PROJECT PROPOSAL

Project Title : VISUALIZATION OF FLIGHT DELAYS IN USA

Team Members:

First Name	Last Name	University ID
Prathima	Godha	2001077682
Vikranth	Nallapuneni	2001077680
Manasa	Gudise	2001077681

Abstract:

There are lots of flights that travel within different cities of states of USA to help people to travel in short time. And few times the people who had travel plans suffer due to delay or cancellation of flights which disturbs the travel plans of many people. So a visualization with previous delays information would help in planning ahead to handle the situations in case of delay. Our project is to visualize the 2015 delay of flight arrivals and departures of different airways at different airports. These visualizations help people in planning the trip accordingly or in choosing a better airways based on the requirement while booking a flight ticket.

And also there would be many people who suffers due to cancellation of scheduled flights. So we plans to prepare a visualization which shows the 2015 cancellation records of different flights of different airways of USA so that helps in preparing the users to handle the situations in case of cancellation of any flight.

Along with this visualizations of delay and cancellation records of different airways at different airports of USA, we also plans to visualize the percent of various reasons like weather, security issue etc.. because of which the flights were delays and cancelled in the previous records of data which will help people to decide the time of scheduling the travel plan that is they could plan travelling earlier if there is any chance of bad weather.

We also discuss the relevance and reasons for choosing specific visualization methods over others by comparing existing related works.

Introduction and Motivation:

Now a days everybody is travelling around the world and the primary means of transportation people choose is flights since it is the fastest means of transportation. In case of emergency or other preferences like short journey, comfortable journey etc.. flights are the best means of transportation.

But in some cases journeys might not go as expected. For example, We have experienced flight delay which further lead to flight cancellation during our first USA visit at JFK airport while travelling to the Indianapolis and we later found out that the flight delays are very common from JFK airport to Indianapolis. So, our main motivation in this project is to help other people know in advance which airlines have maximum and minimum delays, which months there are more delays and in which months there are less delays, which airlines have more flight cancellations and which airlines have less flight cancellations. These visualizations helps people to plan their journey or vacation in advance regarding the month to travel for vacation or which airline to choose for departure and arrival and may help them plan their stay if the flight delay time is more or if a particular airline has more frequent flight cancellations than others.

We would like to visualise the 2015 data of arrival and departure delays of the flights (US National Flights) of different airways based on origin and destination airports, different reasons for delay, reasons for cancellation of flights, average delays of different airlines, delays of different airlines in different months.

