DSC 365/465: Data Visualization
Final Report
Group Name: Speed Demons

Members: Mashall Jahangir, Erik Wolf, Christian Marcol, Wilson Wu, Amanuel Petros

Appendix:

1. Group Technical Report

1a. Abstract:

Big Data is such a popular and in-demand term in today's world. Many companies and organizations are looking for data scientists to use the voluminous information to settle on sound business choices so as to amplify list potential and limit misfortune. The Bureau of Transportation Statistics is constantly analyzing real-time data based on several airlines. We want to determine obvious trends and patterns to understand what factors and variables cause airline delays. With the use of advanced visualizations, we will attempt to make beneficial observations. Each individual chose a different technique to analyze the dataset and answer the research questions.

1b. Introduction:

On the air travel passenger complaints by the US today, the top five problems are baggage, cancellation, reservations and ticketing, customer service and Delay. Our goal is to analyze an aggregated Airline Dataset from the year 2018 to see what airline has experienced maximum delays and what caused those delays. And to do this, we plan to produce visualizations from the variables including the month of the year, UniqueCareer Code, Tail Num, Origin, Destination, Arrival time, Departure time and etc. Our dataset includes 690,979 observations with 35 variables. Table 1.1 shows all our variables for the study. Moreover, to plot maps and the routes for each individual airline we decided to use another data set with the Latitudes and Longitudes of each destination airport.

Table 1.1

Month	The month of 2018. Values from 1-12 to indicate.
OpUniqueCarrier	Unique Carrier Code.
OpCarrierAirlineID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
TailNum	Identification Number painted on the aircraft
OpCarrierFlNum	Flight Number
OriginAirportSeqId	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of

	time. Airport attributes, such as airport name or coordinates, may change over time.
OriginCityMarketId	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Origin	Origin Airport
OriginCityName	Origin Airport, City Name
OriginStateAbr	Origin Airport, State Code
OriginStateNm	Origin Airport, State Name
DestAirportSeqId	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
DestCityMarket	Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Dest	Destination Airport
DestCityName	Destination Airport, City Name
DestStateAbr	Destination Airport, State Code
DestStateNm	Destination Airport, State Name
CRSDepTime	CRS Departure Time (local time: hhmm)
DepTime	Actual Departure Time (local time: hhmm)

DepDelay	The Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DepDelayNew	The Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DepDelay15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
CRSArrTime	CRS Arrival Time (local time: hhmm)
ArrTime	Actual Arrival Time (local time: hhmm)
ArrDelay	The difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDelayNew	The difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ArrDelay15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
Cancelled	Cancelled Flight Indicator (1=Yes)
Diverted	Diverted Flight Indicator (1=Yes)
Distance	Distance between airports (miles)
CarrierDelay	Carrier Delay, in Minutes
WeatherDelay	Weather Delay, in Minutes
NASDelay	National Air System Delay, in Minutes
SecurityDelay	Security Delay, in Minutes
LateAircraftDelay	Late Aircraft Delay, in Minutes

1c. Source of Data:

The dataset we will be working on is obtained from the Bureau of Transportation Statistics.
https://transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On$20-Time
https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=

1d. Dataset Description:

Dataset Name: Reporting Carrier On-Time Performance (1987-present)

Airport Longitude and Latitude for each Airport State

1e. Technology group plans to use for Project:

Our group will be using R, Tableau, and Python for visualizations.

1f. Exploratory Analysis:

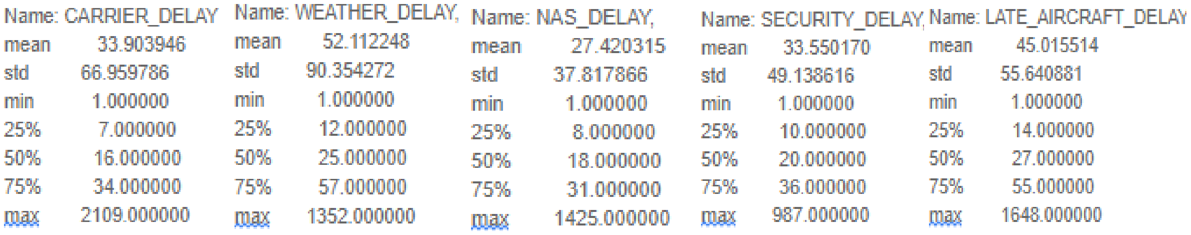
We focus on the Air Carrier, Extreme Weather, National Aviation System, Late-arriving aircraft, and security as our main variables.

1. Air Carrier delay - the cause of delay due to the maintains or airline service.
2. Extreme Weather delay - the cause of delay due to the bad weather condition.
3. National Aviation System delay - the cause of delay due to airport operation or airline traffic.
4. Late-arriving aircraft delay - the cause of delay due to previous flight with the same aircraft arrived late.
5. Security delay - the cause of delay due to security violations or long lines passengers in screening areas.

We used python to clean the null data in each column and select the Top four airlines.

1. (WN) Southwest Airlines, 132251331 passengers, has a 20% market share in domestic
2. (DL) Delta Air Lines, 106062211 passengers, has a 16% market share in domestic
3. (AA) American Airlines, 99857863 passengers, has an 11% market share in domestic
4. (UA) United Airlines, 71722425 passengers, has an 11% market share in domestic.

Then, we combined all the twelve months' data for the year 2018 and removed all the canceled flight data to focus mainly on the delay caused for the flights which were not canceled.
As the first step, using the bar chart (**Fig 1.1**) to count the occurrence of each of the delays we found out that the Carrier delay, Late-arriving aircraft, and National aviation are the top three causes for delay.
Second, we analyzed the count of delay by each Carrier. From this, we conclude that Delta Air Line's performance is pretty good compared to American and Southwest Airlines. (**Fig 1.3**)
Third, using a histogram to check each Delay's distribution, we used a logarithmic scale on each delay count with a bin of 20. From this, we found that the histogram of the distribution (**Fig 1.4 & Fig 1.5**) for each delay is right-skewed.



```
Name: CARRIER_DELAY   Name: WEATHER_DELAY,   Name: NAS_DELAY,       Name: SECURITY_DELAY,  Name: LATE_AIRCRAFT_DELAY
mean     33.903946    mean     52.112248     mean     27.420315     mean    33.550170      mean     45.015514
std     66.959786     std     90.354272      std     37.817866      std    49.138616       std     55.640881
min      1.000000     min      1.000000      min      1.000000      min     1.000000       min      1.000000
25%      7.000000     25%     12.000000      25%      8.000000      25%    10.000000       25%     14.000000
50%     16.000000     50%     25.000000      50%     18.000000      50%    20.000000       50%     27.000000
75%     34.000000     75%     57.000000      75%     31.000000      75%    36.000000       75%     55.000000
max    2109.000000    max   1352.000000      max   1425.000000      max   987.000000       max   1648.000000
```

From the above statistics, we can see how the mean delay for all delays is within the range of 27 to 52 minutes where NAS Delay caused a minimum delay of 27 minutes on an average. We excluded all the zero values to have an accurate measure across.

