**Flight Delay Analysis**

**BAIS:6050 Data Management and Visual Analytics**

**Professor Patrick Fan**

Group 3:

Brooke Arnold, Brooks Rodish, Deanna Kwon

Munju Kam, Sydney Bosch

**Problem and Dataset**

US airlines often experience flight delays or cancellations during busy travel seasons whether it would be for vacation, holiday, or business. Simple questions such as “when is an optimal time to purchase flight tickets,” “which carriers are reliable for transporting customers safely,” and “which flight will most likely be on time?” triggered us to dive deeper into analyzing this dataset.

Our project utilizes the Flight Delay dataset, SQL code, and Tableau visualizations in order to answer the following analysis questions about flights between years 2004-2019:

Which airlines have the highest percentage of delayed flights?

Which airlines have the highest percentage of flights on time?

What is the most common reason for delay?

What is the most common reason for delays? – categorized by airline

Which airports experience the most delayed flights?

Which year has the most flight delays?

Which month has the most flight cancellations?

The Flight Delay dataset consists of three separate csv files, each containing information on flight carriers, airports within the United States, and details about arrivals by each carrier in each airport. The dataset was directly retrieved from the official website of Bureau of Transportation governed by the United States Department of Transportation.

The “Carrier.csv” file consists of 34 unique carriers with the following columns:

Carrier ID - the specific ID or code for specific airline

Carrier Name - the name of the airline

The “Airport.csv” file consists of 410 unique airports in United States with the following columns:

Airport ID - the specific ID or code for specific airport

Airport Name - the name of the airport

City - the city which the airport is located

State - the state which the airport is located

The “Arrivals.csv” file consist of 283,668 rows and contains information about arrival events aggregated by the carrier, airport, and date of flights. The file has 19 unique columns listed below:

Year – the year which flight occurred

Month - the month which flight occurred

Arr Flights - total count of flights that arrived

Arr Del15 - total count of flights that were delayed for more than 15 minutes

Carrier CT - total count of flight delays occurred due to air carrier or airline

Weather CT - total count of flight delays occurred due to unfavorable weather condition

NAS CT - total count of flight delays occurred due to National Aviation System

Security CT - total count of flight delays occurred due to security reasons (e.g., delays in screening procedures, aircraft security maintenance)

Late Aircraft CT - total count of flight delays occurred due to delay of same aircraft on previous flight

Arr Cancelled - total count of flight cancellation

Arr Diverted - total count of diverted flights

Arr Delay - total time in minutes of all delayed flights

Carrier Delay - total time in minutes of flight delays occurred due to air carrier or airline

Weather Delay - total time in minutes of flight delays occurred due to unfavorable weather condition

NAS Delay – total time in minutes of flight delays occurred due to National Aviation System

Security Delay - total time in minutes of flight delays occurred due to security reasons

Late Aircraft Delay - total time in minutes of flight delays occurred due to delay of same aircraft on previous flight

**E-R Diagram and Relational Schema**

The specific relationship between the three entities is indicated below under *Figure 1.1* The Entity-Relationship Diagram and *Figure 1.2* Relational Schema.

***Figure 1.1*** *Entity – Relationship Diagram*

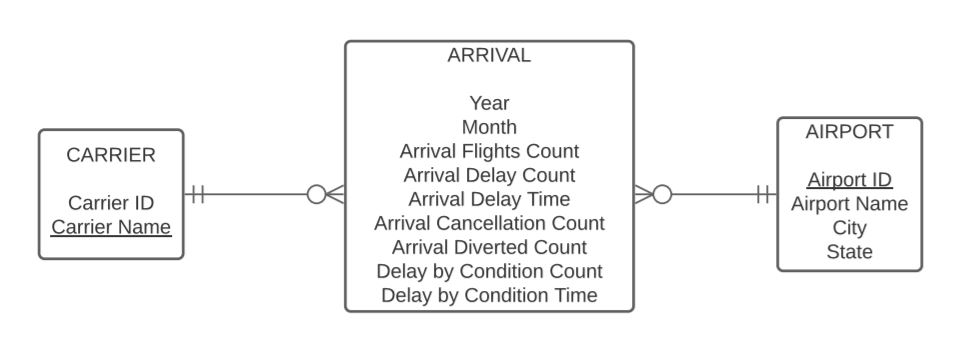


Figure 1.1 is the Entity-Relationship Diagram also known as ERD of the dataset. ERD indicates the relationship between the three entities by assignment with each other. The three csv files are separated as entities in the following dataset. Arrival plays a role of associative entity which include information linking Carrier and Airport information in the entire dataset. As shown in Figure 1.1, Carrier and Arrival data have one-to-many relationship, with mandatory-one to optional-many cardinality constraint. This means, each arrival data is assigned with at least and only one carrier, but carrier information could be assigned to multiple or no arrival events at all. Additionally, Airport and Arrival data also have one-to-many relationship, with mandatory-one to optional-many cardinality constraint. This means, each arrival data is assigned with at least and only one airport information, but airport information could be assigned to multiple or no arrival information at all.

