SPRINGBOARD DATA SCIENCES CERTIFICATION PROGRAM CAPSTONE PROJECT

MICHAEL ENGLISH

INTRO TO M.E

- Michael English
- Double Major in Mathematics and Music
- Student Researcher
- Violinist/Violist

WHAT'S ALL THIS ABOUT?

HARTSFIELD-JACKSON INTERNATIONAL AIRPORT (ATL)

Ask-A-Librarian | A-Z Index

Bureau of Transportation Statistics

Topics and Geography | Statistical Products and Data | National Transportation Library | Newsroom

Select a month: March 2019 (The month selection does not apply to on-time data.)

Select an airport: Atlanta, GA: Hartsfield-Jackson Atlanta Inte Submit

Show all airports (by state)

Atlanta, GA: Hartsfield-Jackson Atlanta International (ATL)

Scheduled Services except Freight/Mail

BTS Data as of 7/26/2019

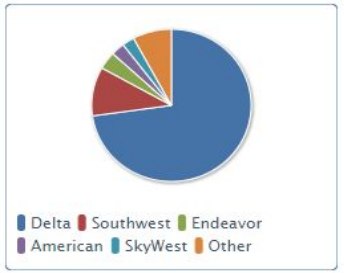
Summary Data (U.S. Flights Only)

Passengers*	2018**	2019**	%Chg	Rank***
Arrival	44,750k	46,242k	3.33%	1
Departure	44,705k	46,184k	3.31%	1
Scheduled Flights				
Departures	388,662	395,013	1.63%	1
Freight/Mail (lb.) (Scheduled and Non-Scheduled)				
Total	662m	664m	0.29%	17
Carriers				
Scheduled	23	24	4.35%	

* Scheduled enplaned revenue passengers.
** 12 months ending April of each year.
*** Among 786 U.S. airports, 12 months ending April 2019

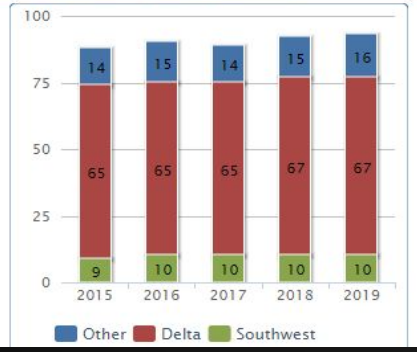
Carrier Shares for May 2018 - April 2019

Carrier	Passengers	Share
Delta	67,269	72.78%
Southwest	9,593	10.38%
Endeavor	3,434	3.71%
American	2,558	2.77%
SkyWest	2,311	2.50%
Other	7,261	7.86%

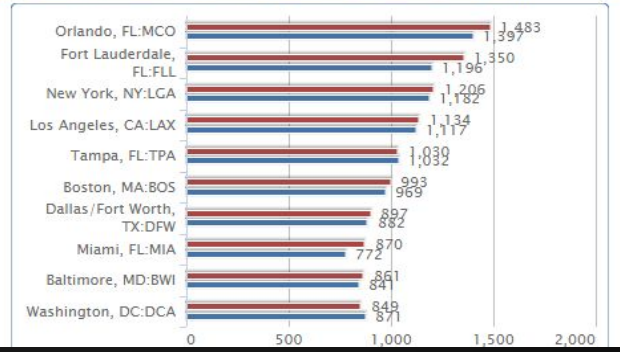


Based on enplaned passengers(000) both arriving and departing.

Total Passengers (U.S. Flights, in millions)



Top 10 Destination Airports (U.S. Only, Passengers (000))



Source: https://www.transtats.bts.gov/airports.asp?pn=1

WHAT IS THE QUESTION?