1g. Analysis and Discussion:

In our group project, we used many different visualization techniques to display airline delays including bar graphs, choropleth, line and area chart, and interactive destination map. The type of delays we focused on are the weather, national aviation system, security, and late aircraft delays. The method of incorporating both line and area chart relates to and complements each other. The use of scale is the same in all four types of delay, the visualization shows similar trends among the four types of delays. The bar graph and choropleth used flight origin and destination of the data to show which city as well as which state has the highest and the lowest delays. To show a different side of the data which is not categorical like with bar graph we used a line and area chart. Using quantitative data, this technique showed us a simple visualization of a sequence of values and helped us to see trends over time.

In general, the lowest cause of delays are weather delay and security delay, California is the busiest city in December followed by Texas, summer is where most delays occur in all types of delays, the highest delay for security occurred in September, and in our interactive destination map Kentucky has the highest delay when the origin of the flight is from Chicago. We were also able to see that DFW, ATL, MDW, and ORD were the most utilized by AA, DL, WN, and UA respectively. AA operates out of DFW, DL operates out of ATL and UA operates out of ORD. Interestingly WN does not operate out of MDW, this is because unlike the previous three airlines, Southwest does not utilize a hub and spoke system, instead of using a point to point, “rolling-hub” model for their operations. Chicago is simply the busiest city for these four airlines and Southwest operates exclusively out of MDW in the Chicago area.

For analyzing cities and state in each delay relationship, we compare two types of graphs, bar charts (**Fig 1.6**) and choropleth maps (**Fig 1.7**). In the bar chart with the top ten cities in each delay, Bismarck (ND) and Fayetteville (NC) both have significant Weather Delay and Median late-arriving aircraft Delay. In choropleth maps to perform State in each delay, For the Median Weather Delay by City, Alaska has the highest median of Weather Delay, 40.05 min and Second high is New Hampshire, 40 min. The Weather hazards reason for Alaska might explain. Alaska is readily affected by volcanic activity, earthquakes, floods, and severe snowstorms and New Hampshire has thunderstorms, hurricanes, and Snowstorms. For the Median National Aviation System (NAS) Delay by City, New Jersey has the highest median of National Aviation System Delay, 26 min and Second high is New York, 23 min. However, the top busy airport is in Atlanta International Airport, so actually, Atlanta has a good performance on airport operation or traffic, and John F. Kennedy International Airport is the fifth busy airport. Moreover, We are interested whether distance really affects on arrival delay and each airline performance, so we create a contingency table with the top ten count route (**Fig 1.9.**). For our discover, the airport distance doesn’t really matter to the arrival delay, because the ORD_SFO has the largest distance, 1846miles, but the delay times is less than ATL_LGA or ORD_LGA, 762 miles.