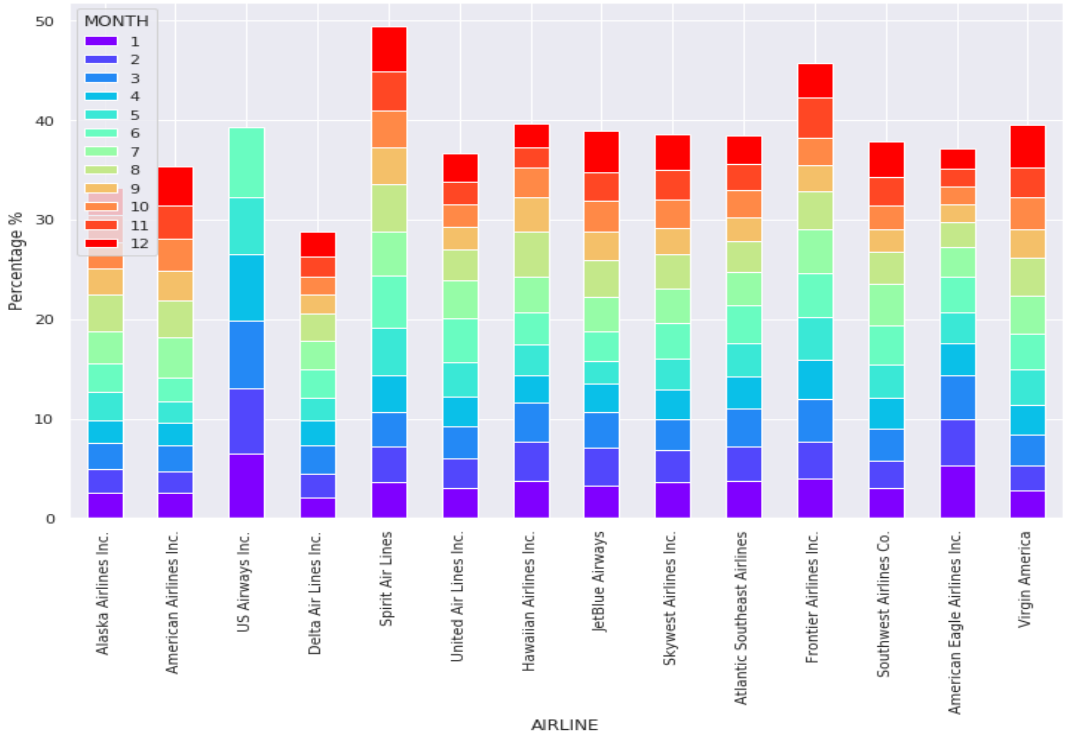
To plot the above visualizations, we obtained the data from US Bureau of Statistics (source of dataset). From below data analysis over years from 2013 to 2022. We can identify that there is pretty reasonable percentage of flight delays and few cancellations. So in that situations its necessary for travellers to make required arrangements. We considered a data of flights of one particular year (2015) to plot the mentioned visualizations. The data contains information like scheduled departure and arrival time, actual departure time and arrival time, reasons for delay, reason for cancellation if the flight is cancelled and the arrival ad departure cities of flight, airline to which the flight belongs to. These data pretty represents the whole information which need to be visualized in order to easily understand the pattern or trends between different airlines and also major causes of delay and cancellation.

Year	Ontime Arrivals	Ontime (%)	Arrival Delays	Delayed (%)	Flights Cancelled	Cancelled (%)	Diverted	Flight Operations
2013	2,892,534	77.33%	775,845	20.74%	62,971	1.68%	9,026	3,740,376
2014	2,538,538	74.39%	766,539	22.46%	98,017	2.87%	9,489	3,412,583
2015	2,651,094	77.74%	681,843	19.99%	67,641	1.98%	9,652	3,410,230
2016	2,655,798	80.97%	571,198	17.41%	44,130	1.35%	8,794	3,279,920
2017	2,592,615	78.39%	660,224	19.96%	46,298	1.40%	8,142	3,307,279
2018	2,652,750	79.31%	632,220	18.90%	51,722	1.55%	8,210	3,344,902
2019	3,328,044	77.52%	854,765	19.91%	98,200	2.29%	12,358	4,293,367
2020	2,345,752	80.88%	282,315	9.73%	267,567	9.22%	4,828	2,900,462
2021	2,646,155	82.33%	509,173	15.84%	50,611	1.57%	7,996	3,213,935
2022	2,957,098	75.71%	822,661	21.06%	116,583	2.98%	9,687	3,906,029

During the initial research, users will have a misconceptions on airport and airline carrier reliability based on general user reviews on websites etc. But with help of these above mentioned visualisations plotted one can come to know about the probability of delay of the different flights of different airways, know the probability of delay at certain period of year and the most likely reasons for delay and cancellations which helps the users in choosing an airline which has less percentage of delays or cancellations from the required location of departure at the desired period of year or make travel plans at certain period of time that is when there is less chance for the situations that are the reasons for cancellation of flights and also helps users in choosing based on some other preferences. Thus these visualizations helps travellers in making decisions appropriately based on the overall data rather than depending on few random user reviews in websites.

