DATA ANALYSIS AND DECISION SUPPORT

FLIGHTS DELAY ANALYSIS

Using MongoDB, Tableau & R

Santosh Pawar (29085683)

1 TABLE OF CONTENTS

2 INTRODUCTION 1
3 FLIGHTS DELAY DATASET 1
3.1 DATASET CLEAN-UP & ISSUES 1
3.1.1 BLANK DATA FOR AIRPORT LAT/LONG 1
3.1.2 FAULTY DATA VALUE FOR SEPTEMBER MONTH 1
3.1.3 DATE FIELD FORMAT 2
3.1.4 WEEKDAY 2
3.1.5 NA-VALUES 2
3.1.6 EXTRANEOUS COLUMNS 3
4 ANALYSIS 3
4.1 OVERVIEW OF BUSINESS VALUE 3
4.2 SUMMARY STATISTICS 3
5 ANALYSIS METHODOLOGY 5
5.1 TIME BASED HEAT MAP ANALYSIS 5
5.2 TIME BASED TAXI IN – TAXI OUT DISTRIBUTION 7
5.3 AIRLINES ANALYSIS 7
5.3.1 AIRLINE VS DELAYED FLIGHTS % 7
5.3.2 GEOGRAPHICAL DISTRIBUTION OF DELAYED FLIGHTS VS AIRLINES 8
5.3.3 AIRLINE VS ODD TIME DELAY 9
5.3.4 DELAY CAUSES 9
5.4 TOP 10 ANALYSIS 10
5.5 FORECASTING 10
6 CONCLUSION 13
7 REFERENCES 14

1 INTRODUCTION

The report presented discusses the statistical and visual analysis of the Flight's Delays in United States for the year 2015. The analysis have used the storage solution of NoSQL MongoDB Community version database server. The state of the art tool for statistical data analysis-Rstudio(R Programming) & the Industry wide recognized BI & Visualization tool Tableau. In addition to these tools, Studio 3T is used for Mongo database browsing and as SQL wrapper tool for query building.

The report briefly discuss the Dataset selected for the analysis, the programming environment, discussion of the analysis towards the Business Question and concludes with the future scope.

2 FLIGHTS DELAY DATASET

The Dataset selected for this final assignment collection of Domestic flights and airlines data. It is originally generated and maintained by the U.S. Department of Transportation's Bureau of Transportation statistics. The Dataset contains the flights and their respective airline details along with departure times, arrival times, departure delays and arrival delays and so on. The Dataset has Airport data of 322 different airports across USA and neighboring countries with their IATA code. The Airline data contains the top Air travel providers and their IATA code.

2.1 DATASET CLEAN-UP & ISSUES

For any data analysis and the exploration cleaning the data and getting the data into the standard format is essential. Cleaning data ensures that there are no faulty data, which might affect the analysis and disturb the patterns in the data that we are after.

Few of the issues faced during the cleaning of the dataset are listed below

2.1.1 Blank Data for Airport Lat/Long

The Airports data with 322 entries had three entries, which did not have latitude and longitude values.
Solution – Based on the other fields such as airport name and city, the Lat/Long retrieved from online and updated manually.

2.1.2 Faulty data value for September month

The Data of all other months except September had the Origin Airport ID & Destination Airport ID as Numeric. While the Dataset of Airports had only character codes. Due to this while, joining the two files the details for the flights in September was getting lost.

Solution: To avoid this loss of data & under the assumption that the pattern of delays stays same across the year the original data of month September was removed and filled with random sample data from other months.

2.1.3 Date field format

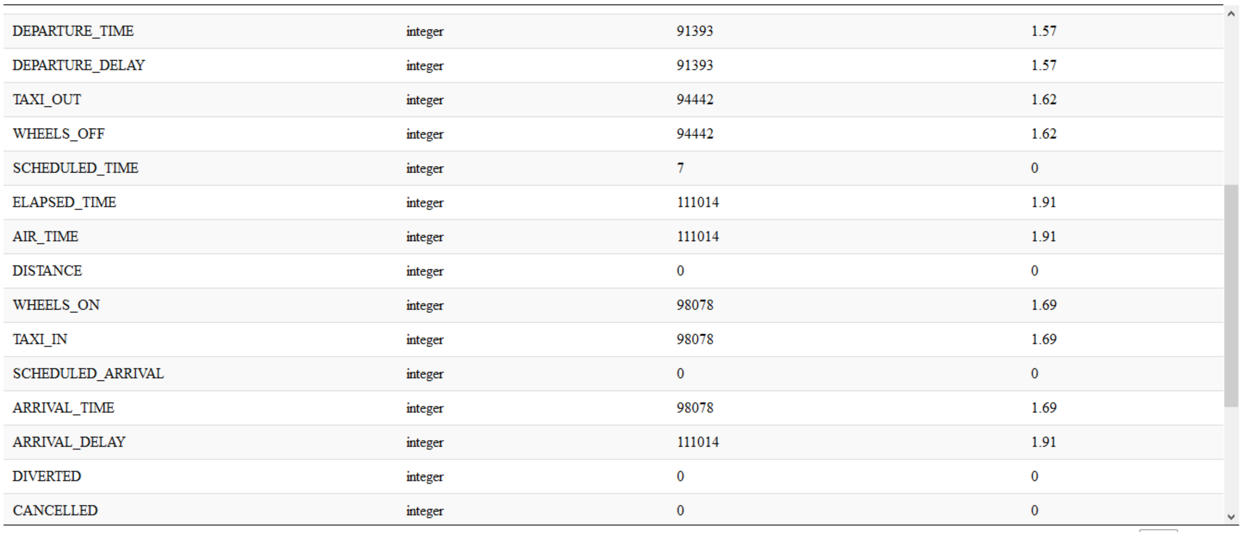
The original data had the Date, Month and the Year in different columns.
Solution: The different columns merged to get the date in one column as standard Date format.

2.1.4 Weekday

The format of the weekday column was numeric.
Solution: The numeric codes converted to respective day of the week. e.g. 1 = Monday, 2= Tuesday so on

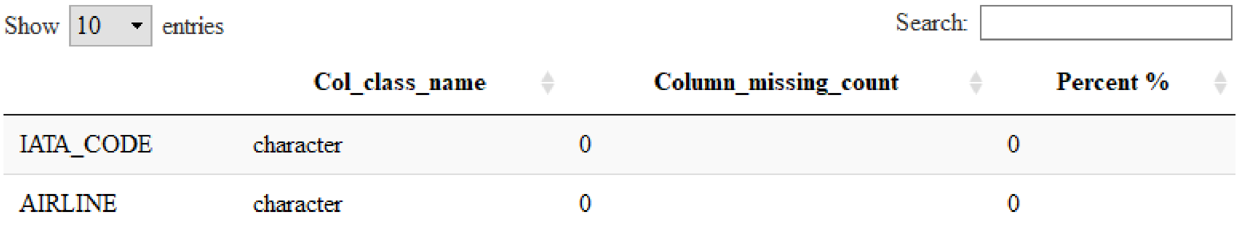
2.1.5 NA-Values

The Missing value analysis was done specially to identify the columns with missing values and either remove them or replace with the default values. Below are the screenshots of the Missing value analysis of three files.



