Syracuse University
School of Information Studies

IST 652: SCRIPTING FOR DATA ANALYSIS
Final Project Report

Instructor: Professor D. Landowski

Student Name

Academic Year 2018 – 2019

I. INTRODUCTION

For this final project, I decided to analyze the **2015 Flight Delays and Cancellations** dataset (available on Kaggle at *https://www.kaggle.com/usdot/flight-delays*) which was originally downloaded from the Department of Transportation Statistics website (available at *https://www.transtats.bts.gov/ > Data Finder[By Mode [Aviation]] > Airline On-Time Performance Data > On-Time Performance >* [all months from 2015 can be downloaded from this page]).

I used the Kaggle version of the dataset as it was slightly better curated than the raw data on the Department of Transportation Statistics website, though it still required a great deal of cleansing and preprocessing. The data consisted of a large folder (about 592 MB) with 3 .csv files:

- The first file, **airlines.csv** (2 columns), is a short table summarizing all the airlines contained in the main dataset of the folder (flights.csv). It lists each airline's unique identifier code (IATA) in addition to the full name of the company.
    - e.g.

IATA_CODE	AIRLINE
UA	United Air Lines Inc.

- The second file, **airports.csv** (7 columns), is a table summarizing all the airports contained in the main dataset of the folder (flights.csv). It lists each airport's unique identifier code as well (IATA), the airport's full name, as well as key information regarding the geolocalization of the airport (i.e. city, state, country, latitude, and longitude.)
    - e.g.

IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919

- The third file, **flights.csv** (31 columns), is the main dataset of the folder and contains all the flights data and information.
    - e.g. [please go to https://www.kaggle.com/usdot/flight-delays#flights.csv for a preview of full flight.csv dataset.]

The third dataset was the main focus of this analysis, while the first two allowed me to combine them together to eventually import the airlines' and airports' full names into the flights.csv dataset. Indeed, the main flights.csv dataset did *not* include any of that information and only provided airlines' and airports' IATA codes for brevity (as there were already many columns.)

Below you will find a quick overview of the metadata. The documentation on the metadata can be found on Kaggle and the Dept. of Transporation website. However, some of the variable definitions were not clear enough, so I made sure to improve and refine them after carefully working on the data and looking up key concepts relevant to the field of aviation.

METADATA:

- **YEAR** year of the given flight

- **MONTH** month of the given flight (e.g. 1 = January, 9 = September, etc.)
- **DAY** day of the given flight (e.g. 1, 24, 28, etc.)
- **DAY_OF_WEEK** day of week of the given flight (e.g. 1 = Monday, 4 = Thursday etc.)
- **AIRLINE** airline unique IATA identifier
- **FLIGHT_NUMBER** flight identifier
- **TAIL_NUMBER** aircraft identifier
- **ORIGIN_AIRPORT** departure IATA airport identifier for a given flight
- **DESTINATION_AIRPORT** arrival IATA airport identifier for a given flight
- **SCHEDULED_DEPARTURE** official planned departure time for a given flight
- **DEPARTURE_TIME** result of WHEEL_OFF - TAXI_OUT
- **DEPARTURE_DELAY** total delay on departure (in minutes)
- **TAXI_OUT** time duration elapsed between departure from the origin airport gate and wheels off
- **WHEELS_OFF** time point that the aircraft's wheels leave the ground
- **SCHEDULED_TIME** planned time amount needed for the flight trip
- **ELAPSED_TIME** result of AIR_TIME+TAXI_IN+TAXI_OUT
- **AIR_TIME** time duration between wheels_off and wheels_on time
- **DISTANCE** distance between both departure and arrival airports for a given flight
- **WHEELS_ON** time point that the aircraft's wheels touch on the ground
- **TAXI_IN** time duration elapsed between wheels-on and gate arrival at the destination airport
- **SCHEDULED_ARRIVAL** official planned arrival time for a given flight
- **ARRIVAL_TIME** result of WHEELS_ON+TAXI_IN
- **ARRIVAL_DELAY** result of ARRIVAL_TIME-SCHEDULED_ARRIVAL (in minutes)
- **DIVERTED** whether aircraft has been routed from its original arrival destination to a new, typically temporary, arrival destination (dummy variable, 1 = diverted)
- **CANCELLED** whether a flight has been cancelled (dummy variable, 1 = cancelled)
- **CANCELLATION_REASON** reason for cancellation of a given flight:
    - A - Airline/Carrier
    - B - Weather
    - C - National Air System
    - D - Security
- **AIR_SYSTEM_DELAY** delay caused by air system (in minutes)
- **SECURITY_DELAY** delay caused by security (in minutes)
- **AIRLINE_DELAY** delay caused by the airline (in minutes)
- **LATE_AIRCRAFT_DELAY** delay caused by aircraft (in minutes)
- **WEATHER_DELAY** delay caused by weather (in minutes)

II. BUSINESS QUESTIONS

For this project, my first objective was to perform an exploratory analysis of the data and to answer key questions regarding flights, delays, distance ranges, cancellations, airports, etc. and airlines and several combinations of those variables. My second goal was to use machine learning to see if there would be any sort of predictive power in my final flights.csv dataset.

Below is a list of the key questions I sought to answer in this analysis.

Descriptive analysis:

1. Which airlines typically have significant flight delays?
2. Which airports typically have significant flight delays?
3. Which time of year, month, and day typically have the most flight delays?
4. Which day(s) of the week typically have the most flight delays?
5. Which flight distance range(s) typically involve the most flight delays?
6. What are the statistics of the most common delay reasons? What are the top cancellation reasons?

Predictive analysis:

- Can a Random Forest classifier learn to predict flight delays based on the dataset?
- What is the accuracy reached?

III. METHODOLOGY

a) Data importation

The multi-purpose programming language *Python* was used for this project. The code was developed in Jupyter Notebook and Jupyter Lab. The very first step consisted in importing all 3 datasets into Python using the .read_csv command, and looking at their dimensions.

```
AIRLINES DATASET OVERVIEW:
rows: 14
columns: 2

AIRPORTS DATASET OVERVIEW:
rows: 322
columns: 7

FLIGHTS DATASET OVERVIEW:
rows: 5819079
columns: 31
```

As you can see, the main flights.csv dataset was very big and had 5,819,079 rows. It also needed extensive cleansing and preprocessing.

Please see code for more details.

b) Data cleansing and preparation

In order to prepare the dataset for the analysis, I had to cleanse the data and make several decisions as to what to keep and delete, how to deal with missing records, etc. All those decisions have been included in the code attached to this report. Below is an overview of the major steps.

