**Prediction Analysis of Canceled and Delayed Flights of the US Airlines (2018)**

*Aparna Biswas (21060641008), Sumantra Kumar (21060641051), Sumeet Gaikwad (21060641052) and Vidya Mallick (21060641056)*

**1. Introduction**

**1.1 Situation overview**

The major functions of an airport contain operational services (all coordinational activities), traffic handling services (scheduling) and commercial services (hospitality to travellers within the airport, customers to shops, and customer service). The importance of understanding flight delays and cancellations is crucial to this functioning.

**1.2 Key Stakeholders**

There are two perspectives: Management and Customers

**Management:** Air traffic control and management is rather complex and needs detailed planning. The planning process relies on accurate prediction of delays, cancellations and on-time flights. The resource availability would be arranged according to these predictions. Moreover, security and safety are a top priority during flight operations. A flight cannot be landed anywhere/anytime and stopped mid-air. This means knowing flight routes (**origin** and **destination**), which flights would be on those routes for what **distance** and **time** become important along with other factors such as Weather Conditions, National Aviation System, Security etc. An appropriate prediction model would **increase operational efficiency** and **reduce costs** due to the wastage of material due to last-minute changes. These become crucial especially during high-traffic days/time, for smooth functioning.

**Customers:** Travelers are the backbone of the air industry and contribute significantly to the economy in terms of business, tourism etc. A daily flight needs to have multiple availabilities across airlines because even if one airline is not available, an alternate should be scheduled immediately. A model that can predict the **delay time** well in advance would help reschedule such flights, ensuring that travelers reach their destination on time. Secondly, air traffic is high during the holiday season and weekends with tourists and families. Amenities to be provided would rely on these prediction models.

Other activities are baggage loading and unloading mechanisms, baggage delivery conveyor belts, and passenger processing systems, which are also an important part of the management which again is crucial to the **actual** and **scheduled time of flights.**

**1.3 Objective:** This report aims to discuss the prediction analysis of cancelled and delayed + delay time for flights of 28 US Airlines. Furthermore, using this analysis to provide the necessary information to the management for proper scheduling of flights for efficient and smooth functioning of the airport operations. The goal is to understand the pattern of the same and reduce the delay by atleast 5%.

**1.4 Questions Addressed:**

1. Which Airlines have the highest number of delayed or cancelled flights?
2. How accurately we can predict these delays and cancellations?
3. The reasons for these delays and cancellations and how it impacts the business.
4. What measures can be taken to control this?

**1.5 About the dataset:**

The data contained in the compressed file has been extracted from the Marketing Carrier On-Time Performance (Beginning January 2018) data table of the "On-Time" database from the TranStats data library.

There are 61 variables and 1048575 observations. The majority of the variables are different time records or variables that are observed as part of activities in the airport.

There are 6 variables associated with calendar entries of the flight date including year, quarter, month, day of the month, day of the week and flight date.

14 variables discuss the various marketing and operating IDs associated with each airline, these also include the tail number which is unique to each airplane.

18 variables describe the origin and destination IDs associated with the airport.

Further, data on estimated and actual flight timings (arrival and departure timings), flight landing (*wheels-on and wheels-off*), runaway taxi durations, airtime, distance, flight diversion, flight cancellation were provided which account for the remaining 23 variables.

**2. Discussion:**

**2.1 Pre-processing:**

* Cleaning for Cancelled and Delayed Flights:
  1. In theory, a cancelled flight should not have an arrival or a departure time. Corresponding rows containing arrival or departure time were removed (accounting for 500 observations).
  2. Wherever CRS arrival and Arrival time were equal, 0 was imputed in order to remove null values, Same was done for CRS departure and departure time.
* Categorical Variables:
  1. Since one hot encoding or dummy variables would blow up the dimensions of the dataset due to the presence of a large number of unique values in each column.
  2. The idea was to trace back each and every airline, origin, destination etc which accounted for the highest proportion of cancelled flights, i.e. a rank of 1 was assigned to the airline, origin, destination etc. having the highest proportion of cancelled flights and so on.