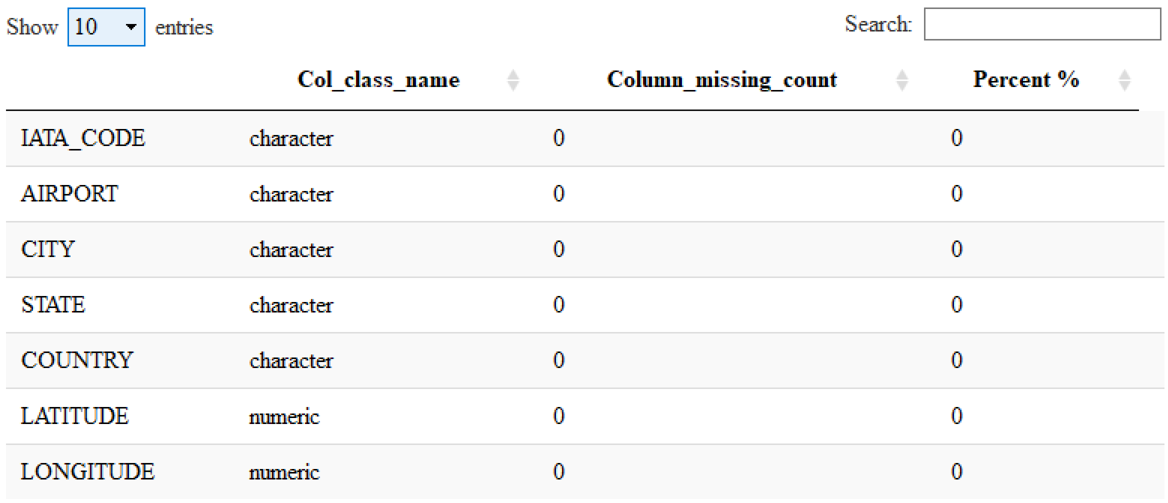
DEPARTURE_TIME	integer	91393	1.57
DEPARTURE_DELAY	integer	91393	1.57
TAXI_OUT	integer	94442	1.62
WHEELS_OFF	integer	94442	1.62
SCHEDULED_TIME	integer	7	0
ELAPSED_TIME	integer	111014	1.91
AIR_TIME	integer	111014	1.91
DISTANCE	integer	0	0
WHEELS_ON	integer	98078	1.69
TAXI_IN	integer	98078	1.69
SCHEDULED_ARRIVAL	integer	0	0
ARRIVAL_TIME	integer	98078	1.69
ARRIVAL_DELAY	integer	111014	1.91
DIVERTED	integer	0	0
CANCELLED	integer	0	0

Figure 1 Flights Missing Data Analysis



	Col_class_name	Column_missing_count	Percent %
IATA_CODE	character	0	0
AIRLINE	character	0	0

Figure 2 Airlines Missing Data Analysis



	Col_class_name	Column_missing_count	Percent %
IATA_CODE	character	0	0
AIRPORT	character	0	0
CITY	character	0	0
STATE	character	0	0
COUNTRY	character	0	0
LATITUDE	numeric	0	0
LONGITUDE	numeric	0	0

Figure 3 Airports Missing Data Analysis

As the Size of the dataset is 5.7 million rows the 1% to 2% missing data was neglected and removed from the source dataset. The NA values for Delay Types was logically correct. Because if the Delay type has value NA that means the flight was on time. Hence, zero replaced all types of delay, which had NA values.

2.1.6 Extraneous columns

The extra and unnecessary columns that might not be helpful in deriving the insights from the data were removed.

3 ANALYSIS

3.1 Overview of Business Value

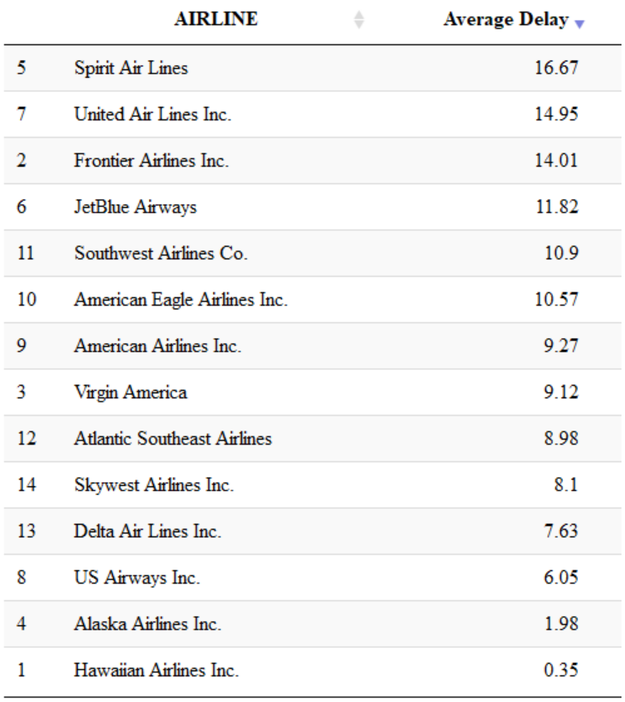
The Business Question for the analysis of the dataset is concentrated towards the customers of the airlines services. All the airline providers boast their efficiency in travel industry and use marketing tricks to dupe the customers. On the other hand, customers do not have tools to validate their claims.

Hence, the analysis is focused towards letting the customers make informed decision about their selection of Airline Providers when they are planning to travel within USA.

Apart from this main business question, the analysis to some extent will also help Airline providers' to find insights, which could be useful to identify the patterns of inefficiency in their service and improve on it.

