Ban Qu

**Final Report:**

**U.S. Flights Delay Analysis**

**Problem Statement**

Flight delays and cancellations are always a major topic in air travel service. The Bureau of Transportation Statistics (BTS) has been recording airline customer data to help decision makers better understand statistics on transportation. To be specific, the goal is to improve operational efficiency, to examine emerging topics, and to create relevant and timely information that foster understanding of transportation and its transformational role in society.

The present study contains a 2015 dataset with 28 features that could be used to predict a potential flight delay in U.S. By using and analysing the data via Python and Tableau, various visualization and machine learning models are deployed to find patterns and better predict unseen delays.

The final dashboard of Tableau enables an overview summary of delayed information that involves an interaction among cities, airlines, and time. Machine learning techniques involves K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes, and Gradient Boosting. The best performances were among KNN, Random Forest, and Gradient Boosting, and the latter two agree on the most predicting feature of arrival delay of U.S. flights as departure delay. Air System delay is the second main reason for causing arrival delays.

**Data Wrangling**

The raw dataset involves 3 different files: airlines, airports, and flights. Airline data contains 14 unique airlines and their codes, and the table contains no missing data. Airports data contains 322 rows and 7 columns that are related to locations of airports. There are 3 missing values on latitude and longitude, respectively, and therefore removed from this dataset.

Flights data contains 5819079 rows and 31 columns. “ARRIVAL\_TIME” and “SCHEDULED\_ARRIVAL” are removed because “ARRIVAL\_DELAY” simply summarizes these two features. Similarly, “SCHEDULED\_DEPARTURE” and “DEPARTURE\_TIME” are removed because “DEPARTURE\_DELAY” summarizes the two. Columns not interested in the current study are also dropped: “TAXI\_OUT”, “TAXI\_IN”, “WHEELS\_OFF”, “WHEELS\_ON”, “AIR\_TIME”, “SCHDULED\_TIME”, and “ELAPSED\_TIME”. Missing values under “TAIL\_NUMBER” are dropped since missed aircraft identifier rows are useless.

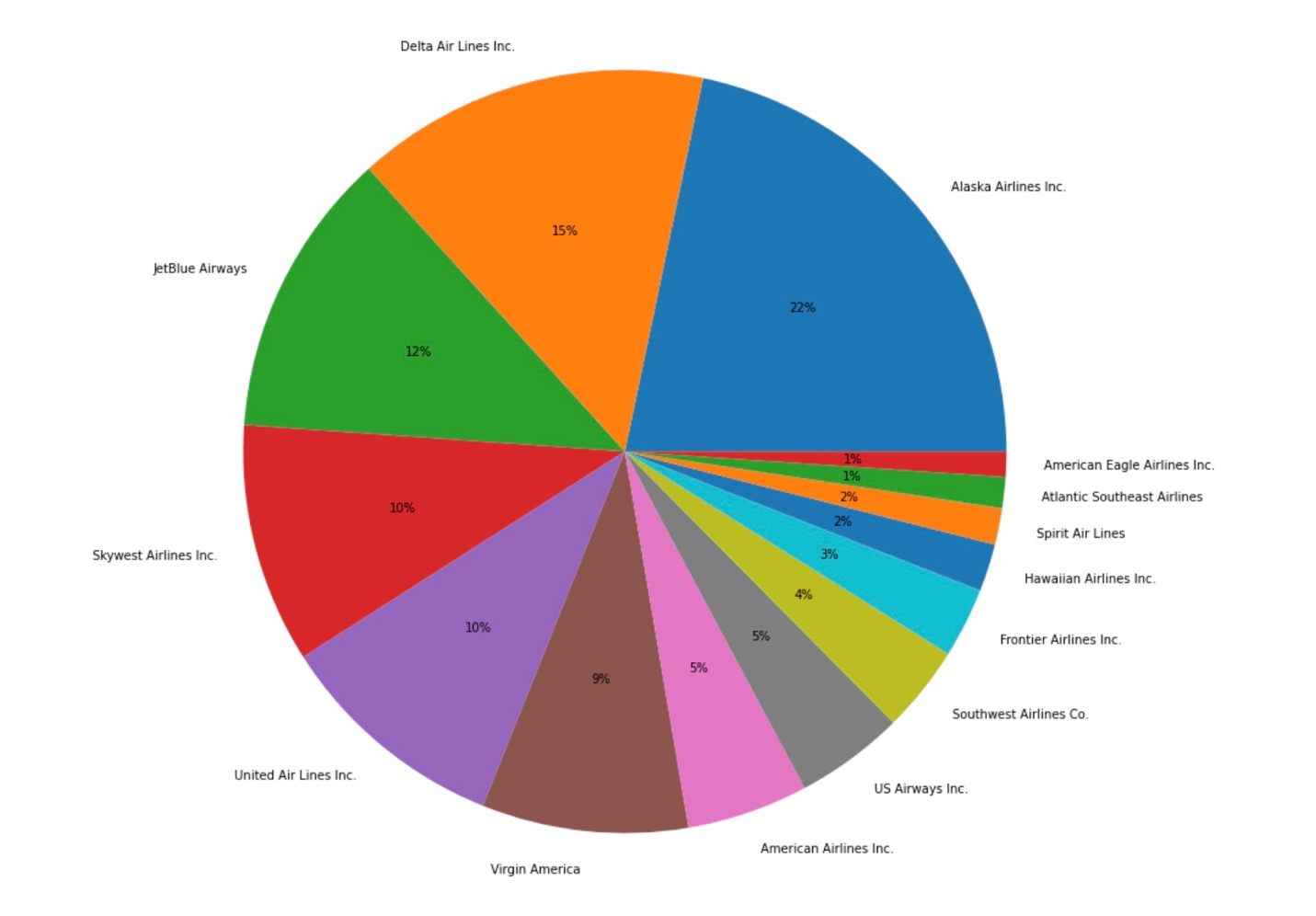
A large portion of missing values are found in “CANCELLATION\_REASON”, “WEATHER\_DELAY”, “LATE\_AIRCRAFT\_DELAY”, “AIRLINE\_DELAY”, “SECURITY\_DELAY”, and “AIR\_SYSTEM\_DELAY”. In this case, missing values under these columns are filled by zeros because a closer look at when some delays equal to zeros indicates that there’s no impact on the flight by these data that causes a delay. For “CANCELATION\_REASON”, a relabelling is made that transforms “A” (Airline/Carrier), “B” (Weather), “C” (National Air System), “D” (Security) into “1”, “2”, “3”, and “4” respectively.

