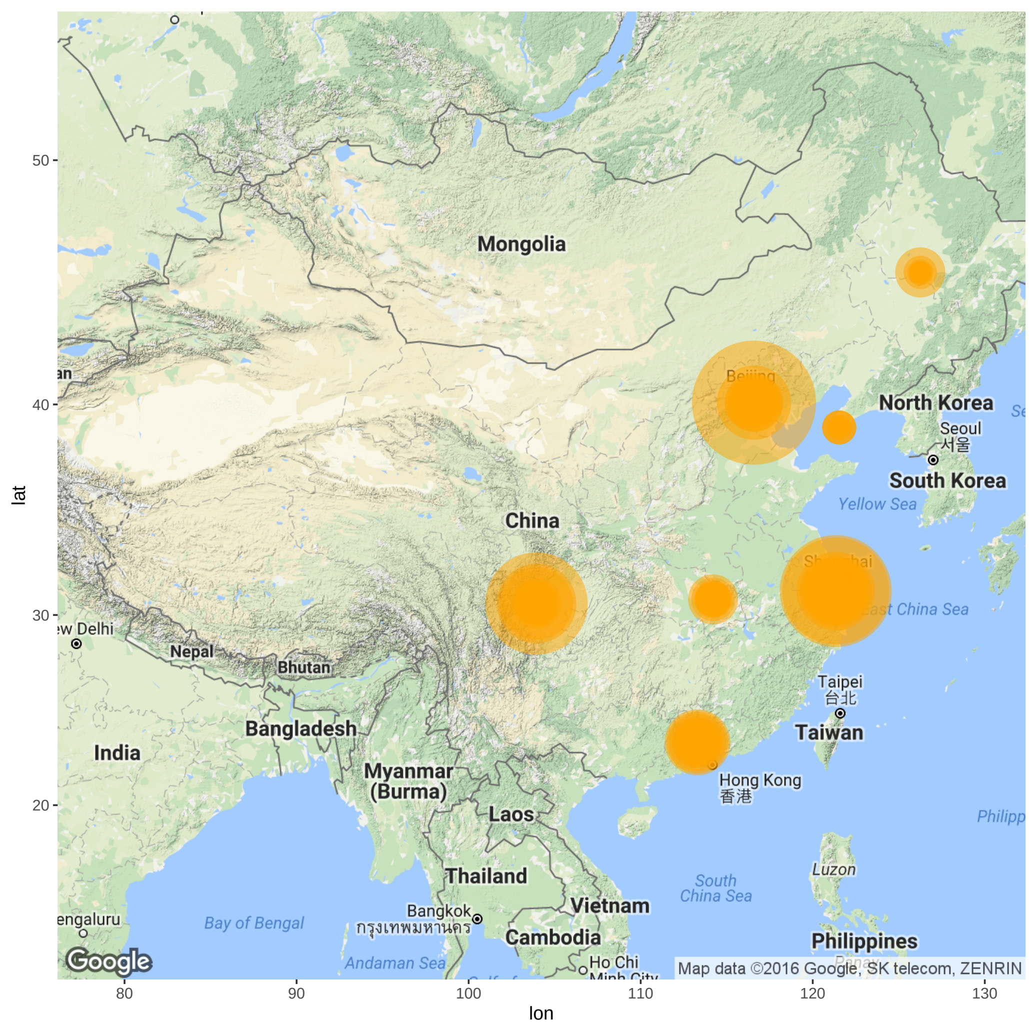


Figure 6.10: HeatMap to visualize the size of delay at each airport - China Case Study

6.4.2 Heat Map For Australia Case Study

We create a heat map in order to visualize the delay size of each airport. As the Figure 10.11 shows, most of the flight delays happen in Sydney, Perth, and Melbourne. This is due to the large size of these cities and the large number of flights in their airports. As a result, this is a good indication that the airports factor plays a significant role on the flight delay problem.

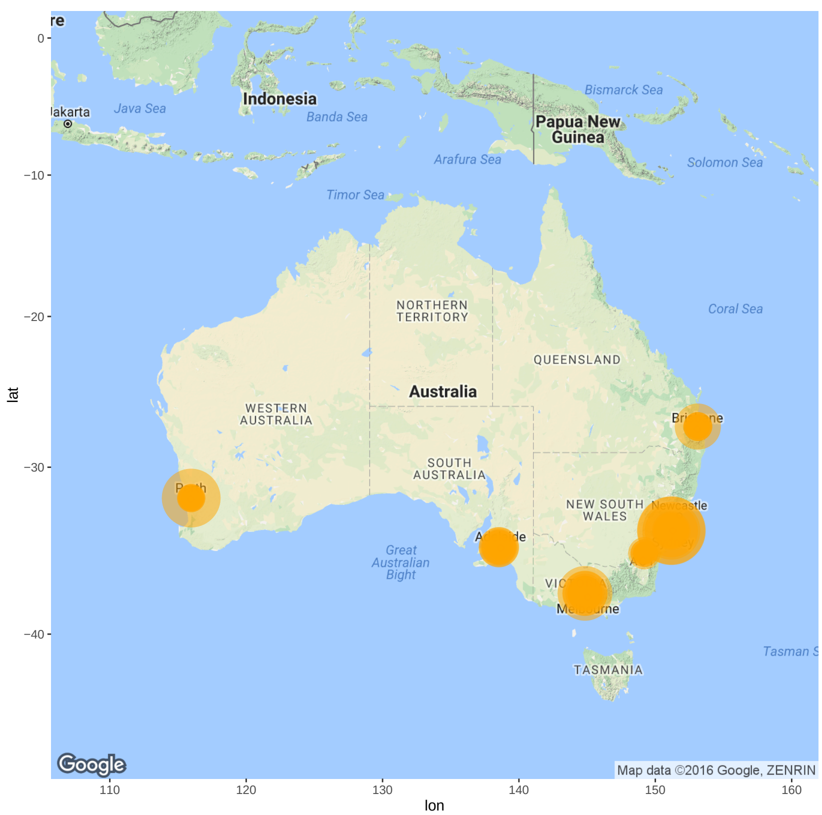


Figure 6.11: HeatMap to visualize the size of delay at each airport - Australia Case Study