- More consumers utilize air travel than before to fly around the world. Record inbound and outbound flights at Hartsfield-Jackson International Airport (ATL) have lead to ATL consistently being named the world's busiest airport. Since the concentration of people is dense at ATL there is potential to study the social and fiscal trends associated with the steady increasing flow of people through checkpoints. Online resources collect various consumer and aviation data that can be organized to illuminate the notion of using flight routes to determine highly populated times and days at ATL.

- This research will utilize various data sets that include passenger throughput, maximum number of passengers per flight, aircraft load data, and popular flight routes from Atlanta to predict the population of passengers in the airport. Results from this research determine when ATL is overcrowded through the comparison of passengers enplaning that are traveling to the six most populous flight routes from Atlanta: Fort Lauderdale, FL; Orlando, FL; Tampa, FL; New York, NY; Los Angeles, CA; and Boston, MA.

HOW WE DO IT

- Data Wrangling :: Cleaning and Organizing Data for Efficient Use
- Exploratory Data Analysis :: After Cleaning, a comprehensive analysis of the data and structure
- Statistical Analysis and Data Visualization :: Visual charts, plots, and graphs which allow for statistical inference.
- Machine Learning :: Applying regression models to datasets to determine actively "predict" what is happening, mathematically

THE DATASETS

To develop a model for predicting the population at Hartsfield-Jackson Airport (ATL) on a particular day, specific information is retrieved and merged from a list of publicly available databases.

- Atlanta Routes Database [1]: In this database, the total number of passengers that enplaned in ATL in March 2019 are listed.
- BTS Database [2]: This database provides information regarding all flights enplaning at ATL for the month of March in 2019.

	A	B	C	D	E	F
1	a	FL_DATE	TAIL_NUM	ORIGIN_CITY_N	DEST_CITY_NAME	
2	2019	3/1/2019	N343FR	Atlanta, GA	New York, NY	
3	2019	3/1/2019	N117HQ	Atlanta, GA	New York, NY	
4	2019	3/1/2019	N443YX	Atlanta, GA	New York, NY	
5	2019	3/1/2019	N535NK	Atlanta, GA	Fort Lauderdale, FL	
6	2019	3/1/2019	N525NK	Atlanta, GA	Fort Lauderdale, FL	
7	2019	3/1/2019	N515NK	Atlanta, GA	Fort Lauderdale, FL	
8	2019	3/1/2019	N948UW	Atlanta, GA	New York, NY	
9	2019	3/1/2019	N7722B	Atlanta, GA	Fort Lauderdale, FL	
10	2019	3/1/2019	N560WN	Atlanta, GA	Fort Lauderdale, FL	
11	2019	3/1/2019	N479WN	Atlanta, GA	Fort Lauderdale, FL	
12	2019	3/1/2019	N469WN	Atlanta, GA	Fort Lauderdale, FL	
13	2019	3/1/2019	N553WN	Atlanta, GA	Fort Lauderdale, FL	
14	2019	3/1/2019	N487WN	Atlanta, GA	New York, NY	
15	2019	3/1/2019	N441WN	Atlanta, GA	New York, NY	
16	2019	3/1/2019	N944WN	Atlanta, GA	New York, NY	
17	2019	3/1/2019	N565WN	Atlanta, GA	New York, NY	
18	2019	3/1/2019	N8317M	Atlanta, GA	New York, NY	
19	2019	3/1/2019	N706JB	Atlanta, GA	Fort Lauderdale, FL	
20	2019	3/1/2019	N531JL	Atlanta, GA	New York, NY	
21	2019	3/1/2019	N613JB	Atlanta, GA	Fort Lauderdale, FL	
22	2019	3/1/2019	N592JB	Atlanta, GA	New York, NY	
23	2019	3/1/2019	N656NK	Atlanta, GA	Fort Lauderdale, FL	
24	2019	3/1/2019	N656NK	Atlanta, GA	Fort Lauderdale, FL	
25	2019	3/1/2019	N660NK	Atlanta, GA	Fort Lauderdale, FL	
26	2019	3/1/2019	N952UW	Atlanta, GA	New York, NY	
27	2019	3/1/2019	N410YX	Atlanta, GA	New York, NY	
28	2019	3/1/2019	N417YX	Atlanta, GA	New York, NY	
29	2019	3/1/2019	N907DN	Atlanta, GA	New York, NY	
30	2019	3/1/2019	N358NW	Atlanta, GA	New York, NY	
31	2019	3/1/2019	N363DN	Atlanta, GA	Fort Lauderdale, FL	
32	2019	3/1/2019	N673DL	Atlanta, GA	Fort Lauderdale, FL	
33	2019	3/1/2019	N339DN	Atlanta, GA	Fort Lauderdale, FL	
34	2019	3/1/2019	N683DA	Atlanta, GA	Fort Lauderdale, FL	
35	2019	3/1/2019	N556NW	Atlanta, GA	Fort Lauderdale, FL	
36	2019	3/1/2019	N6714Q	Atlanta, GA	Fort Lauderdale, FL	
37	2019	3/1/2019	N320DN	Atlanta, GA	Fort Lauderdale, FL	
38	2019	3/1/2019	N554NW	Atlanta, GA	Fort Lauderdale, FL	
39	2019	3/1/2019	N347DN	Atlanta, GA	Fort Lauderdale, FL	
40	2019	3/1/2019	N659DL	Atlanta, GA	Fort Lauderdale, FL	

