Airline Arrival and Departures

Carolyn Morikawa

ckm65

May 14th, 2019

Abstract:

Traveling from place to place can be extremely challenging, time consuming, and stressful, especially when a scheduled flight is cancelled or delayed. Many airline flights are delayed for a variety of different factors including: weather, technical issues, carrier crew, security, air systems, and late incoming aircraft. This article reviews, analyzes and presents information on the performance of airplane arrival and departures. As a result, my hypothesis examines if there is a correlation between the arrival of delayed airplanes and late departure airplanes. This article looks specifically at airline arrival and departure delays based on aircraft model, location, and specific days of the week. This analysis examines the following questions:Is there is a specific aircraft that arrives constantly arrives and departs late? Is there a specific location that consistently has late aircraft departures? Is there a specific location where airplanes typically have delayed arrivals? Is there a specific day of the week when airplanes are delayed? To answer these questions, *Airline Arrival and Departure* utilizes sophisticated data science tools to merge three unique datasets for further analysis.

1. INTRODUCTION

It's 7 pm on a Monday night, I am waiting in Chicago, IL International airport and my flight has been delayed for five hours and counting. I am traveling to New York City for a business meeting with a client on Tuesday at 8am. I anxiously begin to shift back and forth in the uncomfortable airport seats checking the time on my Apple watch. *Will I make it on time? What will my boss say?* These thoughts fill my mind as I try and take deep breaths. I am surrounded by screaming toddlers and other angry passengers waiting for the gate attendants to inform us that the flight is delayed for the fourth time. I make a mental note to never fly on this airline again. It is now 11pm and we are finally starting to board, 9 hours later, I am now angry, stressed and in a furious mood.

1.1 Purpose

The purpose of this analysis is to identify trends in delayed flights looking specifically at airplane departures and arrivals. In a real world context, flights are constantly delayed based on a variety of different factors. Machines and algorithmic processes are regularly viewed by humans to "work perfectly." However, machines make mistakes and malfunction which results in delays. Airplane flight delays are a real world occurrence that are not only frustrating, but also all too common today. It is important when traveling to understand and analyze common trends in typical delayed airline departures and arrivals before booking travel itineraries.

1.2 Key Objectives:

The key objectives of this analysis is to look at airline arrival and departure data to determine if:

1. One type of aircraft typically arrives and departs late to specific destinations? If so, which aircraft model? What is the significance?
2. A specific location constantly has delayed departures and receives delayed arriving flights? If so, which location and why do you think that is? What is the significance?
3. A specific day of the week has more delayed departing and arriving flights? What is the significance?

1.3 Significance

The significance of this article is to provide the viewer, frequent traveler, or upcoming traveler with information regarding delayed flights. The information is meant to inform the traveler about specific correlations in order to better prepare and avoid the agony of flight delays in the future.

2. DATA COLLECTION

In order to reach my conclusions, three datasets were merged to link the necessary variables for analysis.

2.1 On Time Performance Data

This dataset was created to map the scheduled and actual departure and arrival times of airplane flights reported by certified US air carriers. The dataset was created to measure the performance of specific airplanes in 2019. The data was collected by the Office of Airline Information and the Bureau of Transportation Statistics. This large dataset contains 583,986 data entries with 110 columns associated with specific flights for the month of January 2019. To clean this dataset, I parsed specific column names in order to identify the key fields required for my analysis.The field entries I extracted included:

- Tail Number = Aircraft Tail Number
- DayOfWeek = Day of Week
- DestCityName = Destination Airport, City Name
- OriginCityName = Origin Airport, City Name
- ArrDelayMinutes = Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
- DepDelayMinutes = Difference in minutes between scheduled and actual departure time. Early departures set to 0.
- ArrDel15 = Arrival Delay Indicator, 15 Minutes of More (1=Yes)
- DepDel15 = Departure Delay Indicator, 15 Minutes of More (1=Yes)
- AirTime = Time of arrival