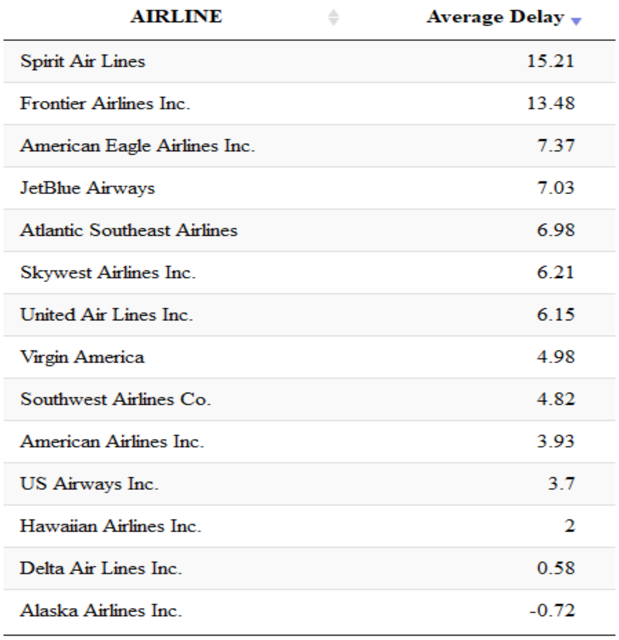
3.2 Summary Statistics

Upon basic exploration of the data through mongo, queries in R below are few basic summary statistics were observed.



	AIRLINE	Average Delay
5	Spirit Air Lines	16.67
7	United Air Lines Inc.	14.95
2	Frontier Airlines Inc.	14.01
6	JetBlue Airways	11.82
11	Southwest Airlines Co.	10.9
10	American Eagle Airlines Inc.	10.57
9	American Airlines Inc.	9.27
3	Virgin America	9.12
12	Atlantic Southeast Airlines	8.98
14	Skywest Airlines Inc.	8.1
13	Delta Air Lines Inc.	7.63
8	US Airways Inc.	6.05
4	Alaska Airlines Inc.	1.98
1	Hawaiian Airlines Inc.	0.35

Figure 4 Average Departure Delay by Airline



AIRLINE	Average Delay
Spirit Air Lines	15.21
Frontier Airlines Inc.	13.48
American Eagle Airlines Inc.	7.37
JetBlue Airways	7.03
Atlantic Southeast Airlines	6.98
Skywest Airlines Inc.	6.21
United Air Lines Inc.	6.15
Virgin America	4.98
Southwest Airlines Co.	4.82
American Airlines Inc.	3.93
US Airways Inc.	3.7
Hawaiian Airlines Inc.	2
Delta Air Lines Inc.	0.58
Alaska Airlines Inc.	-0.72

Figure 5 Average Arrival Delay by Airlines

These summaries and more insights are explored in the further analysis of the data to help the customers in selection of the airlines for their travel. The departure or arrival delay are key points to help customer reach to the decision.

However, we drill down further for more insights. Below is the screenshot of the distribution of overall flights by month so it can be one of the key point while selecting to travel by air by the distribution we can say that February has the less chances of being delayed during air travel and if combined by appropriate Airline service it can reduce even more.

So on the foresight with just two key points we can say that a combination of certain airline in particular month may affect the delay time.

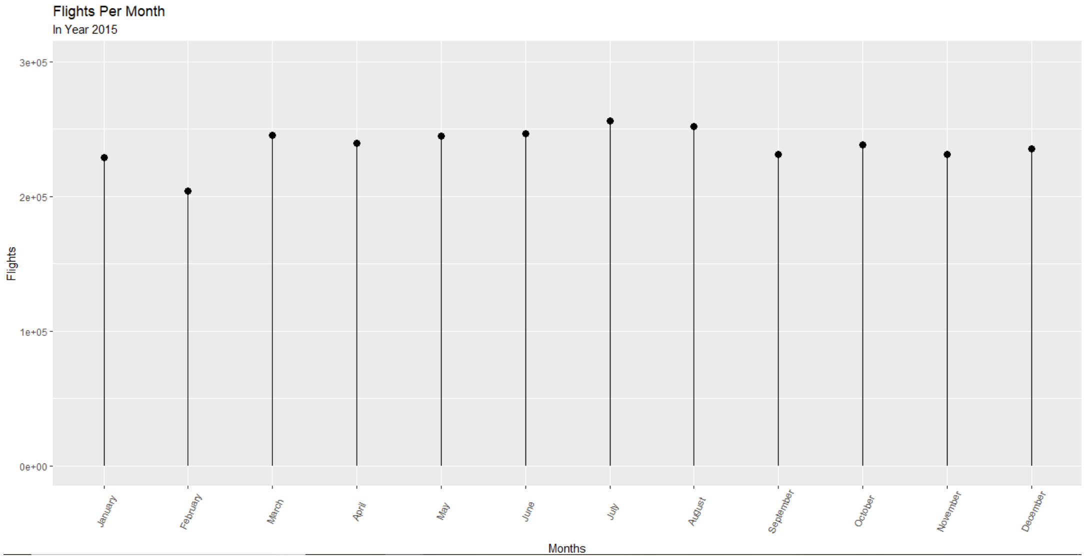


Figure 6 Delayed Flights per month

4 ANALYSIS METHODOLOGY

At the first glance of the data & with the summary statistics, we can go further into analysis to find more key points. To go further with analysis the Rstudio used extensively along with the MongoDB database to get the sample data which might have potential for an insight. This subset of data then plotted into meaningful visualization to find patterns. Below are the few types of methods used for the analysis of the data to find the insights.

4.1 TIME BASED HEAT MAP ANALYSIS

The following heat map show the distribution of delayed flights over the one year and density of the monthly delayed flights are highlighted by either red if more and green if less. Drilling down even further we observed the distribution of flights over week with similar heat map giving detailed pattern of weekly and daily delays.