[1]

12756	WN	425.2	Southwest Airlines Co.	Atlanta, GA	New Orleans, LA	2019	3
12399	DL	413.3	Delta Air Lines Inc.	Atlanta, GA	Providence, RI	2019	3
12333	WN	411.1	Southwest Airlines Co.	Atlanta, GA	Washington, DC	2019	3
12250	WN	408.3333333	Southwest Airlines Co.	Atlanta, GA	Los Angeles, CA	2019	3
11902	AA	396.7333333	American Airlines Inc.	Atlanta, GA	Los Angeles, CA	2019	3
11698	DL	389.9333333	Delta Air Lines Inc.	Atlanta, GA	Albany, NY	2019	3
11598	WN	386.6	Southwest Airlines Co.	Atlanta, GA	Nashville, TN	2019	3
11597	AA	386.5666667	American Airlines Inc.	Atlanta, GA	Philadelphia, PA	2019	3
11268	DL	375.6	Delta Air Lines Inc.	Atlanta, GA	Madison, WI	2019	3
11188	DL	372.9333333	Delta Air Lines Inc.	Atlanta, GA	Des Moines, IA	2019	3
11037	DL	367.9	Delta Air Lines Inc.	Atlanta, GA	Wichita, KS	2019	3
10963	DL	365.4333333	Delta Air Lines Inc.	Atlanta, GA	Columbia, SC	2019	3
10756	WN	358.5333333	Southwest Airlines Co.	Atlanta, GA	Kansas City, MO	2019	3
10486	WN	349.5333333	Southwest Airlines Co.	Atlanta, GA	Austin, TX	2019	3
10356	DL	345.2	Delta Air Lines Inc.	Atlanta, GA	Tallahassee, FL	2019	3
10330	DL	344.3333333	Delta Air Lines Inc.	Atlanta, GA	Tucson, AZ	2019	3
10314	9E	343.8	Endeavor Air Inc.	Atlanta, GA	Knoxville, TN	2019	3
9923	WN	330.7666667	Southwest Airlines Co.	Atlanta, GA	Jacksonville, FL	2019	3
9713	DL	323.7666667	Delta Air Lines Inc.	Atlanta, GA	Tulsa, OK	2019	3
9517	AA	317.2333333	American Airlines Inc.	Atlanta, GA	Phoenix, AZ	2019	3
9470	WN	315.6666667	Southwest Airlines Co.	Atlanta, GA	Richmond, VA	2019	3
9441	DL	314.7	Delta Air Lines Inc.	Atlanta, GA	El Paso, TX	2019	3
9270	DL	309	Delta Air Lines Inc.	Atlanta, GA	Albuquerque, NM	2019	3
9238	NK	307.9333333	Spirit Air Lines	Atlanta, GA	Baltimore, MD	2019	3
9163	NK	305.4333333	Spirit Air Lines	Atlanta, GA	Detroit, MI	2019	3
8755	WN	291.8333333	Southwest Airlines Co.	Atlanta, GA	Raleigh/Durham, NC	2019	3
8714	DL	290.4666667	Delta Air Lines Inc.	Atlanta, GA	Harrisburg, PA	2019	3
8709	WN	290.3	Southwest Airlines Co.	Atlanta, GA	Phoenix, AZ	2019	3
8690	WN	289.6666667	Southwest Airlines Co.	Atlanta, GA	Indianapolis, IN	2019	3
8602	YV		Mesa Airlines Inc.	Atlanta, GA	Houston, TX	2019	3
8567	WN	285.5666667	Southwest Airlines Co.	Atlanta, GA	Columbus, OH	2019	3
8454	OO		SkyWest Airlines Inc.	Atlanta, GA	Chattanooga, TN	2019	3
8429	9E	280.9666667	Endeavor Air Inc.	Atlanta, GA	Augusta, GA	2019	3
8390	OO		SkyWest Airlines Inc.	Atlanta, GA	Montgomery, AL	2019	3
8329	9E	277.6333333	Endeavor Air Inc.	Atlanta, GA	Gulfport/Biloxi, MS	2019	3