***Figure 1.2*** *Relational Schema*

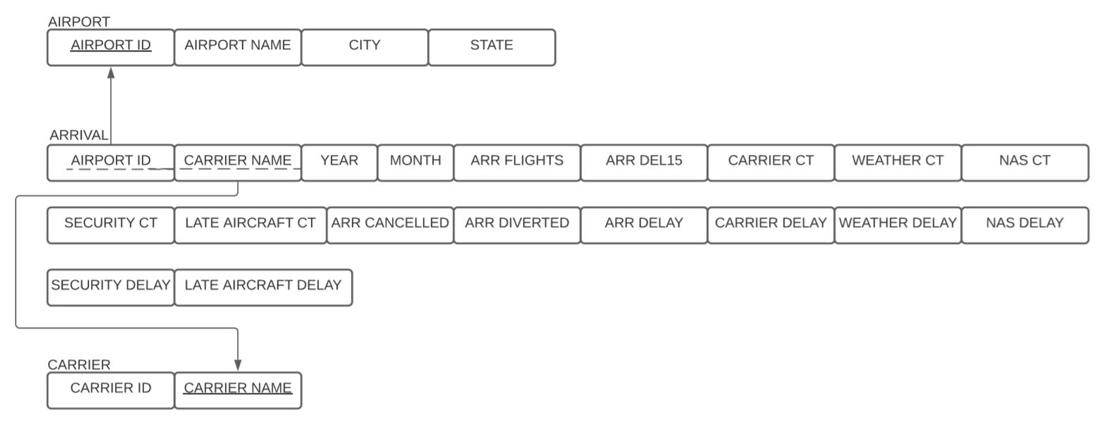
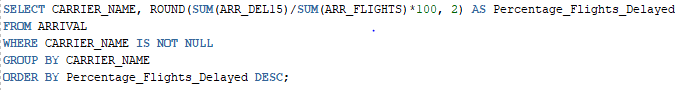
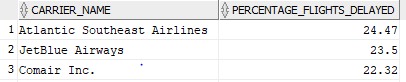


Figure 1.2 is the Relational Schema of the dataset. Relational Schema indicates the relationship and linkage between the three entities by the linking factors or attributes, also called the primary and foreign keys. As shown in Figure 1.2, the three csv files are separated as entities in the following dataset. Each block indicates the specific columns or attributes that are included as part of that certain data. Airport\_ID is the primary key in Airport data that is linked in Arrival data as a foreign key. Therefore, the same Airport information could be found by primary key within Arrival data. Carrier Name is a primary key in Carrier data, that is a foreign key in Arrival data. This means that same Carrier information could be found in Arrival data by searching the Carrier name.

The following analysis was proceeded from assuming the above shown Figure 1.1 Entity-Relationship Diagram, and Figure 1.2 Relational Schema as the baseline.

**Which airlines have the highest percentage of delayed flights?**



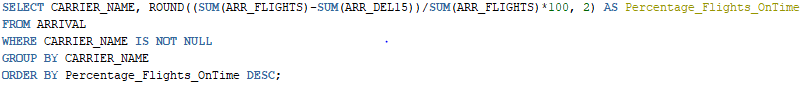


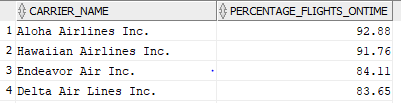
Chart, treemap chart

Description automatically generated

Since not all airlines have similar sample sizes, we decided that the best way to compare delays across airlines was to look at the delays in percentages. The three airlines that experience that highest percentage of delays are Atlantic Southeast Airlines (24.47%), JetBlue Airways (23.50%), and Comair Inc. (22.32%).

**Which airlines have the highest percentage of flights on time?**



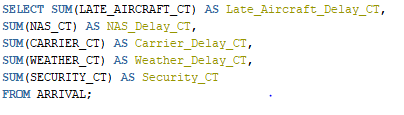


Chart, treemap chart

Description automatically generated

Again, we compared percentages on flights on time because not all airlines are similar in size. The three airlines that have the highest percentage of flights on time are Aloha Airlines Inc. (92.88%), Hawaiian Airlines Inc. (91.76%), and Endeavor Air Inc. (84.11%). We also wanted to point out that major airline Delta Airlines is fourth overall. It appears that of the common, major airlines, Delta is the most reliable airline when it comes to your flight arriving on time.