6.4.3 Heat Map For Europe Case Study

We create a heat map in order to visualize the delay size of each airport. As the Figure 10.12 shows, most of the flight delays happen in large Paris, Madrid, and Berlin. This is due to the large size of these cities and the large number of flights in their airports. As a result, this is a good indication that the airports factor plays a significant role on the flight delay problem.

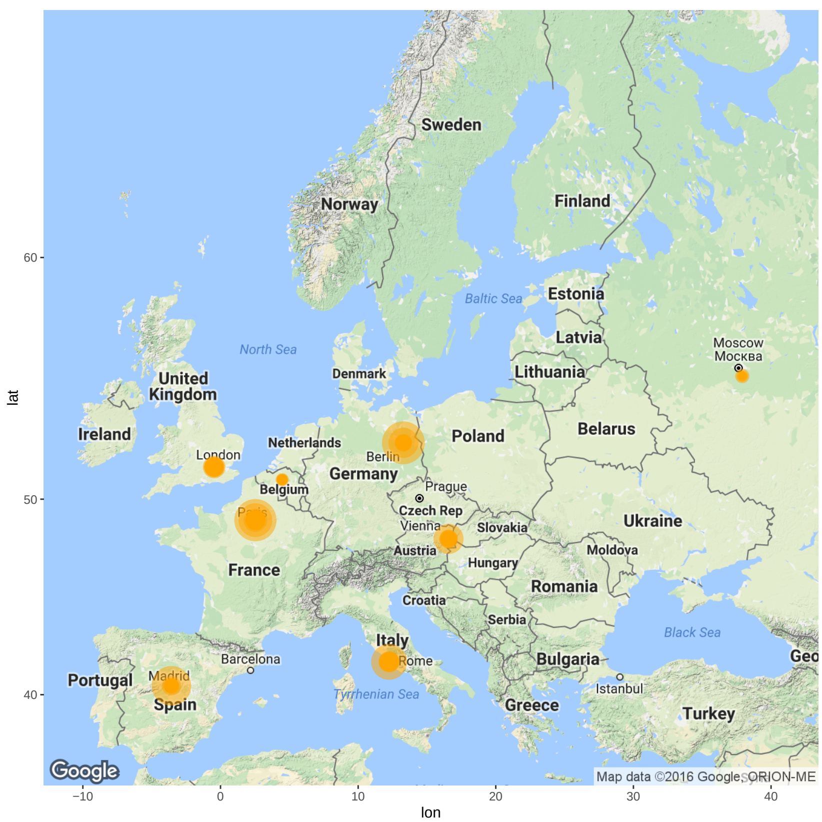


Figure 6.12: HeatMap to visualize the size of delay at each airport - Europe Case Study

Chapter 7

Modeling

7.1 Predictive Model

The purpose of studying the correlation among the features in the IoT data sources in chapter 10 is to determine their impact on the flight delays. When we identify how each contributes to this problem, we will be able to create a predictive model using some machine learning methods as we will see in the subsequent sections. The predictive models that we will build have to classify and predict the flight delays correctly. We want when we pass a given flight ID along with the time of the flight, our model should classify the flight whether delayed or on-time. In addition, our model should determine the delay time for each flight because that will increase the awareness of the user if he/she wants to accept that delay or not.

Therefore, after we have investigated the correlation among variables in our cases studies, we can proceed to create our models. However, we need to set up a unified procedure for building the models. The following section shows the procedure of creating the machine learning models.

7.1.1 The Proposed Model

In this section, we explain the prediction models that can classify the flights and predict the delay time of a flight. In order to build the prediction model, we need to identify the variables that are statically significant to delay at departure (DAD) to conduct the analysis and create the classification and prediction models. For building the model, we perform the following procedure:

1. The first step is to initially create a full model that contains all the variables from the three aforementioned data sources;
2. The second step is to conduct a variable selection process using the **Akaike Information Criterion** (AIC) approach. So, we remove unnecessary independent variables where its p-value larger than 0.05.
3. The third step is to eliminate the variables with correlations through the **multicollinearity** test and update our prediction model;
4. The Final step is to test our final prediction model by conducting the residual analysis on the normality and **homoscedasticity** of the prediction model.

7.2 Flight Delay Classification and Prediction Model for China

We develop two types of models: multiple logistic regression and multiple linear regression. The first is to classify the flights whether they are on-time or delayed. The second is to predict the delay time for each flight. In the following sections, we discuss the variables selection step and the models creations for China Case Study.

7.2.1 Variable Selection

Before we create our models, we need to identify the variables that are correlated to the delay at departure DAD. We use AIC method for this purpose. We present the statistically significant variables using correlation plot. This plot has X sign on the variables that are not statistically significant to the DAD output. That means the variables that are not crosses with X sign will be used in the our models in China case study.