[2]

DATA WRANGLING

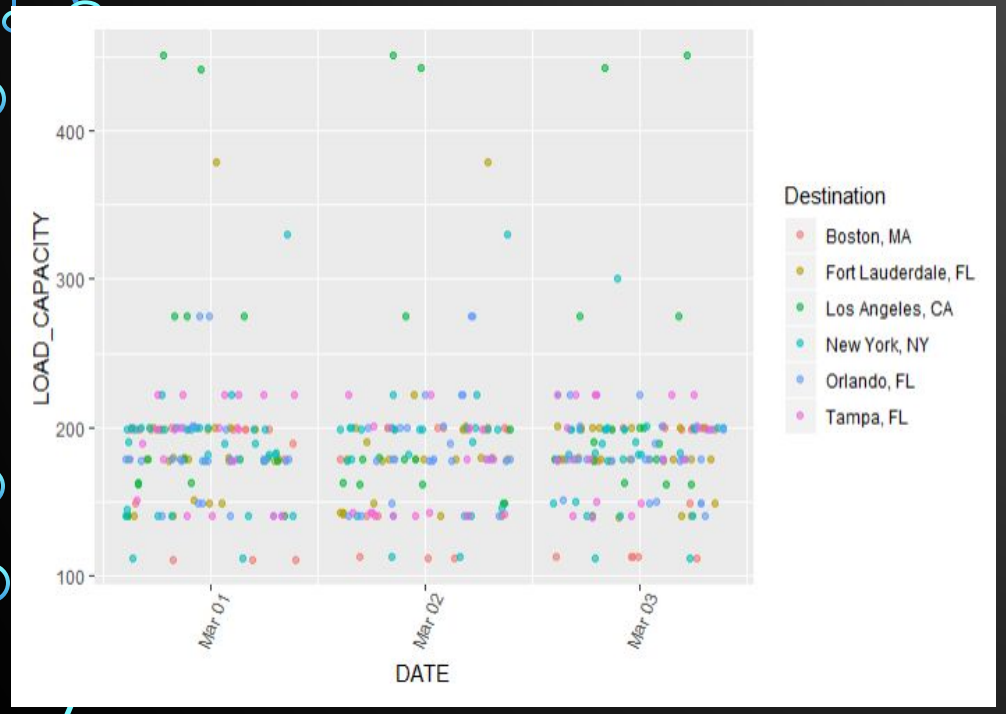
Cleanliness may be defined to be the emblem of purity of mind.
- Joseph Addison