**What is the most common reason for delay? ​**





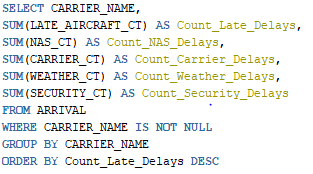
Chart, pie chart

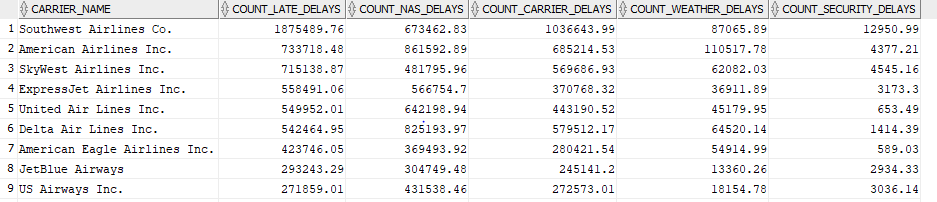
Description automatically generated

The Flight Delay data set contains 5 different reason for delay: Late Aircraft Delay, National Aviation System Delay, Carrier Delay, Weather Delay, and Security Delay. Late Aircraft Delay represents delayed flights due to a previous flight using the same aircraft being late. National Aviation System Delays (displayed as NAS\_DelayCt) refer to delays caused by a variety of possible concerns such as airport operations, heavy traffic volume, air traffic control issues, and even non-extreme weather conditions. Carrier Delay represents delays due to air carrier issues. Weather Delay represents delays due to severe weather and other unpredictable elements. Security Delay represents delays due to security issues at the airport. To explore the most common reasons for delay, we used SQL to sum the count of each of the five reasons for delay. The picture above is our query and result. This suggests that the most popular reason for delay is due to late aircrafts, followed by the National Aviation System and carrier issues.

In order to visualize the reasons for delay, we created a pie chart in Tableau. By using a pie chart, we can effectively compare part-to-whole relationships between variables. When totaling the top three reasons for delay, we see that they account for 96 percent of the reasons for delay. We assume that Late Aircraft Delay is at the top because a late aircraft at one airport will cause the next flight to be delayed at another airport, causing a chain reaction that will impact the entire flight schedule.

**What is the most common reason for delay? - categorized by airline​**



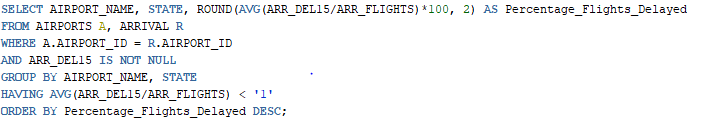


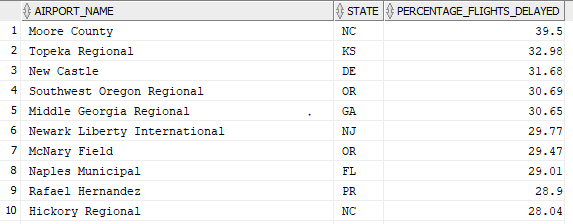
Chart, bar chart

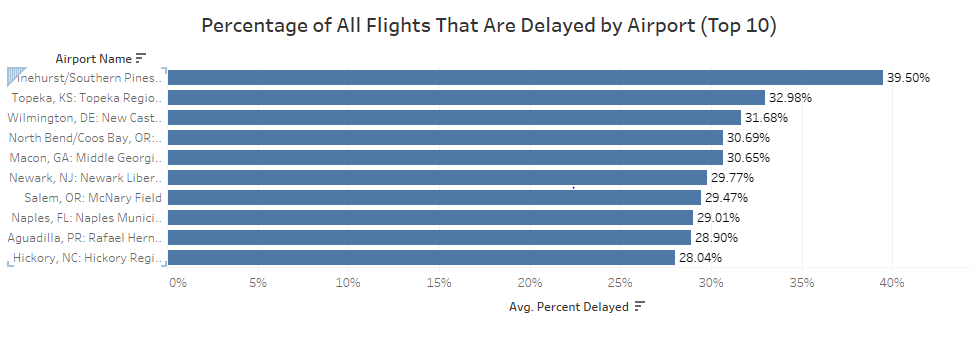
Description automatically generated

We broke down the Tableau and query result again by carrier to see the common reasons of delay by different carriers. Because there are a lot of numbers in the table, we were able to enhance our understanding of the data by creating a bar chart. From the graph, we noticed the pattern that most airlines have the same most common reason for delay (Late Aircraft Delay), and the same least common reason for delay (Security Delay). While most of the carriers follow this trend, two exceptions especially noticeable are Southwest Airlines and Comair. Southwest Airlines’ National Aviation System Delay count (673,462) is much lower than its Carrier Delay count (1,036,643). Likewise, Comair’s Weather Delay count (73,060) is visibly higher than its Late Aircraft Delay count (17,835).