2.2 Aircraft Registration Master File

This dataset contains the records of all U.S Civil Aircrafts maintained by the FAA, Civil Aviation Registry, and Aircraft Registration Branch. This dataset was collected by the U.S Department of Transportation and the Federal Aviation Administration.The master file is a combination of six different registry datasets related to different registry information including: Aircraft Registration Master file, Aircraft Document Index file, Aircraft Reference file, Deregistered Aircraft file, Engine Reference file, and Reserve N-Number file. For my key objectives, I utilized the Aircraft Registration Master file to view the entire registry data as a whole. The master file contains 292,718 total entries with 35 total columns related to registered US aircrafts. The master elements that I extracted to analyze include:

- N-Number = A unique number assigned to an aircraft
- MFR MDL CODE = A code assigned to the engine manufacturer and model.

Positions (46-48) - Manufacturer Code Positions (49-50) - Model Code Positions (51-52)

2.3 Aircraft Reference File (ACTREF)

This dataset contains registry information about various aircraft models. It is a subset of the registry mastery file used in Section 2.2. This dataset was collected by the U.S Department of Transportation and the Federal Aviation Administration. This dataset is relatively small compared to the other two containing 85,777 total entries and 11 column headings. From this dataset, I extracted the various column headings:

- CODE = A code assigned to the aircraft manufacturer, model and series.
- MFR = Name of the aircraft manufacturer.
- MODEL= Name of the aircraft model and series.

3. DATA CLEANING

3.1 Merging Performance and Master Dataset on Tail_Number

In order to examine the correlation between types of aircraft and late arrivals and departures, I was required to link On Time Performance data with the Aircraft Registration Master file. Unfortunately, the tail numbers, a unique number used to identify an aircraft, were labeled differently in the two datasets. In the On Time Performance dataset, the field was labeled as "Tail_Number" and in the Aircraft Registration Master file, the field was labeled "N-Number". In order to link the files, I needed to clean the data in both files. From the On Time Performance dataset, I noticed that for the majority of the values, the first element of the "Tail_Number" began with an "N". For all "Tail Numbers" that did not begin with an "N", I concatenated an "N" as the first character in the field. In addition, for the Aircraft Registration Master file, I needed to add an "N" to each entry since the field did not contain an "N" as the first digit of the number. Since there was no common variable name to merge the two datasets on, a new column was added to the Aircraft Registration Master dataset that was named "Tail_Numbers" as a string. The total number of "Tail_Numbers" from the On Time Performance data including repetitions was 583,985. However, there were only 5,447 individual unique "Tail_Numbers", which was indicated when the "Tail_Number" was set(). The repetitions indicate the recording of airplanes leaving and arriving at different destinations. The "Tail Numbers" and the "N_Number" are matched utilizing the intersection method to discover 5,408 matching. This indicates that for 5,408 "Tail_Numbers" there is a corresponding "N_Numbers". From the number of matchings, the On Time Performance and the Aircraft Registration Master datasets were merged. The variables mentioned in Section 2.1 and 2.1 were the variables included in the merge. Therefore the new dataset, called performance_master was created and cleaned containing the matched "Tail_Numbers" with the corresponding variables from both datasets:

“MFR MDL CODE “, “Tail_Number”, “Year”, “ArrDelayMinutes”, “DepDelayMinutes”,“DayOfWeek”, “DestCityName”,”OriginCityName”.

3.2 Merging Performance_Master and ACTREF

In order to identify the specific type of aircraft and model, the “MFR MDL CODE” variable from the performance_master new dataset, was matched with the “CODE” variable in the ACTREF dataset. The “MRF MDL CODE” contains 571,847 entries and 161 unique entries. The most common “MRF MDL CODE” is ‘1384404’ containing 56,318 reptitions. This makes sense because the “MRF MDL CODE” is the very popular BOEING 737- 7H4 model which is flown by many US airlines. It would be expected due to the high frequencies of flights per day that the “MRF MDL CODE” would be recorded numerous times. The “CODE” variable from the ACTREF dataset has 85,776 unique entries. Since the variable names do not match up a column to the performance_master dataset was added containing the variable name “CODE” which was equal to the values in the “MRF MDL CODE”. The number of times the “MRF MDL CODE” intersected with the “CODE” was 161 entries. The two datasets were merged to create the performance_master_aircraft dataset containing a unique dataset from three different datasets. The variables of the final dataset include: “MFR MDL CODE “, “Tail_Number”, “Year”, “ArrDelayMinutes”, “DepDelayMinutes”,“DayOfWeek”, “DestCityName”,”OriginCityName”. “CODE”,”MODEl” “MFR” , “ArrDel15”, “DepDel15”

