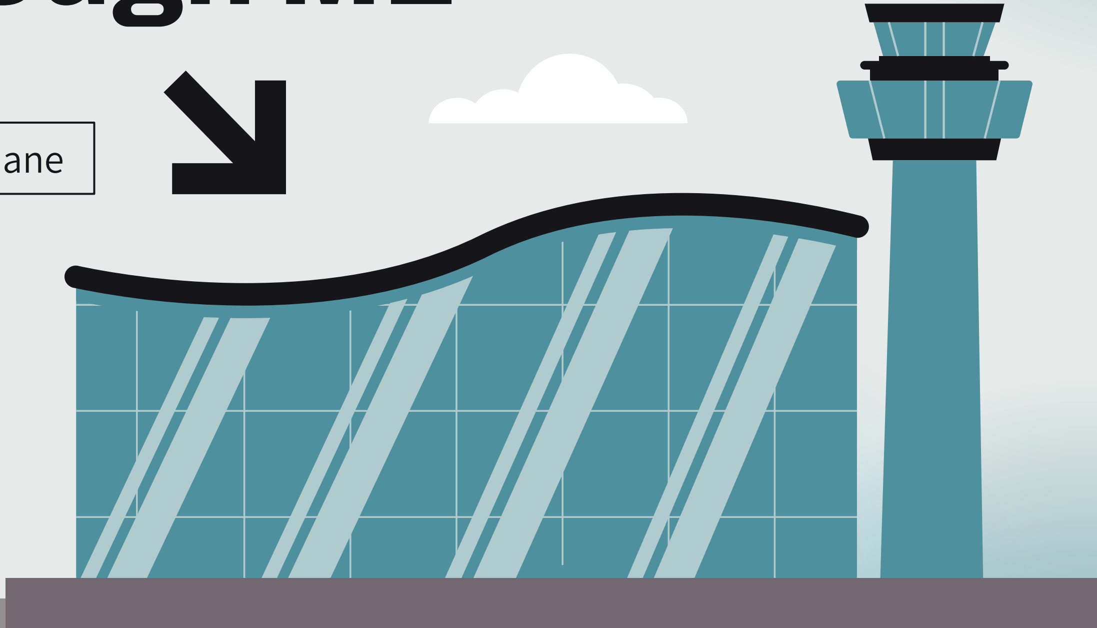
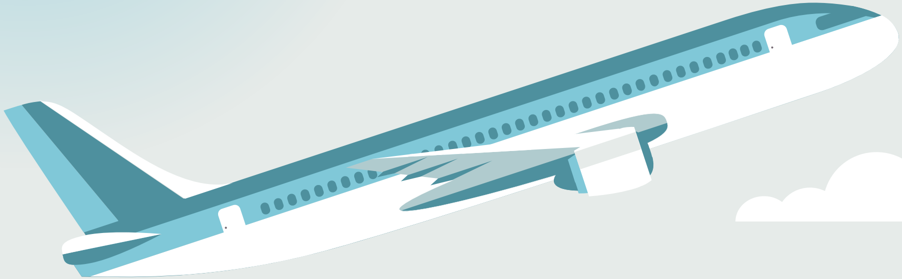
Predicting Flight Delays through ML

Team 4 - Mandy, Maggie, Aiden, Dramane





Introduction

COVID-19 once stalled much of airline businesses as travel became scarce. In recent months, especially since it has become warm and people are going on summer holidays, airline travel has resumed. However, we've noticed that there are more and more delays in flights that often mess up people's schedules. We are curious to know in more detail which airlines have more delays, which airports, and which flights (origin - destination) are causing more of the delay data.

Our *Data* Source

We found an airline dataset with 539383 instances and 8 different features on Kaggle and thought it was a good dataset to utilize for our analysis. It includes:

- Delayed or not
- Airline
- Time
- Flight
- Departing airport
- Arriving airport
- Length of flight

“Delayed or Not Delayed?”