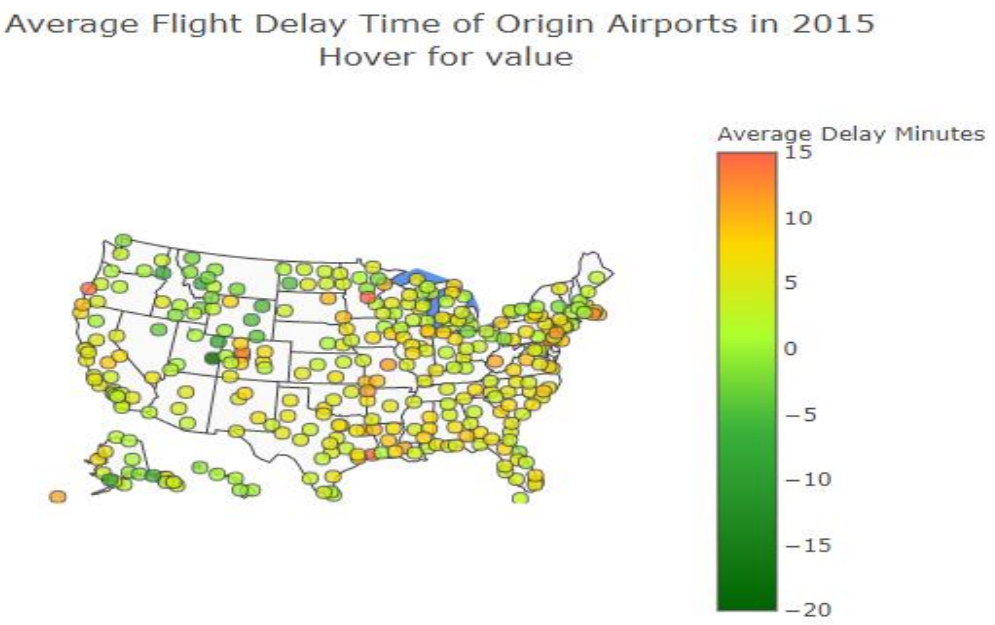
Background or Existing Visualizations:

Source of Visualization 1:



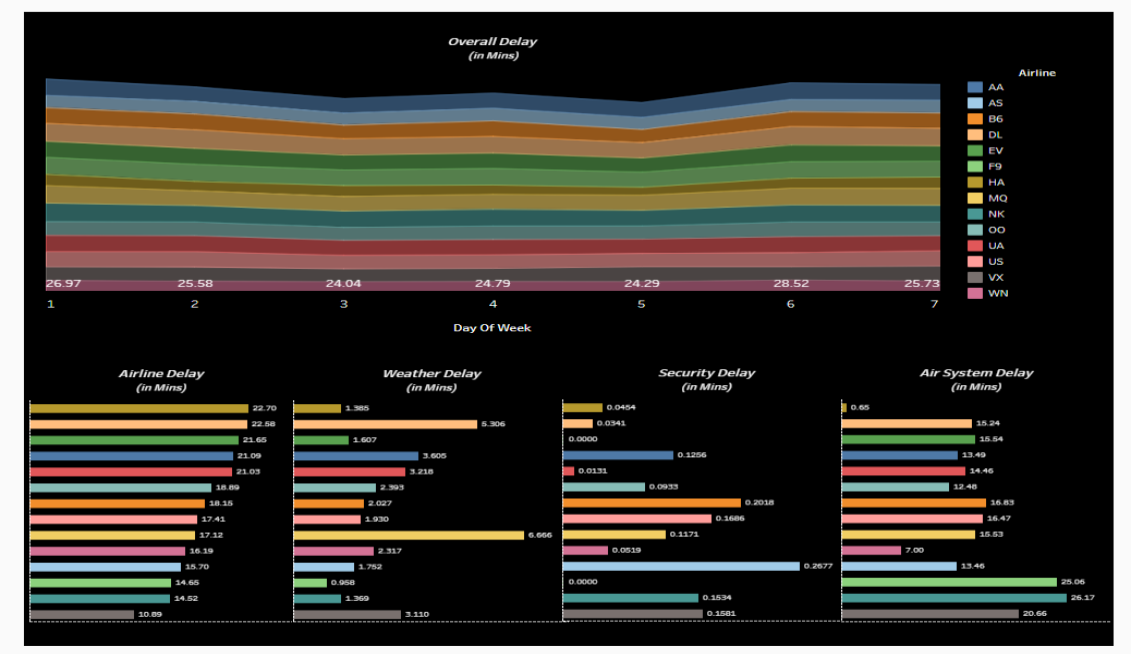
This is one of the visualisations which represents the percent delay of flights of different airlines in USA in different month and this is plotted using the bar plot method. In the above graph it is very easy and useful in identifying which have high or low percentage of delays in overall year, but it is difficult to compare the percentage of delay of an airline with other in one particular month since exact percentage values are difficult to calculate without labelling. For example if the price of ticket of "Virgin America" and "Frontier Airlines Inc" is same and a user wants to compare the percent of the flights that are delayed for "virgin America" and "Frontier Airlines Inc" in month 7, Here we can't predict which airline has high percentage of delay of flights. And also to represent the different months they used colour similar to heat map representation instead of choosing some instinct colours which gives an impression that there is some other variable which represents the intensity of colour. So I would like to visualise this