3.3 Filters

The final step to cleaning datasets was creating specific filters to remove excess or unneeded information. For specific variables such as the “ArrDelyMinutes” and”DepDelayMinutes” some of the entries were blank or equivalent to ‘NaN.’ ‘NaN’ automatically overrides values and in order to prevent errors in my data, it was necessary to remove everything that had “NaN” by creating filters.Thankfully, this filter did not result in a huge loss of data. Before the filter, the length of the dataset was 571,847. After the filter the dataset was shortened to 556,483 results. I utilized the isnull() pandas packages that analyzes the data frame checking for “NaN” values and returning them as True

- For example:
  performance["ArrDelayMinutes"][performance["ArrDelayMinutes"].isnull()]
  non = performance[~performance["ArrDelayMinutes"].isnull() &
~performance["DepDelayMinutes"].isnull()]

4. DATA ANALYSIS RESULTS + SIGNIFICANCE

Looking back at the key objectives from 1.2. The following sequences of data analysis will display the answers to these questions through data visualization graphs and coding techniques.

4.1 Aircraft on Arrival and Departure

Coding

The goal of this section is to calculate the average number of minutes delayed for arrival flights of the top 10 most frequent aircrafts. A boxplot will display the average delay in minutes for each aircraft model for arriving flights. The same will be done for the minutes delayed for departing flights.

Utilizing the "CODE" from the dataset performance_master_aircraft which is the merging of three datasets, I was able to create a Counter for the 10 most common codes numbers that link specifically to the type of Aircraft. I used **clustering** of the "CODE" field to determine the 10 most frequent aircrafts in the dataset

[('1384404', 56318), ('1390008', 40615), ('1390016', 37865), ('3260415', 35033), ('1390015', 25651), ('13844CB', 24716), ('3940032', 19467), ('3940004', 18754), ('3260212', 17227), ('1383700', 16917)].

From there I was able to match that specific CODE(aircraft identification number) to the MODEL and MFR of the top most common identification code numbers:

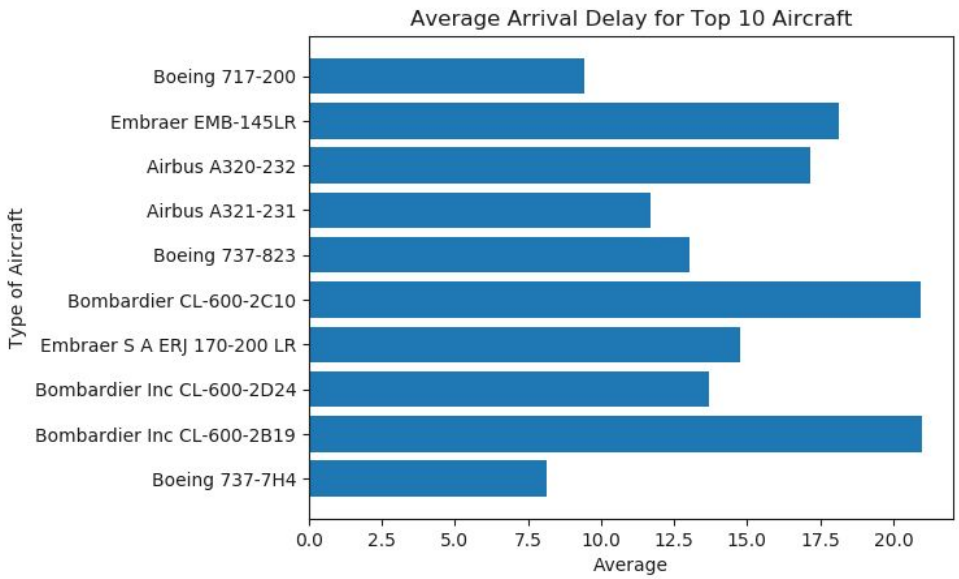
1384404 = BOEING 737- 7H4
1390008 = BOMBARDIER INC CL-600-2B19
1390016 = Bombardier Inc CL-600-2D24'
3260415 = EMBRAER S A - ERJ 170-200 LR
1390015 = BOMBARDIER INC CL-600-2C10
13844CB = BOEING 737-823
3940032 = Airbus A321-231
3940004 = Airbus A320-232
3260212 = Embraer EMB-145LR
1383700 = Boeing 717-200