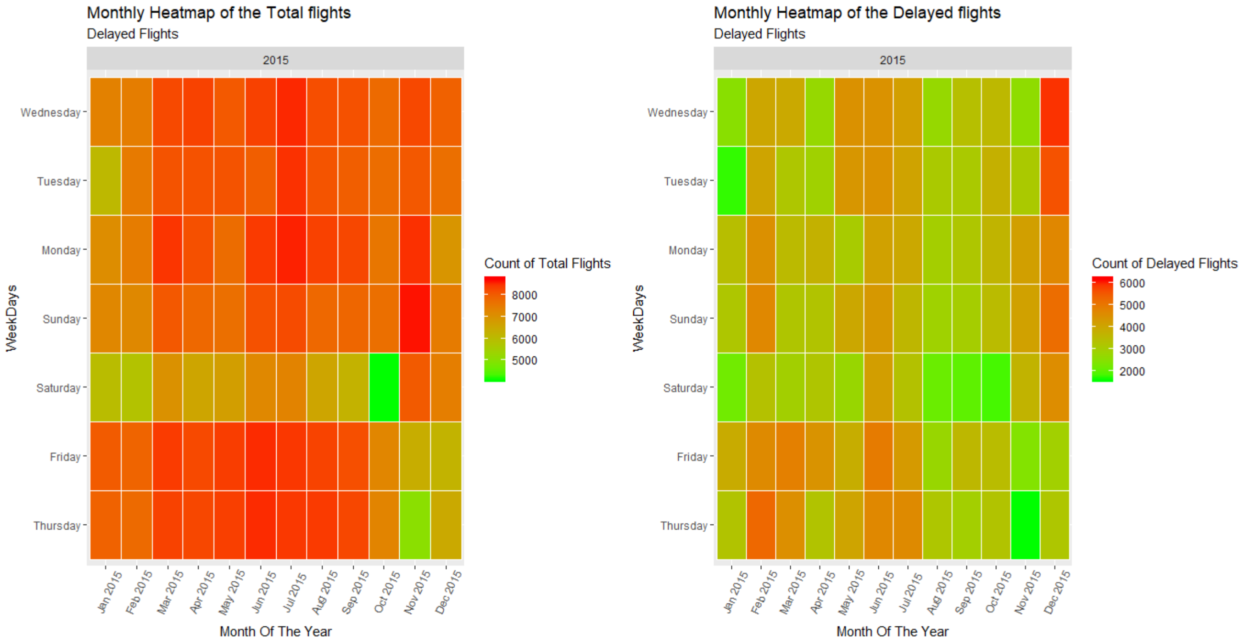


Figure 7 Monthly Heat Map for Delayed Flights

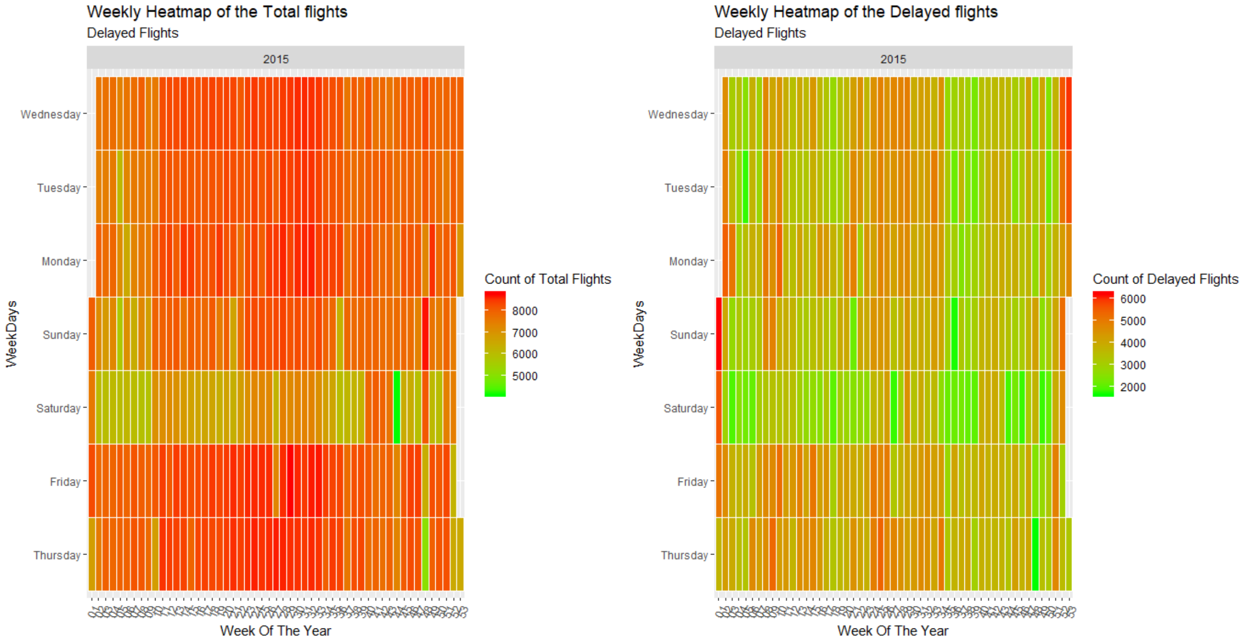


Figure 8 Weekly Heat Map of Delayed Flights

4.2 Time based Taxi in – Taxi out Distribution

The taxi out time is the time calculated from the airline leaves the gate and takes off. While the taxi in time is the time, calculate from the landing of the aircraft to reaching to the gate. This time usually less but it is a contributing factor for the overall delay. Below distribution and vertical pair of plots of Departure - Taxi out & Arrival – Taxi In we can co-relate the pattern.