**Which airports experience the most delayed flights? ​**



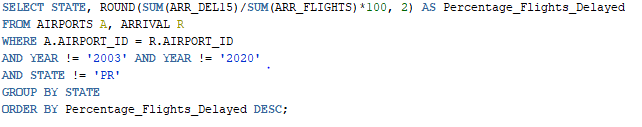


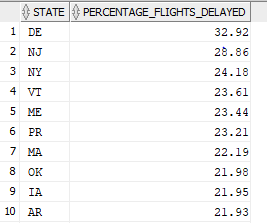


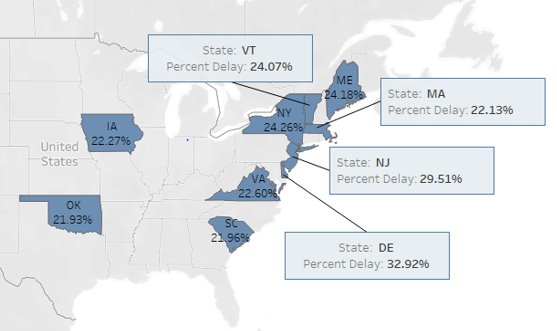
For our SQL query, we decided to find the airports with the most delays based on the percentage of flights delayed at each airport, ordered highest to lowest. We then took the top 10 airports for our Tableau analysis to show a clearer image of where the most delays took place. We decided to exclude one airport as it was an outlier. This airport, Youngstown-Warren Regional, only had one row of data with two total flights recorded. Both flights were delayed, therefore reporting 100%, and would have skewed our understanding of the rest of the data.

The Tableau horizontal bar chart above illustrates the differences between the top 10 most delayed airports. We found that the top three were Moore County Airport in Pinehurst/Southern Pines, North Carolina, Topeka Regional Airport in Topeka, Kansas, and New Castle Airport in Wilmington, Delaware. As shown in the graph, the first airport differs slightly from the last nine, indicating there could be significant problems at Moore County Airport. The other nine airports only vary within 4% of each other. This shows that outside of the top delayed airport, delays follow a consistent arrangement.

**Which states experience the most delayed flights? ​**



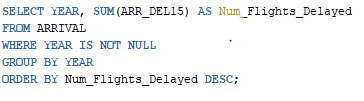


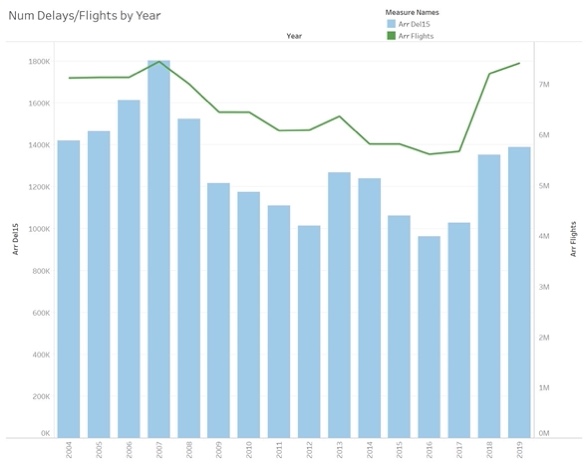


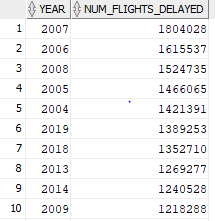
To examine the states with the highest percentage of delayed flights, we created a SQL query that calculates this percentage by taking the sum of all delays divided by the total number of flights. We additionally excluded Puerto Rico, PR, from our SQL query as to keep our results within the US states. Our SQL query results differ slightly from our Tableau U.S. map because within Tableau it aggregated our data. The percentages of the states are off by a few points due to this, but they still flow in a similar sequence.

