Title: US commercial flight delays and cancellation analysis

Name: Yizheng Fang

Student ID: 29407699

Unit: Data Exploration and Visualization

Tutor: Debbie

Tutorial: Friday 10 am

# Introduction

# In the United States, there are thousands of airports existing in the different cities even small size of counties. By developing the airline traffic, people can travel from state to state within a limited time. However, as the result of many unpredictable factors, more and more flights are delayed or cancelled. This project is mainly exploring the influences of flight delays and cancellations, also attempting to find out if there are any patterns and trends in given period. In addition, making reasonable prediction and forecast of the possible delays or cancellations in the future. Moreover, the project explores the possible ways for travellers to avoid the possible delay flights ahead and to eliminate the cost of time with airlines often get delays or cancellations.

# Data Wrangling

# The flight delayed, and cancellation data is from Kaggle and was collected and published by the U.S. Department of Transportation's Bureau of Transportation Statistics. It tracks the on-time performance of domestic flights operated by large air carriers. The flights dataset contains 31 variables and over five million of observations. The data is across from January 2015 to December 2015, and the information of flight delayed, or cancellation is shown in each day is given. Meanwhile, the information of airports and airlines are given along with all airports' longitude and latitude.

The original dataset of flights includes possible factor variables that may influence our target variable departure delay such as scheduled departure time, scheduled arrival time, arrival delay. However, some factor variables are not necessary to be included in our analysis, such as tail numbers, wheels off, wheels on etc.

First, making a selection of related variables and create a correlation matrix using corrplot in R to see if there is any clear linear relationship between departure delay and other variables.

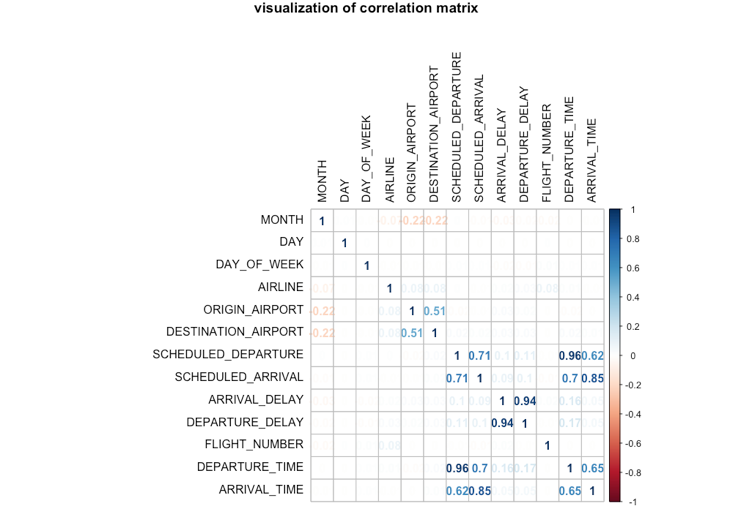


Figure 1.1: The correlation matrix of flight data shows some correlations between selected variables. There is a clearly strong linear relationship between departure delay and arrival delay. Besides, we assume that the departure and arrival time and scheduled departure and arrival time are correlated, which is making sense in the matrix, but departure delay is not quite linearly related to departure time.

The figure is shown above only represents part of variables correlations and may need more pieces of evidence to improve our conclusions. In fact, the flights dataset contains some variables sounds reasonable such as weather delay, airline delay, security delay etc. Unfortunately, these variables are full of missing values; over 90 percent of data in these columns are missing. The existence of these missing values may reasonable as airlines may not report the real reasons that cause flight delays or cancellations. In order to make our analysis and further prediction more sense to make decisions, we decide to exclude these variables from our analysis. A quick overview of these variables shown in box plot created in R as follows

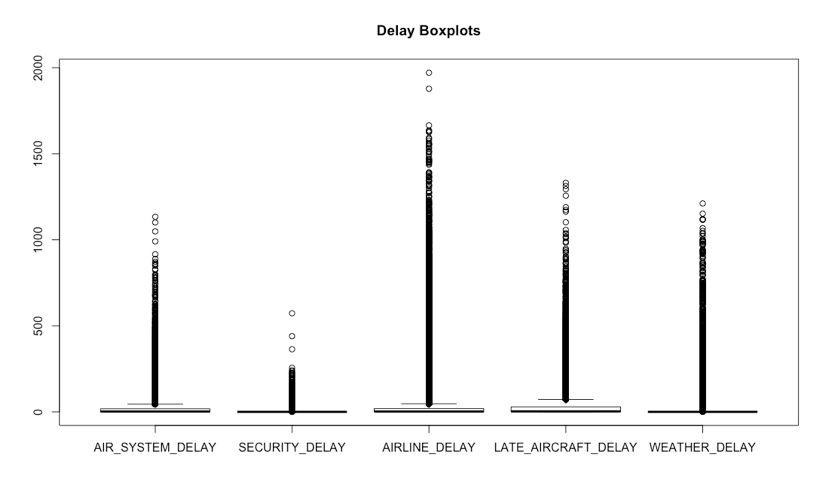


Figure 1.2: The boxplot of five delay reasons for delay time shows an unusual pattern in this data, and a lot of data distributing seems as outliers, but it is quite possible for one flight delay over 1000 minutes which is around 16 hours because of airline or bad weather conditions. However, it is obvious that the airline delay has longer delay time in minutes than other reasons. The unusual long delay time may cause some ambiguous average values for further exploration.

After overview the data correlation, we select some of variables and union flight dataset to itself in Tableau, so a map view of flight route fit in our analysis. Moreover, we create calculated fields as new variables named route id to identify the flight route from original airports to destination airports and route order is the order of flight between airports. Meanwhile, inner join airports data in tableau with flights union data so that the geographic data like longitude and latitude will be included. Here is a preview of data that used to create a route map in tableau.