Three datasets are then merged based on an inner join of “IATA\_CODE” that are unique identifiers of three tables. A check of duplicated columns indicates that “COUNTRY\_x”, “COUNTRY\_y”, “STATE\_x”, “STATE\_y”, “IATA\_CODE\_x”, and “IATA\_CODE\_y” are duplicated, and thus removed. Some column names are renamed and all column names are made in lower cases.

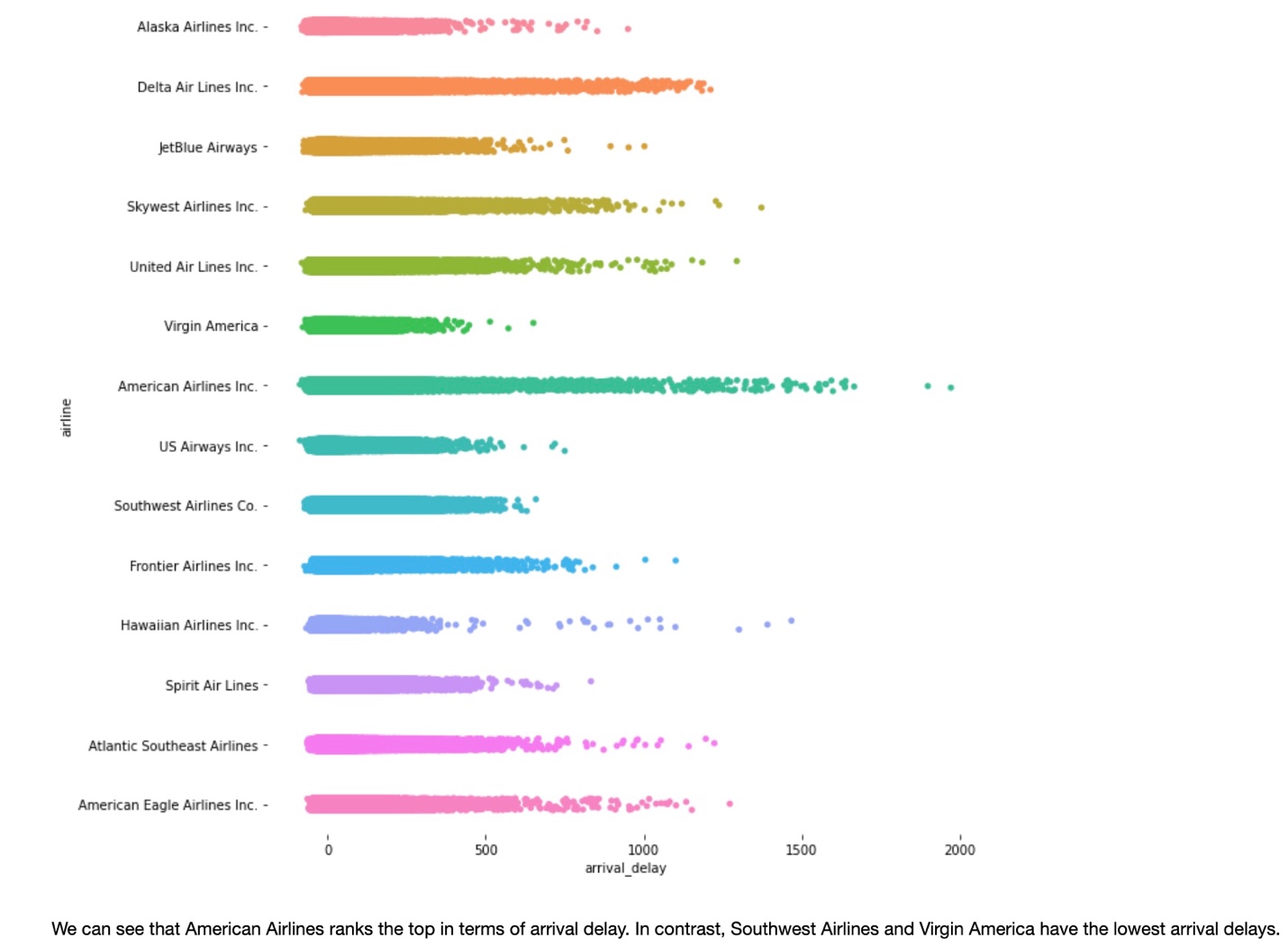
The final cleaned dataset contains 5309350 rows and 28 columns.

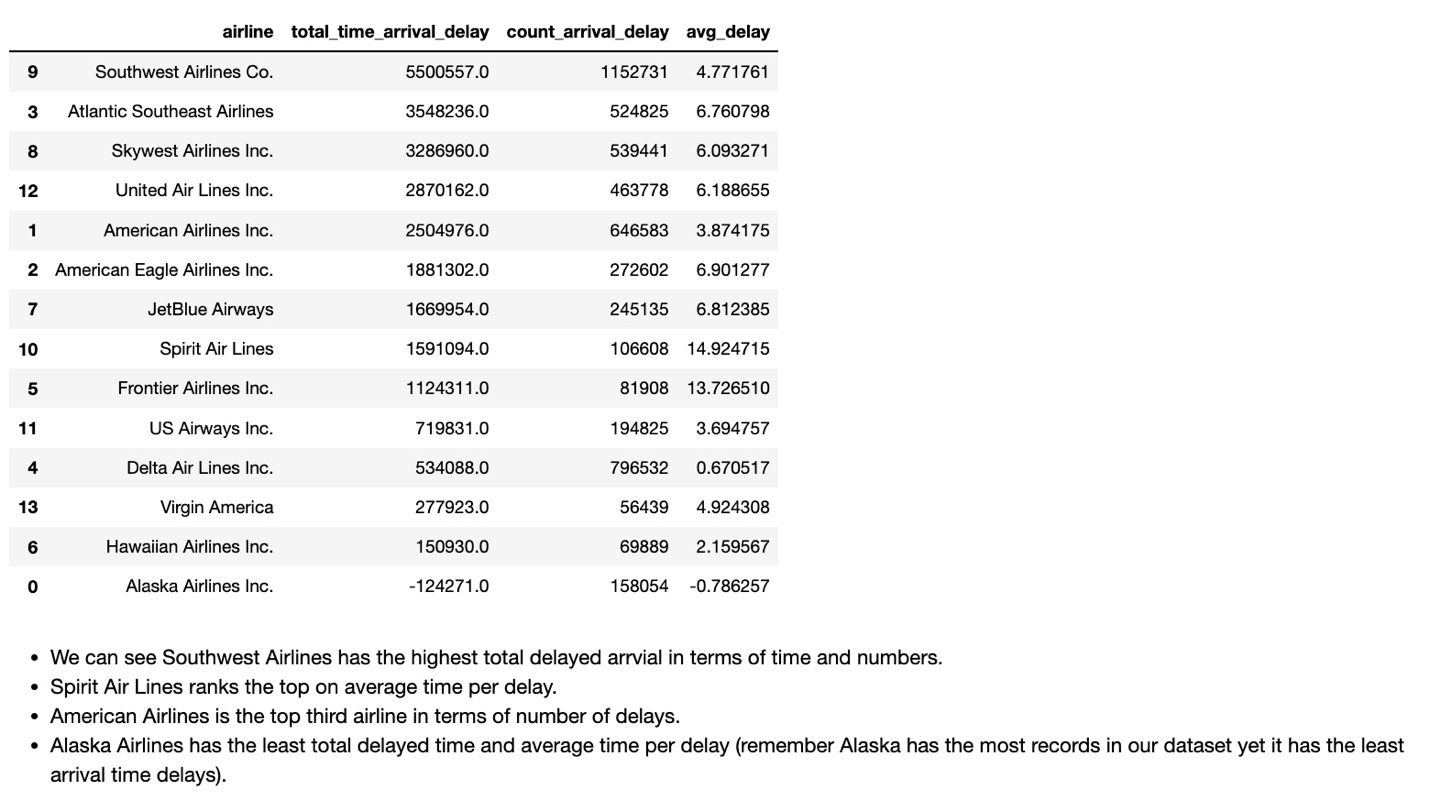
**Exploratory Data Analysis**

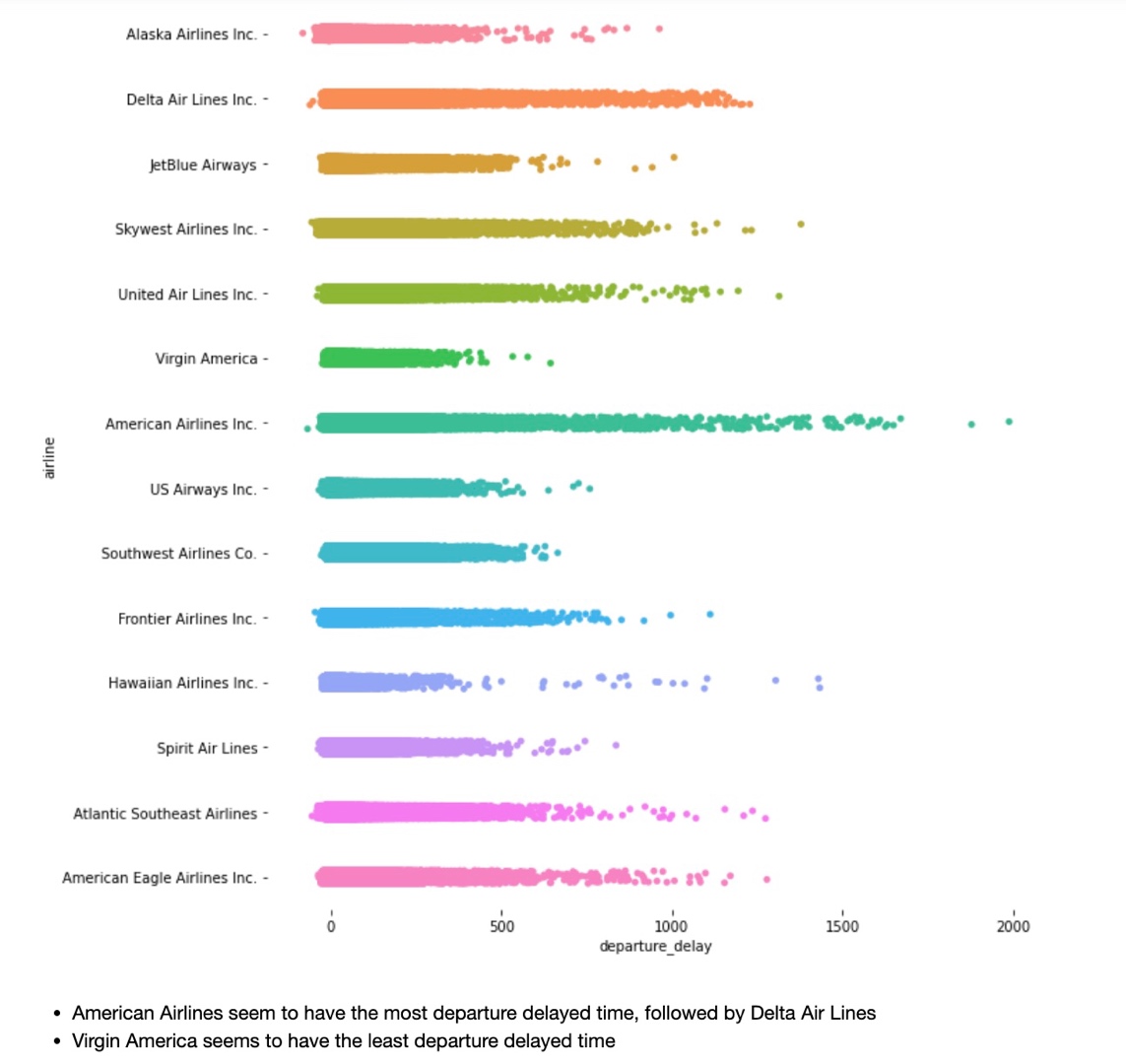
Using Seaborn, Matplotlib, and Tableau, some findings are worth noticing.

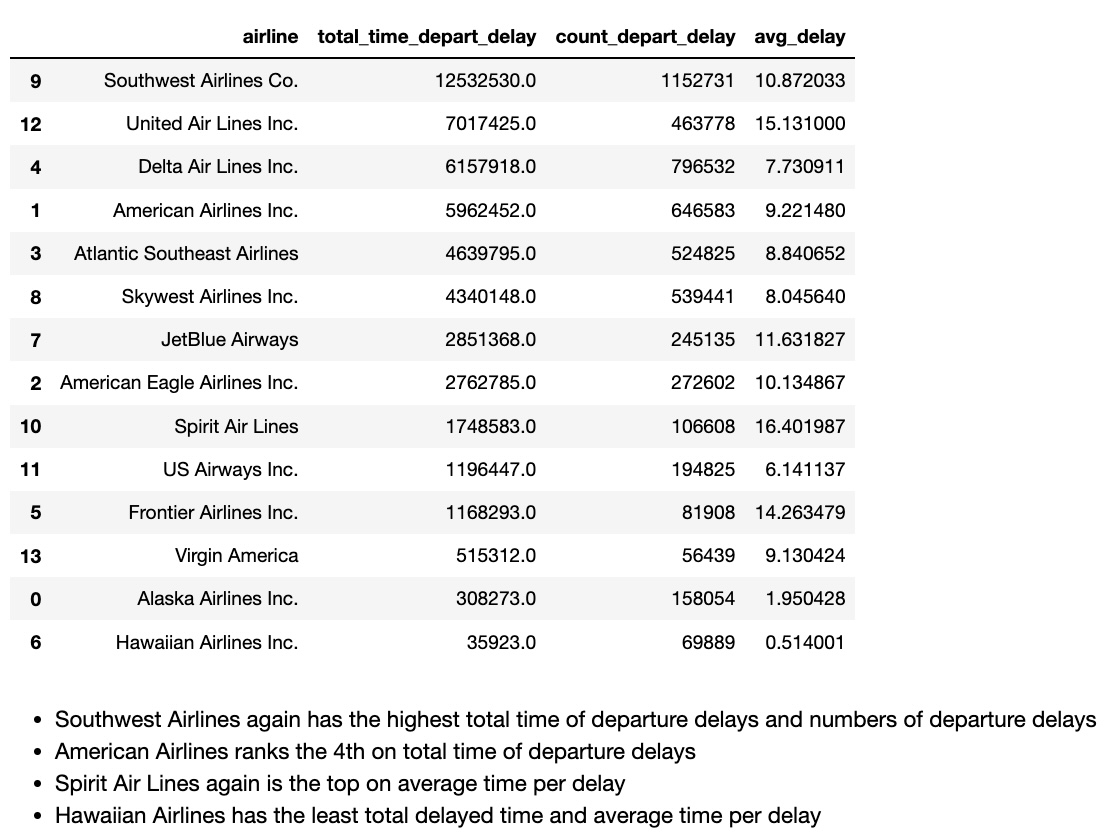
­

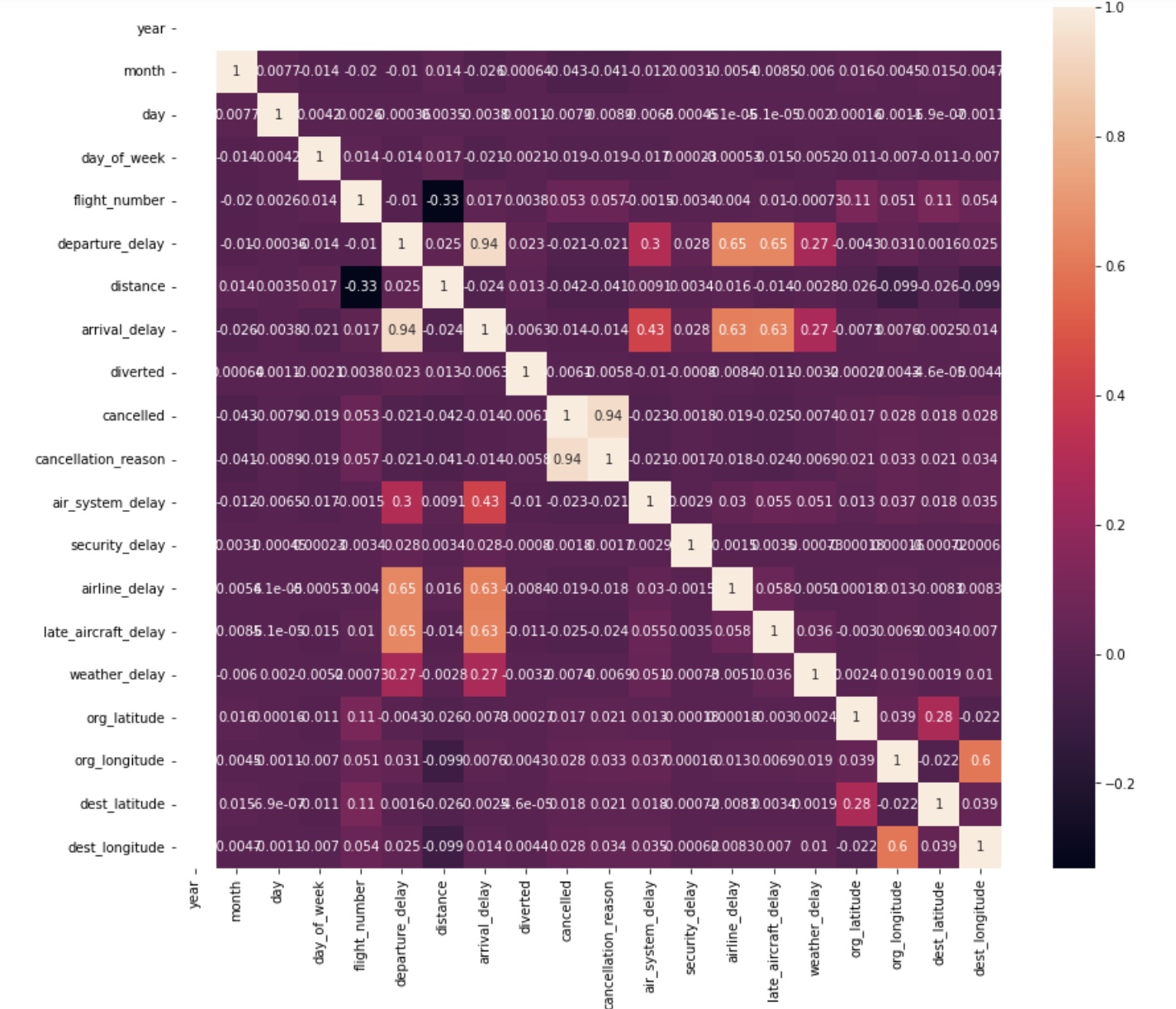
*Airline Distributions*

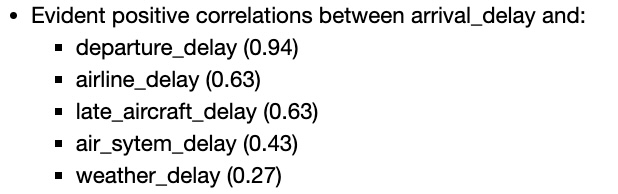


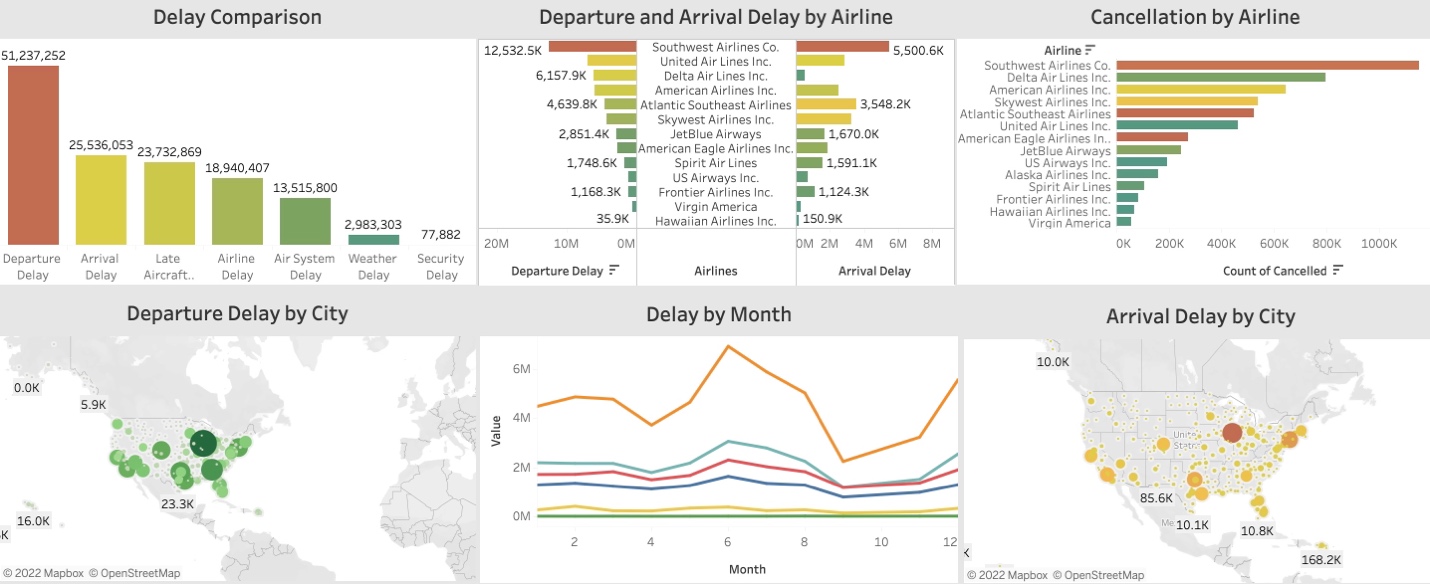












*Please see Tableau workbook/dashboard for more details*

**Pre-Processing**

A check of features that related to delays suggests that these features are not normally distributed. Therefore, a MinMaxScaler is applied to these features: “air\_system\_delay”, “airline\_delay”, “weather\_delay”, “late\_aircraft\_delay”, “security\_delay”, “arrival\_delay”, “departure\_delay”, and “distance”. In comparison, “month”, “day”, and “day\_of\_week” are all normally distributed, and thus they are scaled via StandardScaler. The target feature “delayed” is also converted to binary values (0 and 1).

Since all predicting features are numeric, an ANOVA test is conducted for feature importance. It suggests that “distance” is not important and thus removed from our analysis.

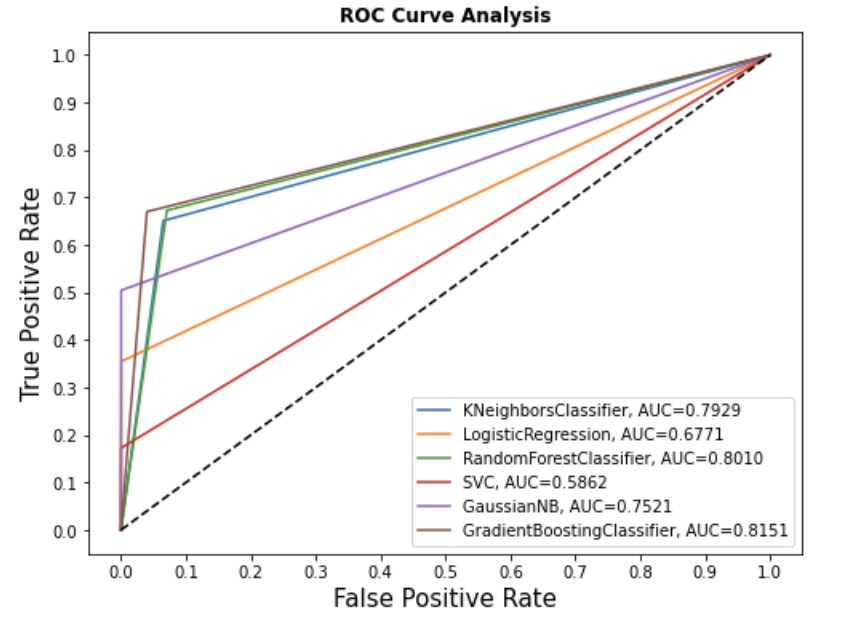


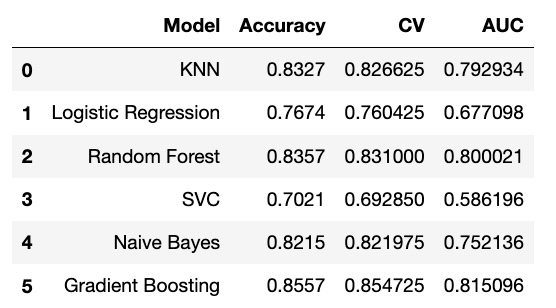
Since the dataset is huge, a sample of 50000 is selected for modelling. Final features for modelling are “month”, “day”, “day\_of\_week”, “air\_system\_delay”, “airline\_delay”, “weather\_delay”, “late\_aircraft\_delay”, “departure\_delay”, and “security\_delay”.

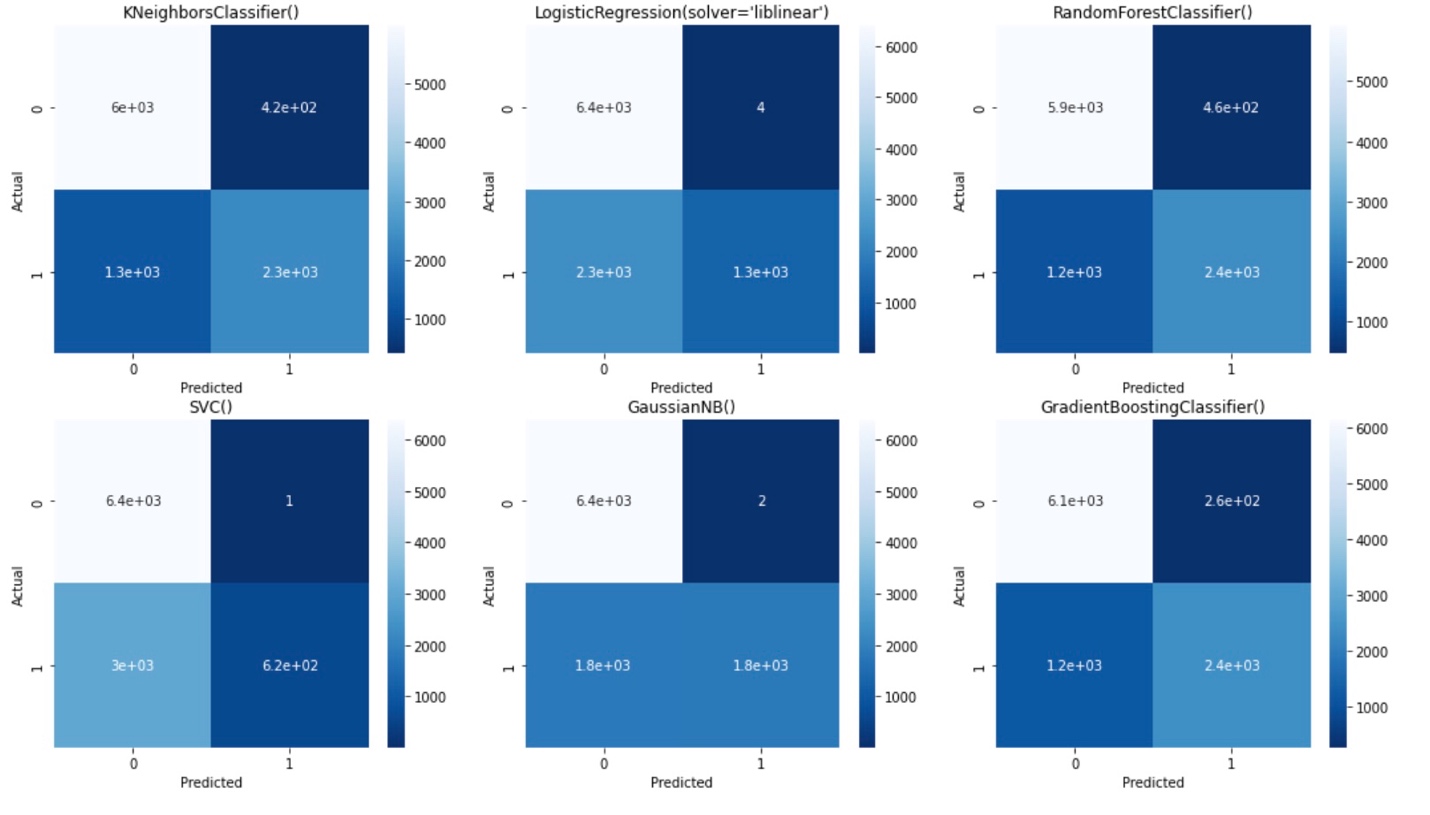
Finally, training data and testing data are split by 80% and 20%: X\_train: (50000, 9); X\_test: (40000, 9); y\_train: (50000, ); y\_test: (40000, ).

**Modelling**

There are 9 predictive features and 1 target feature (delayed: 0 and 1). Since this is a classification problem, the following machine learning models are selected: KNN, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes, and Gradient Boosting. At first, a pipeline is built with default parameters for all models. The evaluation metrics used in this study are Accuracy Score, CV Score, ROC-AUC Score for model performance.





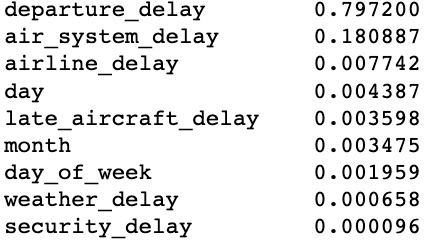


From the above analysis, we can notice that the Gradient Boosting has the highest accuracy score and AUC score, followed by Random Forest. Logistic Regression and SVC seem to be not doing well. KNN and Naïve Bayes are acceptable discrimination. But remember we only train each model with its default parameters; model performance might be different if applying different parameters. So, the next step is hyperparameter tuning.

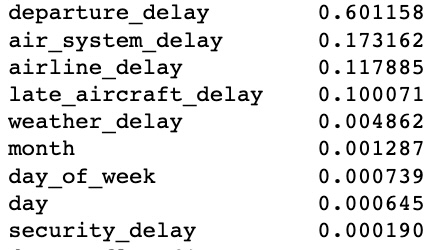
To optimize model performance, three models are selected for hyperparameter tuning: Random Forest, Gradient Boosting, and KNN. For Random Forest and Gradient Boosting, “max\_depth” and “n\_estimators” are selected from GridSearchCV; for KNN, “n\_neighbors” is selected. Models are trained again using optimized parameters.

|  |  |  |
| --- | --- | --- |
|  | Accuracy Score (Default) | Accuracy Score (Tuning) |
| Gradient Boosting | 0.8557 | 0.8560 |
| Random Forest | 0.8357 | 0.8550 |
| KNN | 0.8327 | 0.8344 |

From the results above, we can see that hyperparameter tuning clearly improves our model performance. The highest accuracy score is Gradient Boosting (0.8557).



*Feature importance (Gradient Boosting)*

**

*Feature Importance (Random Forest)*

Gradient Boosting and Random Forest both agree on the most predicting feature of delay: “departure\_delay”.

We can see that departure delay is the leading factor for U.S. airlines delay in 2015. It is also clear that weather, month, day, day of week are not that important when predicting delays. Departure delay would directly make airlines stay on air for more time, and thus consuming more fuel, leading to an extra cost of the company.

In order to cut the fuel cost and increase revenue, U.S. Department of Transportation (DOT) should focus on making measures to reduce the departure delay. To be specific, DOT should pay more attention to Southwest Airlines, which has the highest total departure delay time, arrival delay time, and the number of departure and arrival delays.

In the meantime, DOT could also consider how to handle delay caused by air system, as it is the second top reason causing arrival delay.