**2.2 Exploratory Data Analysis:**

1. **Origin Airline Distribution:** The distribution of the origin of airlines is as observed, more than 50% of the overall airline distribution is majorly covered by just three airlines namely Southwest Airlines co., Delta Air Lines Inc. and SkyWest Airlines Inc. Southwest has the maximum flights originating (departing) with 25.8 %. Followed by Delta Airlines Inc. with 18.9 %. And 12.8% of SkyWest Airlines Inc. United Air Lines Inc and American Airlines come close to the majorly originated Airlines by 11.7% and 7.98% respectively.
2. **Origin State Distribution:** Illinois and Georgia are the two highest, with around 60 thousand flights. Followed New York, and Colorado. with 40 thousand flights. The rest of the states have a range of around 28 - 35 thousand flights.
3. **Number of Flights (day of the week):**  It is observed that the first three days of the week, Monday, Tuesday and Wednesday have a significantly higher number of flights which are around 1.7 lakh flights as compared to other days of the week. On Thursdays and Fridays, the number of flights is 1.4 lakh. Now for the weekend days which are Saturday and Sunday. There are the least number of flights on Saturday among all other days. And on Sunday the number of flights is 1.35 lakh.
4. **Air Traffic by cities:** Based on the day of the week, the first three days that are Monday, Tuesday and Wednesday (Chicago, IL and Atlanta GA), have maximum air traffic which is around 10 thousand flights, whereas the remaining days have a range of 7 - 9 thousand flights. New York NY, Denver CO, and Houston TX, have relatively less air traffic in the range of 5 - 7 thousand. Los Angeles, CA and Washington, DC comparatively have the least Air traffic among other cities in the range of 6 - 4 thousand flights.
5. **Cancelled Flights:**

* On average 1.6% of flights of every airline get cancelled.
* **Airlines:** The top 5 airlines with the highest percentage of cancelled flights are: Peninsula Airways Inc. (11.04%), Alaska to Alaska, Commutair Aka Champlain Enterprises, Inc. (4.04%), Eastern America, Endeavor Air Inc. (3.6%), Eastern America, ExpressJet Airlines Inc. (3.4%), 80% of the States and Mesa Airlines Inc. (3.4%, 80% of the States).

The lowest cancellations were from Cape AIr (0%), Hawaiian Airlines Inc., Compass Airlines, Allegiant Air and Alaska Airlines Inc. with an average of 0.37% cancellations.

* **States:** The top 5 States with the highest cancellations are: Vermont, South Carolina, Maine, Massachusetts and New York with an average of 3.59% cancellations.

The lowest cancellations are from Hawaii, Utah, Arizona, Nevada and Montana with an average of 0.6% cancellations.

* **Distance:** The average distance covered by Peninsula Airways Inc. is 792 miles whereas the remaining are within 500 miles. Flights from Vermont, South Carolina and Maine have a distance within 500 miles, and Massachusetts and New York have a distance of (974 and 845 miles) respectively.
* **Day:** Thursdays have the highest number of cancellations.
* **Date:** The maximum number of flights were cancelled (almost 8%) on January 4th which also happens to be a Thursday.
* **Month:** January has more cancellations than October.

1. **Delayed Flights: *Flights with reported time greater than 15 minutes***

* On average 15% of flights of every airline get delayed.
* **Airlines:** The top 5 airlines with the highest percentage of delayed flights are: Commutair Aka Champlain Enterprises, Inc. (29.04%), Frontier Airlines Inc. (25.75%), JetBlue Airways. (23.5%), Trans States Airlines (23.09%) and ExpressJet Airlines Inc. (3.4%), (80% of the States included).

The lowest delays are from Cape Air, Empire Airlines Inc., Hawaiian Airlines Inc., Delta Air Lines Inc, and Spirit Air Lines with an average of 10.04% delays..