To merge flights and airports data, we use primary key IATA code in airport data

and the foreign key of the original airport in flights data. Then two sets of data can be combined together.



Figure 1.3: Each original airport has its identifier code which is IATA code in airports data. The route identifier is a new derived variable that can identify the flight route from the original airport to the destination airports. The route order only has two numbers 1 and 2, 1 represents that it is an original airport, 2 is a destination airport. Therefore, we clearly show the information of flight path on a map.

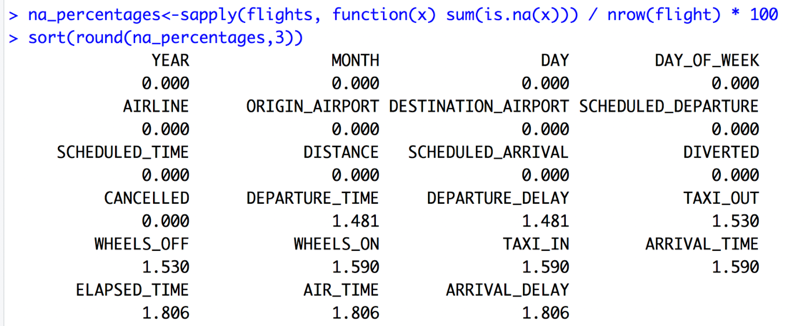
One more important checking is to see if there are any missing values in the dataset. One way to create a summary table, using built-in function in R, the result is shown as follows:

Figure 1.4: The summary table of missing values shows less than 2% in given columns. We use median imputation to replace missing values in the set of data rather than drop rows. After cleaning data, the data seems more meaningful to use for analysis.

In airports data, even though we have missing values for longitude and latitude, by importing data into tableau, the missing values are filled so that can be used properly. And there is no missing value in airline data.

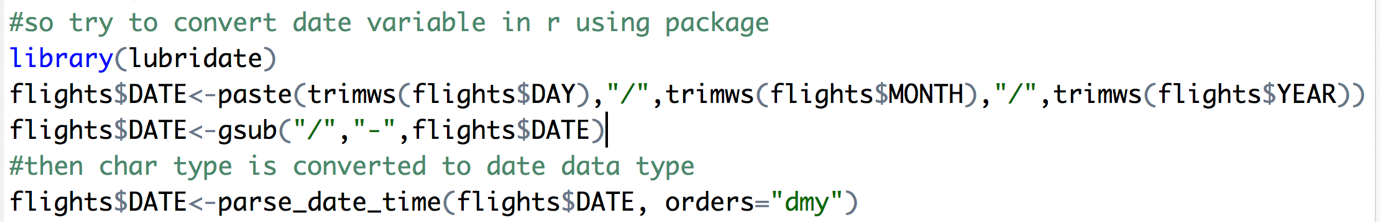
Next, we tried to reformat the date variables, which is the most difficult part to handle with. In flights data, we have year, month and day even day of the week in separated columns, therefore, we tried to combine them in tableau and export as csv file. However, since the extreme large sample size we used, and avoid unnecessary variables included, we tried to use R instead, and paste the date variable using R library lubridate, which has a function can paste year, month and day together then parse a character type variable into date type.

Figure 1.5: The parse\_date\_time is a special function in lubridate package. It can convert character type variable to POSIXct date data type. POSIXct is a special and frequent use date data type in R and can represent in a date format of with specific arguments.

Finally, since the special date variable POSIXct cannot be converted to time series type. We use a simple integer from 1 to 365 to represent as the first day to the last day of 2015. Also selecting target variable column in original data. Therefore, we can build model afterwards in time series.

# Data exploration

After data cleaning and reformatting, we try to fit the data in a linear model using lm function in R. By carefully selecting related variables using correlation., scheduled arrival time, scheduled departure time is included in the model.

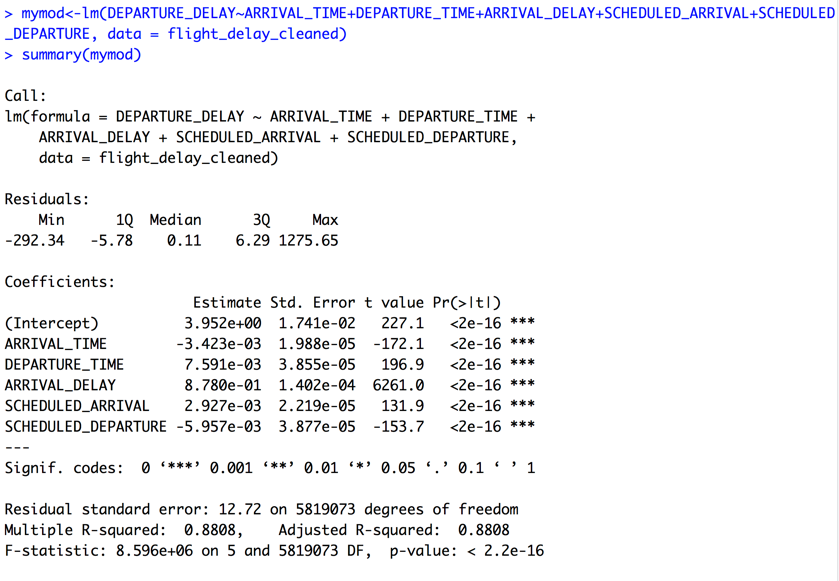


Figure 2.1: The model shown above implies about 88% of variations of target variable departure delay can be explained by given factor variables when looking at adjusted R-squared. Besides, the selected variable has small enough p values, the variables are important to our target variables, we can conclude that it is statistically significant that selected predictors are correlated to the departure delay. And the overall model has a p-value is extremely small, less than 0.001, we can assume that the model is fairly good to use.