Is the question we will try to answer with our Machine Learning Model

Technologies, languages, tools, and algorithms

Technologies: Jupyter Notebook, SQLite3, Tableau

Languages: Python, SQL

Tools: Pandas, Plotly ,Pathlib, Sklearn

Algorithms: Random Forest Classifier

Data Exploration

Data Exploration

```
In [49]: airplane_df = pd.read_csv("Airlines.csv", encoding="ISO-8859-1")
         airplane_df
```

Out[49]:

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay
0	1	CO	269	SFO	IAH	3	15	205	1
1	2	US	1558	PHX	CLT	3	15	222	1
2	3	AA	2400	LAX	DFW	3	20	165	1
3	4	AA	2466	SFO	DFW	3	20	195	1
4	5	AS	108	ANC	SEA	3	30	202	0
...	...	...	...	...	...	...	...	...	...
539378	539379	CO	178	OGG	SNA	5	1439	326	0
539379	539380	FL	398	SEA	ATL	5	1439	305	0
539380	539381	FL	609	SFO	MKE	5	1439	255	0
539381	539382	UA	78	HNL	SFO	5	1439	313	1
539382	539383	US	1442	LAX	PHL	5	1439	301	1

539383 rows × 9 columns

- Downloaded into Jupyter Notebook
- N/A (null) rows removed
- Identified count of 18 unique Airlines
- Identified count of 293 Airports (all domestic flights)
- Changed DayOfWeek numerical values to corresponding day
- Create new DataFrame and export to CSV and DB using SQLite3

Data Exploration (cont.)

jupyter Airlines_ML Last Checkpoint: 10 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Logout

Run Code

Out[17]:

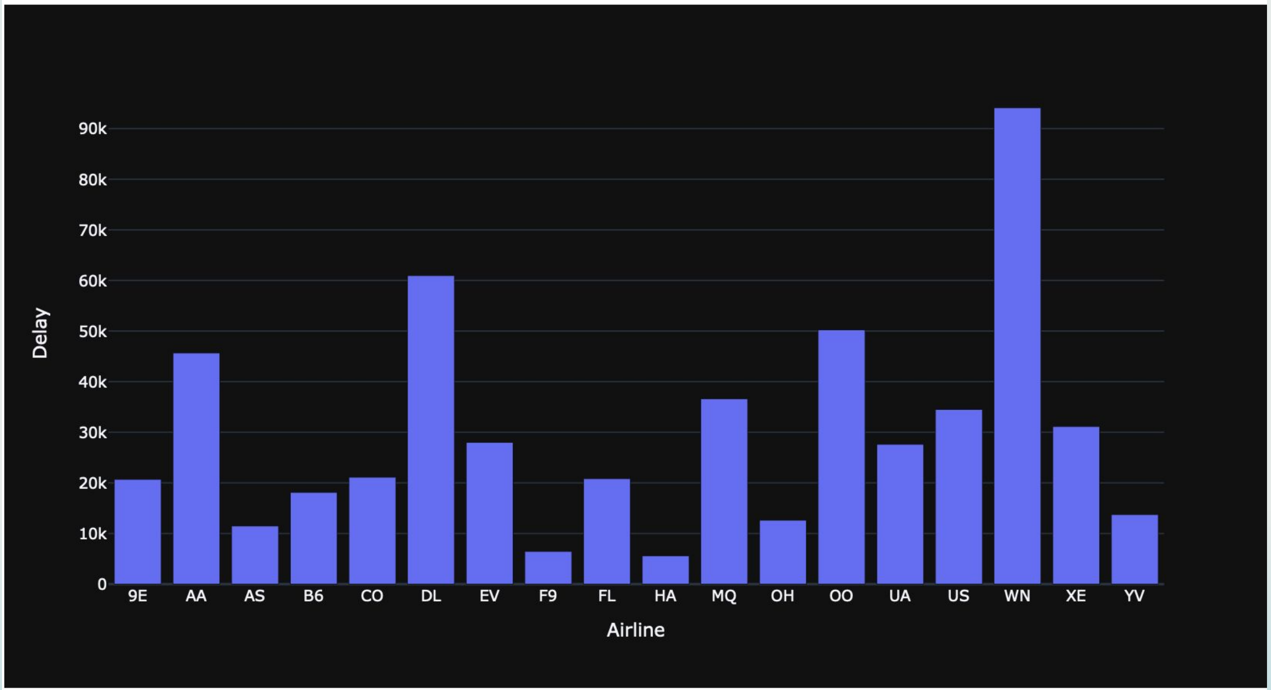
	Airline	Delay	AirportFrom	AirportTo	DayOfWeek	Length
0	CO	1	SFO	IAH	Thursday	205
1	US	1	PHX	CLT	Thursday	222
2	AA	1	LAX	DFW	Thursday	165
3	AA	1	SFO	DFW	Thursday	195
4	AS	0	ANC	SEA	Thursday	202
...	...	...	...	...	...	...
539378	CO	0	OGG	SNA	Saturday	326
539379	FL	0	SEA	ATL	Saturday	305
539380	FL	0	SFO	MKE	Saturday	255
539381	UA	1	HNL	SFO	Saturday	313
539382	US	1	LAX	PHL	Saturday	301