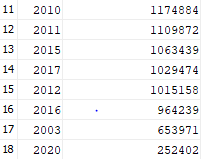
Shown above, we created a highlighted map to represent our results of the top 10 states with the most delays. From our results, we found that the top 3 states were Delaware, New Jersey, and New York. However, the other states within the top 10 followed relatively close behind. We also noted that the majority of the top flight delays occurred in the North Eastern region of the United States. We attributed the delays here to be either due to non-extreme weather conditions or air traffic control. From research, we found that the North Eastern region of the US is subject to diverse climate and weather conditions that could explain the delays in this area. Also, air traffic control deals with regulations and restrictions that force air traffic control to reroute and delay flights due to weather conditions or jet streams.

**Which year has the most flight delays?**



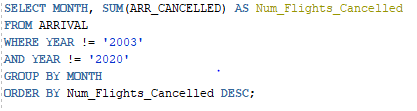
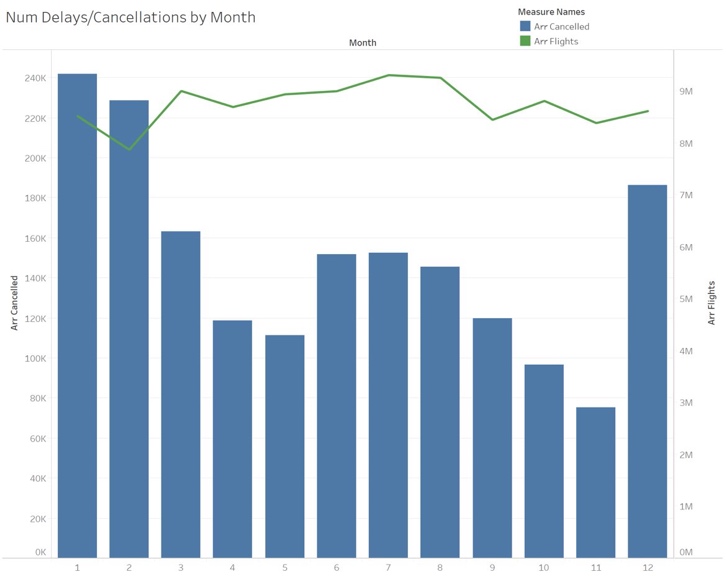


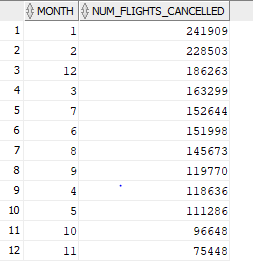




From the SQL and Tableau output, we can see that the year 2007 has the highest count of flight delays. According to CBS News, this was the second highest on record. Much of this can be attributed to “rising passenger demand coupled with congestion in the skies" and a decrease in air traffic controllers. Overall, however, increases/decreases in delays seem to be closely related to the total number of flights during that year. Meaning as the total number of flights increase, so do the total number of delays.

**Which month has the most flight cancellations?**





To determine the months with most flight cancellations, we created the SQL output and Tableau table above. From these visuals, it is clear that the highest number of cancellations take place in months twelve (December), one (January), and two (February). In the United States, these months (the winter months) typically bring bad weather, which is very likely to cause a flight cancellation rather than a delay. It is also evident that there is another smaller peak in months six (June), seven (July), and eight (August), the summer months. We would likely attribute this increase in flight cancellations to an increase in passenger demand during that time of year, causing more backed-up flights and cancellations. We see from the line of total flights in the Tableau visual that it is highest at that time.

**Conclusion**

To summarize our analysis from the past few weeks, the conclusions from our investigation are listed below. It is important to note that these conclusions are drawn from 2004 to 2019 only, and that most of our analysis is made using percentages. This helps us to avoid any bias from sample size differences.

Atlantic Southeast Airlines, JetBlue Airways, and Comair Incorporated have the highest percentage of delayed flights.

Aloha Airlines, Hawaiian Airlines, and Endeavor Air have the highest percentage of flights that arrive on time.

Late Aircraft Delay is the most common reason for delay. This type of delay causes the domino effect, as one delay can cause the next flight to be delayed, and so on down the line. It is not surprising that this is the most common.

Pinehurst Airport in North Carolina, Topeka Regional Airport in Kansas, and Wilmington Airport in Delaware experience the highest percentage of delayed flights.

Delaware, New Jersey, and New York are the top three states with the highest percentages of flight delays.

2007 has the most delays in recent history, and the second highest ever on record. The winter months have the most cancellations, likely due to weather issues.

**References**

<https://www.cbsnews.com/news/2007-airline-delays-second-worst-ever/>

<https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1>

<https://www.bts.dot.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays>