Understanding Airlines Flight Delay

Project

by

Jolly Ogbole

December 11, 2022

Introduction

The United States domestic airlines industry consists of operators primarily engaged in the transportation of passengers and cargo. Top domestic airlines include big names such as Delta, United, American and Southwest airlines. As of 2022, the US domestic airline market has a market size of $144 billion and employs about 545,774 people. Over the most of the five years to 2022, rising levels of disposable income and an increasing number of businesses have precipitated increased demand from both business and leisure travelers. The Federal Aviation Authority's Air Traffic Organization (ATO) provides service to more than 45,000 flights and 2.9 million airline passengers across more than 29 million square miles of airspace every day. That is a significant amount of daily air traffic.

My recent roundtrip flight booking from LAX to SFO for this year's thanksgiving inspired me to leverage this final project to further understand the US domestic airline industry. So I decided that I would focus my project on the US domestic travel industry. I, like millions of other passengers, have had flights delayed multiple times. In this project, I pulled data to help me better understand what factors contribute to these delays and the relationship between them.

Business Framework

In undertaking this project, I assumed the role of a data analyst at the United States department of transportation. And I imagined that my Boss, in my first week on the job, provided me with this data and asked: Jolly, what insights can we gain from these records of flight data in the month of January? To this end, I formulated two analytical questions to elicit insights from the data. These include:

Question 1: What Variables Constitute Most Airlines Flight Delays ?

Airlines are known to have regular delays, but most passengers are not aware that there are multiple delay factors, as I reveal in subsequent pages of this report. If the department of transportation and airlines understand what factors impact delays the most, they may be able to allocate more capital resources to improve the situation. This is why this question is value-adding and relevant for investigation.

Question 2: Which Airports Are The Least Reliable For Passengers ?

Airports are complex locations with factors that operate in many intersecting ways that impact the efficiency of operations of airlines and their flights. Airports suffer from a number of challenges including weather, runway design, congestion and even in rare bizarre cases, animal factors like birds and stray dogs on the run-ways have impacted on-time flight departures and arrivals. In this question, I investigate how airports compare as no one wants to get stuck in an airport on their way to an important trip. Passengers should know which airports to avoid. Hence this question is useful to offer some insights.

Understanding Data

The dataset is a population of 201,664 records (row observations) and 14 initial variables (columns) of United States domestic flights in January 1-15 of 2015, available from the United States department of transportation.

Description of variables and the data types is shown in the table below:

Variable	Description	Data Type
FlightDate	Flight Date (yyyymmdd)	object
UniqueCarrier	Unique Carrier Code	object
FlightNum	Flight Number(Flights on different days may have the same flight numbers)	float64
Origin	Origin Airport	object
Dest	Destination Airport	object
ArrDelay	Difference in minutes between scheduled and Actual arrival time	float64
Cancelled	Canceled Flight Indicator (1=Yes, 0 = No)	float64
CarrierDelay	Carrier Delay, in Minutes	float64
WeatherDelay	Weather Delay, in Minutes	float64
NASDelay	National Air System Delay, in Minutes	float64
SecurityDelay	Security Delay, in Minutes	float64
LateAircraftDelay	Late Aircraft Delay, in Minutes	float64