469504 rows × 6 columns

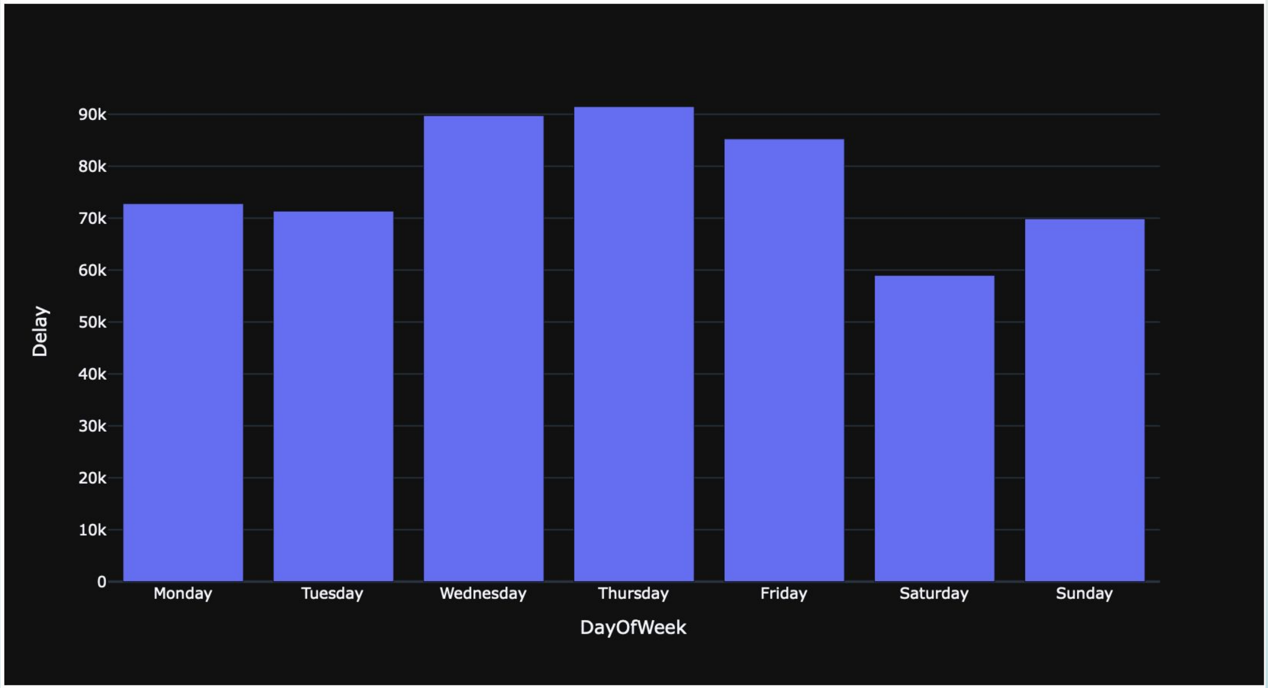
Data Analysis

- Compare delayed frequency between airports
- Compare delayed frequency by Airline
- Create random forest classifier model to predict