Furthermore, we want to have some insights of the overview of the data using a line chart in tableau. The diverging colour can show how the delay time changes across 12 months in 2015.

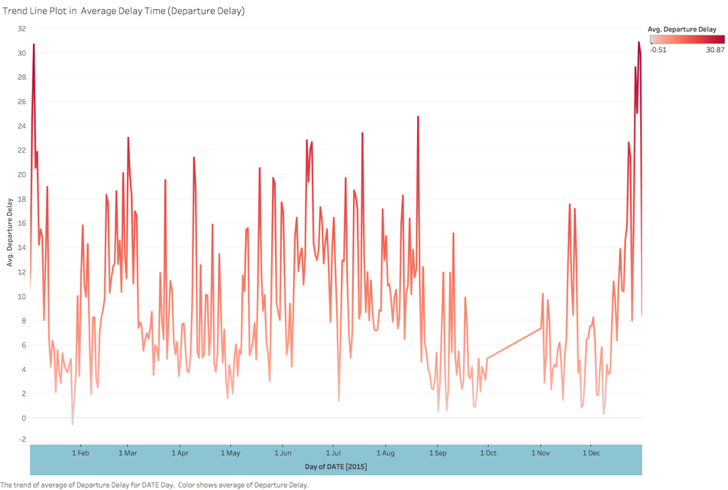


Figure 2.2: The line plot shows an overall trend of average delay time in 2015. The obvious upward trends are centred near the end of the year, which is reasonable as it's Christmas and New Year holiday. Moreover, there is also an upward near June and July, which is summer holiday for American. A large number of travellers may cause airports busy in the period and may lead flight delays.

The project is only exploring the influences that related to flight delays, since the proportion of cancellations in the data is quite small, we show the result in an airline cancellation ranking chart as follows:

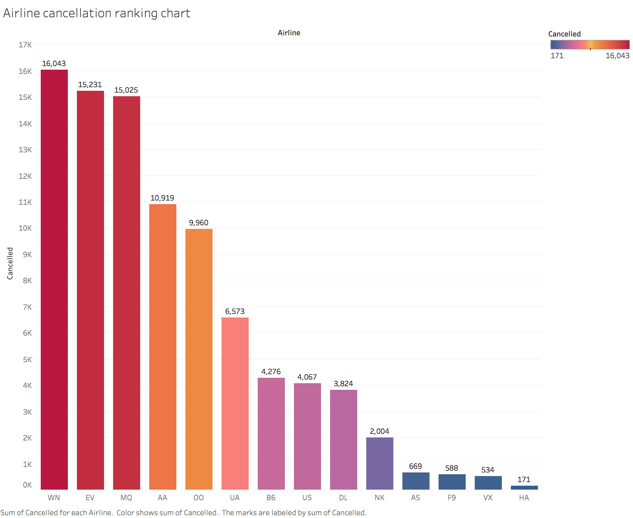


Figure 2.3: In the bar chart, we can see there are 3 top airlines has the most cancellation number in a year, which is WN (Southwest Airlines), EV (Atlantic Southeast Airlines), MQ(American Eagle Airlines). Each of these airlines company has over 15000 cancellations across the year, which is large, but we may want to explore what time those cancellations happen.

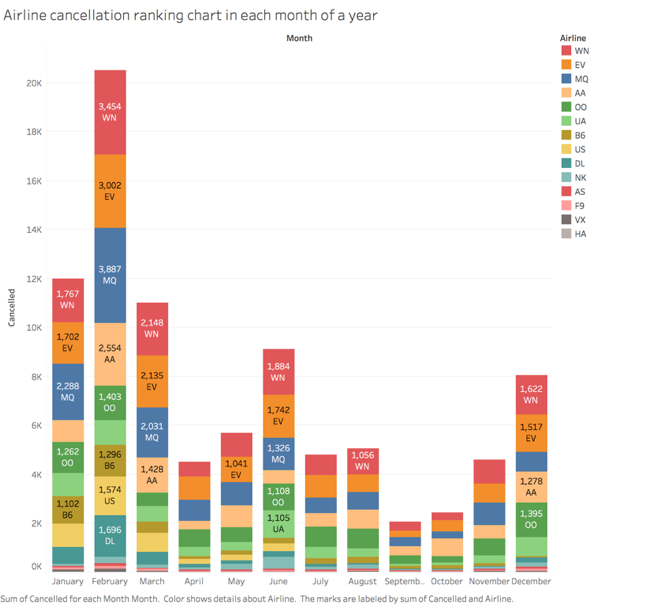


Figure 2.4: The stacked bar chart shows in each month, the three airlines company mentioned before have the most number of cancellations. Also, the cancellation season is consistent with the period that flight delay happens when comparing with the line plot in previous plots. The chart with no labelled number means cancellations number is less than 1000.

Although the number of cancellations seems large in a year, the small proportion of cancellations in the whole dataset may provide little information to our analysis.

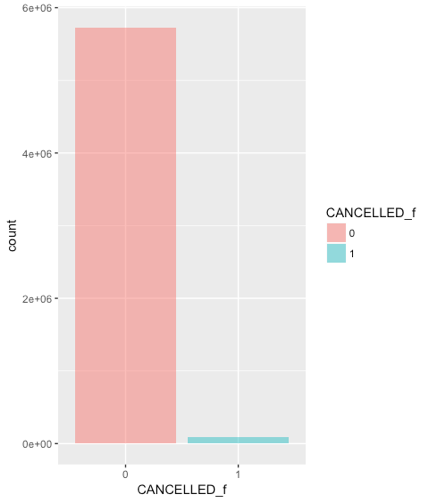


Figure 2.5: The bar plot shows as two colours to represent whether the flight is cancelled or not. The 0 means not cancelled, which means the flight may be delayed instead; and 0 means it is cancelled. In this plot, we can see that the non-cancellation part weighs more than cancellation part, so our main purpose is to predict flight delays.