Utilizing the Pandas mean() function on the variable ["ArrMinutesDelay"] and ["DepMinutesDelay"]. I was able to calculate the descriptive statistics of the average number of minutes delayed for the top 10 aircraft listed above.

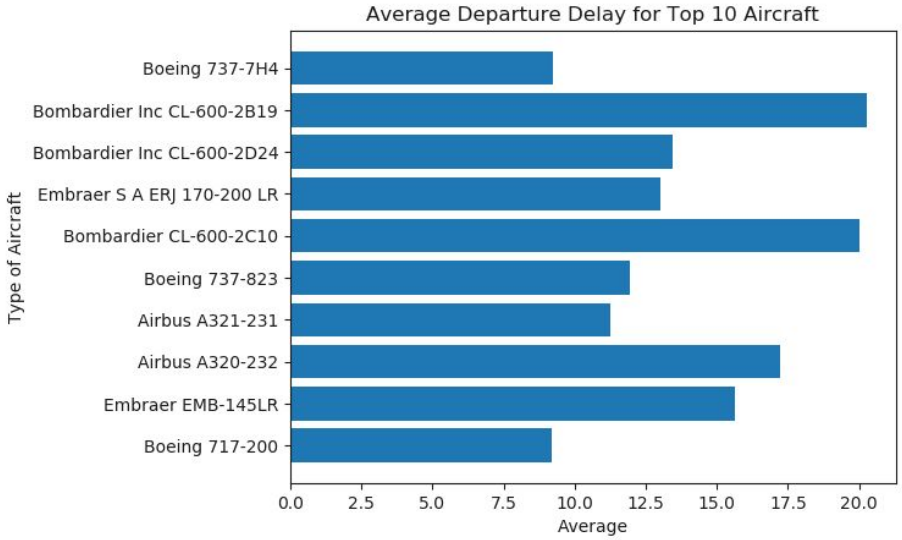
By utilizing descriptive statistics of the mean, and standard deviation, I was able to find the aircraft that had the most arrival delays. The results show that for the top 10 aircrafts used, the 1390008 = BOMBARDIER INC CL-600-2B19 had the highest average of minutes delayed for arrival (20.99 seconds) with a standard deviation of 69.15. The aircraft with the highest standard deviation for arrivals was Bombardier Inc CL-600-2C10.

By utilizing descriptive statistics of the mean and standard deviation, I was able to find the aircraft departing delays. The results show that for the top 10 aircrafts used, the 1390008 = BOMBARDIER INC CL-600-2B19 had the highest average for minutes delayed for departures (20.28 seconds) with a standard deviation of 69.39.

Visuals:



I further constructed a bar chart for the type of aircraft and the average number of minutes delayed for arrival and departure. The significance of the boxplot allows the viewers to comparatively see the average mean of the aircraft type and then average number of minutes.



Significance Testing:

To test if the average time of arrival delays for aircraft number '1390008' is significantly different from the average amount for delay time for a typical aircraft, I examined aircraft

number ‘1390008’ which is the Bombardier Inc CL-600-2B19 aircraft because it consistently arrived late.

The *null hypothesis* is that there is not any significant difference between the average amount in minutes delayed for the ‘1390008’ which is the Bombardier Inc CL-600-2B19 and the average time in minutes of a typical aircraft.