Looking at the structure (str) of BTS, we see that there are an array of columns removed and added:

- YEAR and ORIGIN_CITY_NAME were variables available in both datasets, and thus redundant. The extra set of rows were removed.
- Developed a sub-dataset that refines the collection to 6 points, collective, based on the six cities. The load factors of aircraft, averages based on the flights to the six

YEAR <dbl>	DATE <S3: POSIXct>	TAIL_NUM <chr>	LOAD_CAPACITY <dbl>	ORIGIN_CITY_NAME <chr>	Destination <chr>	Avg_LF <dbl>	PASS_DAY <dbl>
2019	2019-03-01	N927NN	162	Atlanta, GA	Los Angeles, CA	143.6778	3220.633
2019	2019-03-01	N900WN	140	Atlanta, GA	Los Angeles, CA	124.1660	3220.633
2019	2019-03-01	N212WN	140	Atlanta, GA	Los Angeles, CA	124.1660	3220.633
2019	2019-03-01	N8313F	140	Atlanta, GA	Los Angeles, CA	124.1660	3220.633
2019	2019-03-01	N992NN	162	Atlanta, GA	Los Angeles, CA	143.6778	3220.633
2019	2019-03-01	N517NK	179	Atlanta, GA	Los Angeles, CA	158.7551	3220.633
2019	2019-03-01	N7724A	149	Atlanta, GA	Boston, MA	132.1481	3220.633
2019	2019-03-01	N8504G	189	Atlanta, GA	Boston, MA	167.6241	3220.633
2019	2019-03-01	N589JB	200	Atlanta, GA	Boston, MA	177.3800	3220.633
2019	2019-03-01	N206JB	110	Atlanta, GA	Boston, MA	97.5590	3220.633

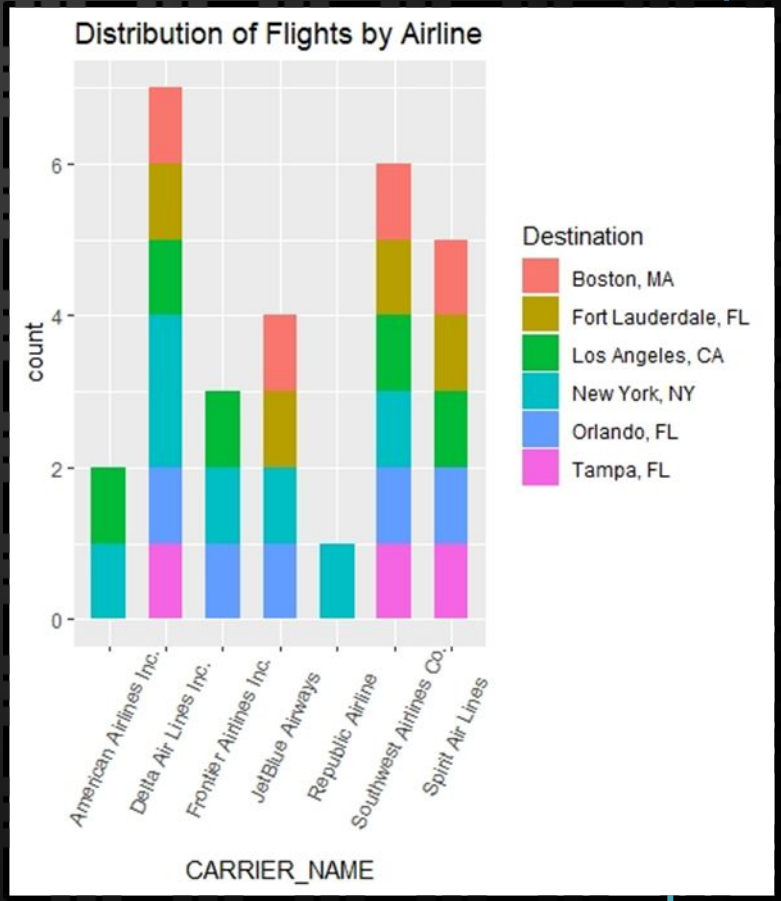
STATISTICAL ANALYSIS



[3]