- Many columns (variables) were removed from the data because they were (1) irrelevant, (2) redundant/repetitive, or (3) contained the same value in every observation. Below is a list of all the variables that were taken out.
    - **YEAR** # was all the same for every row, i.e. 2015
    - **FLIGHT_NUMBER** # would not be helpful in the analysis
    - **TAIL_NUMBER** # would not be helpful in the analysis
    - **DEPARTURE_TIME** # redundant, as the official scheduled time and departure delay variables together would provide the same information
    - **TAXI_OUT** # would not be helpful in the analysis – too detailed
    - **WHEELS_OFF** # would not be helpful in the analysis – too detailed
    - **SCHEDULED_TIME** # redundant, as again a combination of other variables in the data would provide the same information
    - **ELAPSED_TIME** # redundant, as again a combination of other variables in the data would provide the same information
    - **AIR_TIME** # redundant, as again a combination of other variables in the data would provide the same information
    - **WHEELS_ON** # would not be helpful in the analysis – too detailed
    - **TAXI_IN** # would not be helpful in the analysis – too detailed
    - **ARRIVAL_TIME** # redundant, as again a combination of other variables in the data would provide the same information
    - **DIVERTED** # would not be helpful in the analysis
- I created and added a brand-new column ("STATUS") that could have three possible options: on time, delayed, cancelled.
    - In order to create it, I used the "ARRIVAL_DELAY" and "CANCELLED" original variables as a basis.
    - I used the official FAA definition of a delayed flight, i.e. "A flight delay is when an airline flight takes off and/or lands later than its scheduled time. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time."
    - However, I based the answer on the arrival delay and not the departure delay, as I believe that it is important to consider a flight delay as a whole. Given that flight delays can be made up for during flight time, what truly matters is how late the aircraft reaches its destination. The departure delay is less important to determine the overall delay in our case.
- I turned time (hours) data into bins following this pattern: 0-559 --> night, 600-1159 --> morning, 1200-1759 --> afternoon, 1800-2359 --> evening.

- I turned distance data into distance range bins following this pattern: [0-500], [500-1000], [1000-1500], etc.
- Many columns were eventually re-ordered for simplicity.
- Many columns were eventually renamed for simplicity.
- Missing values and NA's, which represented a very small portion of the data, were removed from the data.

```
Data shape BEFORE cleansing:
(5803891, 19)

Data shape AFTER cleansing:
(5714007, 12)
```

Once the flights.csv dataset was clean, I used it to create 3 final datasets, which would be the ones that I would eventually use for the analysis.

- DATASET: DELAYS AND REASONS
- DATASET: CANCELLATIONS AND REASONS
- DATASET: ALL FLIGHTS

The procedure to create and cleanse those 3 brand-new datasets is thoroughly explained in the code, even though their names are already self-explanatory. The goal was to create new datasets that would help me answer my main business questions. Besides, the third dataset (ALL FLIGHTS) became the new main dataset for the analysis. It contained on time and delayed flights only (cancelled flights were removed as they were not the main focus of this analysis.)

At the end of this section, I also took advantage of one of Python's greatest features: dictionaries. I used the original airlines.csv and airports.csv datasets (described above) to store the IATA_CODES (unique identifiers for airlines and airports) codes as keys, and the full names of airlines and airports as values. Thus, a command such as `print(airlineNames['B6'])` would return `JetBlue Airways`, and `print(airportNames['LAX'])` would `return Los Angeles International Airport`. This came in very handy in the descriptive part, when plotting data involving airlines and airports (seeing the IATA codes would not have been very helpful, so they were replaced by the actual airline company name or airport full name).

Please see code for more details.

c) Data descriptive analysis

For the descriptive analysis section, I used several of Python's libraries commonly used for data wrangling and analysis. I used NumPy (a common package for scientific computing with powerful N-dimensional array objects) and Pandas (a very well-known package used for data manipulation and analysis of tabular data and data frames), as well as Matplotlib and Seaborn for data visualization. This section involved a lot of summary statistics, plots, and new datasets. The groupby function was the most useful resource for this part.

All my project descriptive questions were answered in this section. Please see code for more details.

d) Data predictive analysis

Dear Prof. Landowski,
I know that predictions were not a requirement for the project. At first, I was leaning towards not doing machine learning as I still have a lot to learn in that area, but in the end I decided to try anyway so I could get feedback and suggestions from you. I really want to improve and hope I will not be penalized for taking this quite simple approach.
David

For the predictive analysis section, I used Python's Scikit-Learn library, a very well-known machine learning package.

Before training the model, my ALL FLIGHTS dataset was further refined to be optimized for machine learning.

- Since the main dataset was far too big to be used for predictions (especially considering the amount of featured it had after using one hot encoding), I decided to create two new smaller datasets and to use them to create two separate models to see if there would be a significant difference in the performance of the classifications.
    - Model #1: THE 5 AIRPORTS THAT HAVE THE LARGEST NUMBER OF DELAYS vs. ALL AIRLINES
    - Model #2: THE AIRLINE WITH THE LARGEST NUMBER OF DELAYS vs. ALL AIRPORTS
- Some variables were removed from those datasets as they would only confuse the classifier (i.e. DEPARTURE_DELAY, ARRIVAL_DELAY, and DESTINATION).
    - The first two variables that were deleted were already contained and summarized in the 'STATUS' variable, which was eventually set to be the target variable so the classifier would try and predict if a flight was going to be *on time* or *delayed*. Keeping them would have confused the algorithm as they were too precise (integer values, not bins).
    - The third variable that was deleted would have also made it difficult for the classifier to be efficient considering the number of airports there were in the data. Besides, the flight distance range was enough to account for the destination of a given flight on a broader level.

The algorithm used for this classification task was Random Forest, an ensemble learning method typically used for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. My dataset contained many categorical variables, and this algorithm is well-known for performing well on such datasets (provided that they contain some patterns.)
In this part, I also utilized many of the Scikit-Learn built-in functions (i.e. train_test_split, accuracy_score, confusion_matrix, classification_report, cross_val_score.)

Please see code for more details.