flight delay based on airline, departing airport, arriving airport, and day of week with multiple decision trees.
- Sort features by their importance

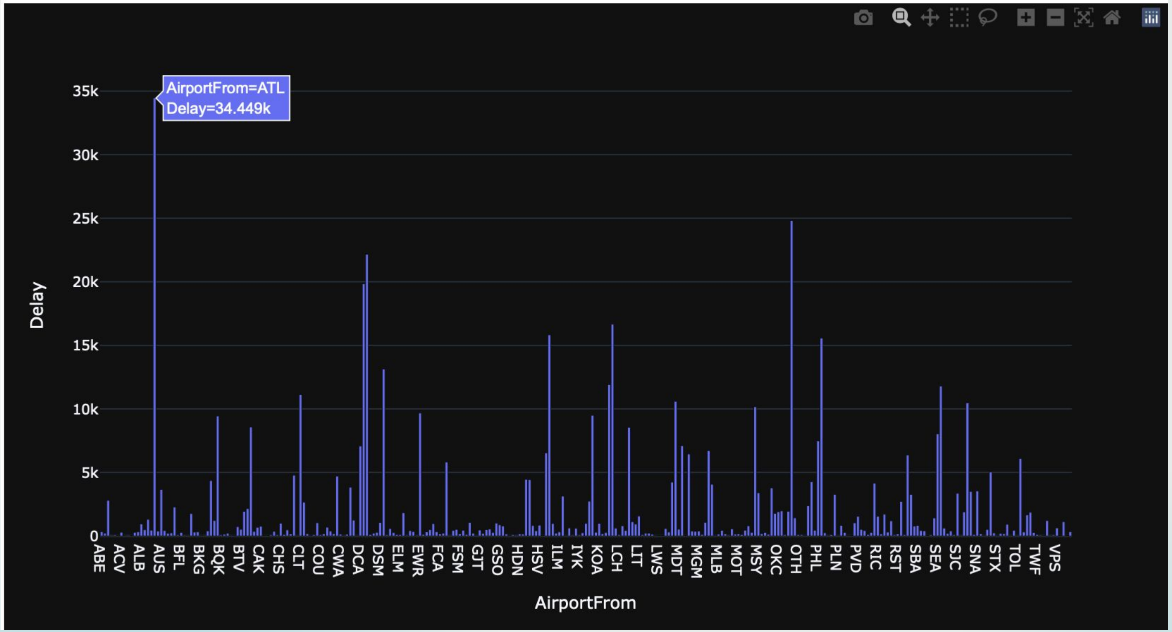
Data Analysis (cont.)



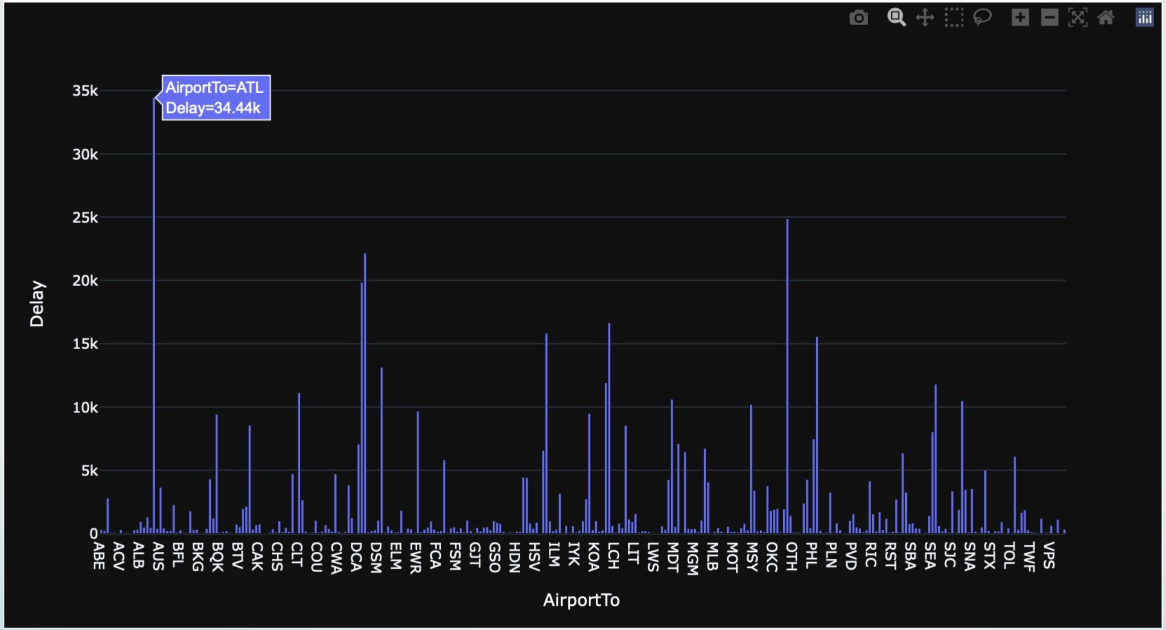
Data Analysis (cont.)



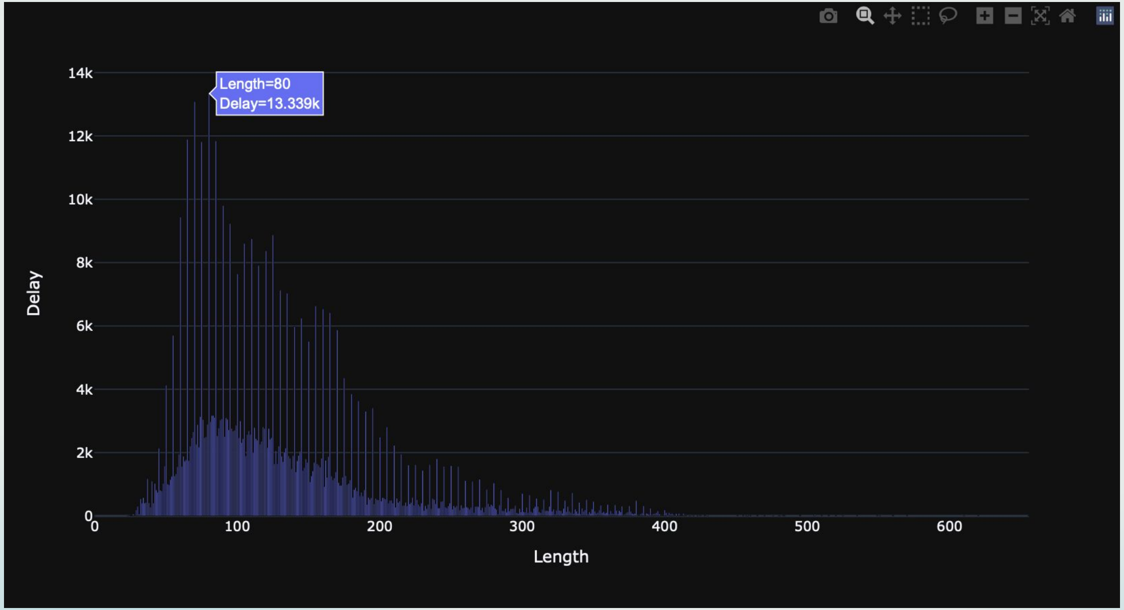
Data Analysis (cont.)



Data Analysis (cont.)



Data Analysis (cont.)



Data Analysis (cont.)

- Created random forest classifier to predict delay or no delay
- Increased n_estimators to 128 in an effort to increase accuracy

```
In [69]: # sample the training data with the BalancedRandomForestClassifier
         from imblearn.ensemble import BalancedRandomForestClassifier
         random_f = RandomForestClassifier(n_estimators = 128, random_state=42)

In [70]: #Fit the model
         random_f = random_f.fit(X_train, y_train, sample_weight=None)

In [71]: # Making predictions using the testing data.
         predictions = random_f.predict(X_test)
         predictions

Out[71]: array([0, 0, 0, ..., 1, 1, 1])
```

Tableau Dashboard

TOOLS THAT WILL BE USEFUL TO CREATE THE FINAL DASHBOARD

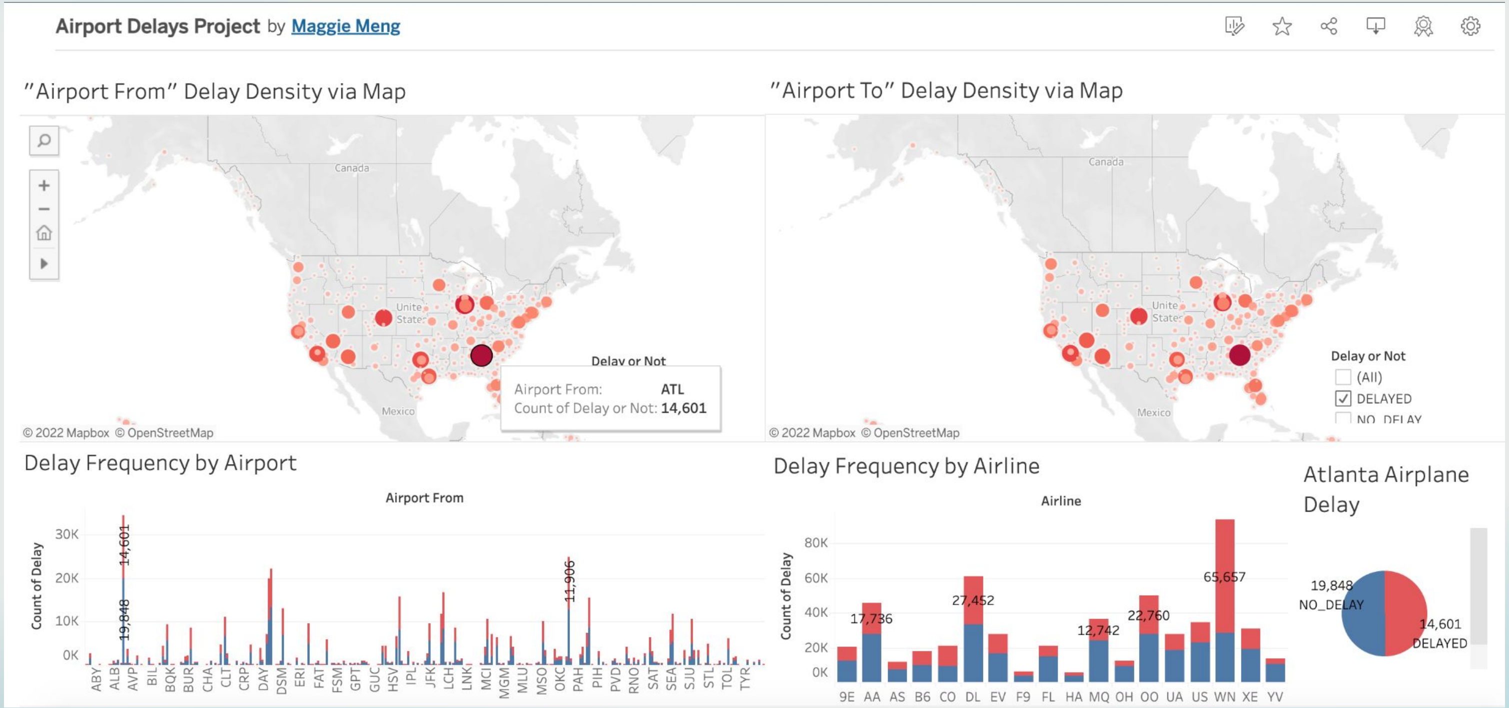
- Tableau
- Machine Learning model

DESCRIPTION OF INTERACTIVE ELEMENT

There will be two to three main components to the *Dashboard*,

- An interactive map of the U.S. where data points/analysis will show when user hovers over certain locations (and when user clicks into the state, they will be shown to a new page(?) of more details and probability of flight delay
- Percentages and table rankings of different metrics that is live and connected to the ML model (i.e. "top airports with most delays", "worst airlines to fly on based on delays within last x days", etc.)

Tableau Dashboard (cont.)



[Airport Delays Project Dashboard]

Result of Analysis

- Original accuracy with n_estimators = 100 was able to obtain 60% accuracy
- Revisited estimators and increased to 128 in an effort to improve accuracy

Confusion Matrix

	Predicted Not_Delayed	Predicted Delayed
Actual Not_Delayed	81424	8466
Actual Delayed	52315	19610

Accuracy Score : 0.6243796928591292

Classification Report

	precision	recall	f1-score	support
0	0.61	0.91	0.73	89890
1	0.70	0.27	0.39	71925