There are two variables about the delay in the flights dataset, one is arrival delay, which is arrival time plus scheduled arrival, and another is departure delay, which is the total delay time. Since these two variables are highly correlated, we decide to choose departure time as my target variable to investigate.

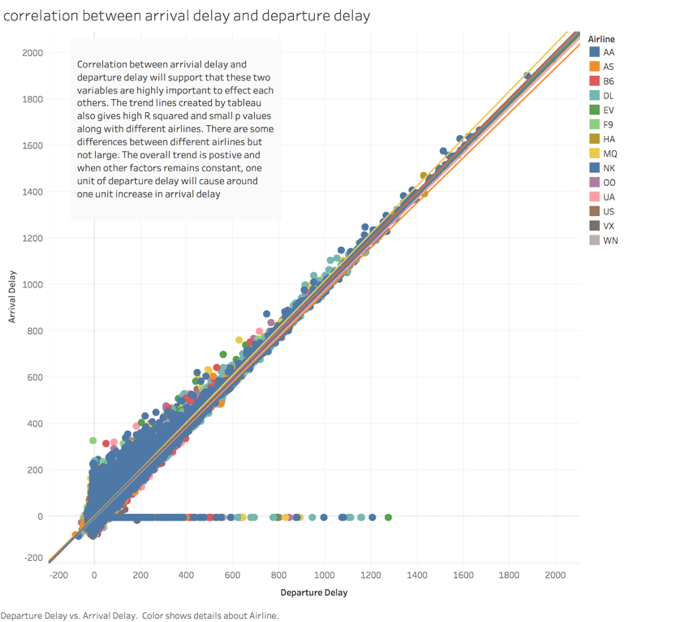


Figure 2.6: Correlation between arrival delay and departure delay will support that these two variables are highly important to affect each other. The trend lines created by tableau also gives high R squared and small p values along with different airlines. There are some differences between different airlines but not large. The overall trend is positive and when other factors remain constant, one unit of departure delay will cause around one unit increase in arrival delay.

For answering the question like how we can avoid the flight delays as a traveller, we also need to consider the locations of the airports and choosing the right airline companies.

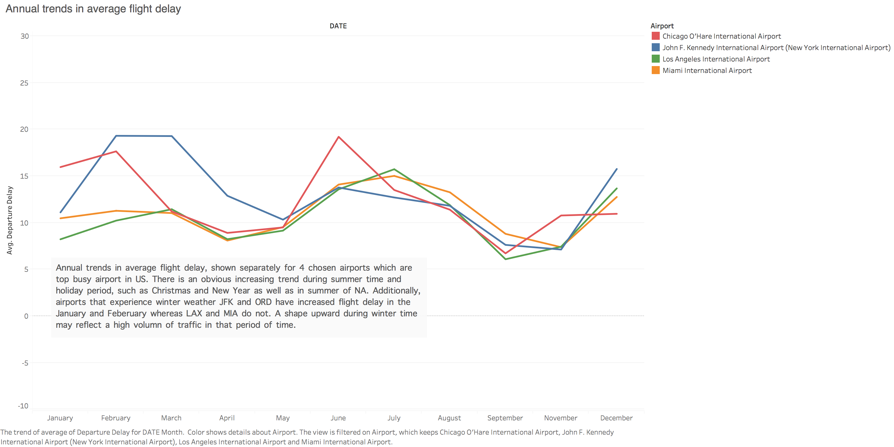


Figure 2.7 Annual trends in average flight delay, shown separately for 4 chosen airports which are the top busy airport in the US. There is an obvious increasing trend during summer time and holiday period, such as Christmas and New Year as well as in summer of NA. Additionally, airports that experience winter weather JFK and ORD have increased flight delay in the January and February whereas LAX and MIA do not. A shape upward during winter time may reflect a high volume of traffic in that period of time.

Also using bar plot again to show the difference in average delay time between different airlines using different colours.

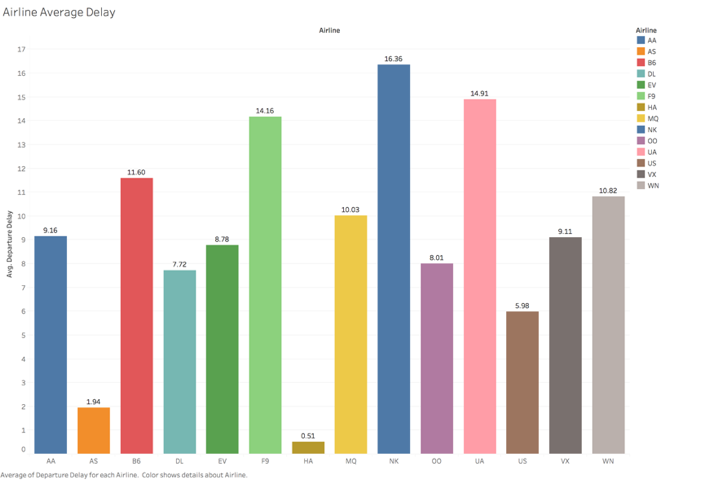


Figure 2.8: Each colour represents an airline company in the US, apparently, NK(Spirit Airlines), DL(Delta Airlines) and UA(United Airlines) are top three airline companies that have the most average delay time across the year and have longer delay time than other airlines.

The merged data can create a map view using tableau, with marked flight routes using selection bar, also together with showing the delay time in each airport for each day using heatmap table. The map and heat map table can update simultaneously.