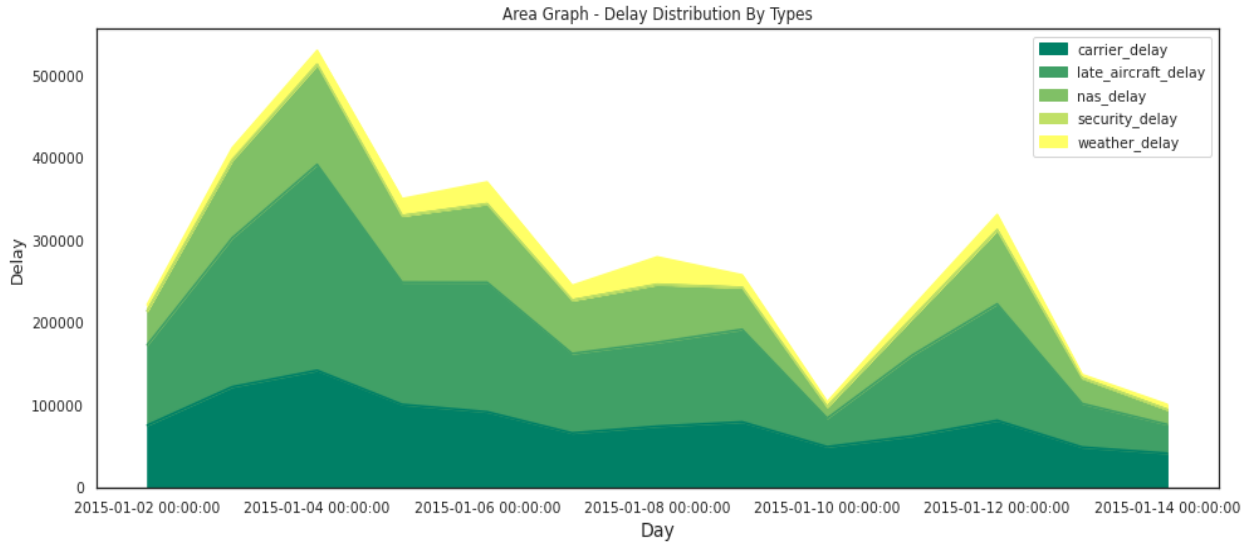
Findings

Question 1: What Variables Constitute Most Airlines Flight Delays?

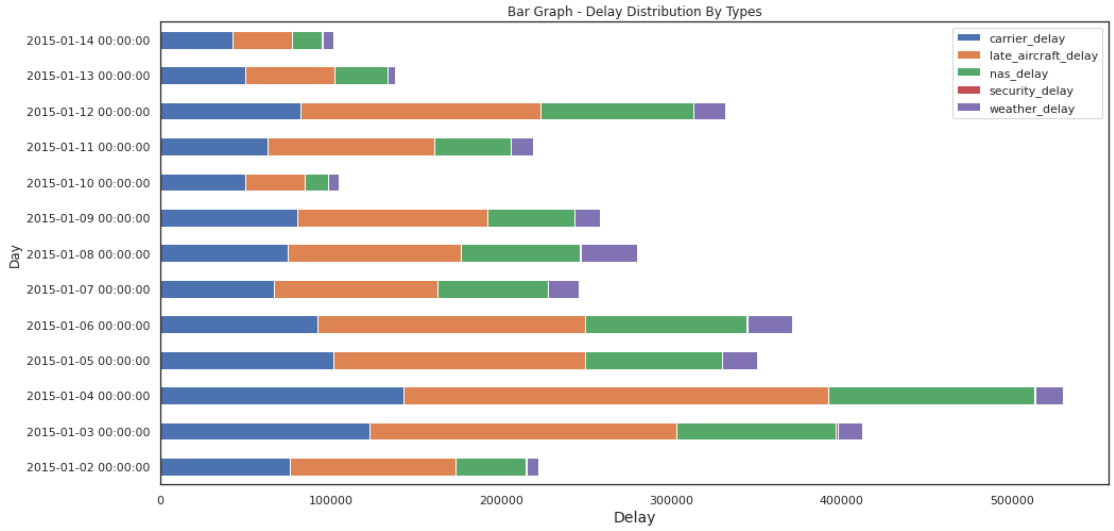
count	58533.000000
mean	17.829532
Standard deviation	43.529473
min	0.000000
max	1444.000000
25th percentile	0.000000

50th percentile	3.000000
75th percentile	19.000000

The table above gives a summary statistics of the "carrier_delay" variable from the dataframe. The average flight delay time is just under 14 minutes. But the max delay is a full day (1440 minutes). And with a standard deviation of almost 50 minutes, I can tell there is high variability in flight delays.



The area graph above captures the relative proportion of the different variables that contributes to flight delays. As seen, late aircraft arrival, followed by carrier delay contributes the most to aircraft delays. This implies that any given flight delay duration is not a good predictor of subsequent or future flight delays duration.