* **States:** The top 5 States with the highest delays are: Massachusetts, New Jersey, Vermont, Illinois, and Maine with an average of 21.23% delays.

The lowest delays are from U.S. Pacific Trust Territories and Possessions, Alaska, Utah, Hawaii and U.S. Virgin Islands with an average of 9.9% delays.

* **Distance:** The average distance covered by Commutair Aka Champlain Enterprises, Inc, Trans States Airlines and ExpressJet Airlines Inc. is less than 500 miles whereas Frontier Airlines Inc., and JetBlue Airways. are 1050 miles. Flights from Vermont and Maine have a distance within 500 miles, whereas Massachusetts, New Jersey, and Illinois have a distance of (690, 961 and 974 miles) respectively.
* **Average Flight Time Duration:** Based on Airlines (1.5 hours) and based on States (1.5 hours).
* **Percentage Delay (Months):** In the month of January, the departure delay is higher than the arrival delay, but not significantly. It is observed that in January 32% of flights got delayed for departure. And, 30% of flights got delayed for arrival at the required destination.

In the month of October, the arrival delay is higher than the departure delay. The same observations were noted in October which is 30% of flights got delayed for arrival. And, 33% of flights got delayed for departure.

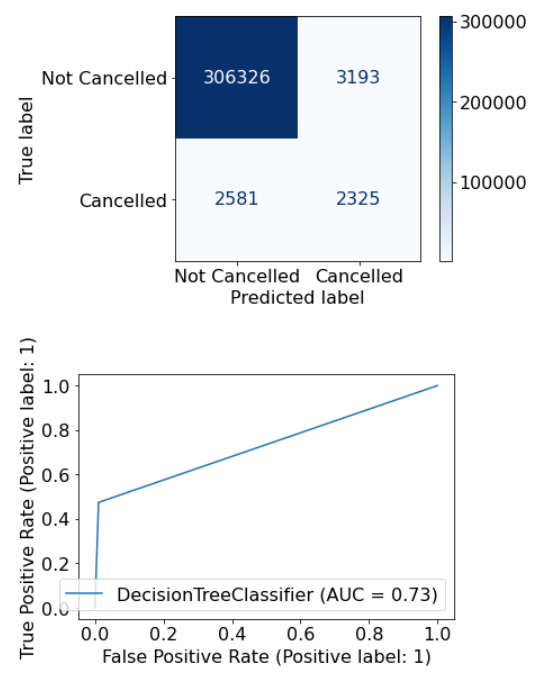
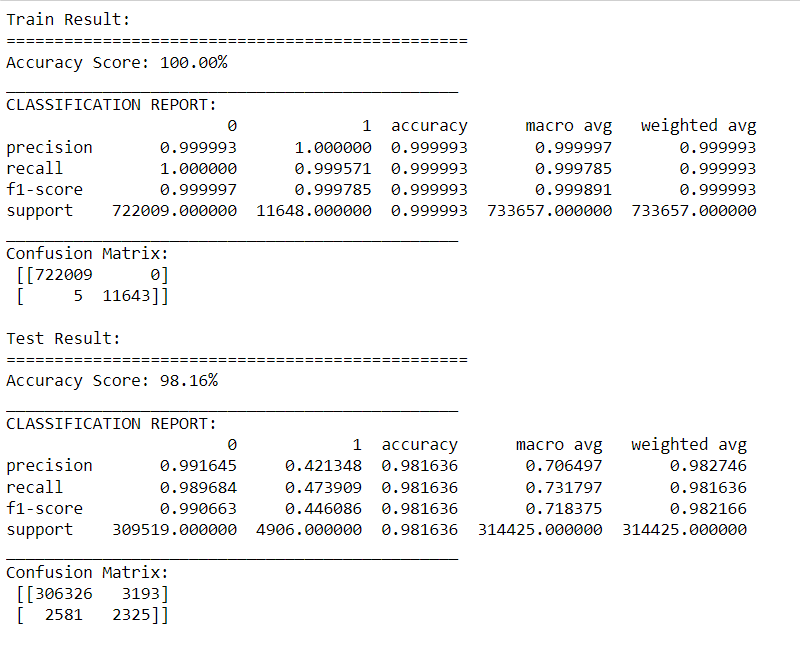
* **Arrival and departure delay by airlines:** Around 11 airlines have arrival delays of more than 5 mins and the rest have arrival delays of less than 5 mins. Airlines such as Champlain Enterprises Inc, Trans State Airlines Frontier Airlines, JetBlue, GoJet Airlines, Mesa Airlines ExpressJet Airlines, SkyWest airlines American Airlines and Peninsula Airlines have higher arrival delay times. The same respective airlines have a higher rate of departure delay which is in the range of 20 mins to 7 mins.
* **Taxi In & Taxi Out:** Average Taxi-out time is 20 minutes based on States and Airlines. Taxi In time is in the range of 8 - 10 minutes. Frontier Airlines Inc. has the highest Taxi In time.