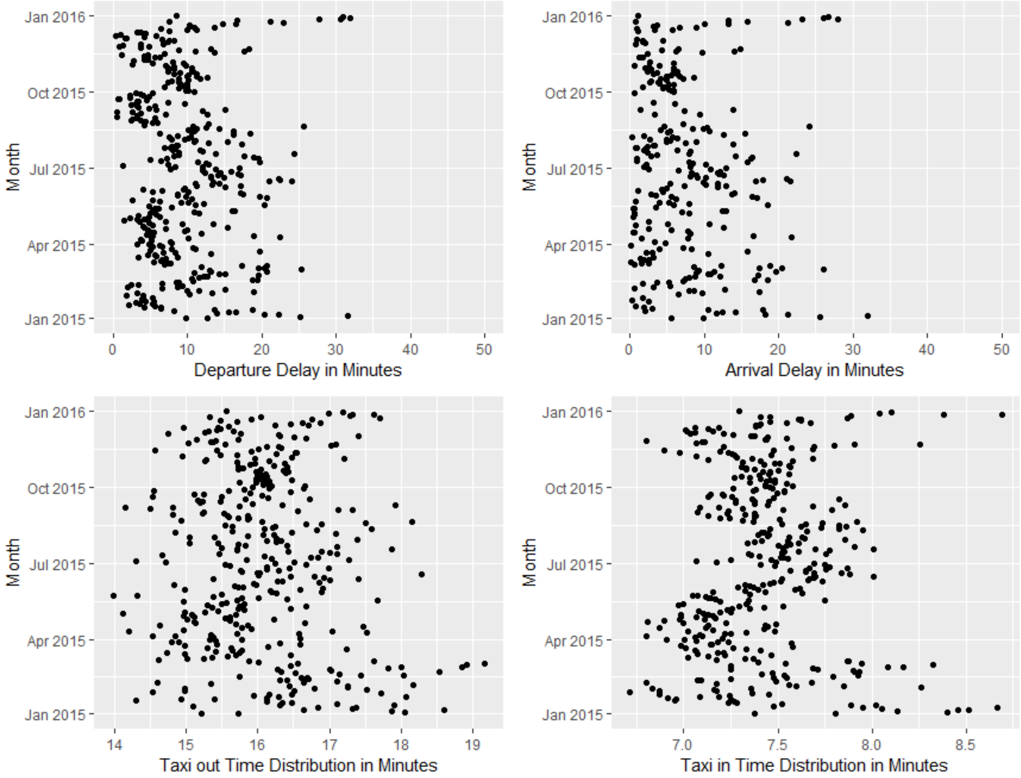


Figure 9 Taxi In Taxi Out Analysis

4.3 Airlines Analysis

4.3.1 Airline Vs Delayed Flights %

The below visualization can help the customer to shortlist the Airline they want as the chart shows the percent value of delayed flights by each airline in year 2015 & the size of the indicator depicts the number of flights. Hence, this could prove to be one key point.

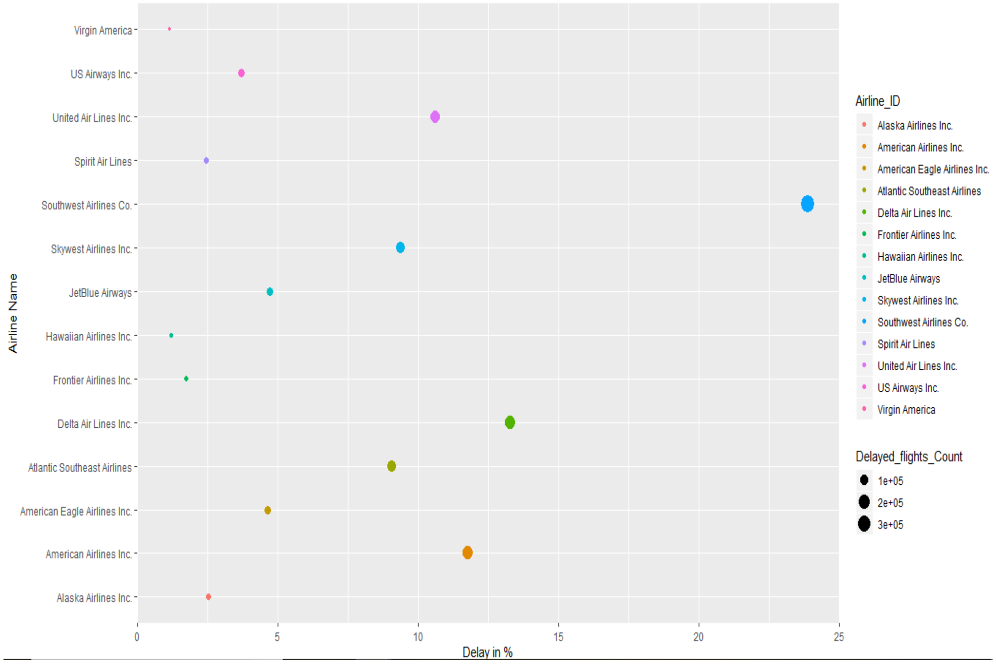


Figure 10 Airline Vs Delay %

4.3.2 Geographical distribution of delayed flights Vs Airlines

In extension to the analysis discussed above with the help of Tableau, we can provide the further insights by including the locations of the delayed flights.

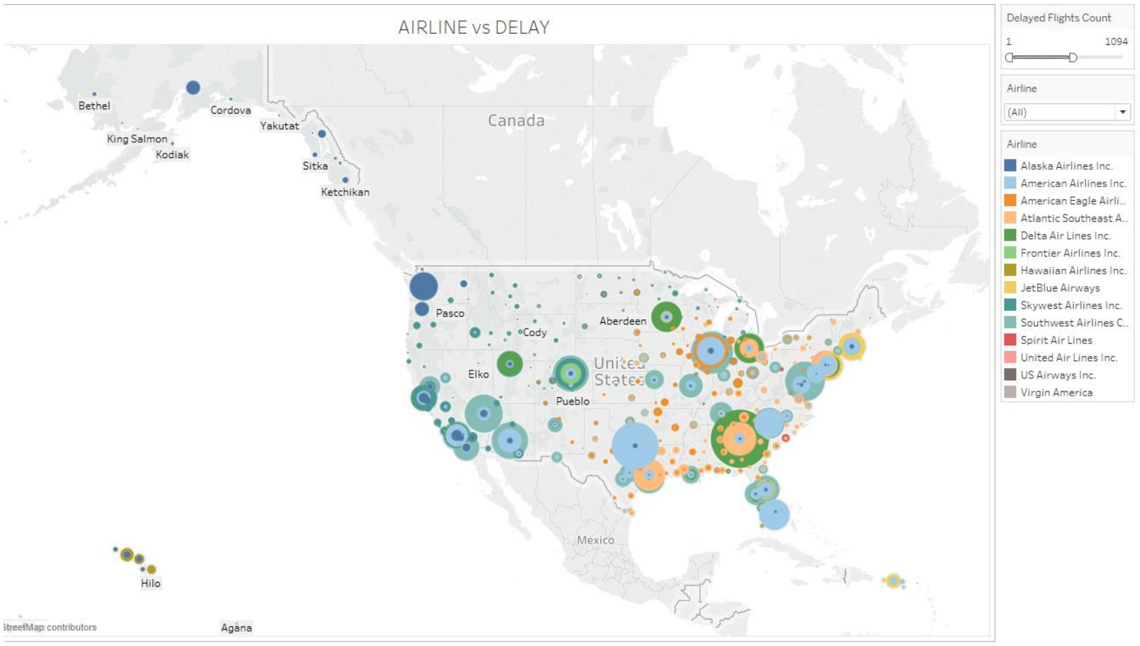


Figure 11 Geographical Distribution of Flights

The analysis distributes the delayed flights by their location and groups them by Airlines. This could help the customers to see which airline-performing better in which region. Therefore, if a customer travelling from New York to the West Coast can certainly make informed decision.

4.3.3 Airline Vs Odd Time Delay

In case of Airlines customers, it is frustrating to see the flights delayed at the odd times like midnight to early morning. The customer might not consider the same Airline next time. As said before it is harmful for service providers to lose the customers. Hence, this is the one chart plotted to show the flights that were delayed during late night and early morning. This chart shows the delayed flights at odd times by different Airlines, So if the customer is travelling by a connecting flight, which falls in these times, can refer to this insights and choose appropriate airline to avoid inconvenience.

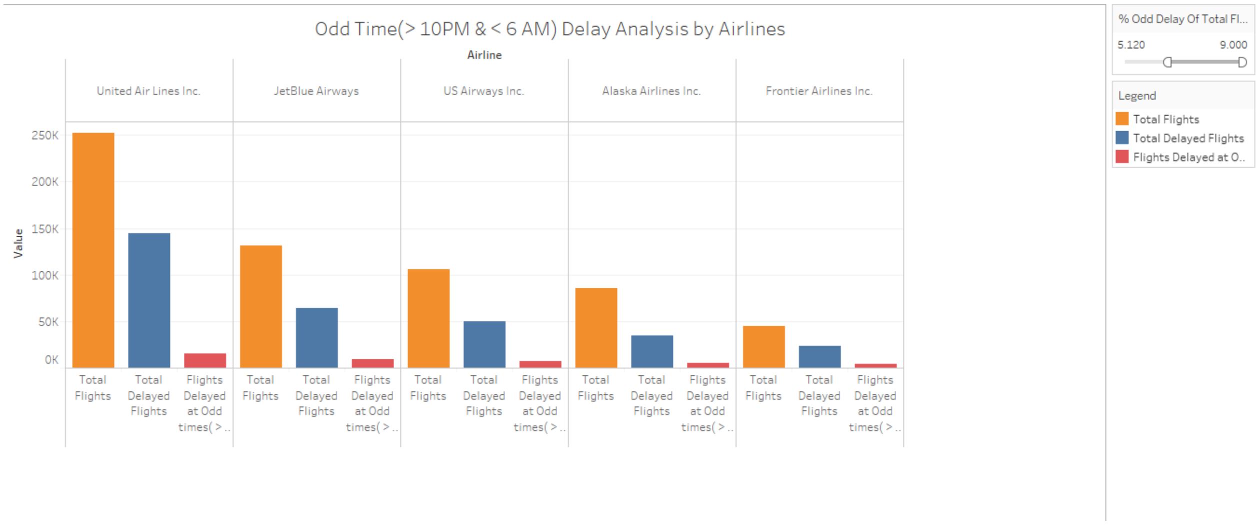


Figure 12 Odd Time Delays Analysis

4.3.4 Delay Causes

As stated before the analysis is done towards helping customers and airline providers. Below chart shows the causes of the delay for the flights. Although weather delays are not under control but the delays related to air system, security checks can be reduced if planned and executed efficiently.

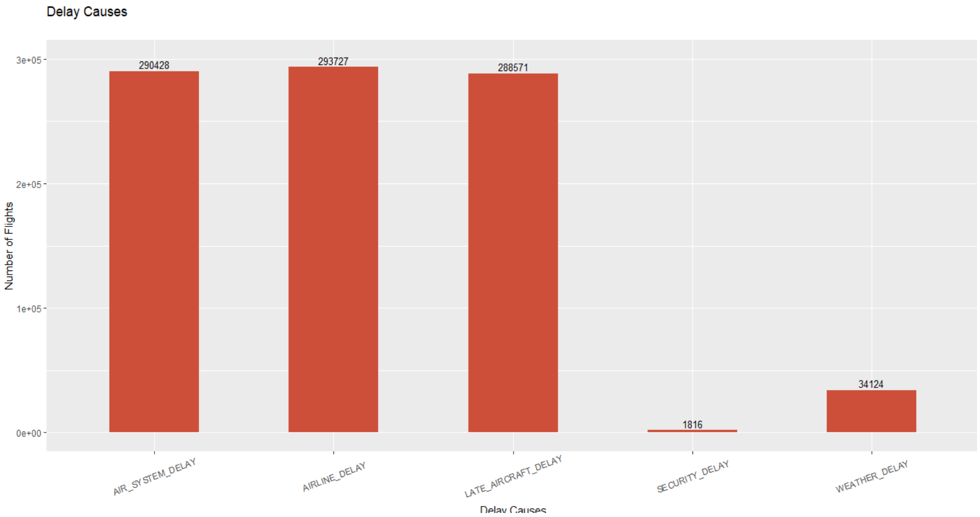


Figure 13 Distribution of Delay Cause for 2015

4.4 Top 10 Analysis

In general, for the customers knowledge below plots could be helpful to show the top 10 busy airports for departure and arrival. With help of previous plots, the customer can choose an appropriate airline to avoid delays in these busy airports too.

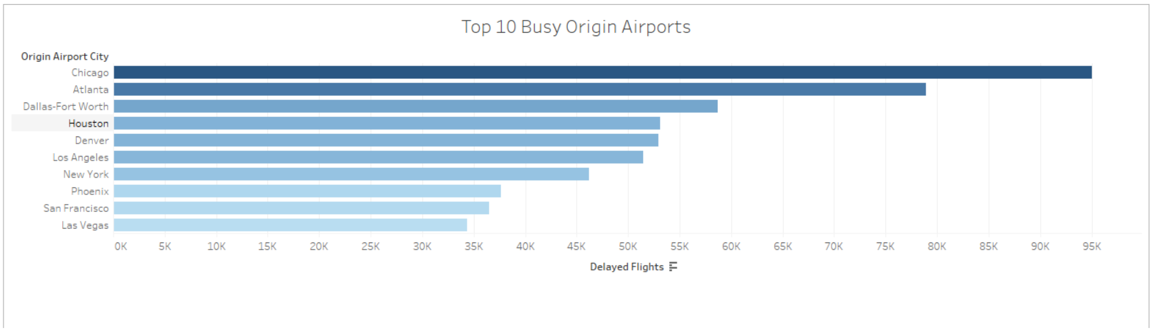


Figure 14 Top 10 Busy Origin Airports

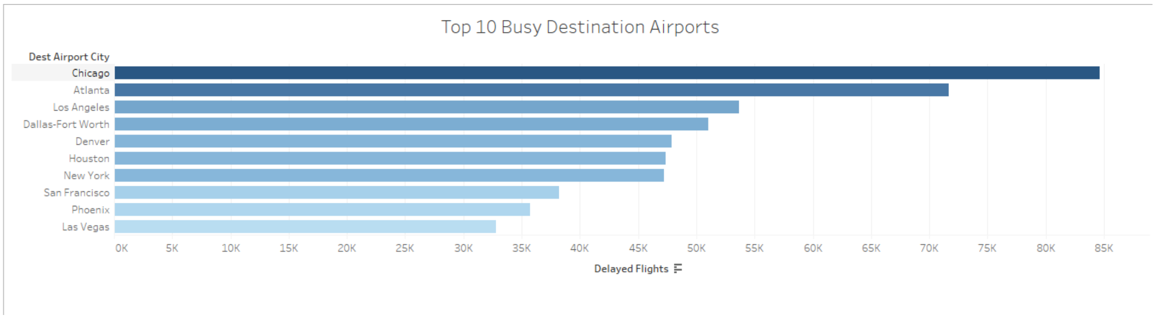


Figure 15 Top 10 Busy Destination Airports

4.5 Forecasting

The forecasting of the departure delay and the arrival delay is the final and important key point towards the decision of the customers while it could be also useful for the service providers to plan ahead and minimize that delay as much as possible.