2. Visualizations:

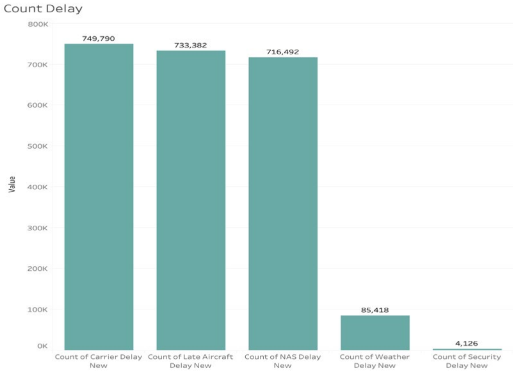


Fig 1.1

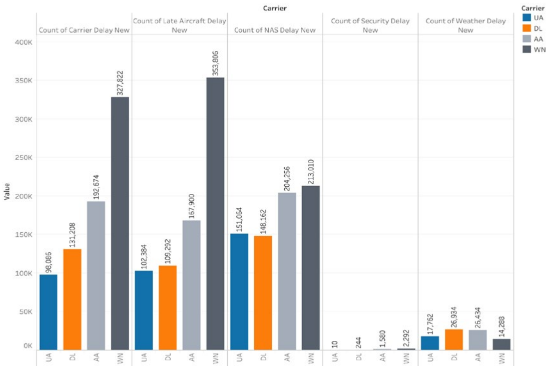


Fig 1.2

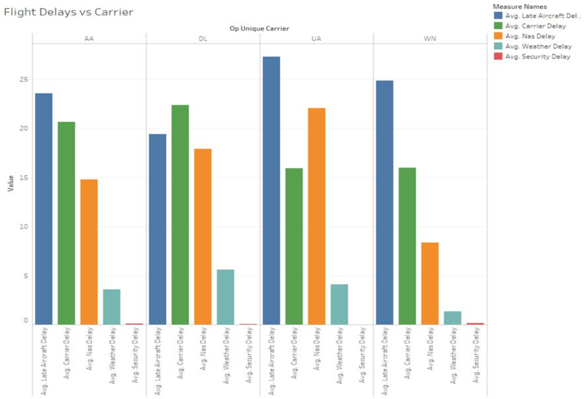


Fig 1.3

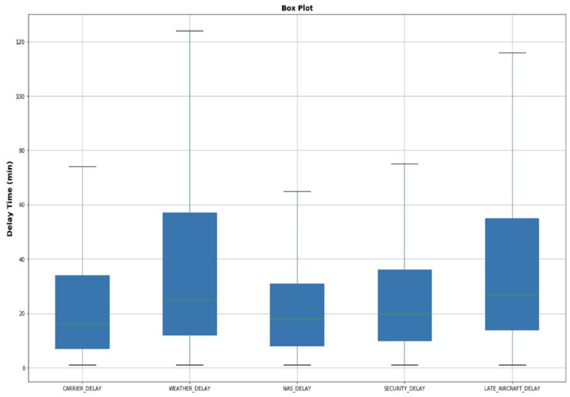


Fig 1.4

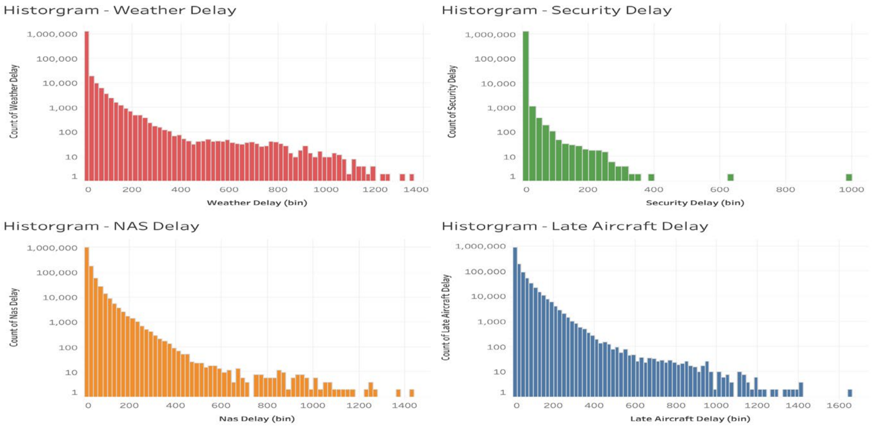


Fig 1.5 The trend of count of Delay for Delay (bins in 20).

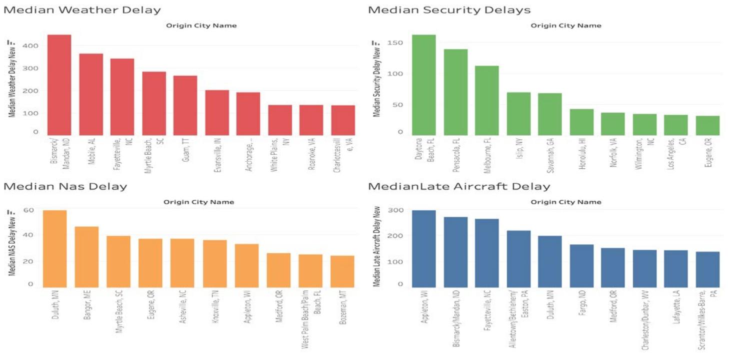


Fig 1.6 Median of Delay excluding all the zero values for different Origin flight City**.**

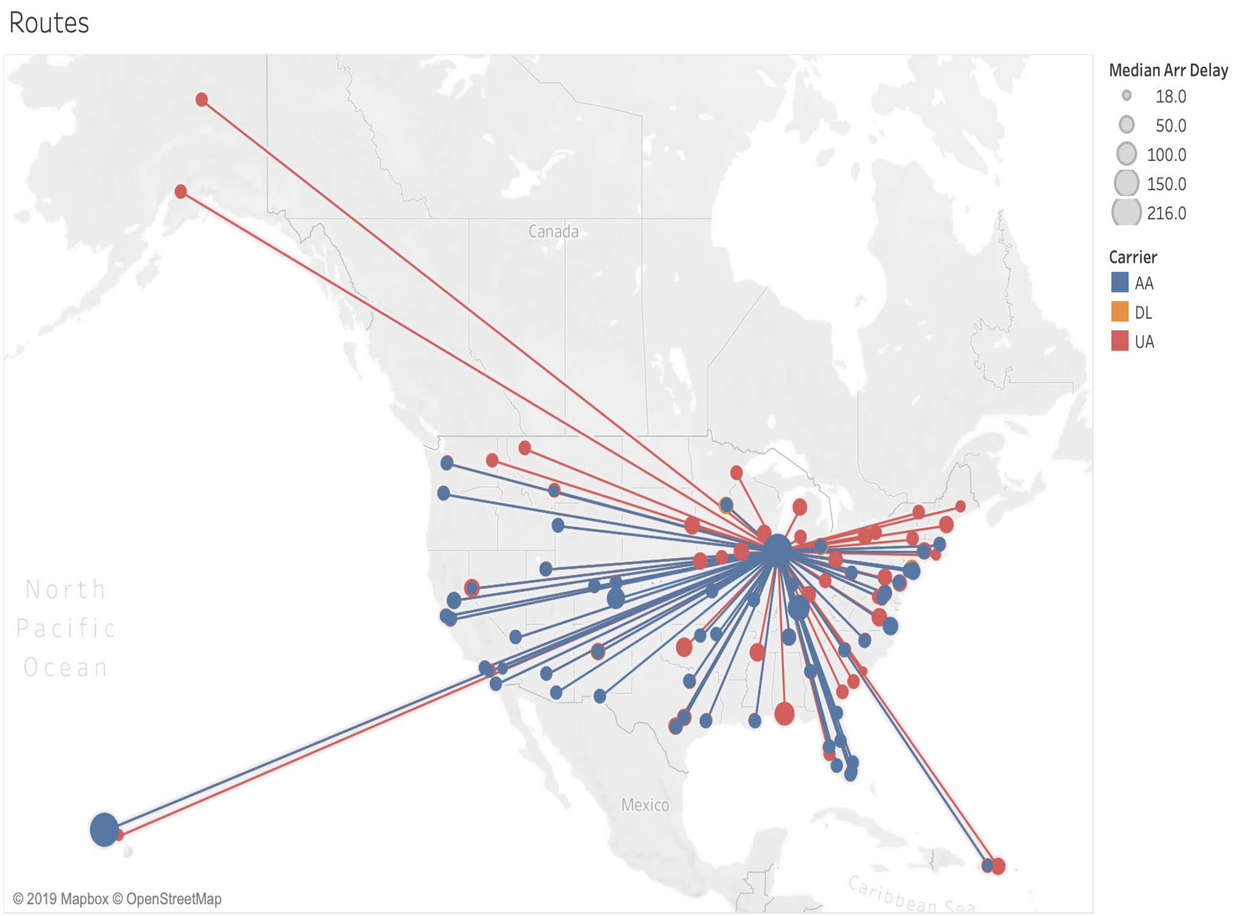


Fig 1.7 Map-based on average of Longitude and average of Latitude and average of Latitude. The color shows details about Carrier. Details are shown for Route Identifier. For pane Average of Latitude (2): Size shows the median of Arr Delay. Details are shown for Route Identifier and Route Order. The data is filtered on Origin, which keeps ORD.