**2.4 Modelling**

**2.4.1 Predicting Cancelled Flights (Decision tree)**

The following (8) variables were considered for modelling flight delays (using Decision Tree):

*Airline, Origin, Destination, Day Of Week, Distance, Month, Day of the month, CRS Departure Time.*

****

The metric chosen for model evaluation is Recall. With cancelled flights labelled as 1 and 0 otherwise, it is very important to correctly predict if a particular flight would be cancelled or not. Thus it is important to avoid false classification of a cancelled flight.

Based on the chosen metric, it was observed that the Decision Tree with 8 variables performed the best among all the combinations of all the variables.

It is really important to check if our model is able to predict cancellation of flights accurately or not, i.e. even though our test accuracy is 98.16% but our recall for predicting ‘1’ came out as 0.48 which is as good as an unbiased coin flip. The reason for this has been further discussed in Data Discrepancies.

**2.4.2 Predicting Delayed Flights (NB Classifier & Random Forest)**

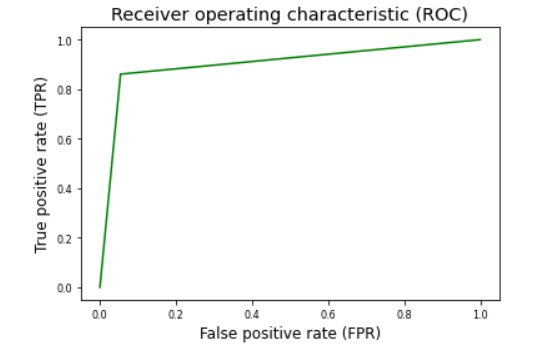
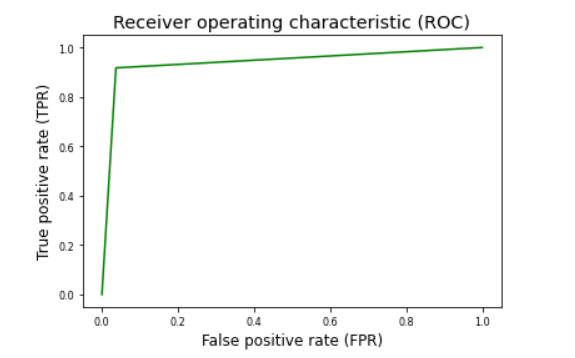
The following (19) variables were considered for modelling flight delays (using Random forest):

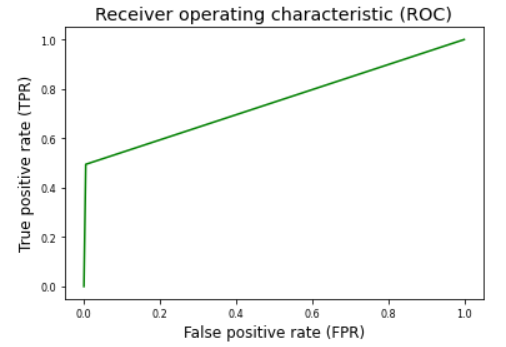
*Airline, Diverted, Arrival time, Air time, Distance, Quarter, Month, Day of the month, Origin airport ID, Destination airport ID, Taxi-In, Taxi-out, Wheels-on, Wheels-Off, Distance group, Departure time bulk group, Arrival time bulk group.*

The second model implemented was a random forest with the above-mentioned variables except for *Diverted, Arrival time, Wheels-on* and *Wheels-Off*.

