Airport Delay Analysis

Submitted by: Charanjit Bambrah, Foroozan Akhavan, Jingwen Shi

(Group I)

Abstract - With the increasing necessity of peoples' daily travel needs, the transportation system plays a huge role in human life. Airlines have become one of the most popular transportation options for people to choose in the world. For those who traveled on an airplane might have experienced the airline delays, it poses a huge inconvenience for travelers regardless of the reason for the delay. The airline delay analysis dataset used in our paper has been taken from Kaggle which contains daily airline information in the United States. In this paper we will analyze the dataset and use visualization to demonstrate the main reason for airline delays and by summarizing the conclusion to give some recommendations and discuss some future work for improving the airline delays.

Index Terms – Interactive Data Visualization, Data Analysis

-----------------◆-----------------

1 Introduction

Air transport has made a huge contribution to the United States' economy. According to the report by IATA Economics called "The Importance of Air Transport to The United States", the air transport sector directly provided about 2.2 million jobs and 1.7 million jobs created by their supply chains in 2017 [1]. On the other hand, in the air transport industry, the airline with its supply chain has supported about $641 billion dollars of GDP in the United States, and the extra $138 billion dollars of GDP with the foreign tourists support, so the total is about $779 billion dollars which equals to the 4.2 percent of whole country's GDP in 2017 [1]. By looking at these data we could realize how the air transport industry is closely related to our life.

Even though the air transport industry benefits us a lot, sometimes there are still some accidents which might cause serious problems if we do not pay more attention to it, for instance the airline delays. Thus, it is necessary for us to figure out the main reasons that caused the airline delays accident and analyze it in order to get a better idea how to prevent or decrease such things happening in the future for those who are working in this industry and what they could improve from it. In this report, we chose the dataset from Kaggle which contains daily airline information covering from flight information, carrier company, to taxing-in, taxing-out time, and generalized delay reason of exactly 10 years, from 2009 to 2019. And we will mainly focus on the data for the recent two years from 2018 to 2019.

We use some tutorials to complete this project. First, we used Jupyter Notebook to do the data preparation and data cleaning and export as a csv file, and after we finished dealing with the data, we used Tableau to do the visualization by creating some charts and a dashboard. Important information is gathered, and some attributes are used to generate the graphs like bar chart or heat map to show what result we got from the dataset , and we combine those charts together by creating the dashboard to display our result.

2 Related Work

To better reach the goal of our project, we searched some relevant reference papers and saw how they analyze the data by some machine learning models and how they visualize their airline delay accidents result. The first research paper named “Prediction of weather-induced airline delays based on machine learning algorithms” by Sun Choi, et al. In this paper, the authors used data mining and machine learning algorithms to predict the airline delays caused by the inclement weather conditions, and they used random forest to perform and they found that the predictions with the actual weather is much better than the predictions with forecast due to the uncertainty in forecast. Thus, from this article, we could realize that weather has a lot of impact on airline delays.

Another research paper named “A machine learning approach for prediction of on-time performance of flights” by Balasubramanian Thiagarajan et al. This paper the authors built a real-time Decision Support Tool to inform the airline and the passengers of the delays before departure.They created an interface which allows the user to search the airline details including delay information, and they developed a model to efficiently predict the flight delays [3]. To summarize it, if the people working in the airline companies could know the delay information in advance, both the companies and the passengers could reduce the losses.

3 Data Collection and Transformation

Here we provided the step to step of preparing the data details.

❖ To start off this project first we needed to find a reliable data source. We found our dataset from Kaggle website which is a well-known website for datasets. You can find the data source at https://www.kaggle.com/sherrytp/airline-delay-analysis.

❖ After looking closely at the datasets provided in this link, we decided to work on two datasets, the years 2018 and 2019. We didn’t choose 2020 because of two reasons, first the dataset was smaller than 2018 and 2019 and the number of columns were less than these two datasets. So, we found this dataset to be less reliable than the past two years. On the other hand, other years had huge amounts of rows which made the dataset really big and hard to work with, since we didn’t have the processing power needed for these datasets to analyze the data. Our 2018 and 2019’s data take a considerable amount of time to be loaded and shown as a dataframe.

❖ To start the rest of the project we created a github repository for our project to access the resources simultaneously with our group members. The link to our repository is https://github.com/Feritaba/Data_visualization.

❖ In the next step we downloaded the datasets mentioned above and took a look at the datasets in the Jupyter Notebook. Jupyter Notebook is a well-known tool to play with data. It is easy to use and there are a huge number of libraries in order to work with the data. In this step, we read the csv file as a dataframe and assigned a name to it. Then, we started to explore the high-level statistics of our data such as histogram of our columns. Also, shape, name of the columns, and the whole look of the dataset was important for us.