IV. DATA ANALYSIS

Program overall description summary

As described above, I first imported the data into Python (all 3 datasets). Then, I cleaned it extensively (reorganized it, tweaked it, renamed it, etc.) and made several decisions as to what to keep and what to get rid of. I took advantage of Python's great features (such as python dictionaries) and eventually created 3 main datasets for the analysis. Then, I performed an exploratory analysis on the data while answering my business questions. Finally, I refined the main dataset a bit more and created two new ones to train a Random Forest classifier. All my findings are reported at the end of this report.

a) Unexecuted code (and detailed program explanation)

Due to the length of the project, please refer to the .ipynb Jupyter Notebook session attached to this report to see the unexecuted code.
Please note that the reasoning behind each step has been explained in the code (in the form of # comments).

b) Executed code

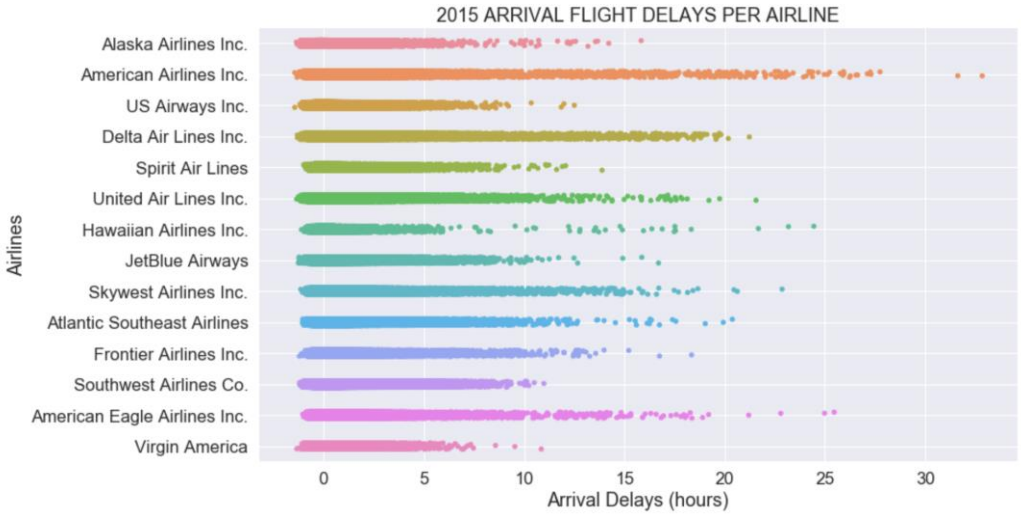
Please see the .ipynb Jupyter Notebook session (containing **the full output of the program**) or its .pdf version attached to this report to see the executed code.
I provided an in-depth explanation for every command in the program for clarity and made sure that the output of each chunk is neatly presented under each code block (with spaces, titles, etc.).

Results

For convenience, the results have been repeated below.

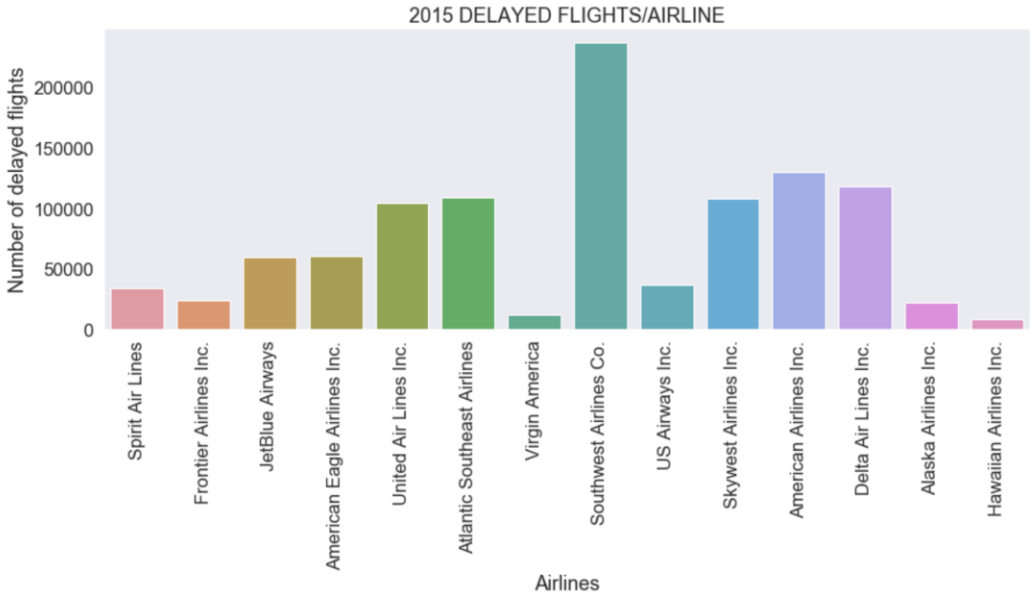
1. Which airlines typically have significant flight delays?

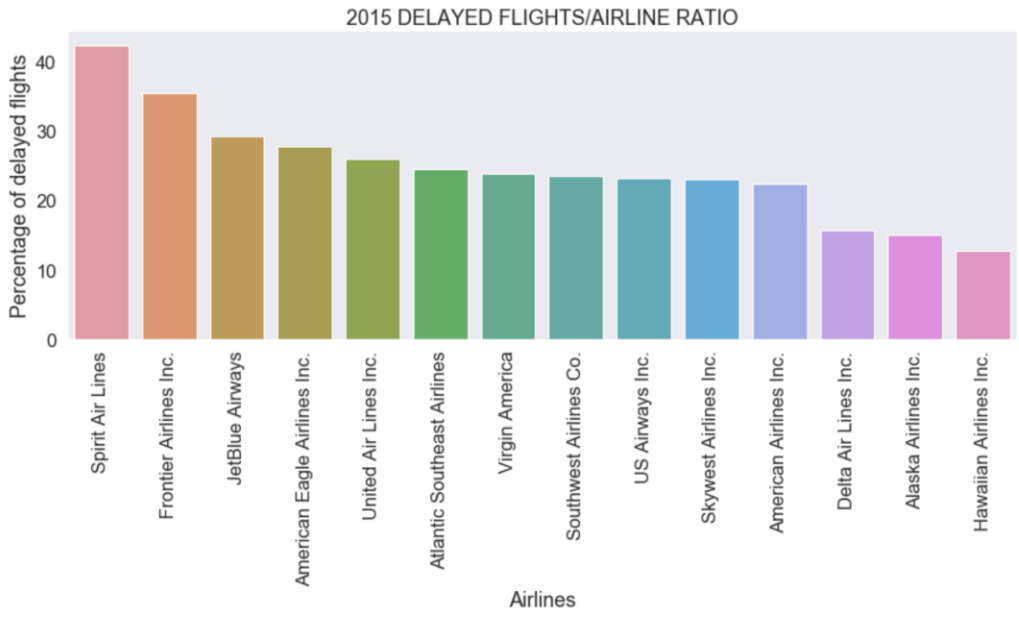
AIRLINE	delayed count	mean	median	min	max
AA	130279	3.451372	-6.0	-87.0	1971.0
AS	22352	-0.976563	-5.0	-82.0	950.0
B6	59175	6.677861	-5.0	-76.0	1002.0
DL	118023	0.186754	-8.0	-79.0	1274.0
EV	109184	6.585379	-4.0	-64.0	1223.0
F9	23569	12.503913	-1.0	-73.0	1101.0
HA	8618	2.023093	-2.0	-67.0	1467.0
MQ	60547	6.457873	-6.0	-63.0	1528.0
NK	34221	14.471800	0.0	-60.0	833.0
OO	107795	5.845652	-4.0	-69.0	1372.0
UA	104722	5.431594	-6.0	-81.0	1294.0
US	36549	3.706209	-4.0	-87.0	750.0
VX	11778	4.737706	-3.0	-81.0	651.0
WN	236626	4.374964	-4.0	-73.0	659.0