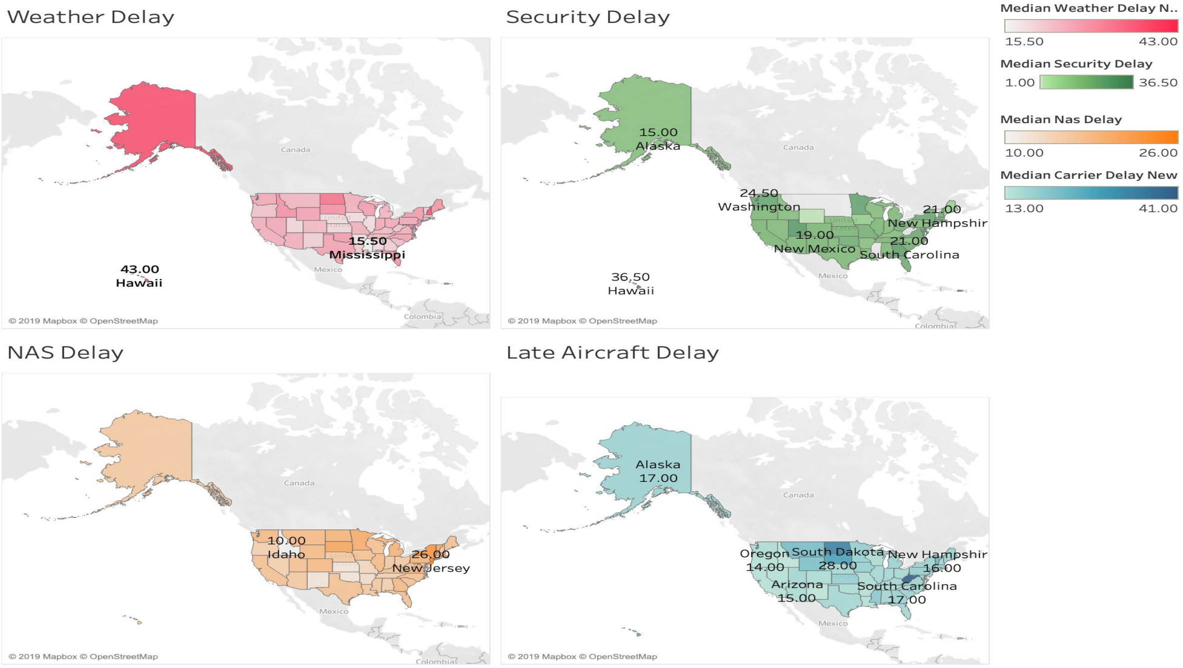


Fig 1.8. Map-based on Longitude (generated) and Latitude (generated). The color shows median of each delay. The marks are labeled by the median of each delay and airport state Name.

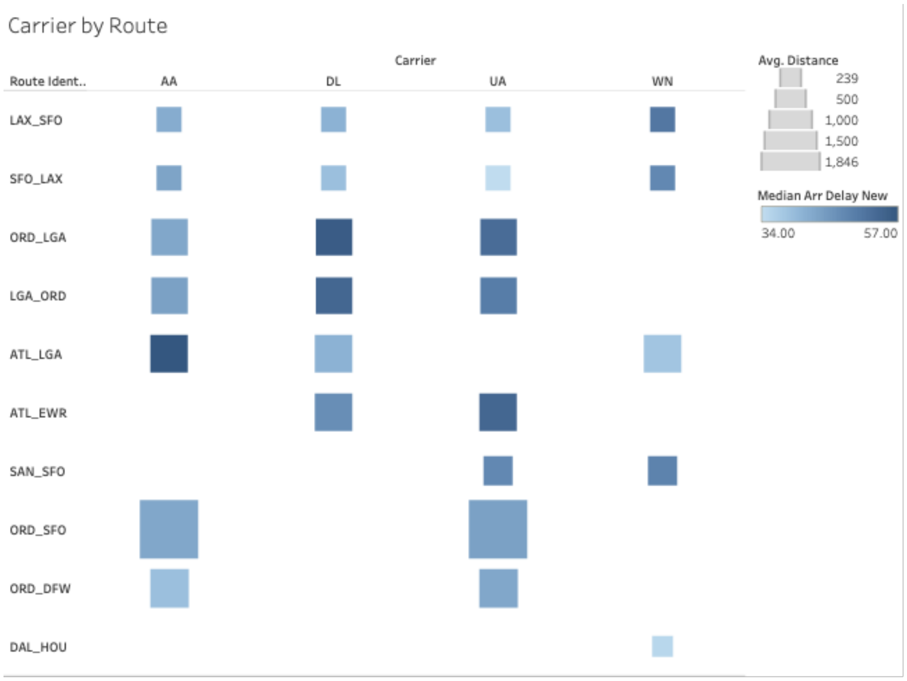


Fig 1.9 Median of Arrival Delay New (color) and average of Distance (size) broken down by Carrier vs. Route Identifier. The view is filtered on Route Identifier, which has multiple members selected.

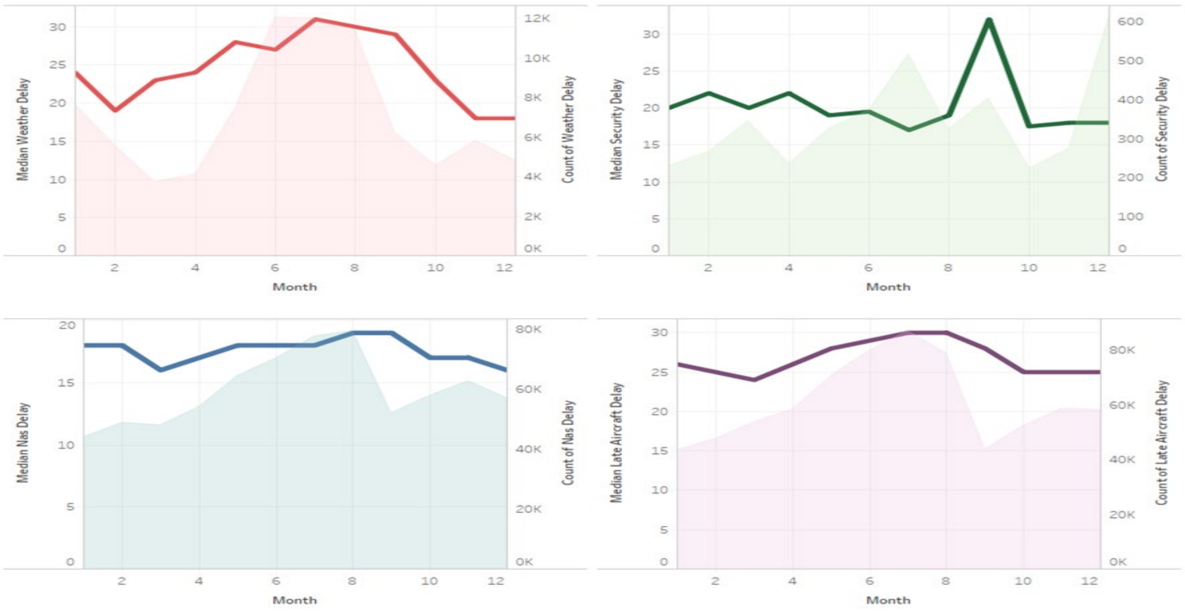


Fig 1.10 The trends of each delay and count of that particular delay for Month. Color shows details about Median Delay and count of Delay. The data is filtered on Delay, which ranges from 0.11 to 987.

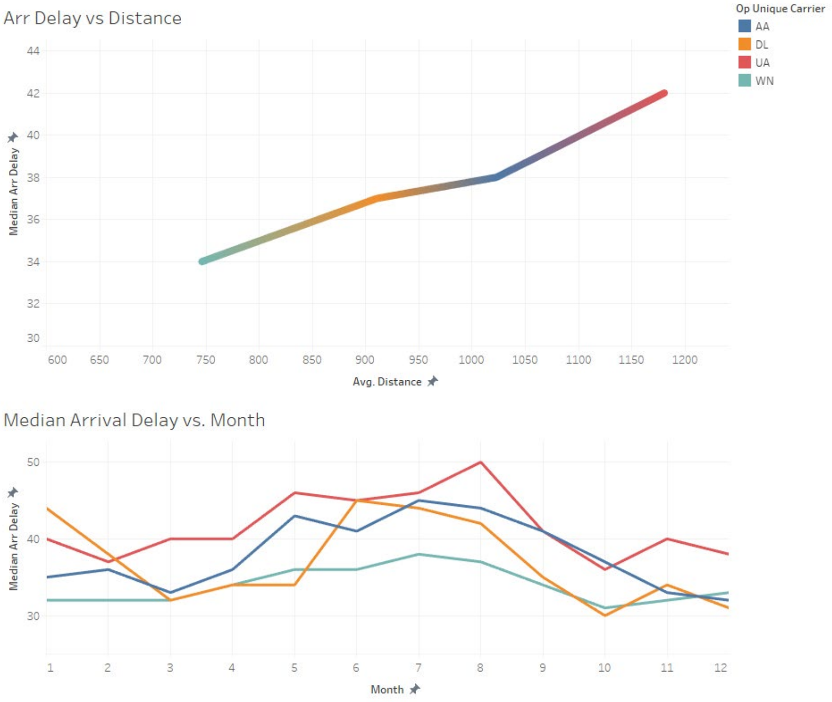


Fig 1.11 Average of Distance vs. a median of Arr Delay. The color shows details about Op Unique Carrier. The trend of the median of Arr Delay for Month. The color shows details about Op Unique Carrier.

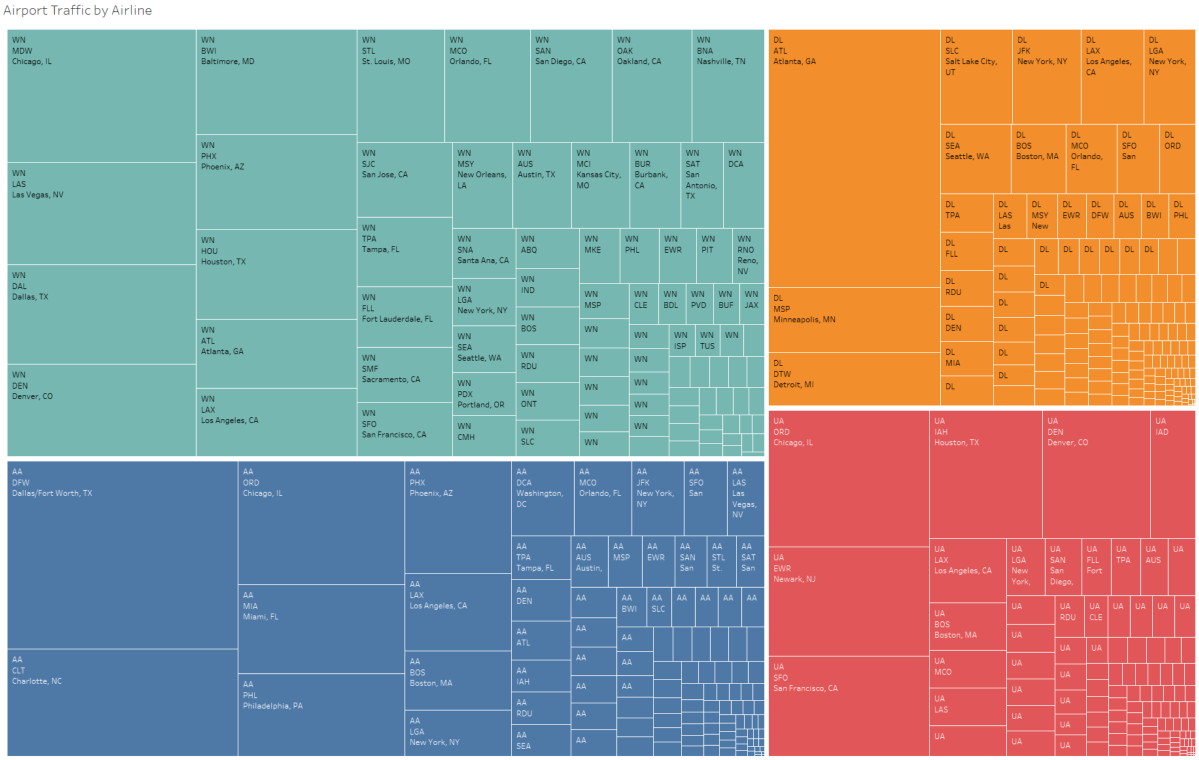
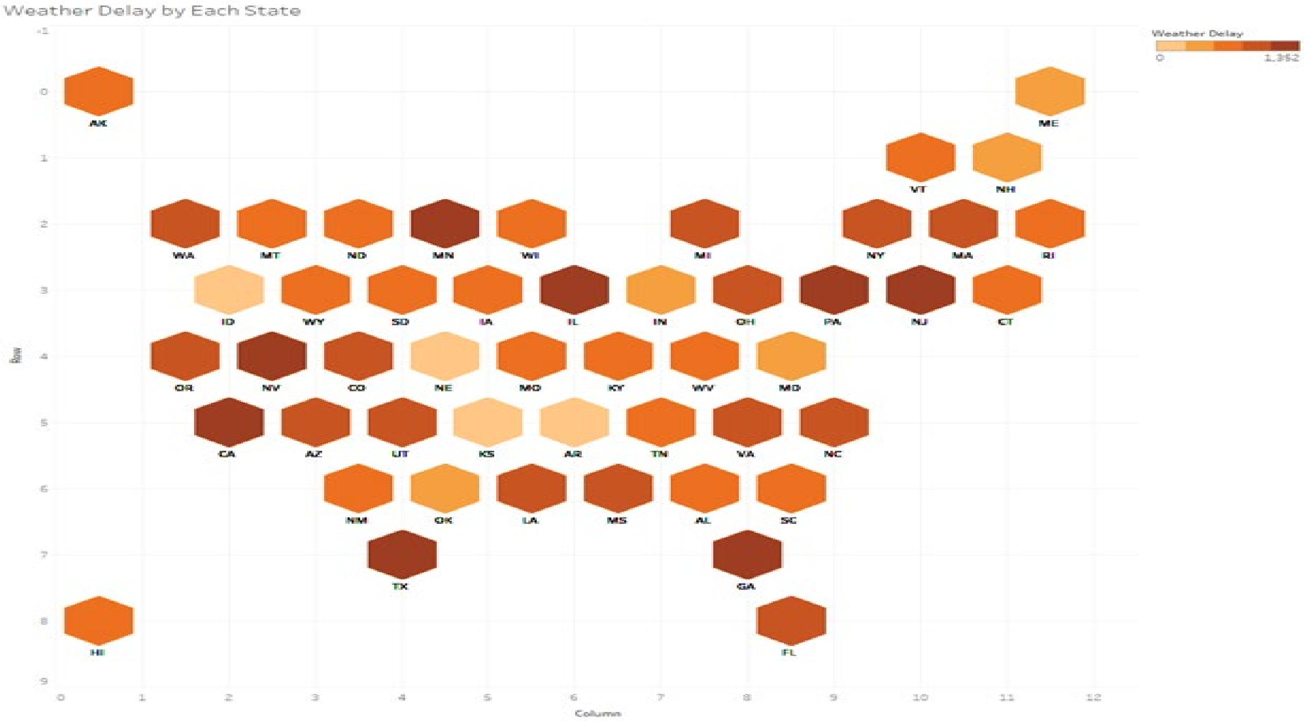


Fig 1.12 Carrier, Origin and Origin City Name. The color shows details about Carrier. Size shows count of Origin. The marks are labeled by Carrier, Origin and Origin City Name



Column vs. Row. Color shows Weather Delay. The marks are labelled by Abbreviation1.

Fig 1.13. Hex Tile Map showing weather delay per state.

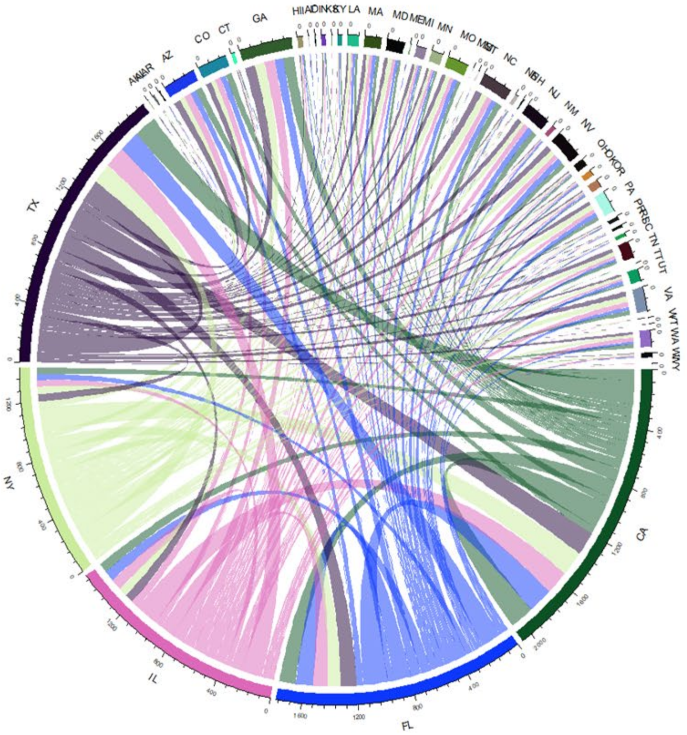


Fig 1.14 Network Graph, Five origin states showing the amount of connections to destination states.

3. Individual Report

3a. Wilson Wu

In my part of the project, I focus on the city and state related to each variation delay, because we analyze is there any pattern between city and state delay and, moreover, what and why cause the states have a huge delay in a different situation. I used python pandas to read and to clean the null data in each column and select specific aircraft company, and, moreover, use matplotlib to do exploratory analysis.

In the first graph, I decided to use the bar chart to perform the top ten cities in each delay (Appendix A.4 Median Delay by City). For the Median Weather Delay by City, the top three cities were Bismarck (ND), Mobile (AL) and Fayetteville (NC) and three of them have over 300 min delay time. For the Median Security Delay by City, the top three cities were Daytona (FL), Pensacola (FL) and Melbourne (FL) and three of them have around 150 to 100 min delay time. Moreover, they all happened in Florida. For the Median National Aviation System (NAS) Delay by City, the top three cities were Duluth (MN), Bangor (ME) and Myrtle Beach (SC) and three of them have around 40 to 60 min delay time. For the Median late-arriving aircraft Delay by City, the top three cities were Appleton (WL), Bismarck (ND) and Fayetteville (NC) and three of them have over 200 min delay time. Furthermore, comparing four bar charts, Bismarck (ND) and Fayetteville (NC) both have significant Weather Delay and Median late-arriving aircraft Delay.

In the second graph, I decided to use choropleth maps to perform State in each delay because the Choropleth map has the ability to show the regional pattern in the data. Map-based on Longitude and Latitude and shades of single-color help to stand out high and low data value. For the Median Weather Delay by City, Alaska has the highest median of Weather Delay, 40.05 min and Second high is New Hampshire, 40 min. The Weather hazards reason for Alaska might explain. Alaska is readily affected by volcanic activity, earthquakes, floods, and severe snowstorms and New Hampshire has thunderstorms, hurricanes, and Snowstorms. For the Median Security Delay by City, Utah has the highest median of Security Delay, 29 min and Second high is Minnesota, 26 min. For the Median National Aviation System (NAS) Delay by City, New Jersey has the highest median of National Aviation System Delay, 26 min and Second high is New York, 23 min. Actually, the top busy airport is in Hartsfield-Jackson Atlanta International Airport, so interestingly that Atlanta has good performance, and John F. Kennedy International Airport is the fifth busy airport 4. For the Median late-arriving aircraft Delay by City, North Dakota has the highest median of Late-arriving aircraft Delay, 26 min and Second high is New jersey, 39 min. For the Median Carrier aircraft Delay by City, West Virginia has the highest median of Carrier Delay, 41 min and Second high is North Dakota, 34 min.

In the design criteria, we unify the color by each delay, weather is red, security is green, Nas is orange and the late aircraft is blue. It helps the audience to keep following the specific data variance. For the exploratory data, Delay Box plot (Fig 1.4), I create in python and exclude all the outlier in the graph, due to the present and clarity, and each delay histogram, I used bin 20 to define the gap and log, due to the detail of the graph and easy to show most of the data in histogram. For the data visualization, on the Bar chart, I filter the data and select the top ten city due to the clarity and erase the redundant and make the chart easy to read. On the choropleth, keep the color the same and also add state and delay value text in the choropleth. Saturation help to define the max to min.

In future work, for the data analysis part, I will combine weather data and crime data by states and cities to focus on the relationship between all of them and find the reason why makes the flight delay. For the data visualization, I wanna work on circular to find the connection between each state and carrier and between each delay and carrier. Moreover, using a contingency table to present the carrier, flight route, arrival delay and distance, four variables, I want to figure out whether distance really effects on delay and each airline performance.

3b. Amanuel Petros

Our dataset is from the Bureau of transportation statistics and my part is to show airline delay trends**.** In datasets such as this one, you would expect to see trends overtime period. Hence the use of a line graph is appropriate because line graphs are used to display quantitative values over a continuous interval or time period. Line Graph is most frequently used to show trends and analyses how the data has changed over time. This visualization also helps us to see trends of different variables of different causes but similar outcomes.

With the addition of area chart, this method of visualization shows both the number of delays (count of delays) and time of delays (median time delay). This means one can see how many delays occurred in contrast to how long the delay took. Combining both line and area charts together as seen in **Fig 1.10** helps to visualize two types of trends over the course of 12 months. I could make two different graphs one for count delays using area chart, and another for time delays using a line graph. However, it is difficult to show patterns, analyze, and compare the two charts separately. Therefore, this method is a rich and deep display of the data for this visualization which incorporates both count and time of delay and I want this graph to be graded.

For the most part of our visualization, we used one unique color for one variable. This means if we used a bar graph for weather delay using a blue color, we must use blue color for weather delay in our area chart visualization. This helps to easily follow the visualization throughout the presentation because it helps avoid the use of multiple colors for one variable all over the presentation. Since we incorporated two charts in a single chart, it is difficult to analyze because the scale of the time delay and count of delay are different measures and needs different scaling technique. To make the chart easier to read, the count of delay is converted to a secondary axis, thus allowing for the data sets to be scaled differently.

From the visualization, we can conclude that the highest delay occurs during the summer also we can conclude that the count of delay is not directly or indirectly proportional to the amount of time delay in minutes. In this visualization, there is a significant change of pattern (high pick) in the security delay of the line graph. This significant change of pattern occurred in September. On the other hand, area chart visualization on the same variable showed an increase of delay but much smaller or lower compared to the line graph. As simple as this visualization, the bar graph could be suggested to visualize these types of flight delays to see which one of them contributes the most towards the overall delay and which one contributes the least. In this project, I get to work on big data and visualization for the first time. With the magnitude of the dataset and type of variables, I had the opportunity to try out different categorical and quantitative visualizations which helped me learn the proper use of graphs.

If I had more time, I might have liked to have developed my visualizations further to an interactive line chart where four types of delays displayed in a single chart to show delay caused by month, as a present of total delay minutes.

3c. Mashall Jahangir

For this project, I decided to work on similar questions and answer what caused the delays and if any, is there any relationship between variables. I used several data visualization techniques on the dataset through the quarter to get a hang of different variables. However, for the final report-I focused more upon Fig 1.7, Fig 1.11 and Fig 1.13. The first one (Fig 1.7- to be graded) is the route map where I used two datasets, one for the airline schedule and the other for the longitude and latitude of all the states to locate the position on the world map. The key message from this visualization was to identify all the routes taken by the top 4 (Southwest, Delta, United and American Airline) have taken from a certain origin. In addition to the routes, it also shows the median air delay caused and the size of that delay. So in short, color shows the details of the carrier, size shows the median of Arrival delay and you can filter data for origin.