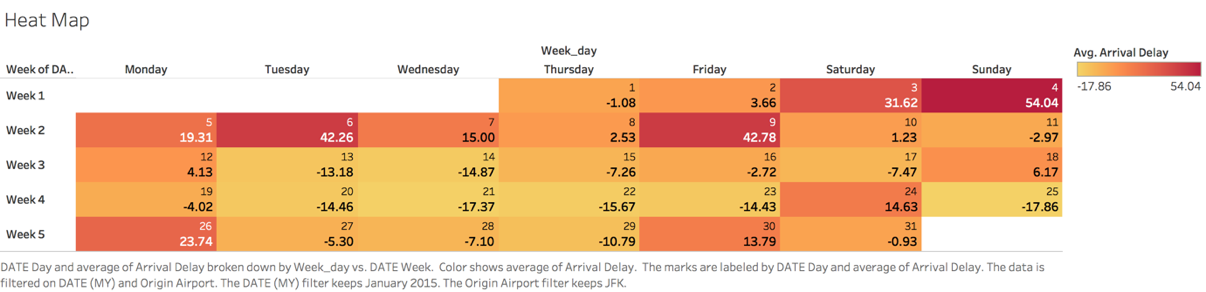


Figure 2.9: The highlighted table also known as heat map table, can show the difference in numbers. The dark red is the largest number which has longer average departure delay time, whereas for light orange is a small number in average delay time. The table can be filtered using tableau and the values can also be updated as choosing different airport on the map.

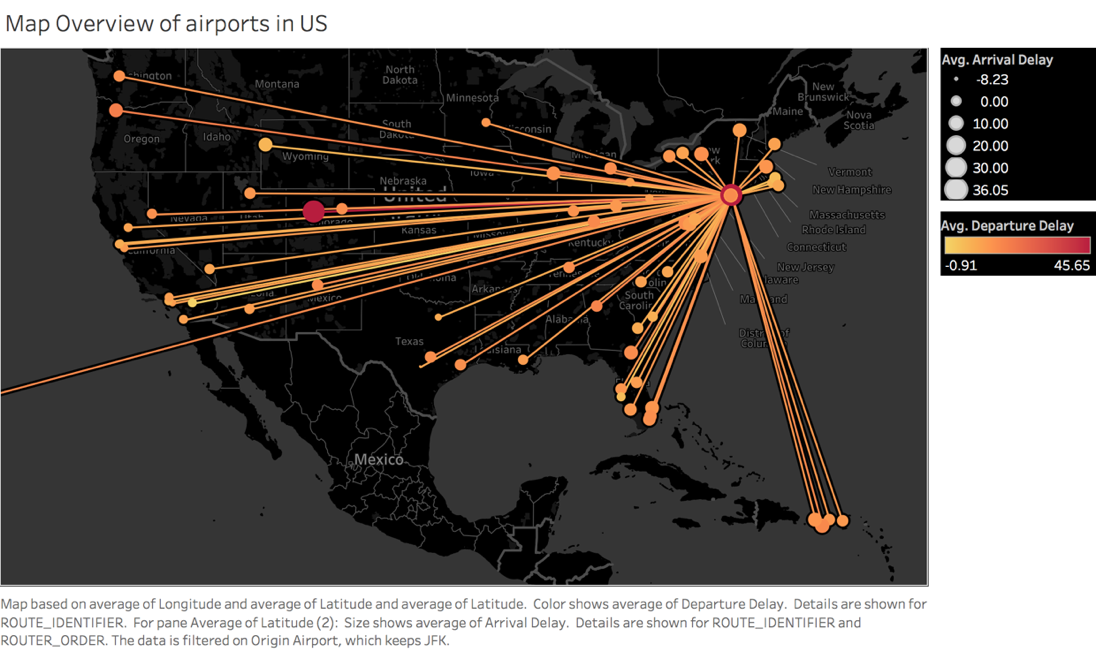


Figure 2.10: The map is created by tableau and using the size of circle to represent the values of average departure delay time. The map shown above is JFK airport, which is one of the most famous airports in the United States, also the top busy airport in the US. The airports are connected using lines and the routes are identified by derived variable route id. In addition, route order shows the order of flight, which can distinguish which is the original airport. The size of points on the map represents the values of average delay time.

Now it is time to build a model for our data. We use time series models to the analysis by aggregating the data and reducing the dimension, because of the extremely large sample size of the data.

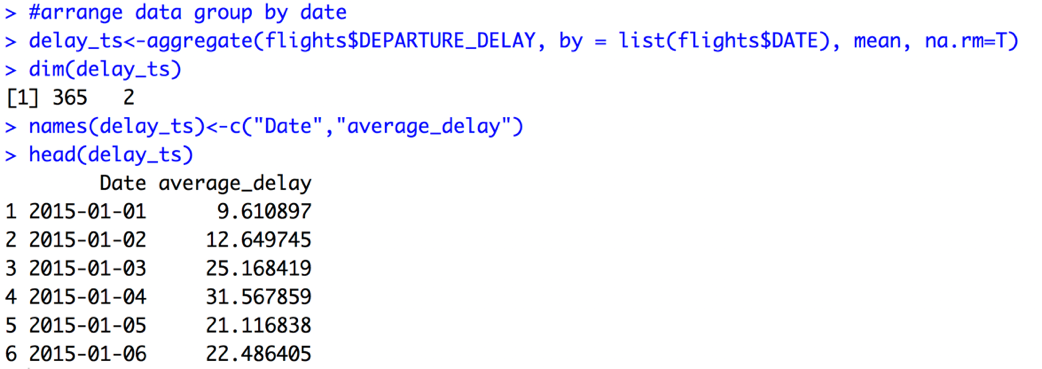
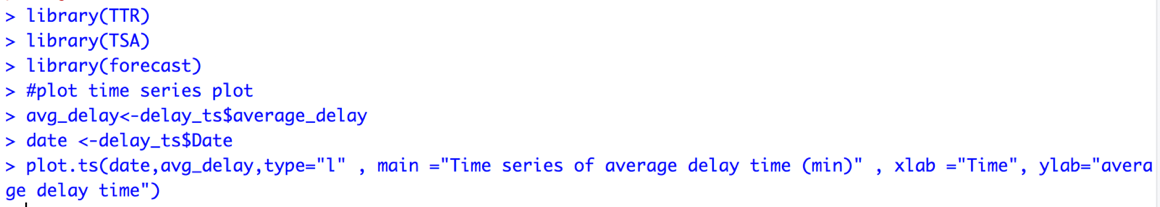


Figure 3.1: The data is aggregated by date, using average value of departure delay time each day.