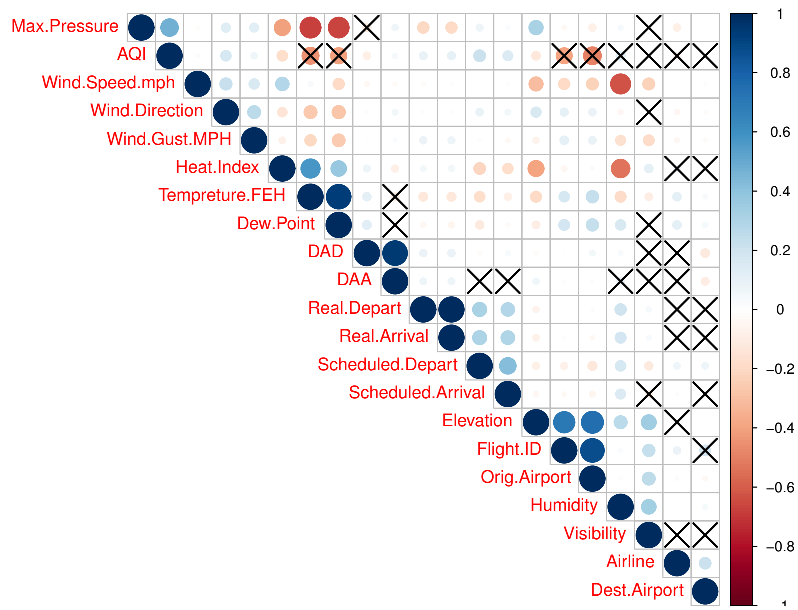


Figure 7.1: Statistically Significant Variables to the Delay At Departure DAD - China Case Study

7.2.2 Multiple Logistic Regression

After conducting the variable selection, we create multiple logistic regression based in order to be able to classify the flights. The model is able to achieves 91.01 % accuracy. We can see the performance of the model using Area Under Cover (AUC) plot. Our model is able to classify the flights correctly. Below is the AUC chart for the multiple logistic regression model.

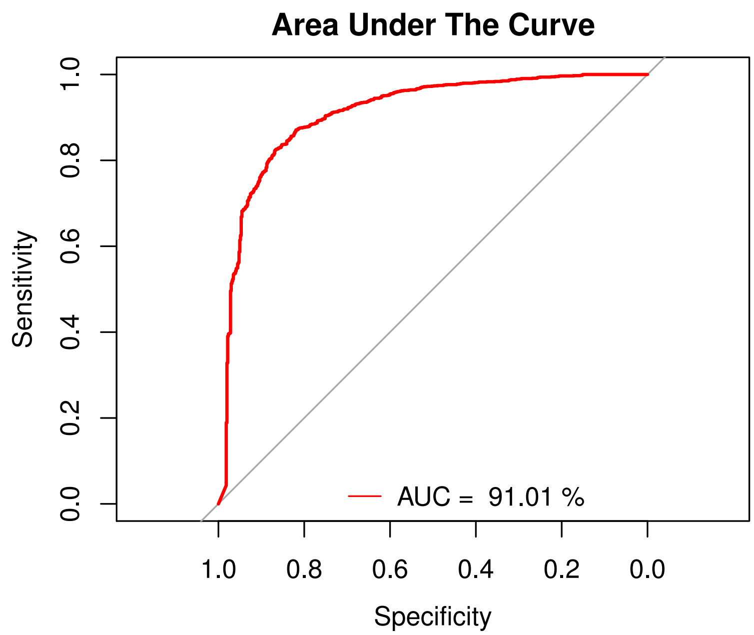


Figure 7.2: Area Under Cover AUC - China Case Study

7.2.3 Multiple Linear Regression

After we are able to classify the flights correctly, we need to determine how much time the delay will be. The departure delay DAD can be predicted using multiple linear regression algorithm. We use the variables we get from the variable selection step. In Table 11.1, the coefficient is a variate for the corresponding variable while p-value is the probability value explaining the significance of the variable. Standard coefficient expresses the power of influence on each variable floating population whereas VIF is the standard value with which to diagnose multicollinearity. The R-squared of this model is 90.59 %. That means the model is able to predict almost the correct value of the delay time DAD for each flight.

Table 7.1: Predictors in Multiple Linear Regression - China Case Study

Variable	Coeff.	Variable	Coeff.	Variable	Coeff.
AQI	-3.749e-02	Dew.Point	1.636e+00	Elevation	-2.908e-03
Wind.Speed.mph	1.287e-03	DAA	9.343e-01	Flight.ID	-4.856e-02
Wind.Direction	-8.175e-04	Real.Depart	4.516e-04	Orig.Airport	3.075e+00
Wind.Gust.MPH	5.779e-04	Scheduled.Depart	6.879e-05	Dest.Airport	2.535e-01
Heat.Index	6.170e-04	Real.Arrival	-1.867e-05	Humidity	-4.331e-01
Tempreture.FEH	-1.202e+00	Scheduled.Arrival	-4.461e-04		

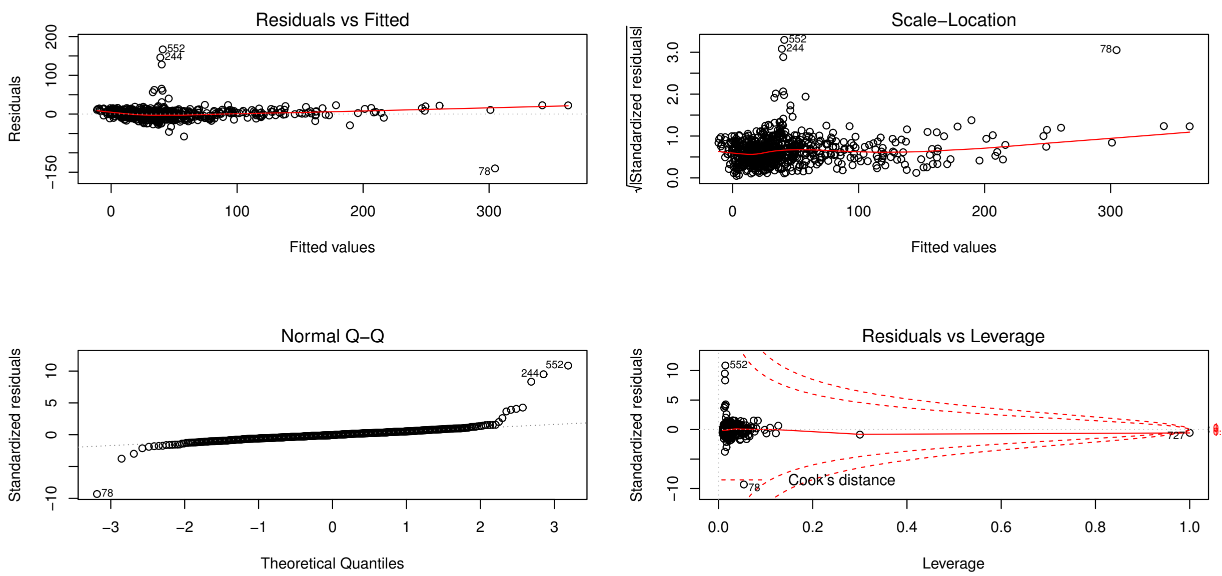


Figure 7.3: Multiple Linear Regression Diagnostic Plots- China Case Study

7.3 Flight Delay Classification and Prediction Model for Australia

We develop two types of models: multiple logistic regression and multiple linear regression. The first is to classify the flights whether they are on-time or delayed. The second is to predict the delay time for each flight. In the following sections, we discuss the variables selection step and the models creations for Australia Case Study.

7.3.1 Variable Selection

Before we create our models, we need to identify the variables that are correlated to the delay at departure DAD. We use AIC method for this purpose. We present the statistically significant variables using correlation plot. This plot has X sign on the variables that are not statistically significant to the DAD output. That means the variables that are not crosses with X sign will be used in the our models in Australia case study.

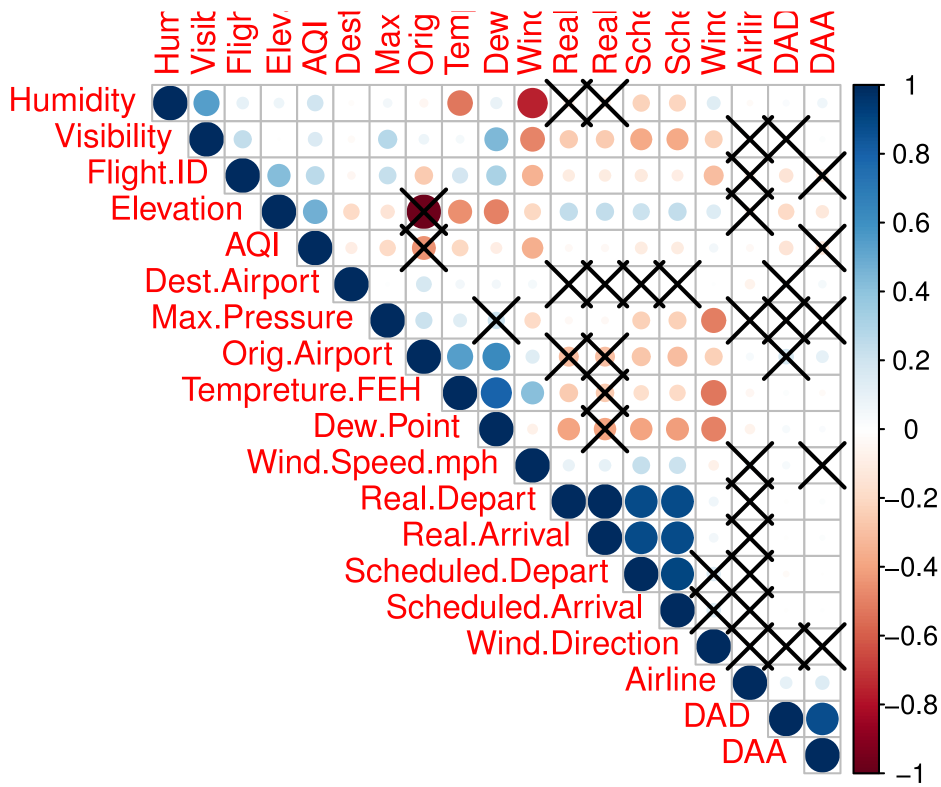


Figure 7.4: Statistically Significant Variables to the Delay At Departure DAD - Australia Case Study

7.3.2 Multiple Logistic Regression

After conducting the variable selection, we create multiple logistic regression based in order to be able to classify the flights. The model is able to achieves 88.63 % accuracy. We can see the performance of the model using Area Under Cover (AUC) plot. That means our model is able to classify the flights correctly.

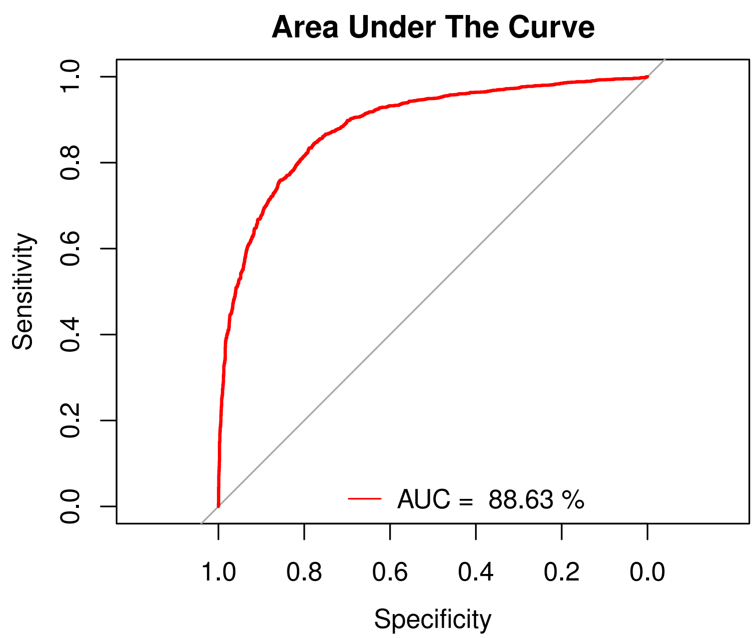


Figure 7.5: Area Under Cover AUC - Australia Case Study

7.3.3 Multiple Linear Regression

The departure delay DAD an be predicted using multiple linear regression. This model consists variables that could explain the delay there. In Table I, the coefficient is a variate for the corresponding variable while p-value is the probability value explaining the significance of the variable. Standard coefficient expresses the power of influence on each variable floating population whereas VIF is the standard value with which to diagnose multicollinearity. The R-squared of this model is 76.97 %. That means the model is able to predict almost the correct value of the delay time DAD.

Table 7.2: Predictors in Multiple Linear Regression - Australia Case Study

Variable	Coeff.	Variable	Coeff.
AQI	5.748e-02	Real.Arrival	-2.054e-03
Wind.Speed.mph	-3.717e-05	Scheduled.Arrival	7.402e-04
Tempreture.FEH	-4.573e-02	Elevation	-1.892e-03
Dew.Point	2.487e-02	Flight.ID	3.023e-03
DAA	7.975e-01	Humidity	2.054e-02
Real.Depart	1.401e-03	Airline	-3.782e-01
Scheduled.Depart	-2.758e-04		

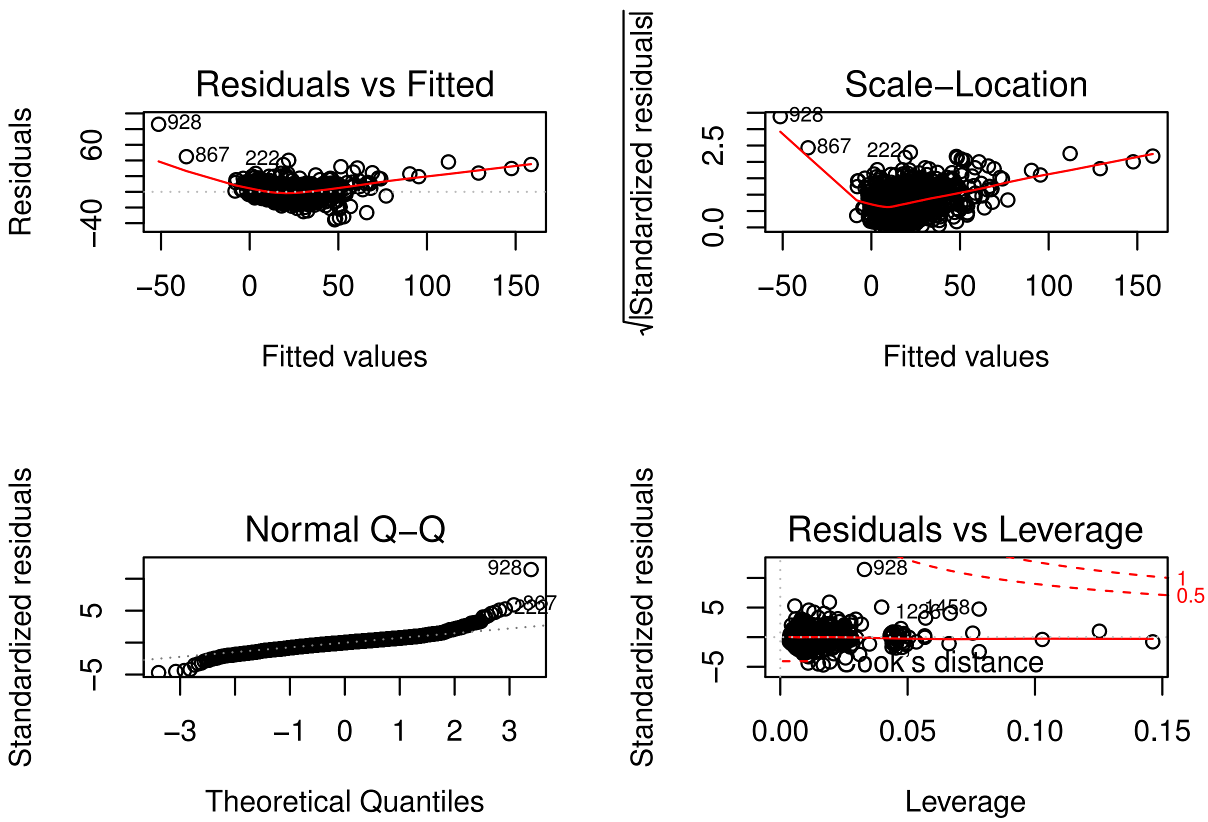


Figure 7.6: Multiple Linear Regression Diagnostic Plots- Australia Case Study

7.4 Flight Delay Classification and Prediction Model for Europe

We use two types of models multiple logistic regression and multiple linear regression. The first is to classify the flights whether they are on-time or delayed. The second is to predict the delay time. The following discusses the variables selection and the models creations.

7.4.1 Variable Selection

Before we create our models, we need to identify the variables that correlate to the delay at departure. We use AIC method for this purpose. We present the statistically significant variables using correlation plot like the correlation matrix we see in the previous chapter. This plot has X sign on the variable that is not statistically significant to the DAD.

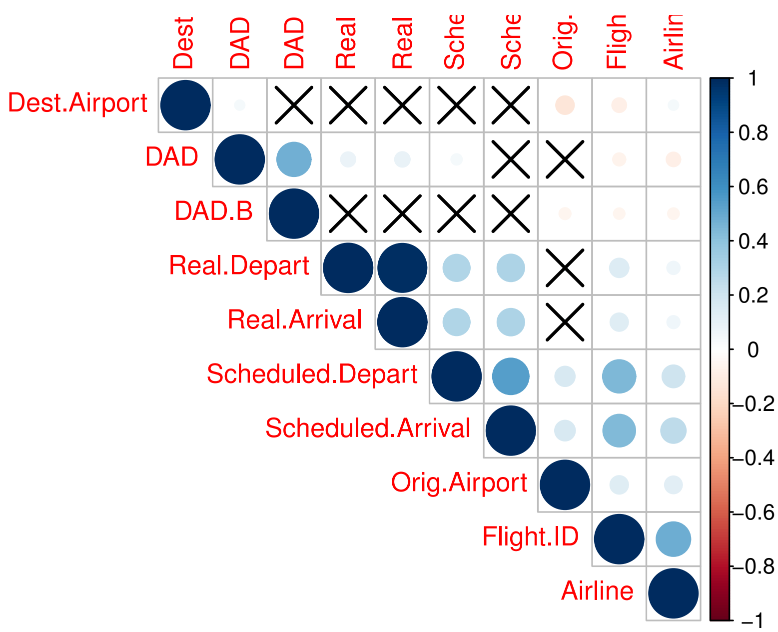


Figure 7.7: Statistically Significant Variables to the Delay At Departure DAD - Europe Case Study

7.4.2 Multiple Logistic Regression

After conducting the variable selection, we create multiple logistic regression based in order to be able to classify the flights. The model is able to achieves 86.45 % accuracy. We can see the performance of the model using Area Under Cover (AUC) plot. That means our model is able to classify the flights correctly.

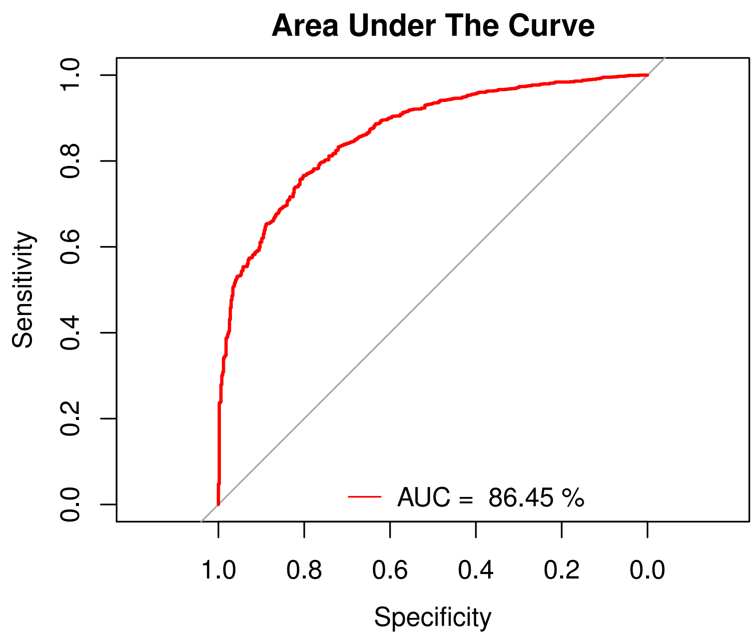


Figure 7.8: Area Under Cover AUC - Europe Case Study

7.4.3 Multiple Linear Regression

The departure delay DAD can be predicted using multiple linear regression. This model consists variables that could explain the delay there. In Table I, the coefficient is a variate for the corresponding variable while p-value is the probability value explaining the significance of the variable. Standard coefficient expresses the power of influence on each variable floating population whereas VIF is the standard value with which to diagnose multicollinearity. The R-squared of this model is 77.68 %. That means the model is able to predict almost the correct value of the delay time DAD.

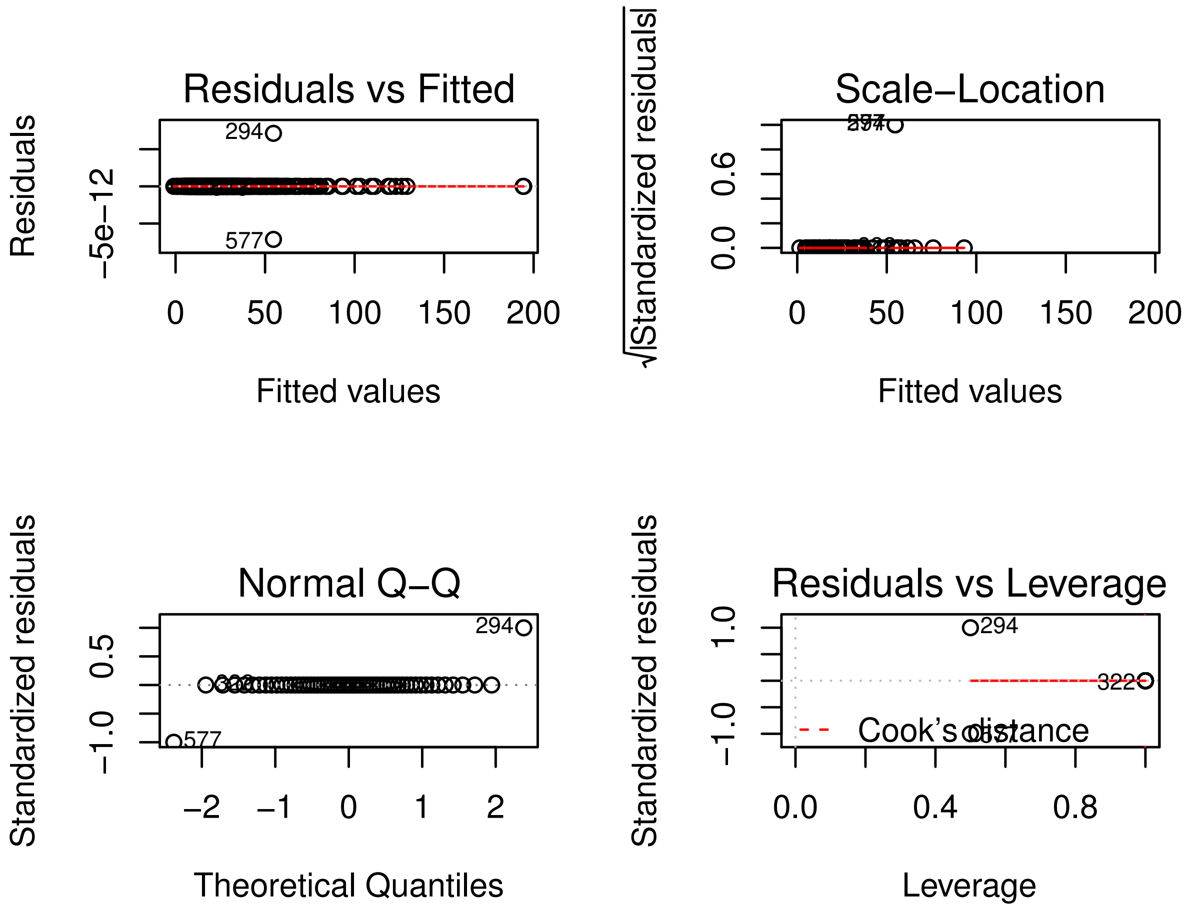


Figure 7.9: Multiple Linear Regression Diagnostic Plots- Europe Case Study

Chapter 8

Conclusion

In this project, we provide a comprehensive investigation and analysis of the flight delays problem. We build a novel service that can predict the flight delay using new and real-time data. We build a crawler to collect data. We analyze the data to see how the data from different data sets are correlated.

Previous studies have addressed the flight delay problem in terms of historical data that were collected by the Bureau of Transportation and the Federal Aviation Administration. These studies were helpful to determine some of the major factors that cause the flight delay. However, studying this phenomenon requires to consider other data rather than depending on the data related to the Air traffic.

In order to tackle this problem, we need to widen our vision and the context of the problem by incorporating some extra data sources that provide real-time data. Therefore, this project will utilize the real-time data provided from the various resources as indicated above. Then it will perform the data mining process to discover more hidden factors that would contribute to the delay. This research will look for the contextual data which is not considered before. This means this study would provide a new step toward investigating this issue.

After we determine the key factors that cause the delay, we come up with a mathematical models that predict the flight delay in advance. These models achieve very high accuracy. That will enable all stakeholders to make the right decisions. This study provides a novel method to discuss the flight delay, and it contributes to allow further investigations by utilizing the contextual data. Furthermore, this study will be significant in both academic and industry. With the emergence of the IoT paradigm, huge amount of data is there. This research is significant because it contributes to put the first brick in order to fill the gap since there is a lack in this field.

We create two types of models in each case study we have. All the models perform well. They achieve high accuracy. In some case studies, we may need in the future to consider other data sources in order to dig deeply into the problem and discover more factors.

Bibliography

[1] Qin, Q. L.; Yu, H. A statistical analysis on the periodicity of flight delay rate of the airports in the US. *Advances in Transportation Studies*. 2014 Special, Issue Special Vol3, p93-104. 12p.

[2] Y. Liu and F. Yang, Initial Flight Delay Modeling and Estimating Based on an Improved Bayesian Network Structure Learning Algorithm, Natural Computation, 2009. ICNC '09. *Fifth International Conference on, Tianjin, 2009*, pp. 72-76.

[3] Y. J. Liu and S. Ma, "Flight Delay and Delay Propagation Analysis Based on Bayesian Network," Knowledge Acquisition and Modeling, 2008. KAM '08. *International Symposium on, Wuhan, 2008*, pp. 318-322.

[4] Tu, Y., Ball, M., Jank, W., 2006. Estimating flight departure delay distributionsA statistical approach with long-term trend and short-term pattern. Working Paper, Department of Decision and Information Technologies, *University of Maryland*.

[5] Xiangyang Xu, Hua Yuan and Yu Qian, "Analyzing the system features of the flight delays: A network perspective," Service Systems and Service Management (ICSSSM),*12th International Conference on, Guangzhou*, 2015, pp. 1-5.

[6] K. Novianingsih and R. Hadianti, "Modeling flight departure delay distributions," Computer, Control, Informatics and Its Applications (IC3INA), 2014 *International Conference on, Bandung*, 2014, pp. 30-34.

[7] Y. J. Liu, W. D. Cao and S. Ma, "Estimation of Arrival Flight Delay and Delay Propagation in a Busy Hub-Airport," Natural Computation,

2008. ICNC '08.*Fourth International Conference on, Jinan*, 2008, pp. 500-505.

[8] Y. Liu and S. Ma, "The Multimode Estimation Modeling for Flight Delay of a Busy Hub-Airport in Flight Chain," Services Science, Management and Engineering, 2009. SSME '09. *IITA International Conference on, Zhangjiajie*, 2009, pp. 557-561.

[9] Juan Jose Rebollo, Hamsa Balakrishnan, Characterization and prediction of air traffic delays, Transportation Research Part C: Emerging Technologies, Volume 44, July 2014, Pages 231-241

[10] Shemshadi et al., ThingSeek: A crawler and search engine for the Internet of Things, *ACM SIGIR 2016*, Pisa, Italy

[11] X. Geng, "Analysis and Countermeasures to Flight Delay Based on Statistical Data," Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 *5th International Conference on*, *Hangzhou*, 2013, pp. 535-537.

[12] Fei Rong, Li Qianya, Hu Bo, Zhang Jing and Yang Dongdong, "The prediction of flight delays based the analysis of Random flight points," *Control Conference (CCC), 2015 34th Chinese*, *Hangzhou*, 2015, pp. 3992-3997.

[13] https://www.flightradar24.com

[14] https://www.wunderground.com

[15] http://aqicn.org/city/beijing/

[16] C. Perera, C. H. Liu, S. Jayawardena and M. Chen, "A Survey on Internet of Things From Industrial Market Perspective," in *IEEE Access*, vol. 2, no. , pp. 1660-1679, 2014.

[17] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans. Intell. Syst. Technol*. 5, 3, Article 38 (September 2014), 55 pages.

[18] J. Cheng, "Estimation of flight delay using weighted Spline combined with ARIMA model," Advanced Infocomm Technology (ICAIT), 2014 IEEE 7th International Conference on, Fuzhou, 2014, pp. 8-20.

[19] L. Qianya, W. Lei, F. Rong, W. Bin and H. Xinhong, ”An analysis method for flight delays based on Bayesian network,” The 27th Chinese Control and Decision Conference (2015 CCDC), Qingdao, 2015, pp. 2561-2565.

[20] Y. J. Liu, W. D. Cao and S. Ma, ”Estimation of Arrival Flight Delay and Delay Propagation in a Busy Hub-Airport,” 2008 Fourth International Conference on Natural Computation, Jinan, 2008, pp. 500-505.

[21] H. Alonso and A. Loureiro, ”Predicting flight departure delay at Porto Airport: A preliminary study,” 2015 7th International Joint Conference on Computational Intelligence (IJCCI), Lisbon, Portugal, 2015, pp. 93-98

[22] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York, 2nd edition.

[23] Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A.A., Zou, B.: Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the united states (2010)

[24] Rosenberger, J.M., Schaefer, A.J., Goldsman, D., Johnson, E.L., Kleywegt, A.J., Nemhauser, G.L.: A stochastic model of airline operations. Transportation science 36(4), 357377 (2002)

[25] Mueller, E.R., Chatterji, G.B.: Analysis of aircraft arrival and departure delay characteristics. In: AIAA aircraft technology, integration and operations (ATIO) conference (2002)

[26] Perera, C., Liu, C.H., Jayawardena, S., Chen, M.: A survey on internet of things from industrial market perspective. IEEE Access 2, 16601679 (2014)

[27] Perera,C.,Zaslavsky,A.,Christen,P.,Georgakopoulos,D.:Contextawarecompuing for the internet of things: A survey. IEEE Communications Surveys & Tutorials 16(1), 414454 (2014)

[28] Rebollo, J.J., Balakrishnan, H.: Characterization and prediction of air-track delays. Transportation Research Part C: Emerging Technologies 44, 231241 (2014)

[29] Tu, Y., Ball, M.O., Jank, W.S.: Estimating flight departure delay distributionsa statistical approach with long-term trend and short-term pattern. Journal of the American Statistical Association 103(481), 112125 (2008)

[30] Real-time Investigation of Flight Delays Based on the Internet of Things Data. Aljubairy, A; Shemshadi, A.; Sheng, Q. Z. In Proceedings of the 12th Anniversary of the International Conference on Advanced Data Mining and Applications (ADMA), Gold Coast, Australia, 2016. to appear.