For the forecasting purposes, we have used the HoltWinters method. It is one of the Exponential forecasting models used for seasonal forecasting of the data.

HoltWinters utilizes a procedure for constantly reviewing a prediction in relation to most recent experience. It assigns exponentially declining weights as the values in time series get older. In short the recent observations are prioritized than the older observations. (Kalekar, 2004)

The results of the Forecast can be observed clearly, The forecast is done quarterly for departure and arrival delay. And the forecast covers next one quarter of forecast which is sufficient enough for customers to plan ahead for travel while avoiding delays and for service providers to plan ahead and put plans in place to avoid the delays.

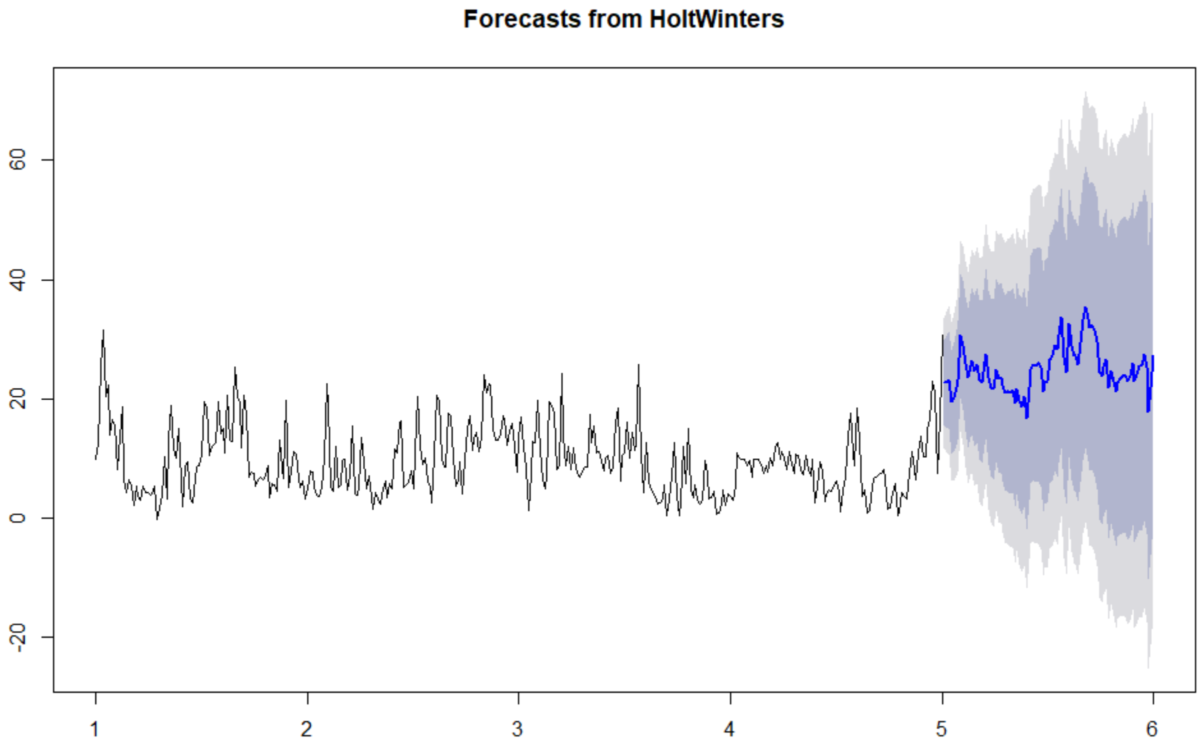


Figure 16 Departure Delay Quarterly Forecasting Version 1

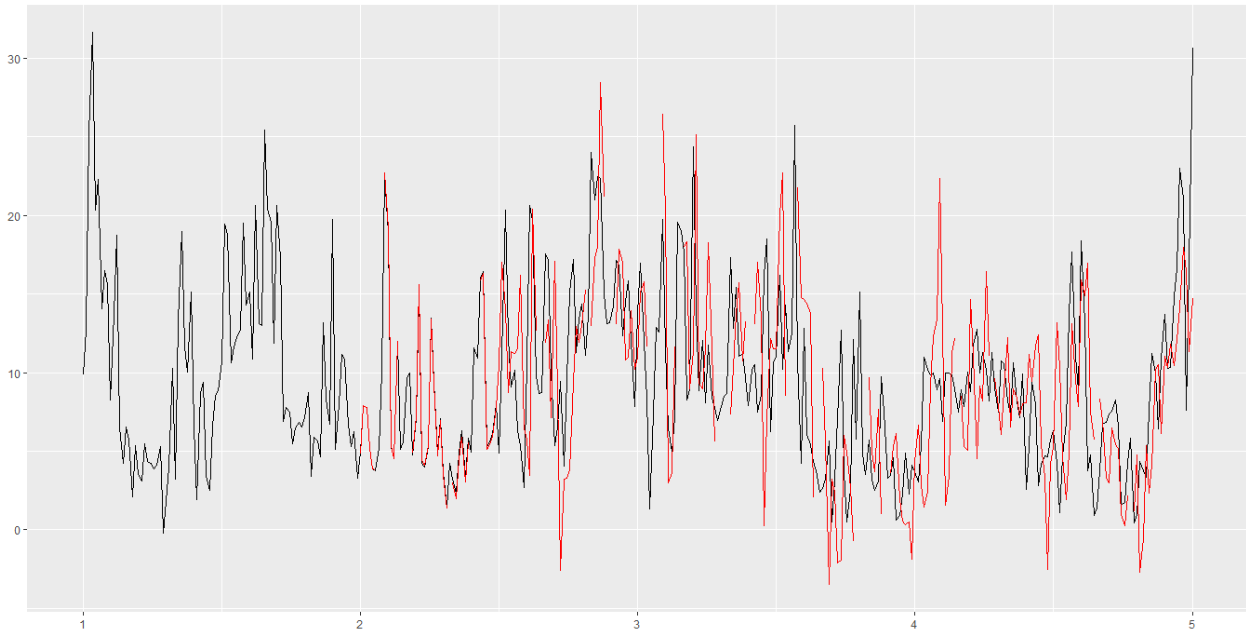


Figure 17 Departure Delay Quarterly Forecasting Version 2

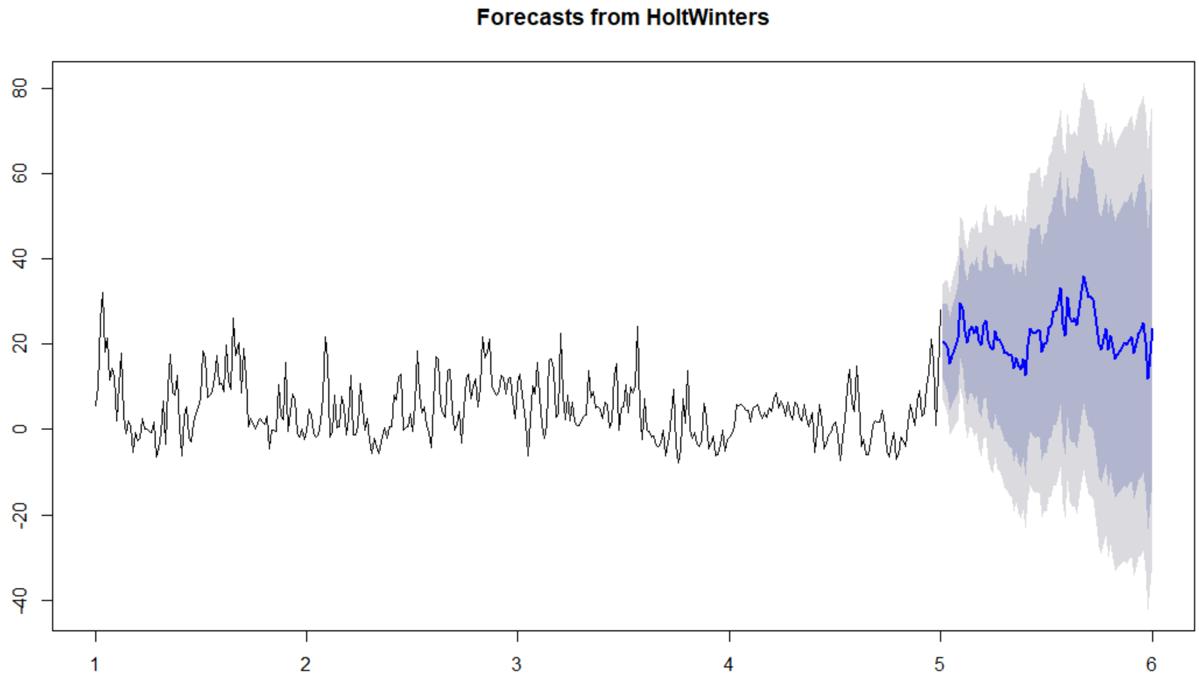


Figure 18 Arrival Delay Quarterly Forecasting Version 1

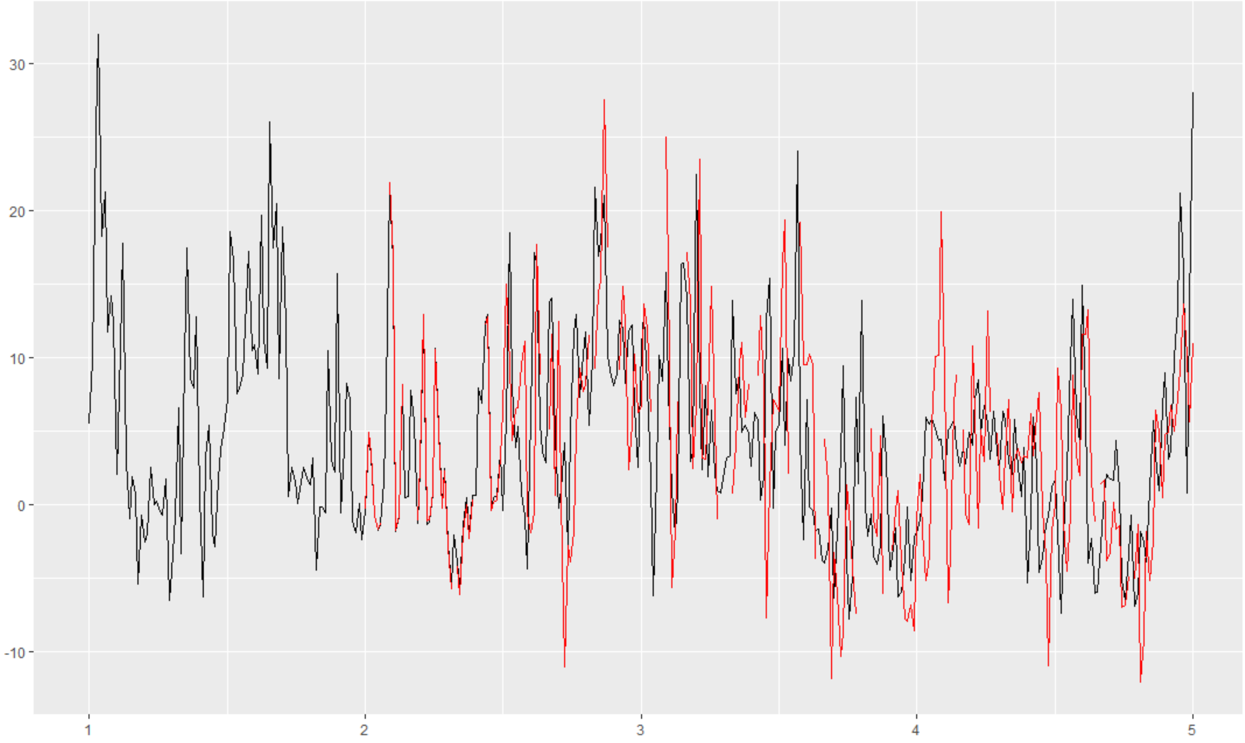


Figure 19   Arrival Delay Quarterly Forecasting Version 2

5 CONCLUSION

To Conclude, All the analysis done during this assignment seems insightful enough for the customer to make an informed decision about selection of the airlines to avoid any potential delay. While the service providers can also look into these results to improve their performance.

6 REFERENCES

Kalekar, P. S. (2004). *Time series Forecasting using Holt-Winters Exponential Smoothing.*