After using functions in R packages, we create a time series line plot, which is quite similar to the colourful line plot as shown at beginning.

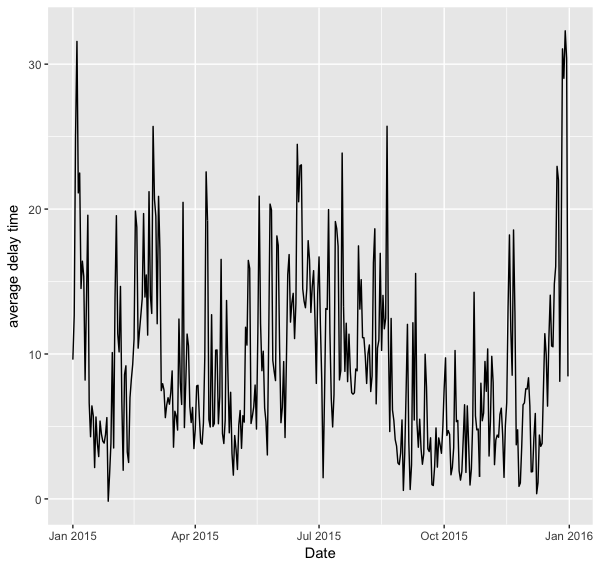
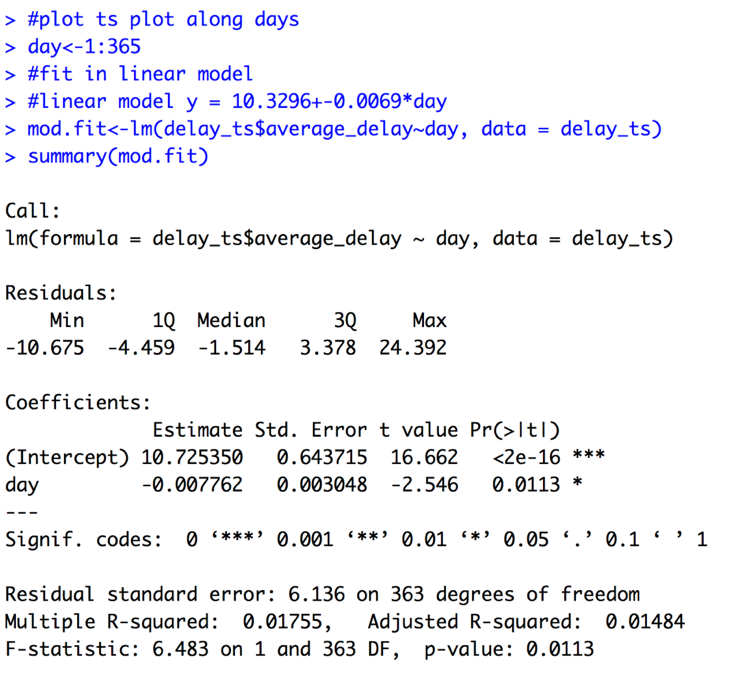


Figure 3.2: The plot is nothing new except the colour of the line, which has exactly same shape and trend create using tableau.

The important part of time series analysis is to fit time in the model, therefore the date variable needs to be relabelled as numerical numbers from 1 to 365 and select column of average delay.



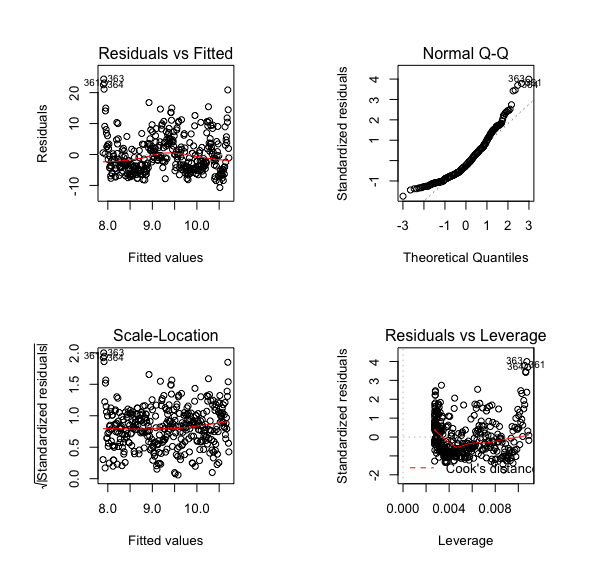


Figure 3.3 & 3.4: The model is a linear regression model fitting average delay with the sequence of number as day. The result shows that the day variable is linearly correlated with average delay, however, since the expected value of residual is not around 0 and Normal QQ plot seems not lie on the straight line, the normality of this model is not good.

Therefore, we need to try higher degree on the day to see if there is any normality.

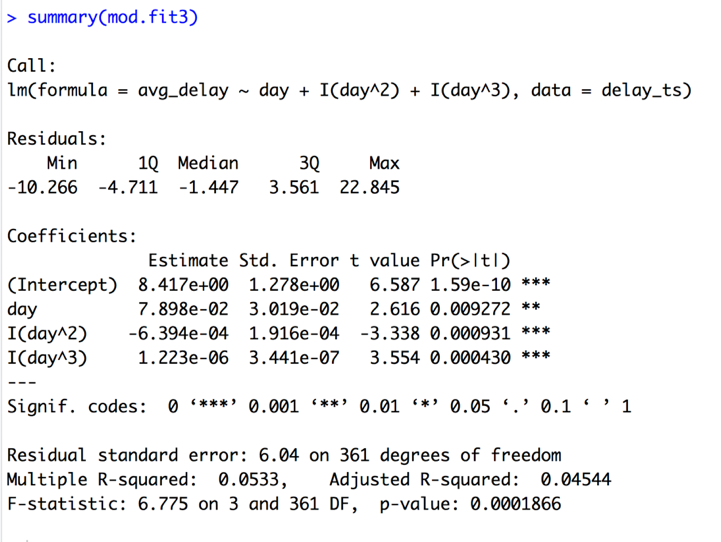


Figure 3.5：The higher degree on day seems to improve our model. The p-value of the overall model is 0.0001866 which small enough for us to conclude that this model is fair good to use even though the R squared value is showing only 4.5% of the variation in average delay can be explained by these day variables.

In order to make our analysis more reliable, we need to start using ACF and PACF plot to analysis our time series data.

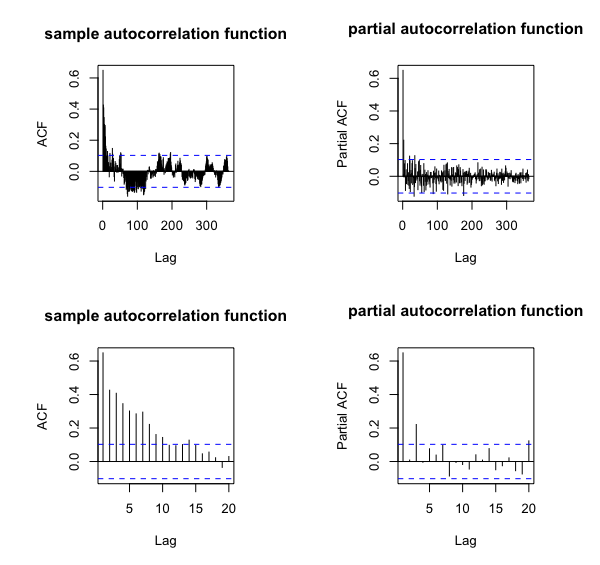


Figure 3.6: The ACF and PACF plot can show whether the data is stationary in time series. The blue line above shows significantly different values than zero. ACF can determine the order in MA(lag) we need to use, whereas PACF can decide the order in AR(lag). Since the plots have too many lags, we specify the lag =20 so we can find the order in MA and AR more clearly. In ACF plot, the order of MA may be 10 and order of AR may be 3.

Moving average and autoregression are two important models used in time series, which can explain the correlation of our target variable and time. In this project, we use some useful functions and tests in R to determine the order of MA and AR also find out the lag in difference. The difference in time series is also very important since it can make our results much meaningful.

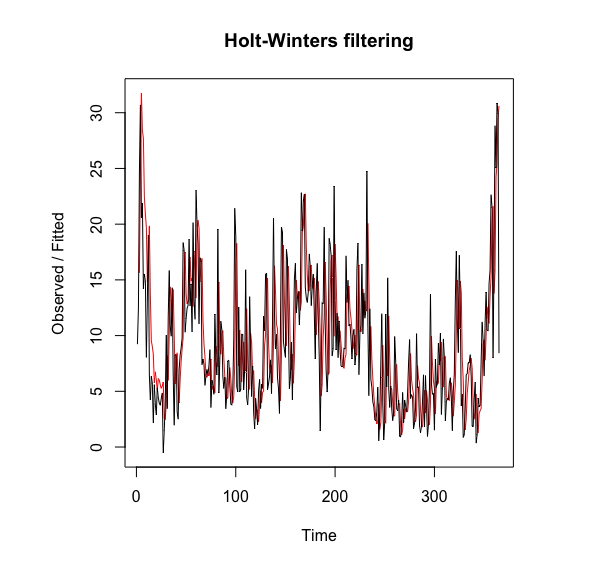


Figure 3.7: The Holt-Winters filtering plot shown above implies this method of forecasting seems well fitting along the real curve in red. Hence, we are using this method to forecast next 100 days in the future.