After calculating a t- test the p value was 4.96e-219. Since our p<value < alpha = .05, we reject the null hypothesis at the 95% confidence level. This means that we cannot conclude that there is no difference between the two means. Looking at the 5 underlying assumptions for t- tests it can be concluded that even with a random sampling through permeation test, it can be expected that the null hypothesis will again be rejected. Therefore, there is a significant difference between the average amount of time in minutes of the Bombardier Inc aircraft which makes sense considering it had the highest average arrival delays.

Overall Significance:

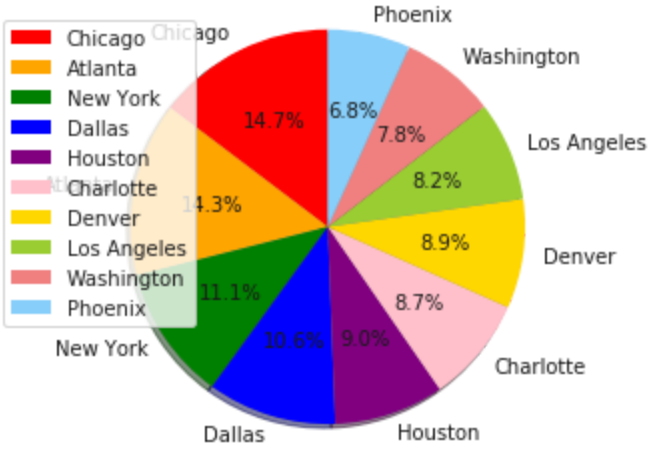
The significance of this data is that is seems as though Bombardier Inc is constantly late and is the least reliable. This matters for planning trips in the future to avoid flying on Bombardier Inc specifically the CL-600 modeled planes. Typically, if an aircraft arrives late it has a high probability of departing late. From the results, Bombardier Inc CL-600-2B19 has a slightly higher higher average of delayed for departures than arrivals which I can infer could be from runway traffic or lack of crew members.

4.2 Location(Origin, Destination) on Arrival and Departure

The goal of this section is to determine if one specific origin or destination city arrives or departs late more frequently than others. Does an airplane from a specific city consistently arrive late? Does an airplane from an origin city typically depart late? I found the top 10, most common origin and destination cities counting the amount of flights going in and out of that specific destination.

Top 10 Origin City and Counts:

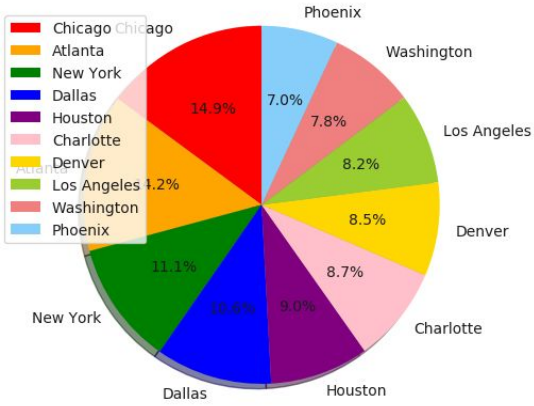
[('Chicago, IL', 31903), ('Atlanta, GA', 30964), ('New York, NY', 24030), ('Dallas/Fort Worth, TX', 22954), ('Houston, TX', 19567), ('Charlotte, NC', 18900), ('Denver, CO', 18429), ('Los Angeles, CA', 17660), ('Washington, DC', 16892), ('Phoenix, AZ', 14703)]



This pie chart represents the frequency of each city for arrival flights. This shows significance in the data that Chicago, IL is the most common destination out of all cities in the US with an average of flights being 14.7%.

Top 10 Destination City and Counts:

[('Chicago, IL', 32624), ('Atlanta, GA', 31151), ('New York, NY', 24365), ('Dallas/Fort Worth, TX', 23078), ('Houston, TX', 19588), ('Charlotte, NC', 19105), ('Denver, CO', 18498), ('Los Angeles, CA', 17977), ('Washington, DC', 17066), ('Phoenix, AZ', 15218)]



This pie chart represents the frequency of each city for departing flights.This shows significance in the data that Chicago, IL is the most common destination out of all cities in the US with an average of flights being 14.9%.

For each of the top 10 cities I ran four different tests on the data.

1) Origin City Name and the Arrival delay minutes.
2) Origin City Name and Departing delay minutes.
3) Destination City name and Arrival delay minutes
4) Destination City name and departing delay minutes

Next, I ran descriptive statistics calculating the mean() and standard deviation from python pandas packages to determine significance in the results.

The conclusions were as follows:
For each of the 4 tests:
1) Origin city Chicago, IL had an average arrival delay of 23.79 minutes which was the largest in comparison to the top 10 most frequent cities.
2) Origin city Chicago, IL had an average departure delay of 21.96 minutes which was the largest in comparison to the top 10 most frequent cities.
3) Destination city Chicago, IL had an average arrival delay of 24.26 minutes which was the largest in comparison to the top 10 most frequent cities.
4) Destination city New York, NY had an average arrival delay of 23.77 minutes which was the largest in comparison to the top 10 most frequent cities.

Significance Testing:

The question I am investigating "Is the average amount of arrival delay time for Chicago, IL significantly different from the average amount for delay time for a specific US city". I utilized location Chicago, IL because it consisted had the highest frequency of departures and arrivals.

The *null hypothesis* is there no significant difference between the average amount in minutes delayed for Chicago, IL and the average time in minutes of any other US city location.

After calculating a t- test the p value was5.9e-308 and had a statistical value of -37.53. Since our p<value < alpha = .05, we reject the null hypothesis at the 95% confidence level. This means that we cannot conclude that there is no difference between the two means. Looking at the 5 underlying assumptions for t- tests it can be concluded that even with a random sampling through permeation test, it can be expected that the null hypothesis will be rejected again.

Furthermore running a significance t-test and culcating the variables for the 4 sections listed above are as follows. As you can see the p values are all very close to 0 rejecting the null hypothesis. Since it rejects the null hypothesis, we can conclude that there is a significant difference which makes sense with our data because Chicago, IL has the highest amount of delayed and arriving flights.

Overall Significance:
The overall significance of this dataset illustrates that Chicago, IL tends to be the central location where most flights are late both arriving and departing.This is most likely because it is a large

city, central hub for international flights and airlines, and often has poor weather. In the real world, this makes sense.

4.3 Day of Week on Arrival and Departure

Code

In this section, I looked at specific days of the week to determine if there was a day of the week that typically had the most delayed flights based on arrival and departures. I utilized the "DayOfWeek" variable which was indicated in terms of numbers 1-7. I set each number to the corresponding day of the week starting with Monday referencing 1 in the dataset.

From the analysis in section 4.2, Chicago, IL was the city that consistently had the highest average number of delays for departing and arriving flights. For Chicago, IL I mapped each flight for each day of the week to determine if there was a day that harbored the most delayed arriving and departing flights. For each day of the week I calculated the maximum, minimum, mean and standard deviation for the minutes delayed for arrival and departures.

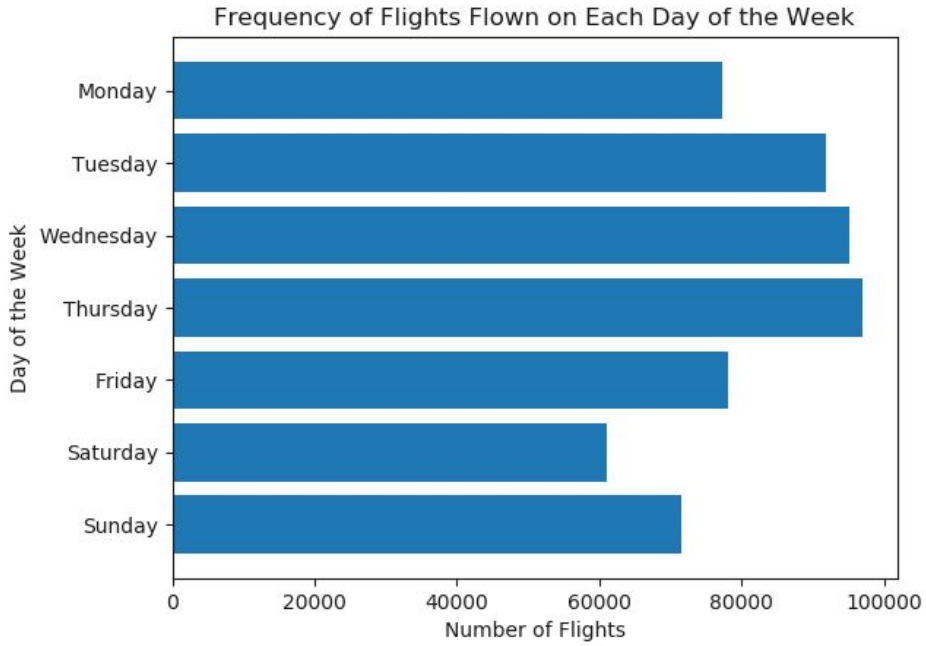
The results concluded that the day of the week where flights from Chicago, IL departed late was Monday.