Fig 1.11 shows the median arrival delay vs the average distance of the flights -with colors to indicate the airline carrier. Here we can see as the distance for the flights increased, the delay increased. Even though the color shows the variation- I think it makes sense because all these flights overlap with their distances. The last one, Fig 1.13 shows the hex map which indicates the weather delay per state. Apart from creating the visualizations, my role as a group member was to plan weekly meetings, build google documents for all the project deliverables, wrote the abstract, introduction, the variables table, appendix, and overall managed this final report while making sure the text, the format looks good.

Over the course, I learned more than what I expected and understood how important it is to be clear and concise in your message. One has to understand their audience and the message they are trying to convey in order to be successful in creating meaningful visualizations. The viz techniques and Softwares we used over time including tableau, R, D3, and Shiny were a lot helpful to create and the visualizations. If I had more time, I would definitely work more on this dataset and find essential patterns and trends to help a customer see what are the main causes of the flight delay and help the airline industry minimize these delays to maximize their profit.

3d. Erik Wolf

I worked on seeing which states have the most airline traffic to and from each state. The visualization is Fig1.14 Network Graph that shows the volume of states leaving and arriving. I aggregated the state by all airports located within each state. The volume on domestic flights leaving California is much larger than any other state. California looks to have the most connectivity to all the other states. The width is number of destination flights. For my R- huCode I used "circlize" library to plot the chord diagram. First I needed to change the states from Factors into categorical type. Once I did that, I loaded the origin/destination into a new data frame, then transformed into an adjacency matrix. After that I loaded the adjacency matrix into chorddiagram() function. The colors chosen for this visualization are a pastel color. This is easier on the eyes when looking at each of the five highest Among the five largest volume of states with traffic, this plot easily identifies that California and Texas have the most flight traffic among the other five states. The edges connect the amount of flights.

I learned a lot of real world applications for using specific visualizations for showing patterns and relationships of categorical data. If I had been present on the our groups first meeting about our visualizations I would've been able to contribute more. I ended up having the same two visualizations as another member. I decided to scrap them and work on the Network Graph.

For future work and final project I want to make this interactive or reduce the dimensionality from 50 states down, possibly grouping by region. A Network Graph with interactivity would make the weights of each connection easier to interpret. That is something that if I had more time I would've liked to implement.

3e. Christian Marcol

My first contributions to the project were towards cleaning and organizing our dataset. Each row in our dataset corresponds to a unique flight across 2018. Many of these samples were cancelled flights. We wanted to restrict our focus only to delayed flights so I used the dplyr library in R to drop the rows in our dataset which were cancelled. The dataset contained flight information from 17 US airlines. We felt using this many carriers would have limited the effectiveness of our visualizations and so we opted to keep only the top four carrier airlines in the US. These were American Airlines (AA), United Airlines(UA), Southwest(WN) and Delta (DL). The rows containing flights from other carriers were dropped. With the final dataset in hand, a second, smaller dataset was created by randomly sampling with a seed from the full dataset. This sample dataset was created to speed up our exploration of visual techniques to represent the data, our final visualizations were created using the complete dataset. The difference between the mean and the median measures of flight delay times suggested the presence of outliers. In this case flights whose delay time was considerably longer than usual, these were very rare. When conducting our analysis we chose to use the median of delay time. Each delay type was itself a feature of the data, as each flight in the dataset was delayed by one delay type, all median value calculations did not include zero values from other delay types.

Primarily, I sought to create visuals to compare airlines and draw conclusions about the activity of each airline from these comparisons. Firstly, I wanted to compare the contributions of each carrier to airport traffic as well as the proportional utilization of each airport by carrier. The dataset had many carriers so a treemap was appropriate for visualizing this data. I used tableau to create the treemap because I felt it was easier to read and more aesthetically pleasing than my implementation in ggplot2. To create the treemap, I created and aggregate feature for the count of each carriers appearance in the dataset, as well as each airports appearance and encoded these into size. Next, I colored them by carrier and added detail metrics for all of the variables. I also wanted to compare carriers by the frequencies of delay types in the top five cities. As this analysis uses a three way contingency table, a mosaic plot was appropriate. For the mosaic plot I used vcd::mosaic in R. As the weather and security delays did not contribute much to total delays, they did not show up very well in a single mosaic plot so I opted to create three mosaic plots for five, two, and one delay type respectively. While I feel the mosaic plot has the potential to convey very descriptive useful information. I did not feel I was able to customize the plot enough to create an effective visualization. For this reason I would like to focus on the treemap visualization.

By creating these visualizations, I was able to better understand why some visualizations are effective and others aren’t. In particular I learned how to create effective visualizations of categorical data. Through the class and the project I feel more comfortable creating plots using ggplot2 and Tableau. I would like to create a more effective mosaic plot and will likely continue to work on that going forward.

4. References:

1. **Hristina Byrnes.December 8,2018. usatoday.com.*The 13 biggest air travel complaints of 2018, from flight delays to discrimination and more.* [online] Available at:** https://www.usatoday.com/story/travel/flights/2018/12/08/the-biggest-air-travel-complaints-of-2018/38691345/ **Accessed November 25, 2019.**
2. **Bureau of Transportation Statistics. March 29, 2019. Bureau of Transportation Statistics.com. *Airline On-Time Performance and Causes of Flight Delays* [online] Available at:** https://www.bts.gov/topics/airlines-and-airports/airline-time-performance-and-causes-flight-delays **Accessed November 25, 2019.**
3. **Christy Rodriguez. 2019. upgradedpoints.com .Which U.S. Airlines Dominate Market Share in North America? [online] Available at:** https://upgradedpoints.com/us-airlines-marketshare-north-america/. Accessed **November 25, 2019.**
4. **Melanie Renzulli. June 26, 2019. The 25 Busiest Airports in the United States[online] Available at:** https://www.tripsavvy.com/busiest-airports-in-the-usa-3301020. **Accessed November 25, 2019.**