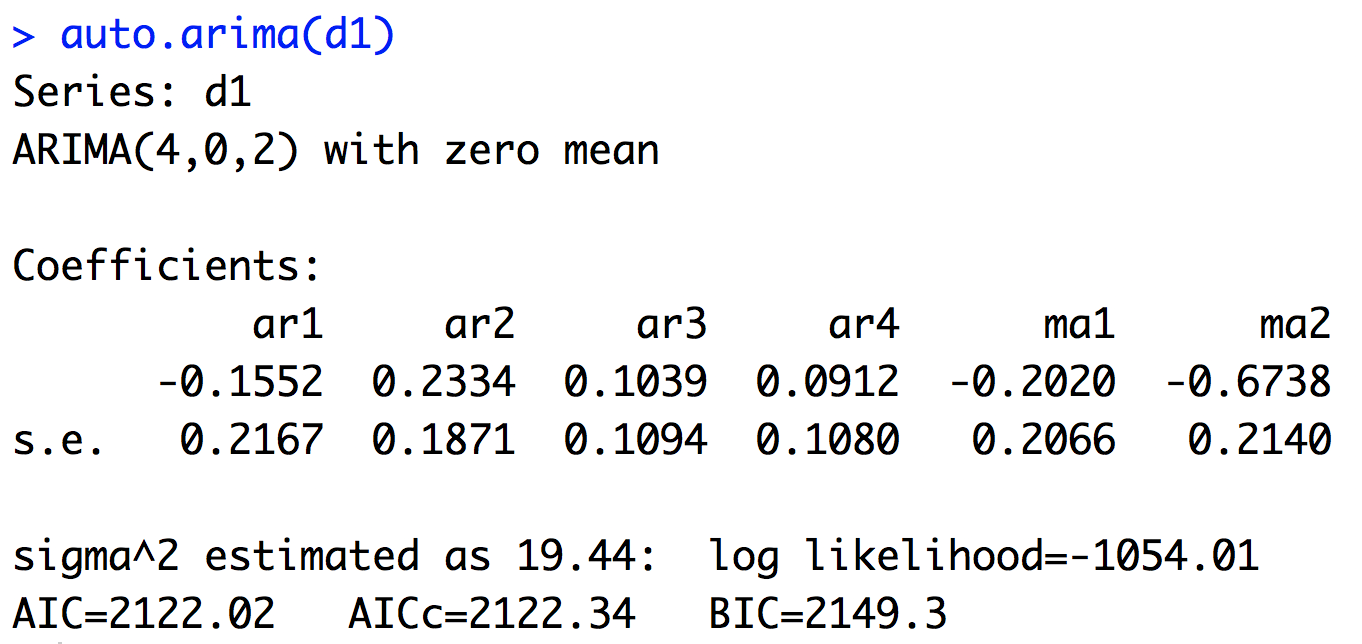


Figure 3.8: The function auto.arima in R, can calculate the order of MA, Difference and AR. The output of auto.arima gives us the coefficient and standard error of each parameter in the model. MA (4) and AR (2) with lag = 0 in difference. The seasonal =FALSE means we remove the seasonality from the model.

Using ARIMA notation introduced above, the fitted model can be written as:

Y = -0.1152\*Yt-1+0.2334\*Yt-2+0.1039\*Yt-3+0.0912\*Yt-4-0.2020et-1-0.6738et-2 +€

where € is a random error in the model and order of difference is 0.

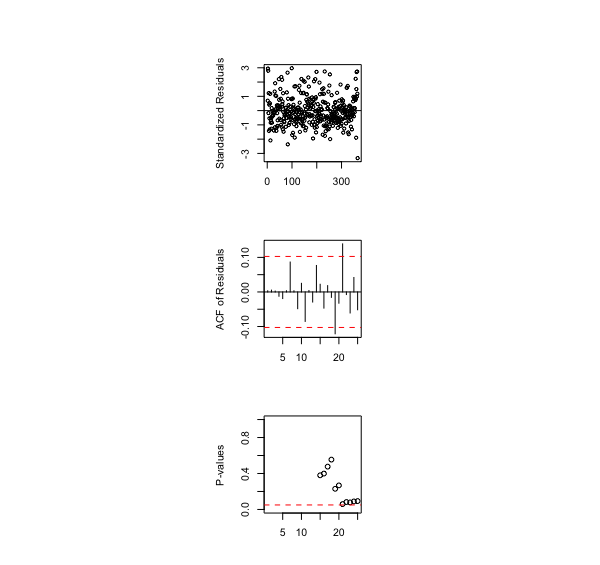


Figure 3.9: The plots above are the plots of residuals in given model, ARIMA (4,0,2). The residuals scatter plot is constant around zero, also in sample autocorrelation function plot, the dashed line covers the most of lines in given lag. Therefore, the model has normality and stationary.

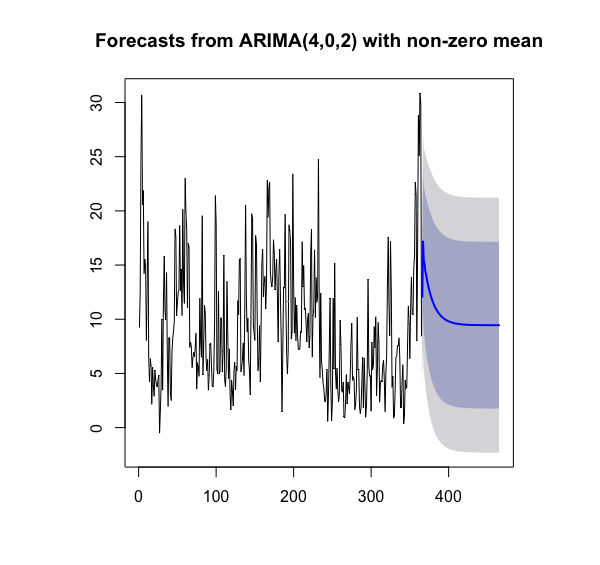


Figure 3.10: The forecast plot provided the prediction for next 100 days in the blue line, the blue shaded area is confidence and grey bound is prediction interval. The bounds are reasonable since the flexibility in average delay.

# Conclusion

In summary, after the exploration of the data, we can find some important patterns in seasonality as we look at the line plot and time series plot. The peak delay time and locations are the holiday time in a year and in the top busy airport in the United States. There are some differences between different airlines, in both cancellations and delays. Travellers may need to choose the right time, such as not a holiday period, to choose a right airport to start from, not in busy airports. Besides, to choose good reputation airlines is always needed.

# Reflection

Data exploration in flight delays and cancellations will give us a general idea of how to avoid some of the flight delays and cancellations in our real life. However, the most related reasons such as weather delays, airline delays are not reported, therefore even the data is released by official data source and large sample size, we cannot be very convinced by results of models using limited and inaccurate data. As long as we are only interested in the trends and predictions, the data used in the project may be good but still need more shreds of evidence for further analysis and explorations.

# Appendix and References

https://www.kaggle.com/usdot/flight-delays

https://en.wikipedia.org/wiki/List\_of\_the\_busiest\_airports\_in\_the\_United\_States

https://www.transtats.bts.gov/homedrillchart.asp