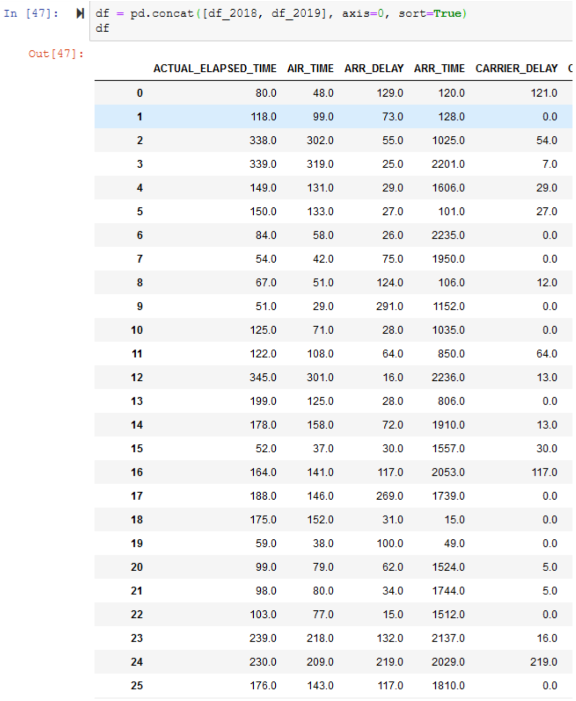
❖ After we got familiarized with our dataset, we found out there is a column called carriers which shows the carrier name of each airline. This column used carrier codes. For example, instead of American Airlines we see "AA". In this step, we took a pause to search for these codes using Google. The list of the codes associated with the airline names are available in the Readme file at https://github.com/Feritaba/Data_visualization. Then, we found out that we don't have data for one of the carriers in 2019's dataset which was available in 2018's dataset. In order to have a consistent analysis we deleted the data from 2018's dataset.

❖ For the next step, we took a huge leap to prepare our data for visualization. Since we needed the delays columns and, in these data sets, the delay reasons were separated to illustrate the details of delays. Each row that has NaN value was useless for us since we didn't have enough data to rely on. We needed to delete the rows that we didn't have enough data and keep the rows with actual values. If the data is not available for delays, how can we make a decision based on "No Information"? The other issue of our data was the "CANCELLED" and "CANCELLATION_CODE" columns, which were entirely "NaN" in our two datasets. Since again we didn't have enough data, we deleted these columns.

❖ We also took minor steps for the final touch ups of our data and made the data types consistent in the entire numerical columns.

❖ In the end, we concatenated the two datasets and made one dataset and converted it to a CSV file.

All of the works related to cleaning and preparing the datasets can be found in this https://github.com/Feritaba/Data_visualization/blob/main/Project%201%20-%20Data%20Visualization.ipynb to the Github page.

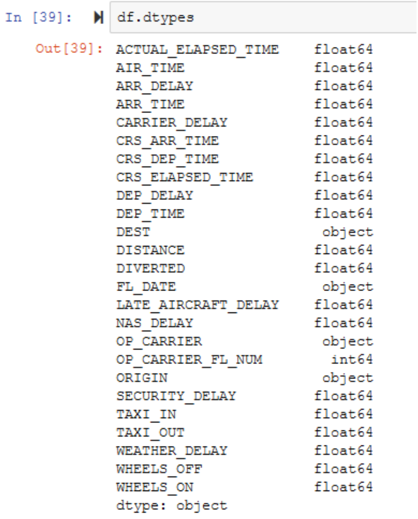


```
In [47]:  df = pd.concat([df_2018, df_2019], axis=0, sort=True)
          df

Out[47]:
```

	ACTUAL_ELAPSED_TIME	AIR_TIME	ARR_DELAY	ARR_TIME	CARRIER_DELAY	C
0	80.0	48.0	129.0	120.0	121.0	
1	118.0	99.0	73.0	128.0	0.0	
2	338.0	302.0	55.0	1025.0	54.0	
3	339.0	319.0	25.0	2201.0	7.0	
4	149.0	131.0	29.0	1606.0	29.0	
5	150.0	133.0	27.0	101.0	27.0	
6	84.0	58.0	26.0	2235.0	0.0	
7	54.0	42.0	75.0	1950.0	0.0	
8	67.0	51.0	124.0	106.0	12.0	
9	51.0	29.0	291.0	1152.0	0.0	
10	125.0	71.0	28.0	1035.0	0.0	
11	122.0	108.0	64.0	850.0	64.0	
12	345.0	301.0	16.0	2236.0	13.0	
13	199.0	125.0	28.0	806.0	0.0	
14	178.0	158.0	72.0	1910.0	13.0	
15	52.0	37.0	30.0	1557.0	30.0	
16	164.0	141.0	117.0	2053.0	117.0	
17	188.0	146.0	269.0	1739.0	0.0	
18	175.0	152.0	31.0	15.0	0.0	
19	59.0	38.0	100.0	49.0	0.0	
20	99.0	79.0	62.0	1524.0	5.0	
21	98.0	80.0	34.0	1744.0	5.0	
22	103.0	77.0	15.0	1512.0	0.0	
23	239.0	218.0	132.0	2137.0	16.0	
24	230.0	209.0	219.0	2029.0	219.0	
25	176.0	143.0	117.0	1810.0	0.0	

Figure 1 - Dataset

The datatype of these attributes can be seen below.



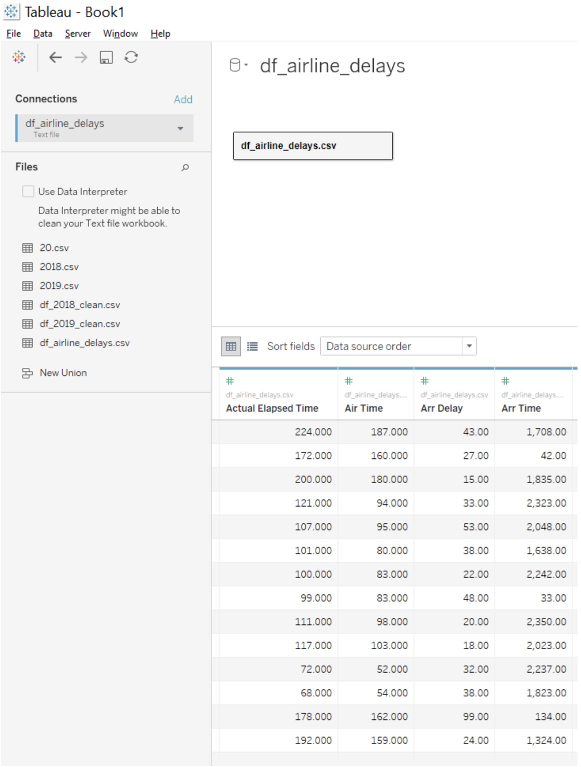
```
In [39]:  df.dtypes

Out[39]: ACTUAL_ELAPSED_TIME    float64
         AIR_TIME               float64
         ARR_DELAY              float64
         ARR_TIME               float64
         CARRIER_DELAY          float64
         CRS_ARR_TIME           float64
         CRS_DEP_TIME           float64
         CRS_ELAPSED_TIME       float64
         DEP_DELAY              float64
         DEP_TIME               float64
         DEST                    object
         DISTANCE               float64
         DIVERTED               float64
         FL_DATE                 object
         LATE_AIRCRAFT_DELAY    float64
         NAS_DELAY              float64
         OP_CARRIER              object
         OP_CARRIER_FL_NUM        int64
         ORIGIN                  object
         SECURITY_DELAY         float64
         TAXI_IN                float64
         TAXI_OUT               float64
         WEATHER_DELAY          float64
         WHEELS_OFF             float64
         WHEELS_ON              float64
         dtype: object
```

Figure 2 - Data type

4 Data Visualization

As seen in the dataset above all the cleaning process has been done via Jupyter Notebook. The only tool we use for our visualizations is Tableau. Here are the steps taken to the end of the analysis and the visualization.

1. Loading the CSV file to the Tableau and press the Automatic Update button.



df_airline_delays.csv Actual Elapsed Time	df_airline_delays.... Air Time	df_airline_delays.csv Arr Delay	df_airline_delays.... Arr Time
224.000	187.000	43.00	1,708.00
172.000	160.000	27.00	42.00
200.000	180.000	15.00	1,835.00
121.000	94.000	33.00	2,323.00
107.000	95.000	53.00	2,048.00
101.000	80.000	38.00	1,638.00
100.000	83.000	22.00	2,242.00
99.000	83.000	48.00	33.00
111.000	98.000	20.00	2,350.00
117.000	103.000	18.00	2,023.00
72.000	52.000	32.00	2,237.00
68.000	54.000	38.00	1,823.00
178.000	162.000	99.00	134.00
192.000	159.000	24.00	1,324.00

Figure 3 - Loading data into Tableau

2. Create the first sheet for our charts.

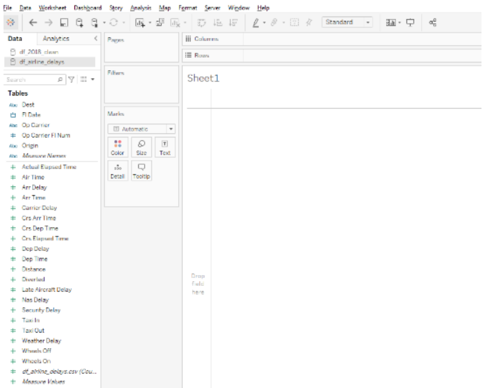
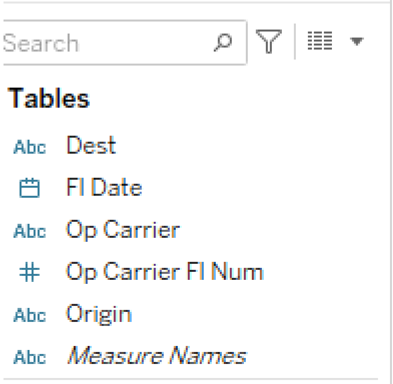


Figure 4 - First sheet in Tableau

3. Dimension columns of our dataset



Search

Tables

Abc Dest

Fl Date

Abc Op Carrier

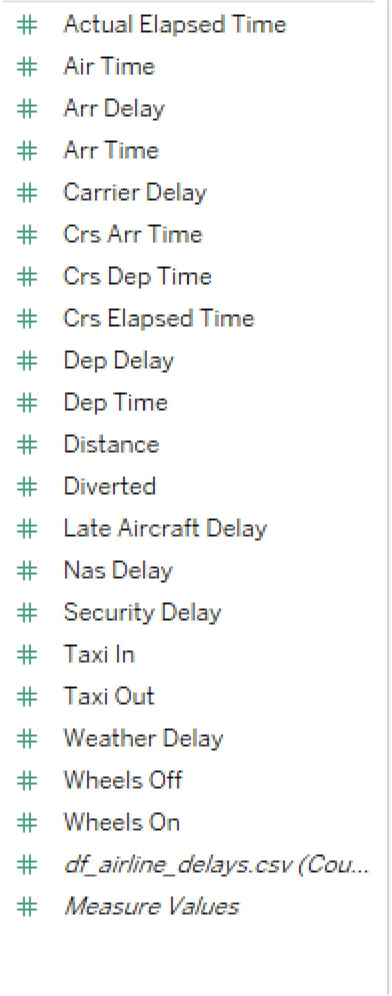
Op Carrier Fl Num

Abc Origin

Abc *Measure Names*

Figure 5 - Dimensions

4. Measure columns of our dataset



Actual Elapsed Time

Air Time

Arr Delay

Arr Time

Carrier Delay

Crs Arr Time

Crs Dep Time

Crs Elapsed Time

Dep Delay

Dep Time

Distance

Diverted

Late Aircraft Delay

Nas Delay

Security Delay

Taxi In

Taxi Out

Weather Delay

Wheels Off

Wheels On

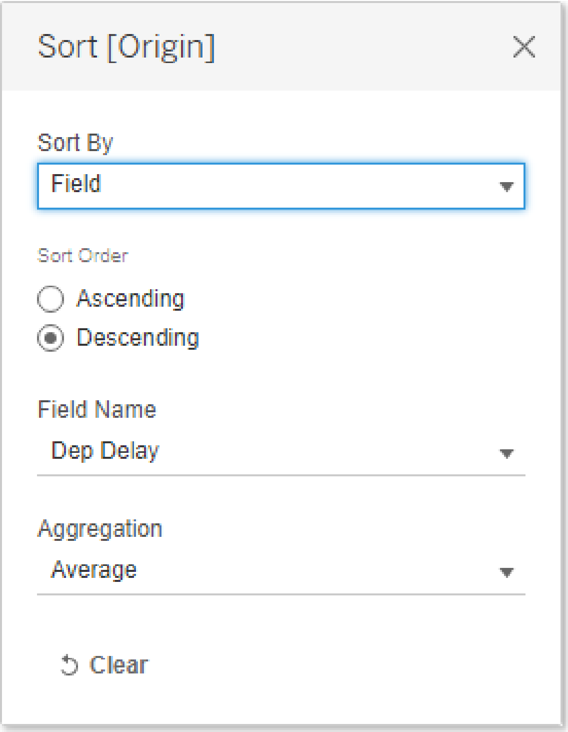
df_airline_delays.csv (Cou...

Measure Values

Figure 6 - Measures

5. Creating the first chart using the tools in the Tableau to analyze in which airport the highest average of departure delays (in minutes) happened.

We created this chart using the Origin column as a column and Avg of Departure Delays as rows. Then we sorted the Origin column based on the avg measures of the field of dep_delay in descending order.



Sort [Origin]

Sort By: Field

Sort Order

○ Ascending

● Descending

Field Name: Dep Delay

Aggregation: Average

Clear

Figure 7 - Sorting in Tableau

The next step we took was to choose only top 10 airports. With using the Edit Filters, choosing By Field, top 10 on avg of Dep_Delay columns we created the final chart of our first step.

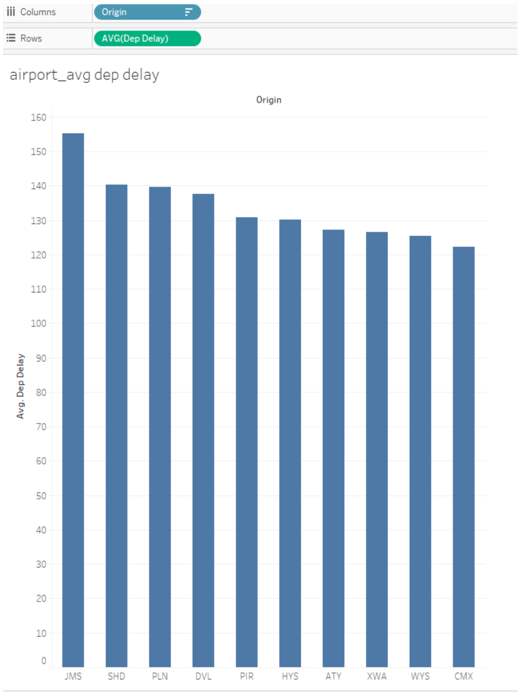
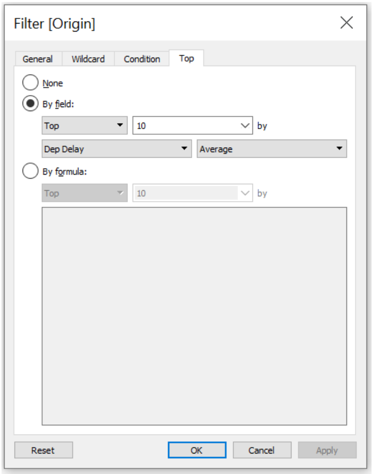


Figure 8 - First Chart

Also, we used a size button to resize the bar charts. We have not changed the color of the chart since blue is the most appropriate color for color blindness issues. Otherwise, we could use the color button of the shown screenshot to change the color of the chart as well.


Filter [Origin]

General | Wildcard | Condition | Top

○ None
● By field:
Top | 10 | by
Dep Delay | Average
○ By formula:
Top | 10 | by

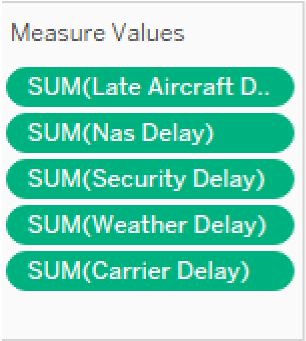
Reset | OK | Cancel | Apply

Figure 9 - Filtering top 10 columns in Tableau

In this chart it is shown that JMS airport in Jamestown city in North Dakota has the highest average minutes of the departure delays. In the next chart we will see which category has the highest minutes for this city.

6. The next try of designing charts as mentioned above was creating a more detailed chart. In the top 10 cities in the previous chart which category is the main reason to blame for the delays.

First, we used the previous chart to help us keep the 10 cities on top of the list. Then we used five delay columns we mentioned earlier as the reasons for the delays. These columns are:


Measure Values

SUM(Late Aircraft D..
SUM(Nas Delay)
SUM(Security Delay)
SUM(Weather Delay)
SUM(Carrier Delay)

Figure 10 - Measures Columns

We used SUM function for these columns because we cannot compare the five averages with the actual average we had and it would be mathematically wrong if we do that. So, the SUM function is our choice for this chart. Then we chose another type of bar chart to have all the data at the same close to each other. In the next step we changed the color of the chart to a colorblindness selection of the Tableau chart. Here is the settings of the colors we used for our chart:

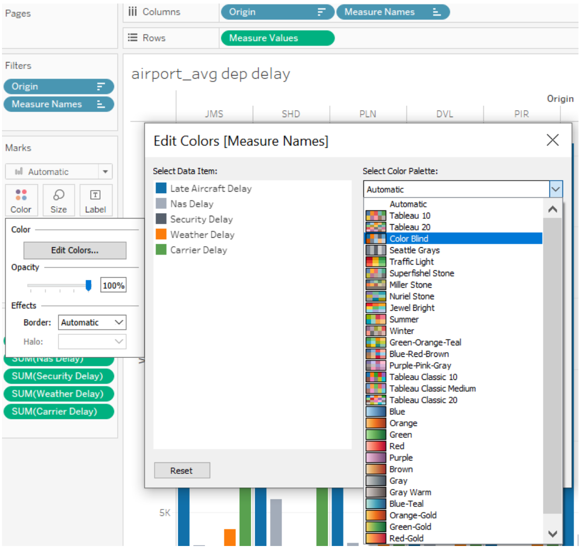


Figure 11 - Choosing Color Blindness Pallet

This way other than creating a different chart we created it more distinguishable.
Here is the result of the chart:

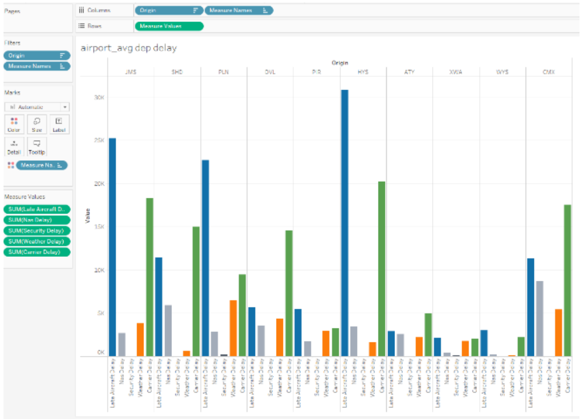


Figure 12 - Aggregation Measures Chart

In the chart above we see why JMS airport had the highest average of delays in minutes among all other airports across the US. First, late aircraft delay which means the airplane arrived at the airport with delay, and second carrier delay which means aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.

In this chart the numbers on the left show the sum of the minutes of departure delays that happened in the top ten airports that had the highest average minutes of delays in just two years.

7. The next question is, can we have a map to see how many delays happened due to the weather delay and NAS delay? (What is NAS delay in the first place? Delays that are within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc. Delays that occur after Actual Gate

Out are usually attributed to the NAS and are also reported through OPSNET.)

Answering this question was a little bit tricky since creating a map needs geological information in Tableau. First, we chose the column and rows we wanted to illustrate the chart with. The Origin on one hand and two other columns, NAS Delays and Weather Delays on the other hand.
Then, from the Dimensions tab on the left side we clicked on the Origin:

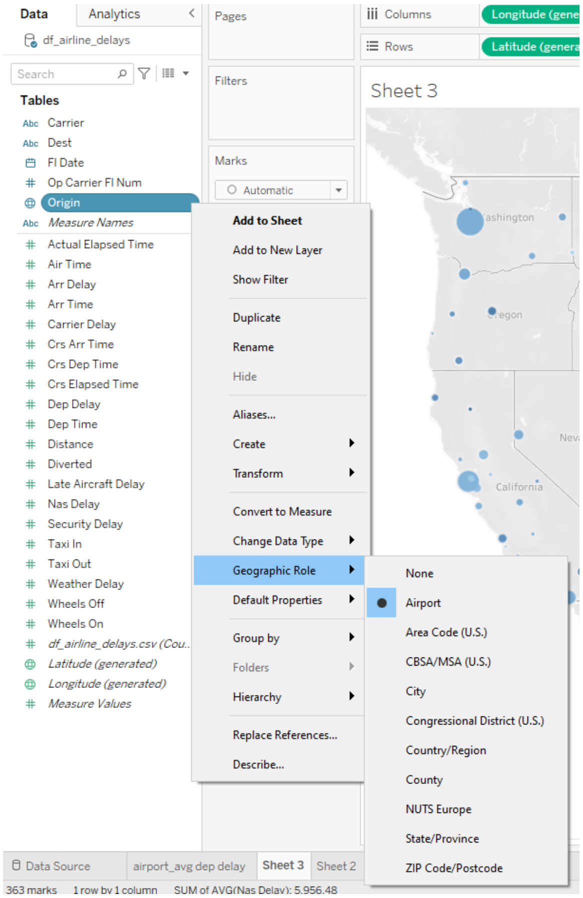


Figure 13 - Geographic Role To Create Map

We chose the Airport for the Geological Role of the origin. Then the columns and rows on top of the Tableau changed to the Longitude and Latitude automatically:

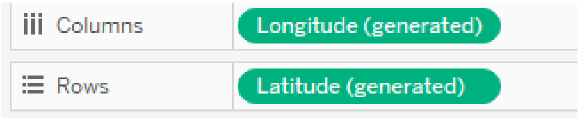
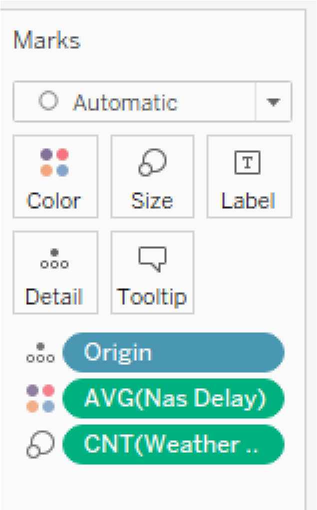

Columns Longitude (generated)
Rows Latitude (generated)

Figure 14 - Long. and Lat. Automatically Appears

On the Marks box in the Tableau there are a couple of items that we can play with to make our chart a little clearer. We chose Avg of NAS delay to show with color differences and Weather Delays with size differences.



Origin
AVG(Nas Delay)
CNT(Weather ..

Figure 15 - The Marks Panel

There are also other bars that helped us to modify our chart better.

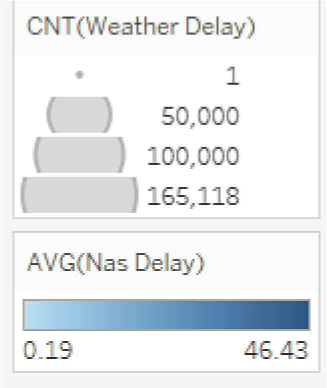


Figure 16 - Change The Color and Size Panel

Each of these can be edited easily from the available menu on the right side of our page.

Another issue was that one of the airports couldn't be recognized by Tableau and we needed to add the Longitude and Latitude of the airport manually which was a great feature of Tableau.

The final chart is shown below:

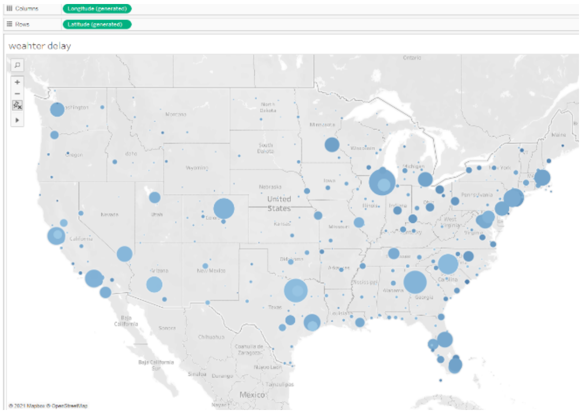


Figure 17 - Map Of Weather and NAS Delays

In this chart we clearly can see on the east coast of the US we have more weather-related delays, also in the Chicago airport, Dallas and Houston we see huge circles that shows the accumulation of weather delays in comparison to other airports.

8. The next question would be what carriers have the most delays in all categories?

To make this chart since we had more experience was easier. choosing the Carrier as column and avg of Dep Delay for the rows it created a simple bar chart for us.

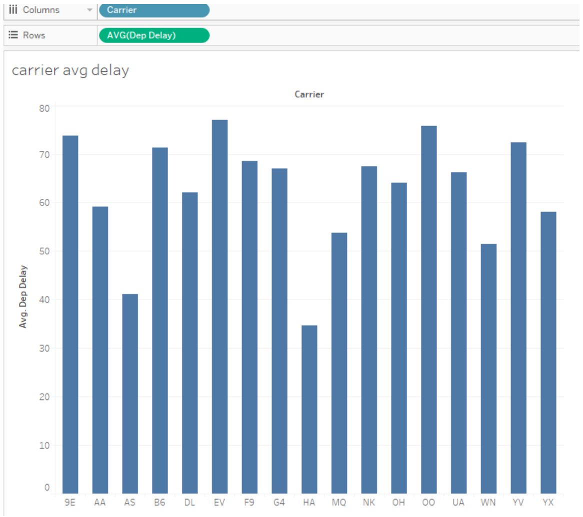


Figure 18 - Departure Delays In Airlines Not Sorted

Of course, this chart has so many problems and it is not legible in one glance. This chart is not sorted. The code names of the carriers are vague and not well-known to the target audience. So we fixed the chart by sorting the Carrier based on the avg of dep delay and chose aliases for the carrier codes. The final chart is:

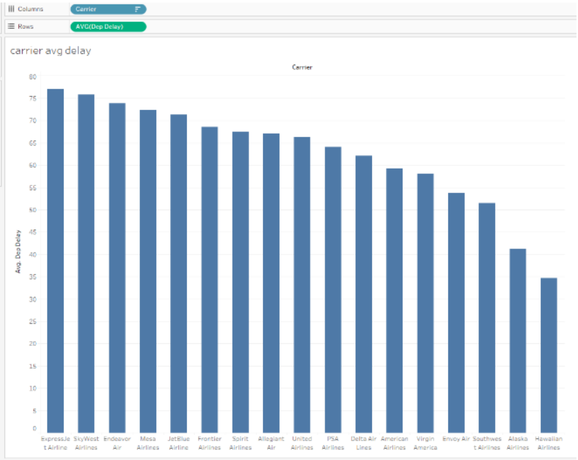
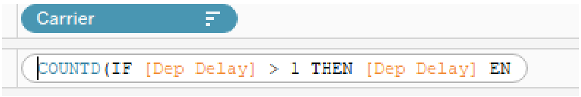


Figure 19 - Sorted Chart

Now with these edits it is obvious that ExpressJet Airline has the highest average minutes of delays among all the carriers and Hawaiian Airlines has the least number of average delays.

In another format we want to optimize the calculation behind the chart to make it more precise. In our dataset there are some rows in the departure delay column that have negative value and it means the flight took off earlier than the actual departure time. So we changed the rows values in the above chart using this line of code:



```
COUNTD(IF [Dep Delay] > 1 THEN [Dep Delay] EN
```

Figure 20 - Piece Of Code

With this line of code, we have departure delays greater than 1 minute and we do not take into calculation the flights that departed sooner than we expected.

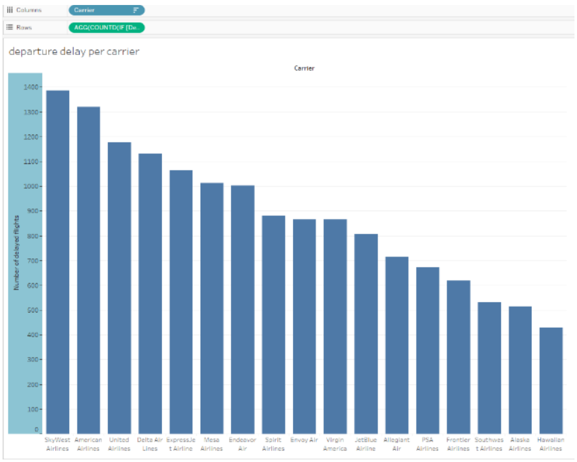


Figure 21 - Chart With New Calculation

So the chart changed to the chart above and we see that the result is totally different from the previous chart. SkyWest Airlines has the highest number of flights with delays in two years and the following leader is American Airlines. Next time you booked a flight with the top airlines above, keep in mind that your flight might have a delay.

9. In the next step we wanted to create a chart to see how much carrier delays are responsible for departure delays for each airline. Since we explained all the details above, we just put the screenshot down below to show the results.

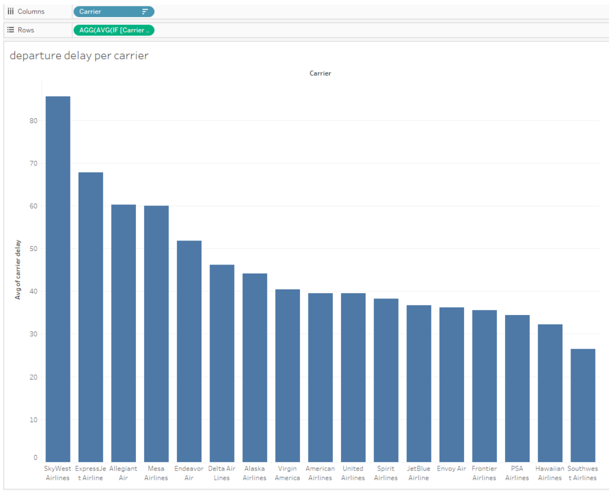
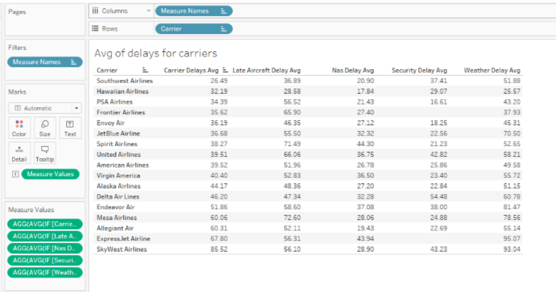


Figure 22 - Carrier Delays Chart Per Airline

In this chart we clearly see that SkyWest Airlines has also the highest avg of carrier delays between all other airlines and the number one culprit of having high avg of departure delays for this airline is carrier delay. We also thoroughly explained what carrier delay is in part 6 of this document.

10. In the last chart we decided to give an overview of all delays to the target audience. That's why we created a table to put all the data in one place to show the overall view. Also, in order to make the data more legible, we hid the rows that had flights less than big companies. In this graph that we will see below there are two airlines ZERO security delays, Frontier airlines and ExpressJet Airlines. Without having the overview look at the data we couldn't find out this information.



Columns: Measure Names

Rows: Carrier

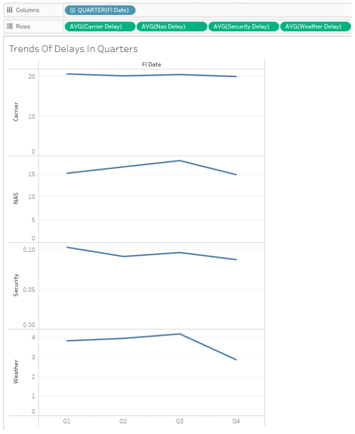
Filters: Measure Names

Avg of delays for carriers

Carrier	Carrier Delays Avg	Late Aircraft Delay Avg	Nas Delay Avg	Security Delay Avg	Weather Delay Avg
Southwest Airlines	26.49	36.89	20.90	37.41	51.88
Hawaiian Airlines	32.19	28.58	17.84	29.07	25.57
PSA Airlines	34.39	56.52	21.43	16.61	43.20
Frontier Airlines	35.62	65.90	27.40		37.93
Envoy Air	36.19	46.35	27.12	18.25	45.31
JetBlue Airline	36.68	55.50	32.32	22.56	70.50
Spirit Airlines	38.27	71.49	44.30	21.23	52.65
United Airlines	39.51	66.06	36.75	42.82	58.21
American Airlines	39.52	51.96	26.78	25.86	49.58
Virgin America	40.40	52.83	36.50	23.40	55.72
Alaska Airlines	44.17	48.36	27.20	22.84	51.15
Delta Air Lines	46.20	47.34	32.28	54.48	60.78
Endeavor Air	51.86	58.60	37.08	38.00	81.47
Mesa Airlines	60.06	72.60	28.06	24.88	78.56
Allegiant Air	60.31	52.11	19.43	22.69	55.14
ExpressJet Airline	67.80	56.31	43.94		95.07
SkyWest Airlines	85.52	56.10	28.90	43.23	93.04

Figure 23 - Measures Table

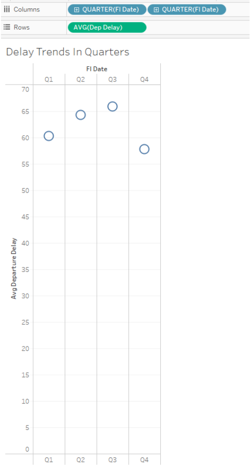
11. The nex charts are created using the knowledge we learned above from the Tableau to show the trends of delays in year quarters.



Columns: QUARTER(Fl Date)

Rows: AVG(Carrier Delay) AVG(Nas Delay) AVG(Security Delay) AVG(Weather Delay)

Figure 24 - Trends Of Each Delays In Quarters Of 2018 And 2019



Columns: QUARTER(Fl Date) QUARTER(Fl Date)

Rows: AVG(Dep Delay)

Figure 25 - Trends Of Departure Delays In Quarters Of 2018 And 2019 Combined

5 Results

The hardest part of this project was deciding on which charts should be shown on the dashboard to give the audience the answer to the questions we first were asked. We decided to put the two bar charts, two trend lines and also one map. But in order to show these charts we chose less information in each chart to make it more simple. For example, in the map we decided to go with the top 10 airports that have the highest delays in weather and NAS delays. This way we could reduce the amount of unnecessary data that has been shown to the audience. The final result of our dashboard is shown in Fig. 26.

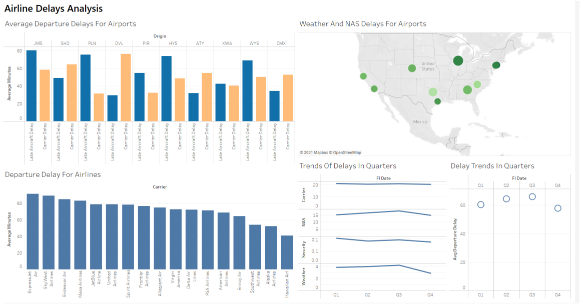


Figure 26 - Dashboard

6 Discussion

With all the illustrations above, it is evident that JMS airport in North Dakota sees the most departure delays while the east coast of the U.S. is prone to weather related delays.

Delays are not only caused by the airport but certain airlines also contribute to the problem. The graphs depict that on an average ExpressJet is known for highest average minutes in delays whereas Hawaiian Airlines have the least delays.

7 Future Work

In this project, we have tried to briefly analyze the various delays taking place at the airports. This analysis can be further improved by applying various machine learning techniques with other analysis software such as Python. It could be integrated with weather forecast information for more accurate prediction of the delays

By making use of regression models, future predictions of airport delays can be another extension to the project. This will allow the airport authorities to foresee any complications and take necessary action to improve the travelling experience of the passengers. It could be integrated with weather forecast information for more accurate prediction of the delays. Although this whole process of modifying might take a lot of resources and time to be truly implemented, the actual benefit behind it could not be underestimated if this model can be more accurately handled.

With all the necessary analysis and visualizations, this project could further develop a web based or mobile application. This will not only help the airport authorities with real time statuses but also helps the extended target audience such as air travelers.

8 Learnings

This research has immensely helped in understanding the basic concepts and method of using Data Visualization to solve real world problems. It allowed us to learn more about the visualization tool, Tableau. Tableau is very effective and easy to use software that focuses on Business Intelligence.

While creating the visual analytics to answer key questions of our research, we found that dashboards must have simple design while conveying the message to a wider audience. As the readability of the graph increases, the message becomes more impactful. Also the spacing between the charts on the dashboard makes it easier to look at a broader picture while making key decisions.

9 References

https://www.iata.org/en/iata-repository/publications/economic-reports/the-united-states--value-of-aviation/

S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 2016, pp. 1-6, doi: 10.1109/DASC.2016.7777956.

B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan and V. Vijayaraghavan, "A machine learning approach for prediction of on-time performance of flights," *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, 2017, pp. 1-6, doi: 10.1109/DASC.2017.8102138.