- In Figure 3, each dot on the scatter plot represents a flight from the BTS database. passenger population at the airport:

Figure 4 represents the breakdown of flights, Southwest Airlines provide a balanced baseline from having the most flights to these major cities. Further, HJ_ATL is the home hub for Delta and Southwest, making them great data banks for enplaning flights.



[4]

MACHINE LEARNING

- In order to provide a more detailed analysis, Machine Learning is incorporated, in which the system conduct analyses based on different regression models or clustering models dictated to it. The exploratory data analysis showed that the number of passengers is related to the number of flights and load capacity available for the given flights.

- Performed a linear regression analysis with LOAD_CAPACITY and Dest (Destination) as the independent variables and PASS_DAY as the dependent variable. Dest, when constructed, had a data type of factor, which was be changed to numeric for use in the regression model:

```
lm(formula = PASS_DAY ~ LOAD_CAPACITY + Dest, data = BTS)

Residuals:
       Min         1Q     Median         3Q        Max 
-5.369e-10  1.110e-12  1.380e-12  1.700e-12  2.670e-12 

Coefficients:
               Estimate Std. Error   t value Pr(>|t|)    
(Intercept)   3.221e+03  6.129e-12 5.255e+14   <2e-16 ***
LOAD_CAPACITY 1.417e-14  2.804e-14 5.050e-01    0.614    
Dest          3.596e-15  3.582e-14 1.000e-01    0.920    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.689e-11 on 400 degrees of freedom
  (33649 observations deleted due to missingness)
Multiple R-squared:  0.5001,	Adjusted R-squared:  0.4976 
F-statistic: 200.1 on 2 and 400 DF,  p-value: < 2.2e-16
```

THE ANALYSIS.

1. p-value, which displays the significance of the model compared to a null model, which is usually a model that displays averages of the dependent variable. It is a matter of laying down a baseline for accuracy of the model. The lower the p-value (ranging from 0 to 1), the more likely the model is more accurate at approximation than the null model, and the null model can be thrown out. Because the p-value is less than 0.05 (and even 0.01 for this regression, being 2.2e-16), the null model can be thrown out and it is concluded that this model is more accurate than the baseline.

2. R^2- value, which determines how closeâ the data points are to the regression (or best-fit) line. The values range from 0 to 1, with values closer to 1 indicating that the data points are â□□closerâ□□ and more tightly correlated to the regression line. In laymanâ□□s terms, the closer the value is to 1, the better approximation of the data points the predictive model will give. If it is closer to 0, then the independent variable may need to be changed as it does not provide enough context or influence on the dependent variable and is thus not useful for the model.

Here, the R^2 value is 0.5001. This is indicative of the values being moderately close to the line, as 0.5 is evenly between 0 and 1. This alsomeans that there is definitely another more refined choice for the best-fit line. With more testing and incorporating more observations, the results will be closer to the regression line.

FUTURE WORK

- In the future, the linear regression model will be expanded to consider additional independent variables.
- Also, the expansion of the data points will allow for a more detailed analysis. The point of approximating “congestion” in different systems such as the airport and overcrowding will be considered.

ACKNOWLEDGEMENTS

- I would like to acknowledge my Springboard mentor, Dr. Goran Milonavanović, my faculty mentor, Dr. Torina Lewis for their expressed support, and the HBCU-UP Implementation Program.

- This research opportunity was sponsored by the National Science Foundation, award numbers 1700408 and 1818682.