The ratios of flight delays per airlines are:

	AIRLINE	delayed count	on time count	percentage of delayed flights
0	Spirit Air Lines	34221	80972	42.26
1	Frontier Airlines Inc.	23569	66520	35.43
2	JetBlue Airways	59175	202867	29.17
3	American Eagle Airlines Inc.	60547	218244	27.74
4	United Air Lines Inc.	104722	403040	25.98
5	Atlantic Southeast Airlines	109184	445568	24.50
6	Virgin America	11778	49470	23.81
7	Southwest Airlines Co.	236626	1005777	23.53
8	US Airways Inc.	36549	157674	23.18
9	Skywest Airlines Inc.	107795	469019	22.98
10	American Airlines Inc.	130279	582656	22.36
11	Delta Air Lines Inc.	118023	752252	15.69
12	Alaska Airlines Inc.	22352	149087	14.99
13	Hawaiian Airlines Inc.	8618	67423	12.78

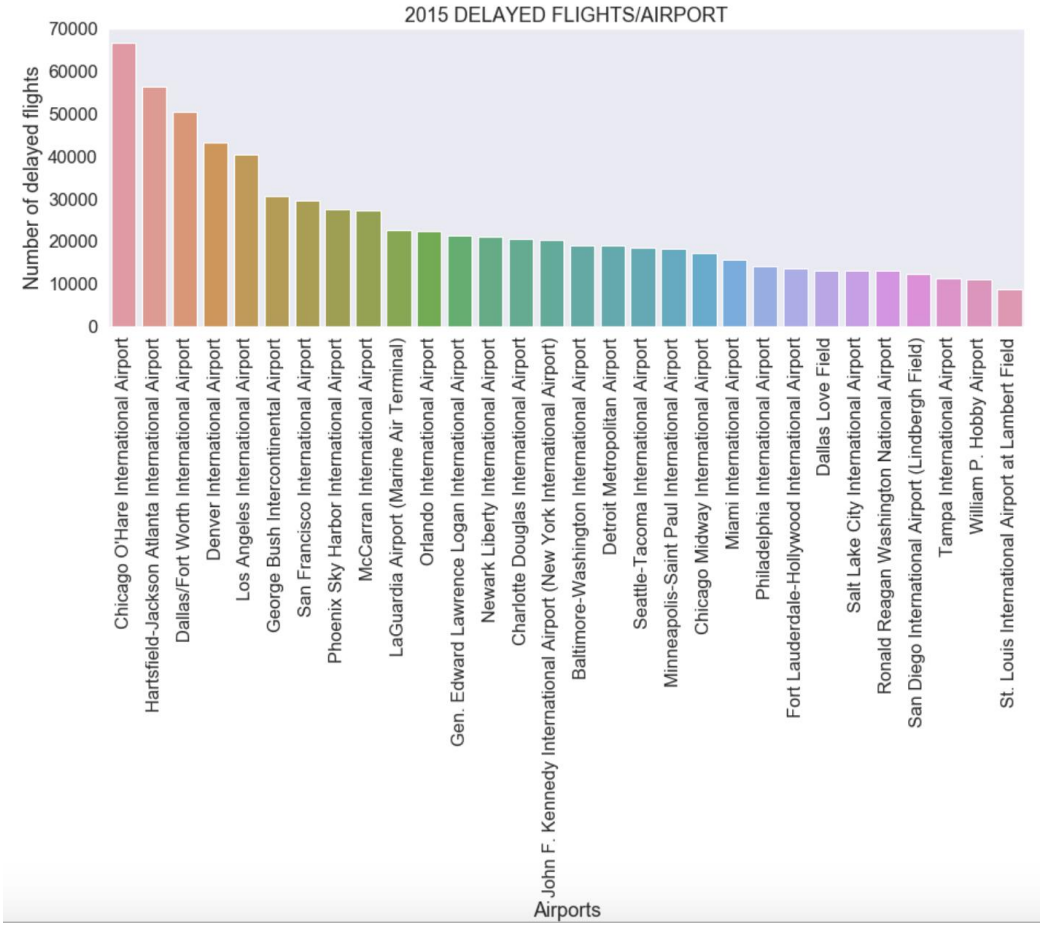




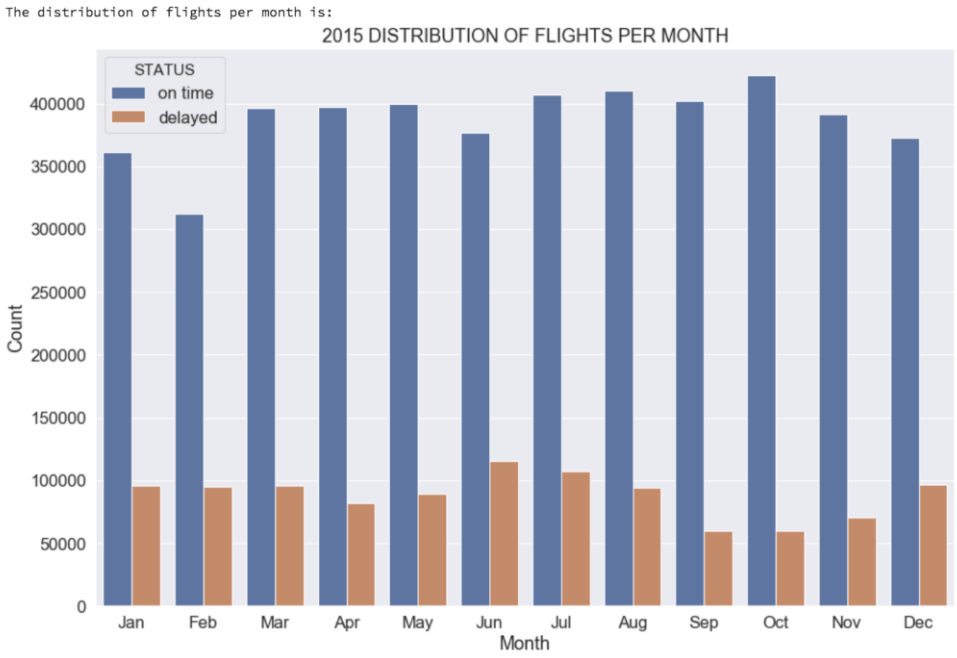
2. Which airports typically have significant flight delays?