The Number in the Dataset For Monday Departing to Chicago arriving late:
3756
The Maximum is:
1383.0
The Minimum is:
0.0
The Sum is:
124529.0
The mean
14.86164646665342
The standard deviation:
53.02855410528787

This makes sense because most people who travel for business, travel on Mondays. The heavy traveler traffic and volume, increases the probability for flight delays. The results concluded that the day of the week where flights from Chicago, IL departed late was Monday.

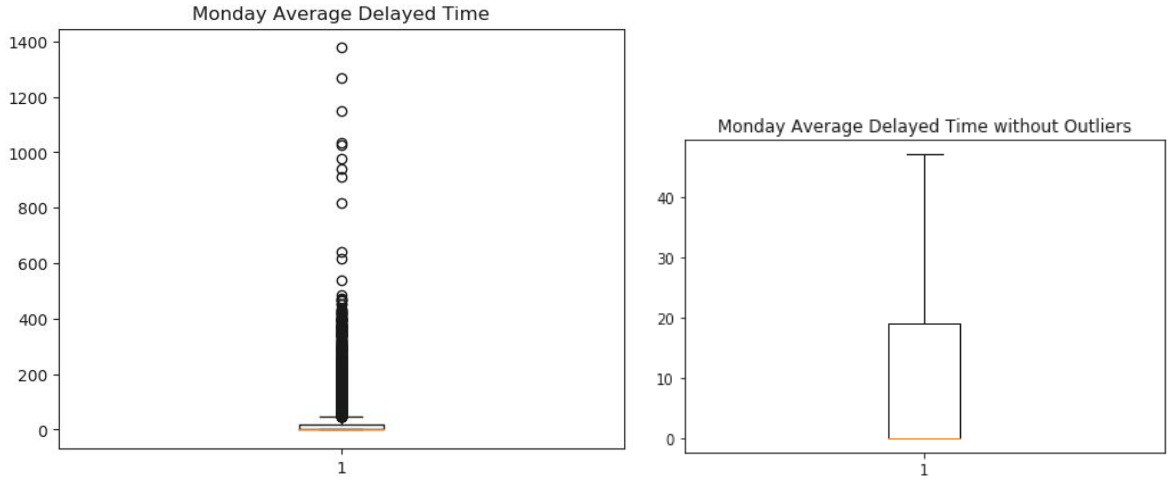
The Number in the Dataset For Thursday Arriving to Chicago arriving late:
4799
The Maximum is:
1259.0
The Minimum is:
0.0
The Sum is:
83038.0
The mean

15.286518161240506
The standard deviation:
46.58794277139517



Visuals

I incorporate whisker plots to accurately represent my data. A whisker plot was the best way to represent my data to incorporate the mean, maximum, and minimum delay values in minutes. As you can see from the graphs there are a lot of outliers which skewed the data. Below is a whisker plot in terms of Monday and Sunday. I created whisker plots for everyday of the week.



Overall Significance:

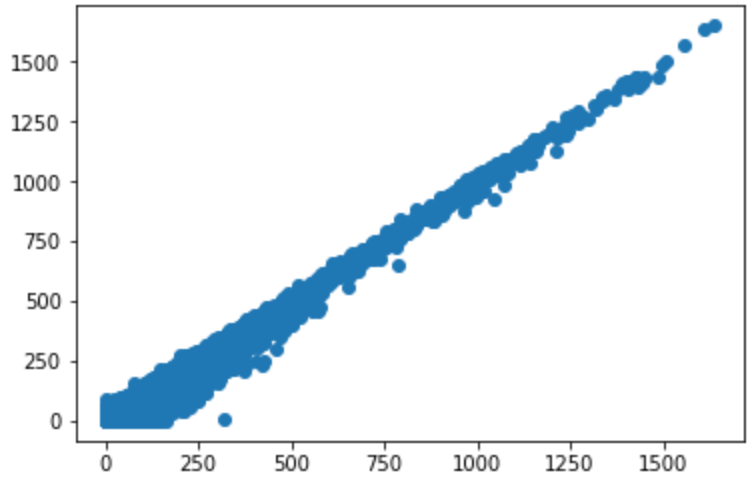
In preparation for an upcoming flight it seems as though Monday harbors the most common delayed flights. After a weekend, many individuals travel for business meetings on Monday resulting in a large number of people flying on that specific day.

4.4 Logistic/Linear Regression

Predictive Analysis

Linear

For my predictive analysis portion of Airline *Arrival and Departures* I ran a linear regression on my original hypothesis that there will be a positive correlation between the arrival delayed minutes and departure delayed minutes. This is reflective that if a flight arrives late it typically departs late. I created a scatter plot and ran a linear regression function to conclude that the correlation coefficient of .97 which is pretty close to 1.



Logistic

I decided to run a logistic regression on the amount of AirTime and if a flight arrived late or not. I used the variables "ArrDel15" which marks a binary flight with 1 being late if arrives more than 15 minutes late if not 0.

From the results of my logistic regression, it can be concluded that there is no correlation between the amount of airtime and delayed flights. The precision value for predicting 0 is .86 while the precision for predicting 1 is .19. Therefore, it seems as though it automatically generates to predict that the aircraft will not arrive more than 15 minutes late. This makes sense since the average for arriving delayed for most aircrafts was around 12-13 minutes. This model is good at guessing when something will arrive on time. There is significant bias towards arriving on time compared to late. Therefore, how long that an aircraft is flying can not be predictive if an airline will arrive late.

	precision	recall	f1-score	support
0.0	0.82	0.61	0.70	68169
1.0	0.19	0.41	0.26	15304
avg / total	0.71	0.58	0.62	83473

5. CONCLUSION

5.1 Summary Findings

After pre-processing, cleaning, filtering, and merging three different distinct datasets I was able to come to conclusions about my key objectives and hypothesis. My hypothesis being that there is a positive correlation between arrival delayed and departing delayed.

Does one type of aircraft typically arrive and depart late to specific destinations? If so, which aircraft model?

It was concluded by my code, visual representations and hypothesis testing that Bombardier Inc aircraft specifically the CL-600 models rarely depart and arrive on time. Specifically the BOMBARDIER INC CL-600-2B19 is the most frequently delayed aircraft.

Does a specific location consistently have a delayed departure and receive delayed arriving flights? If so, which location and why do you think that is?

The data revealed that Chicago IL is the most common location to have delayed departures and arrivals as well as receive delayed flight arrivals. As the origin city Chicago, IL had the most arriving and departing delayed flights. As the departure city, Chicago, IL had the the most arriving delayed flights. However, New York, NY as the departure city, had the most departing delayed flights. This makes sense given that they are two very large cities with heavy aircraft traffic resulting in having a higher frequency of delayed flights than other cities.

Does a specific day of the week consistently have more delayed departing and arriving flights?

It was concluded that Monday was the day of the week that had the most delayed departing and arriving flights.

5.2 Future Analysis

To do further analysis on this data, it is possible to identify the most common reasons for delayed flight arrivals and departures. I could evaluate my data further by using random account to calculate permeation tests for my hypothesis testing. I could also look at logistic regression

comparing the difference between flights being delayed 5, 10, 15 minutes. It would be expected that more flights are delayed when it is 5 minutes or more.

5.4 Thank You!

Thanks so much for a great semester. I hope you enjoyed reading my paper. :)