These variables were dropped since data on these variables would not be available beforehand. Taxi-In and Taxi out are assumed to be standard durations for a given airport.

|  |  |
| --- | --- |
| **Model** | **Recall** |
| Random Forest (19 vars) | 91.76% |
| Random Forest (14 vars) | 86.08% |
| Gaussian Naive Bayes (14 vars) | 49.46% |

****

****

The metric chosen for model evaluation is Recall. With delayed flights labelled as 1 and 0 otherwise, it is very important to correctly predict if a particular flight would be delayed or not. Thus it is important to avoid false classification of a delayed flight.

Based on the chosen metric, it was observed that the Random Forest with 14 variables performed the best.

**2.4.3 Predicting Delayed Flights Duration (MLR & Random Forest)**

The following (9) variables were considered for modelling flight delays (using Random forest):

*Airline, Departure time, Arrival time, Air time, Distance, Day of Week, Day of the month, Taxi Out, Distance Group.*

The second model implemented was a random forest with the above-mentioned variables.

These variables were dropped since data on these variables would not be available beforehand. Taxi-In and Taxi out are assumed to be standard durations for a given airport (i.e. average taxi in and taxi out time).

|  |  |  |
| --- | --- | --- |
| **Metric** | **MLR** | **Random Forest** |
| MAE | 46.16 | 44.49 |
| MSE | 6929.65 | 6479.35 |
| RMSE | 83.24 | 80.49 |

Clearly, Random Forest performed far better in comparison to multiple linear regression but further optimization is possible.

**3. Conclusions:**

**Air Traffic:** State wise, Illinois, Georgia, New York and Colorado,  Airline wise Southwest Airlines co., Delta Air Lines Inc. and SkyWest Airlines Inc, the first three days of the week have the highest air traffic. Colorado, New York and Georgia as well as the three aforementioned airlines have done fairly well in ensuring low canceled and delayed flights despite high air traffic,

**Cancellations:** Peninsula Airways Inc. had the highest number of cancellations (11.04%) and much higher than the average 2-3 % reason for which can be due to cancelled service via Boston to and from Plattsburgh, Bar Harbor and Presque Isle a month early, citing staffing issues, breaking their Essential Air Services contract with the Department of Transportation.

January 4, 2018 had the highest number of cancellations. This can be attributed to the fact that two airplanes collided on the tarmac at John F. Kennedy Airport amid rampant delays caused by [a powerful Northeast winter storm](https://www.cnbc.com/2018/01/06/blizzard-snarls-air-traffic-at-jfk-as-2-planes-collide-and-hundreds-of-flights-are-delayed.html).

The remaining airlines have 3-4% cancellations. It is important to reduce this by at least 2%.

**Delayed Flights:** Commutair Aka Champlain Enterprises, Inc. has the highest number of delayed flights commuting from majorly Eastern States(). Most of the flights are originating from North-East America, which is subjected to strong winds. As the months of January and October are given for our analysis, both months being subject to strong winds can be a reason for high flight delays. It is also interesting to see that the percentage of arrival and departure delay is almost the same overall. But, if we break it down day-wise, the percentage of departure delay is higher. This means, airtime is reduced by increasing the speed of the aircraft and the process.

**Distance:** In terms of cancellation Peninsula Airways Inc.had the largest distance of 792 miles and for delayed flights Frontier Airlines Inc., and JetBlue Airways. Had an average distance of 1050 miles which is significant to note as alternative flights are crucial to be scheduled as alternative transportation to cover such long distances become difficult.

**Modelling:**

For each of the three outcomes: Canceled/Not Canceled, Delayed/Not Delayed and Delay time, the underlying pattern of the data and key influential factors vary significantly. Hence, three different models were deployed. These models can however be layered one on top of the other to create a pipeline which can predict all three outcomes.

The flow of the model would be in the following sequence:

1. Flight cancellation status
2. Flight delay status
3. Flight delay time

**4. Future Scope and Recommendations:**

* Additional data on weather delay, National Aviation System delay, security delay etc and cancellation reasons could provide more information for better and more accurate models.
* Data on air-traffic could improve flight delay forecasts.
* Large data requires greater computational powers. Hence, appropriate resources are required for handling such data.

**5. Data discrepancies**

* Data inconsistency - There were records of cancelled flights which had a departure time associated with them.
* Further, many variables were a linear combination of other variables. These variables don’t add extra information.
* It must be noted that weather is one of the most influential factors for forecasting flight delays/cancellations. Data on weather conditions could improve forecasts.
* After conducting PCA on the cancelled dataset, the variation explained by the data was too less which might be attributed to limited data available for the cancelled dataset, i.e. more data is required to better fit a model to predict cancellation of flights.

**6. Charts and Tables:**

Figure 1 Origin Airline Distribution

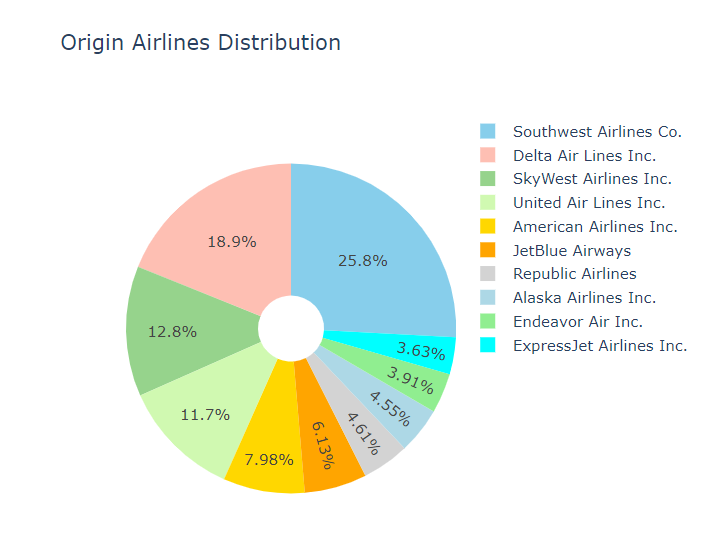
****

Figure 2 Correlation Matrix of City and Day of the Week

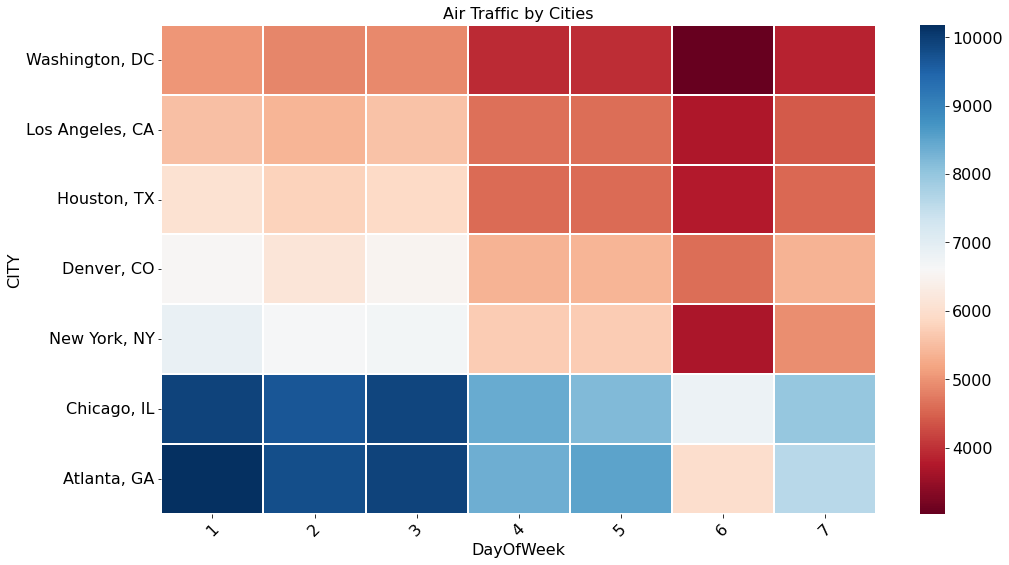
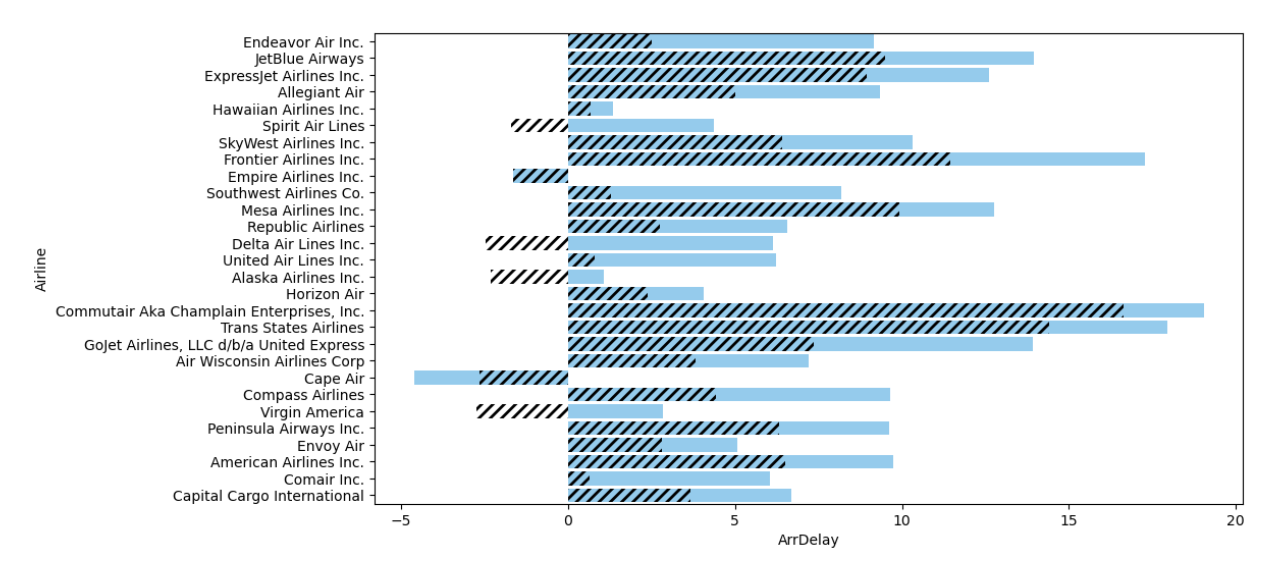
****

Figure 3 Airline and Dalay

****

**Hashes: Arrival Delay and Blue Region is Departure Delay**

Figure 4 Summary table Origin States

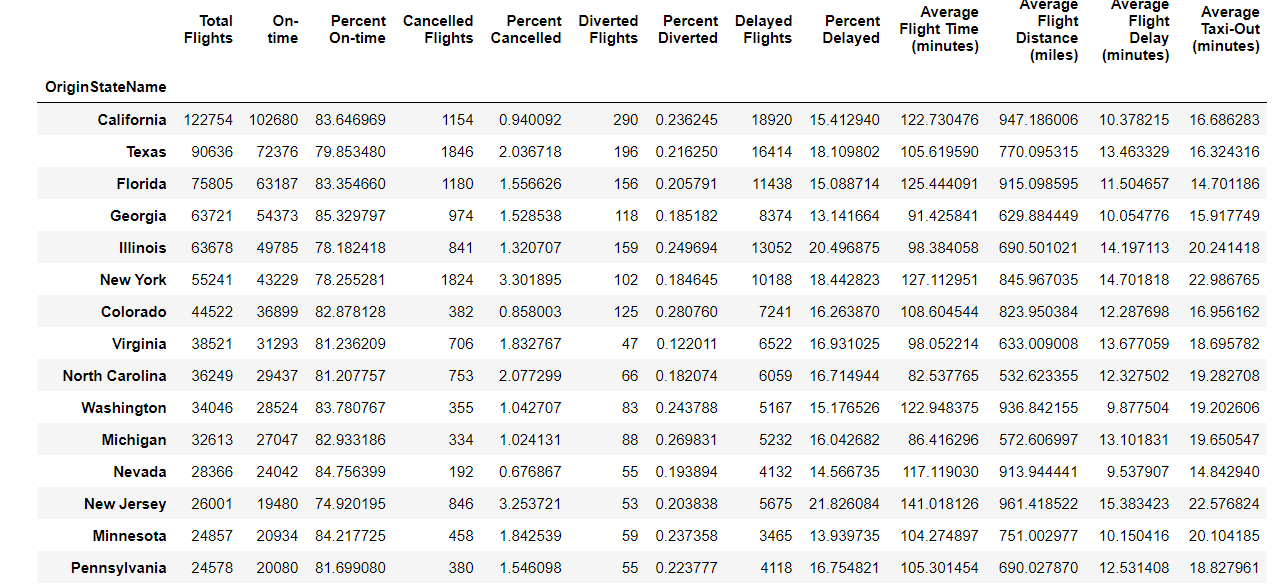
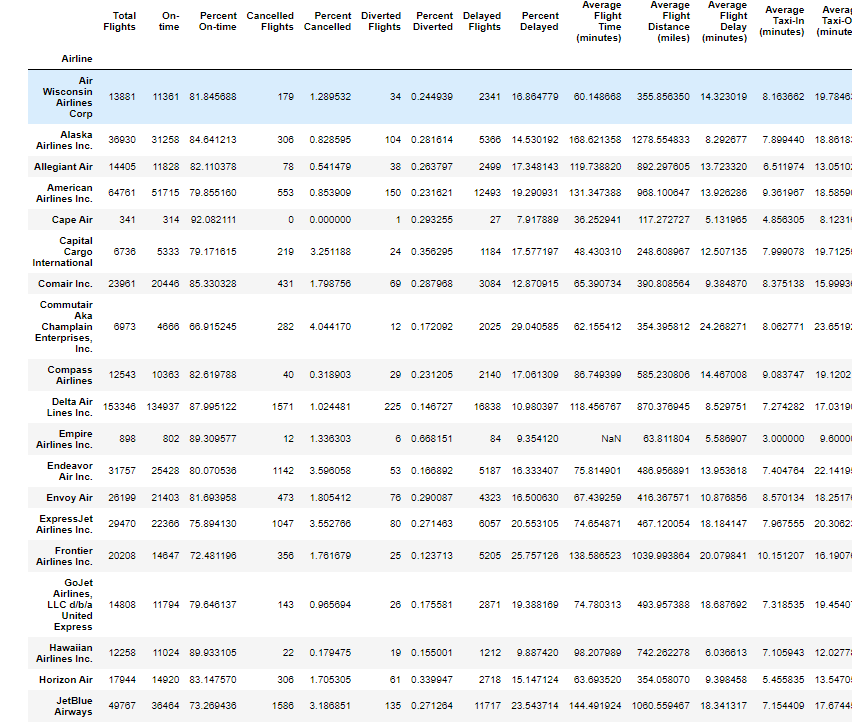
****

Figure 5 Summary table Airlines

****