The airports with the largest number of flight delays are:

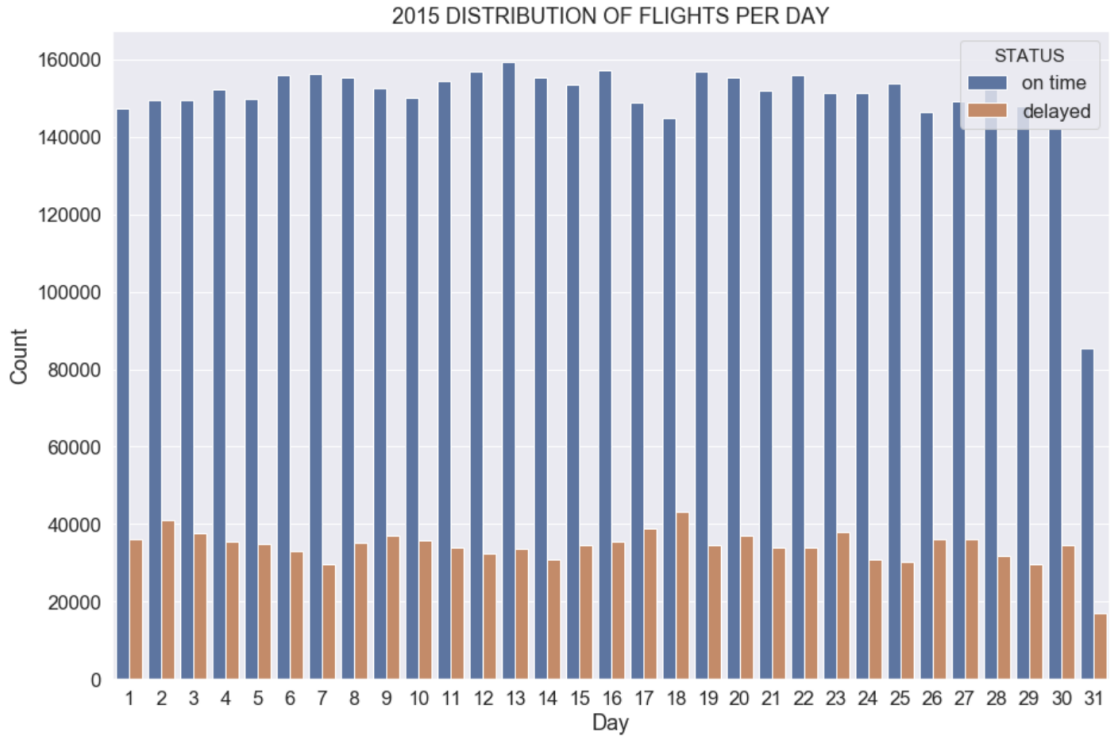
	ORIGIN	STATUS
0	Chicago O'Hare International Airport	66663
1	Hartsfield-Jackson Atlanta International Airport	56462
2	Dallas/Fort Worth International Airport	50478
3	Denver International Airport	43331
4	Los Angeles International Airport	40281
5	George Bush Intercontinental Airport	30690
6	San Francisco International Airport	29534
7	Phoenix Sky Harbor International Airport	27427
8	McCarran International Airport	27225
9	LaGuardia Airport (Marine Air Terminal)	22709
10	Orlando International Airport	22349
11	Gen. Edward Lawrence Logan International Airport	21363
12	Newark Liberty International Airport	21163
13	Charlotte Douglas International Airport	20623
14	John F. Kennedy International Airport (New Yor...	20260
15	Baltimore-Washington International Airport	19049
16	Detroit Metropolitan Airport	18956
17	Seattle-Tacoma International Airport	18554
18	Minneapolis-Saint Paul International Airport	18171
19	Chicago Midway International Airport	17273
20	Miami International Airport	15785



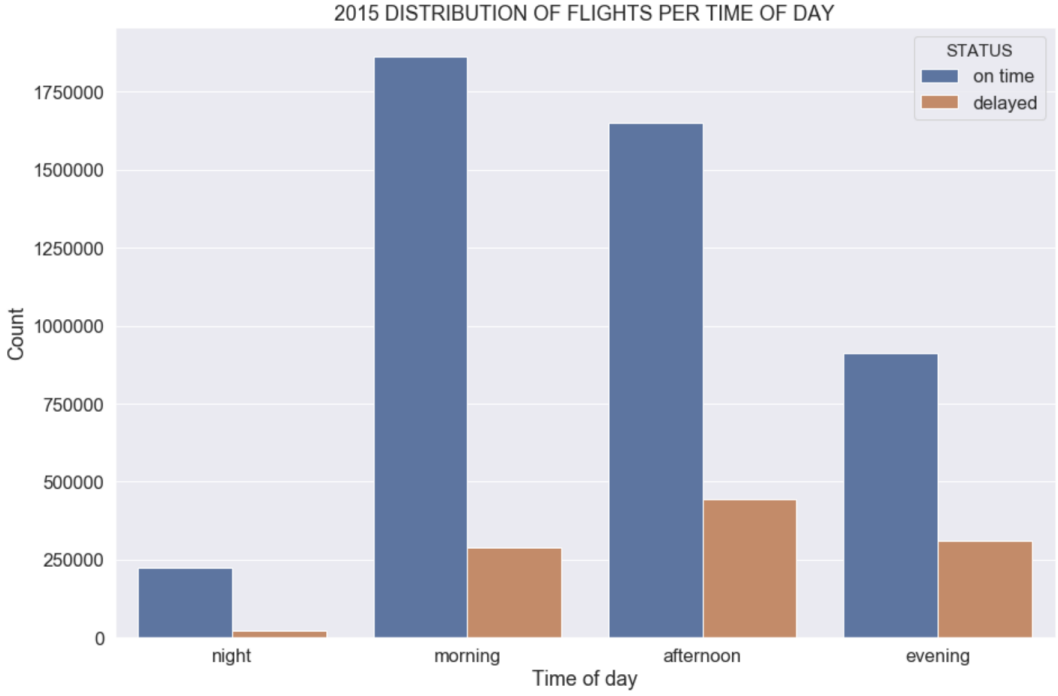
3. Which time of year, month, and day typically have the most flight delays?



The distribution of flights per day is:



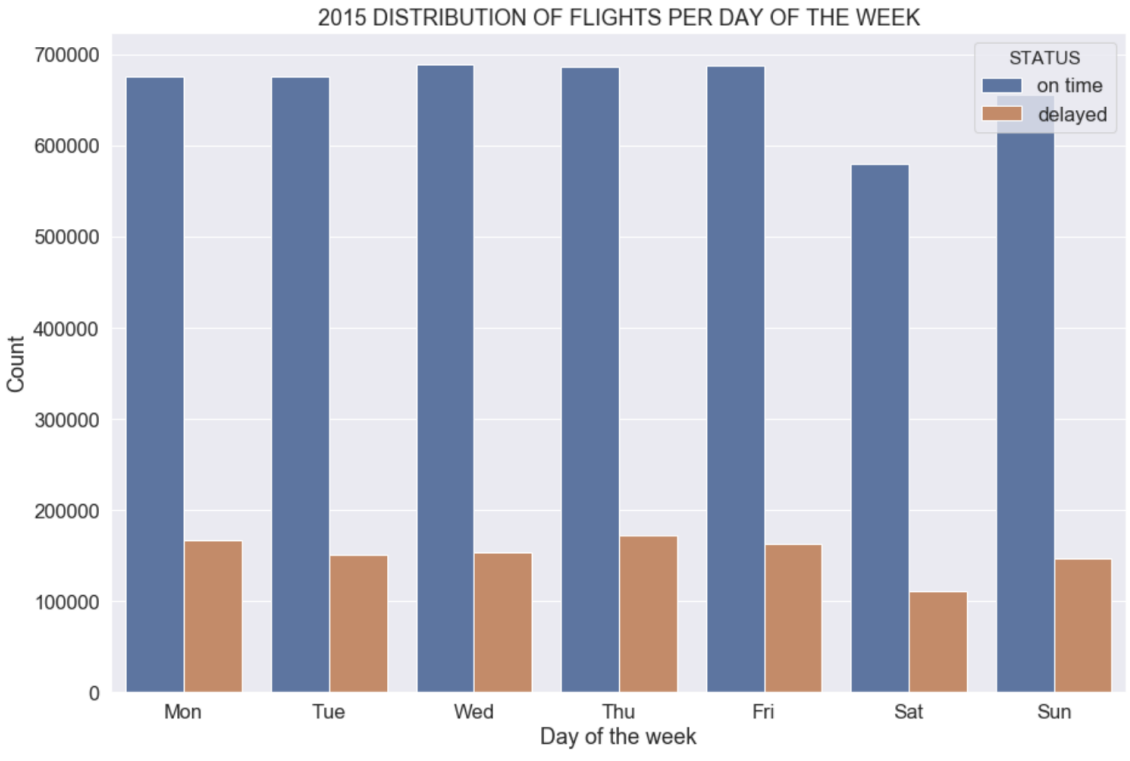
The distribution of flights per time of day is:



4. Which day(s) of the week typically have the most flight delays?

The distribution of flights per day of the week is:

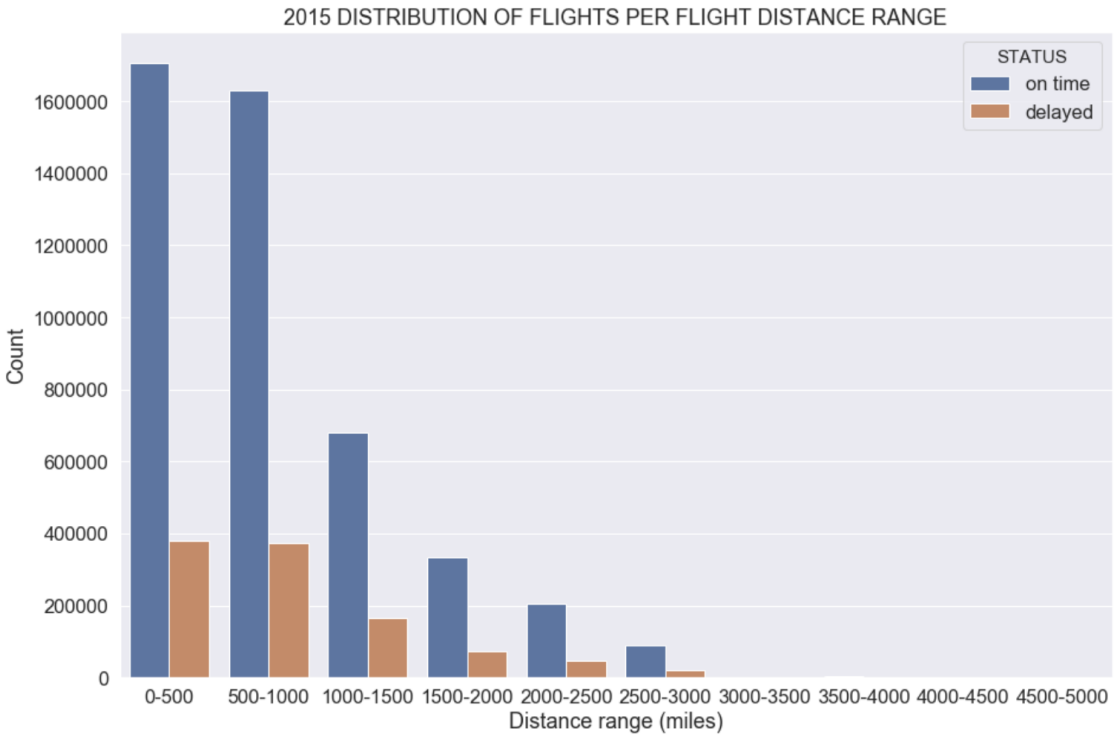
	DAY_OF_WEEK	STATUS	COUNT
0	Mon	on time	675528
1	Mon	delayed	166266
2	Tue	on time	676216
3	Tue	delayed	151183
4	Wed	on time	689638
5	Wed	delayed	153604
6	Thu	on time	686194
7	Thu	delayed	171692
8	Fri	on time	687946
9	Fri	delayed	163441
10	Sat	on time	579430
11	Sat	delayed	110314
12	Sun	on time	655617
13	Sun	delayed	146938



5. Which flight distance range involves the largest number of flight delays?

The distribution of flights per flight distance range is:

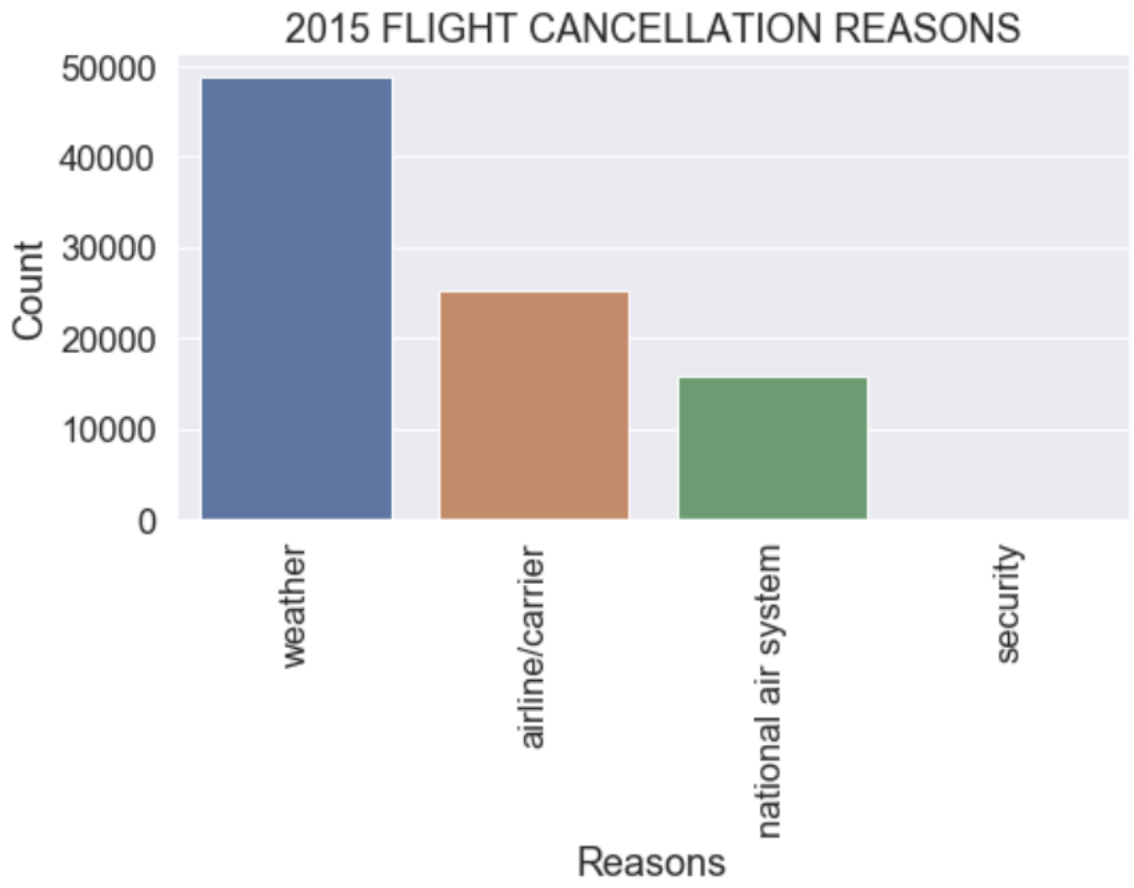
	FLIGHT_DISTANCE	STATUS	COUNT
0	0-500	on time	1707218
1	0-500	delayed	379828
2	500-1000	on time	1629095
3	500-1000	delayed	373799
4	1000-1500	on time	679784
5	1000-1500	delayed	166915
6	1500-2000	on time	332945
7	1500-2000	delayed	74796
8	2000-2500	on time	204490
9	2000-2500	delayed	45749
10	2500-3000	on time	89811
11	2500-3000	delayed	20123
12	3000-3500	on time	1491
13	3000-3500	delayed	457
14	3500-4000	on time	3047
15	3500-4000	delayed	1049
16	4000-4500	on time	717
17	4000-4500	delayed	237
18	4500-5000	on time	1971
19	4500-5000	delayed	485



6. What are the statistics of the most common delay reasons? What are the top cancellation reasons?

```
For delayed flights in 2015, the average air system delay (min) was 13.48
mins.
For delayed flights in 2015, the average security delay (min) was 0.08 mins.
For delayed flights in 2015, the average airline delay (min) was 18.97 mins.
For delayed flights in 2015, the average late aircraft delay (min) was 23.47
mins.
For delayed flights in 2015, the average weather delay (min) was 2.92 mins.
```

	index	CANCELLATION_REASON
0	weather	48851
1	airline/carrier	25262
2	national air system	15749
3	security	22



PREDICTIVE MODEL #1: THE 5 AIRPORTS THAT HAVE THE LARGEST NUMBER OF DELAYS vs. ALL AIRLINES

```
Accuracy: 0.7993748552905765

5-fold Cross Validation: 0.7966188117538201

Confusion Matrix
[[ 24437  52559]
 [ 21960 272478]]

Classification Report
             precision    recall  f1-score   support

    delayed       0.53      0.32      0.40     76996
    on time       0.84      0.93      0.88    294438

avg / total       0.77      0.80      0.78    371434
```

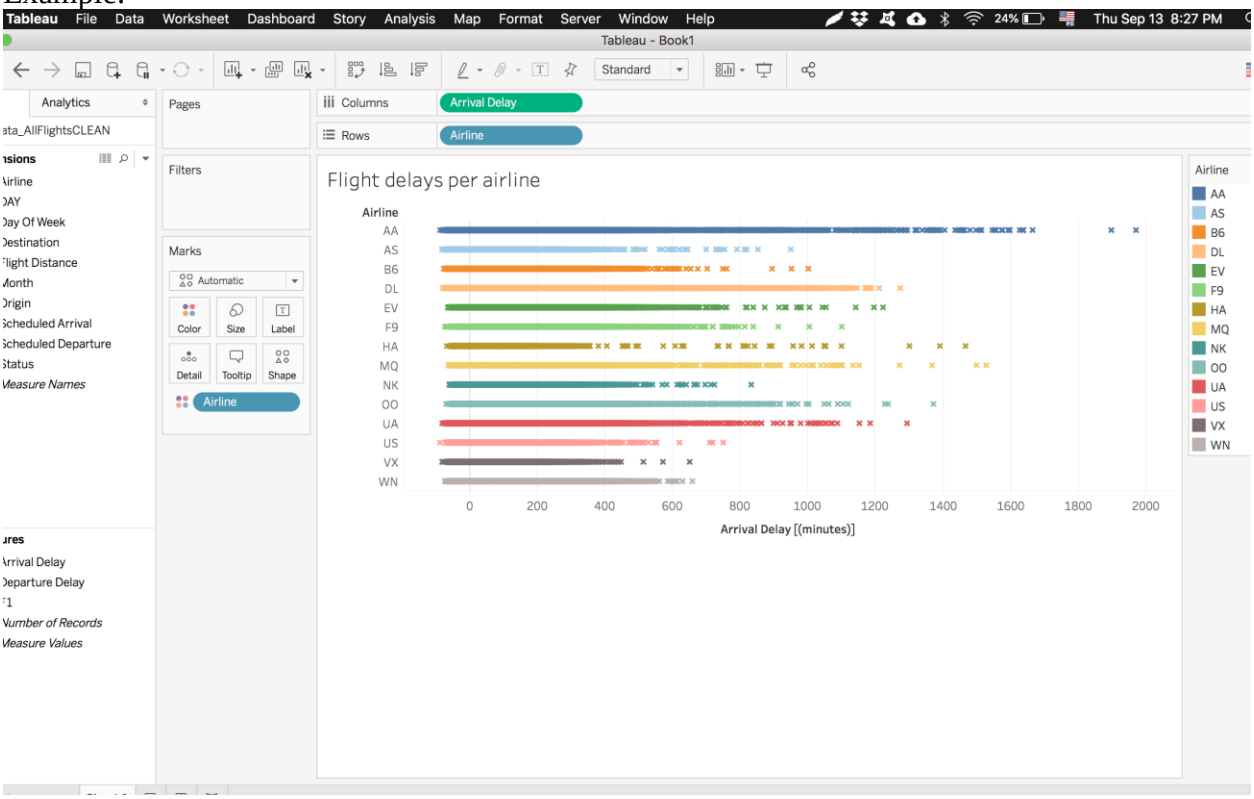
PREDICTIVE MODEL #2: THE AIRLINE WITH THE LARGEST NUMBER OF DELAYS vs. ALL AIRPORTS

```
Accuracy: 0.7974543961837407

5-fold Cross Validation: 0.7945099470585941

Confusion Matrix
[[ 20906  50133]
 [ 25360 276322]]

Classification Report
             precision    recall  f1-score   support

    delayed       0.45      0.29      0.36     71039
    on time       0.85      0.92      0.88    301682

avg / total       0.77      0.80      0.78    372721
```

V. DATA VALIDATION

Data validation is a crucial step of any data science project. To make sure that my results were accurate, I ran similar analyses in ***Tableau*** to double-check my Python code and analysis.

Example:



VI. CONCLUSION

Conclusions

This project has provided us with key insights on flight delays in the United States in 2015. Although 5,819,079 flights were recorded for 14 different airlines operating at 322 airports in total, 5,714,007 flight observations were eventually kept for the analysis due to missing elements in some of the rows of the original dataset. The initial exploratory results showed that the busiest airports in America in 2015 in terms of flight departures (on time, delayed, and cancelled) were the Atlanta, Chicago, Dallas/Fort Worth, Denver, and Los Angeles airports.

After examining each airline's flight delay statistics and visualizing the delays, we have learned that delays greatly vary per airline. For instance, the longest flight delay was recorded for American Airlines (1971.0 minutes, over 30 hours.) On the other hand, Hawaiian Airlines seemed to be the company that had the least delays in the plot. The results also showed that Spirit Airlines was the airline that had the highest *average* flight delay considering all flights (i.e. 14.47 minutes).

On the other hand, when looking at *median* flight delays we notice that they tend to be negative. For instance, Delta Airlines has a median flight delay of -8.0 min, which means that overall, Delta flights tend to arrive 8 minutes early than originally scheduled. Even though some airlines had many more delays than others, it was also relevant to compare the flight delay ratios per airline (considering the total number of flights operated by that same airline) instead of only looking at the count of delayed flights. In doing so, we found that Spirit Airlines was the airline that had the highest ratio of delayed flights (42.26% of Spirit Airlines flights were delayed.)

As for airports, we learned the Chicago O'Hare International Airport, the Hartsfield-Jackson Atlanta International Airport, the Dallas/ Fort Worth International Airport, the Denver International Airport, and the Los Angeles International Airport were the ones with the highest numbers of flight delays, with respectively 66,663; 56,462; 50,478; 43,331; and 40,281 delays. This list almost corresponds to the busiest airports list previously discussed, except that now Chicago is placed before Atlanta when it comes to flight delays, even though Atlanta is a bigger Airport hub overall.

When looking at the time of year, month, and day that typically have the most flight delays on a very surface level, no clear patterns emerged in all analyses. It seemed that the months of June, July, and December showed more delays in the histogram. As for days of a given month, the distribution did not show any clear patterns either.
However, when looking the time of day, clearly the majority of delayed flights happened in the evening, then in the afternoon, and then in the morning, as the histogram suggested. As a result, we may infer that an evening flight is much more likely to be delayed than a morning flight.
When looking at the days of week, no clear patterns emerged even though we could have expected Fridays, Saturdays, and Sundays to show more flight delays due to high passenger traffic.
Similarly, flight distance range did not seem to impact delayed flights too much. In the histogram, the number of delayed flights seem to be proportional to the number of flights in that range.

We learned that for delayed flights in 2015, the average air system delay (= the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control, etc.) was 13.48 mins, the average security delay (= evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas) was 0.08 mins, the average airline delay (= due to circumstances within the airline's control, such as maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.) was 18.97 mins, the average late aircraft delay (= a previous flight with same aircraft arrived late, causing the present flight to depart late) was 23.47 mins, and the average weather delay (= significant meteorological conditions – actual or forecasted – that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane) was 2.92 mins.
Even though weather does not seem to be impacting flight delays as much as other factors, the results showed that weather is nonetheless the main cause for flight cancellations overall, far ahead of airline/carrier reasons, ahead of the national air system.

Finally, the predictive analysis conducted showed that a Random Forest classifier was able to find patterns in the datasets to predict flight delays (with accuracies of approximately 80%), using 5-fold cross validation.

Next steps

The scope of this project had to remain within the scope of the class. However, I truly enjoyed exploring and working on it while using my new skillset revolving around Python. I do not consider this project to be finished – I plan to keep working on it to refine it, to eventually be able to use it in my Data Science portfolio.

Below is a list of some paths I plan to explore.

- I plan to use more advanced machine learning techniques on my clean datasets, as that is the area in which I want to improve the most. I will use other classifiers and compare their performance. My goal in this project was to use machine learning on a surface level just to see if there would be patterns and predictive potential in the data. Since the results suggested that the answer is yes, I definitely plan to explore this further.
- I will try to use the TAIL_NUMBER variable (which I excluded from this project) to retrieve the actual aircraft type (e.g. Boeing 737, Airbus A330, etc.) to see how that will impact the results. I expect this to be a significant contributing factor, as some airplanes tend to be less efficient than others (especially for smaller planes). The year of construction, maintenance information, engine types, number of passenger seats available, etc. could all be very useful information.
- I will also attempt to retrieve weather data for all those flights (e.g. https://www.wunderground.com/weather/api/), which could also greatly impact the results.