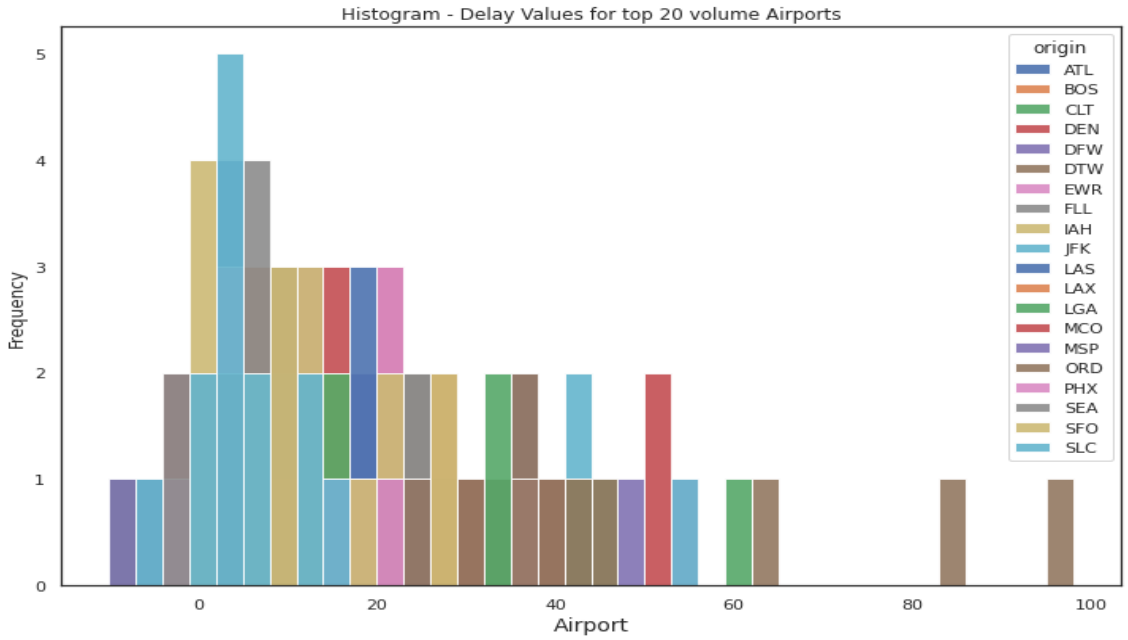
The horizontal bar graph above is another alternative to validate my initial findings that on average late aircraft arrivals and carrier delays are the variables that constitute most flight delays. The findings are consistent.

Question 2: Which Airports Are The Least Reliable For Passengers ?

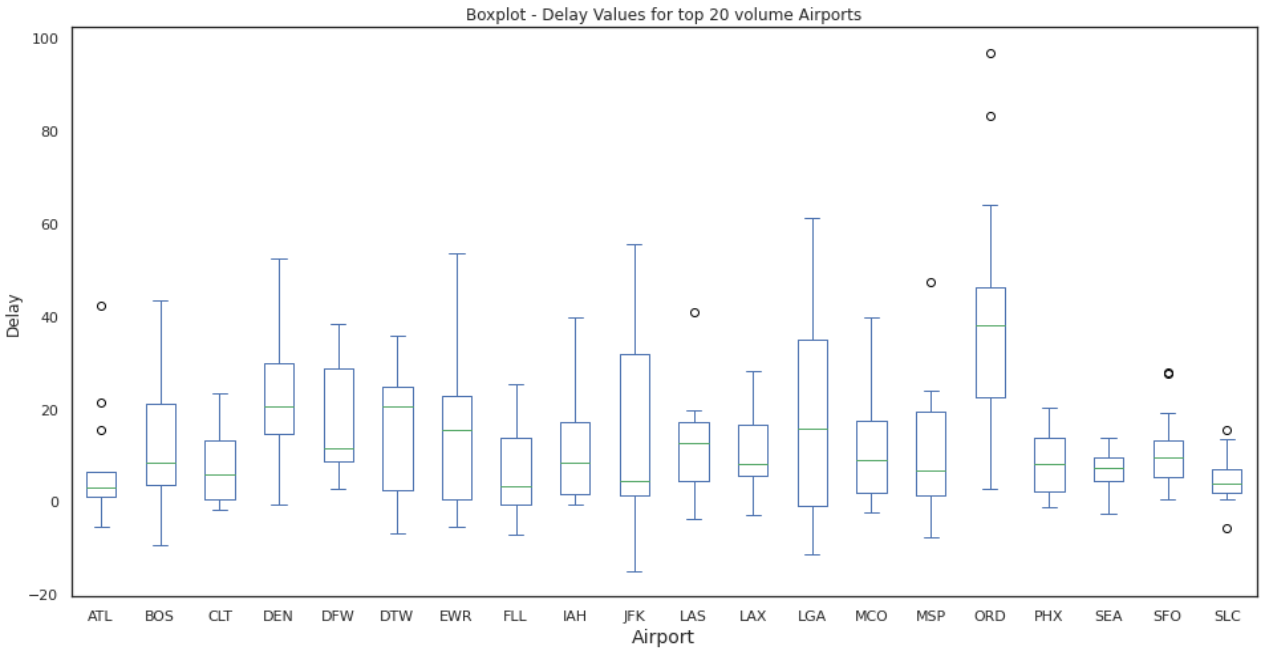
count	13.000000
mean	7.049522
Standard deviation	12.798122
min	-5.473046
max	42.375375
25th percentile	0.944276
50th percentile	3.033099
75th percentile	6.361725

The table above gives a summary statistics of delays from the airport (Atlanta International Airport ~ ATL) with the highest volume of flights in the dataset. Here the average delay was

seven minutes and the dispersion was about 12 minutes. I can tell that ATL was not the worst airport relative to other airports in our dataset, as seen in the visualization below.



The histogram above shows an overlaid distribution of delays by airports. This is tough to understand, so I leverage a better alternative visualization that helps to convey the message.



The box plot above shows a day-to-day distribution of delays at each airport. Airports like JFK had a significant dispersion of delays, while LGA was evenly distributed around the average delay. Chicago airport (ORD), however, was almost twice as delayed all the time, compared to every other high volume airport.

Conclusion

To answer my first question, leveraging the pivot function, I pivoted the date and types of delays, summing the number of minutes of each type of delay. This allowed me to see how the different types of delays stacked up relative to one another. Then I used an area graph to isolate and amplify the proportionality hidden in the pivot table output, as demonstrated in my colab workbook.

To answer my second question, I first filtered the top 20 highest volume airports, then created a boolean index to filter for flight records that originated in the top 20 volume airports. Then I created a pivot table that pivots the flight date on the airport, where the values are the average of flight delays for that day. And last I used a boxplot to clearly convey the insights from the pivot table.

Per recommendation, my findings reveal that airlines and airports have opportunities for growth in the area of operational efficiency optimization. I recommend for both parties to undertake gap analysis in order to discover drawbacks in late arrival and carrier delays because airlines and airports definitely intersect in their operations. Chicago International airport(ORD) is prone to abnormal delays as the boxplot shows and should be avoided as a connection location when possible, especially in January. Lastly, I recommend further and more robust analysis from a panel data approach to consolidate insights from this type of data.

Reflection

Overall, I would say that this project was a rewarding learning experience. To start with, my appreciation for visualization grew by leaps during this project. Visualization is the gold standard for succinctly passing your message across, especially to non-technical audiences. This project gave me the opportunity to apply the materials learnt from this course to a different domain. I came to realize that really analytics is domain agnostic. And I think that is a powerful realization to know that these skills will give you access to any industry you choose. I have gained a better appreciation for the skills that I am developing in this program and this will motivate me to stay consistent and grow my skills for a career in any industry. On data, I find there is more messy than clean data out there , and being able to clean and organize data efficiently might be a near superpower. Analytics is time consuming but promises to be also rewarding. There is much work to be done.

Citation

Air traffic by the numbers. Air Traffic By The Numbers | Federal Aviation Administration. (n.d.). Retrieved December 10, 2022, from https://www.faa.gov/air_traffic/by_the_numbers

Industry market research, reports, and Statistics. IBISWorld. (n.d.). Retrieved December 10, 2022, from https://www.ibisworld.com/united-states/market-research-reports/domestic-airlines-industry/