accuracy			0.62	161815
macro avg	0.65	0.59	0.56	161815
weighted avg	0.65	0.62	0.58	161815

Result of Analysis

```
In [51]: # We can sort the features by their importance.
         sorted(zip(random_f.feature_importances_, X.columns), reverse=True)

Out[51]: [(0.4560371494753297, 'Length'),
          (0.06322096852095473, 'Airline_WN'),
          (0.019320750520881957, 'DayOfWeek_Thursday'),
          (0.01879454960966424, 'DayOfWeek_Tuesday'),
          (0.018278231305831654, 'DayOfWeek_Sunday'),
          (0.017293017502290894, 'DayOfWeek_Monday'),
          (0.015847388603329373, 'DayOfWeek_Wednesday'),
          (0.015030511211531905, 'DayOfWeek_Friday'),
          (0.013028602243875267, 'DayOfWeek_Saturday'),
          (0.005504289602363767, 'Airline_UA'),
          (0.00525865385479413, 'Airline_YV'),
          (0.005045323129703494, 'AirportFrom_ORD'),
          (0.005001438787535601, 'Airline_FL'),
          (0.004881618831607391, 'Airline_US'),
          (0.004337321931338029, 'Airline_CO'),
          (0.004058948469206299, 'AirportFrom_MDW'),
          (0.004044971205961083, 'Airline_DL'),
          (0.0039011624135172175, 'Airline_OH'),
          (0.003651196808092408, 'AirportTo_DFW'),
          (0.003491955691999052, 'Airline_MQ'),
```

Recommendation for Future Analysis

- Decrease maximum number of features
- Utilize Boosting
- Use a data set with clearer features (i.e. Time column)
- Utilize logistic regression to interpret correlation for different features

Things we could’ve done differently

- Explore alternate data sources
  - Scheduled departure time vs actual departure time
  - Scheduled arrival time vs actual arrival time
  - International flights
  - Date of flight (year, month, etc)
  - Cause of delay
- Create List of dictionaries for Airline/Airport abbreviations
- Change length values to bins
- Analyze frequency of flights per day

Thanks