plot in different manner in such a way that we can find which airline is best for the selected month(percent of delayed flights is less).



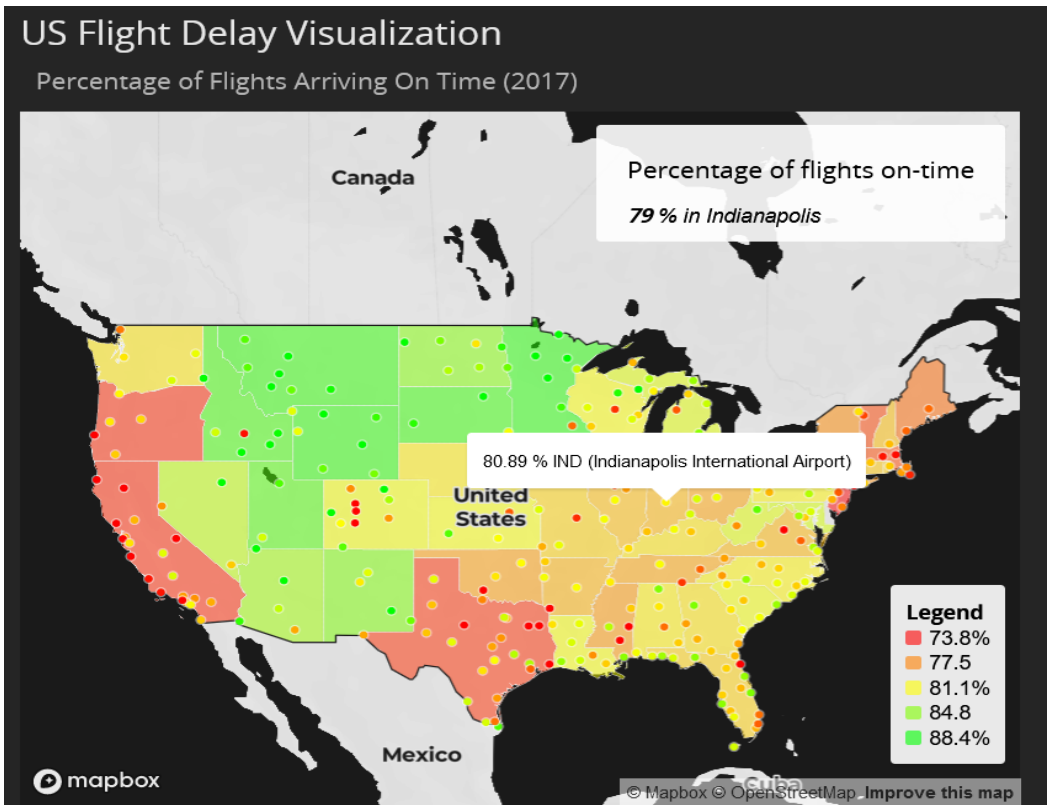
The above visualization of average Flight delay time of Origin airports the heat map representations shows a good visualization of which point has high average delay minutes where each point represents different airport. But if a user is interested to know the average delay of origin of one particular airport and the user will be unaware of the location of the airport until hovers the curser at every point and know the name of airport, then it would be difficult for the user to identify the point and know the average delay minutes of origin. So to eliminate the issue we would like to plot the visualization that makes the user to easily identify any required location without hovering the cursor.

Source of Visualization 2:



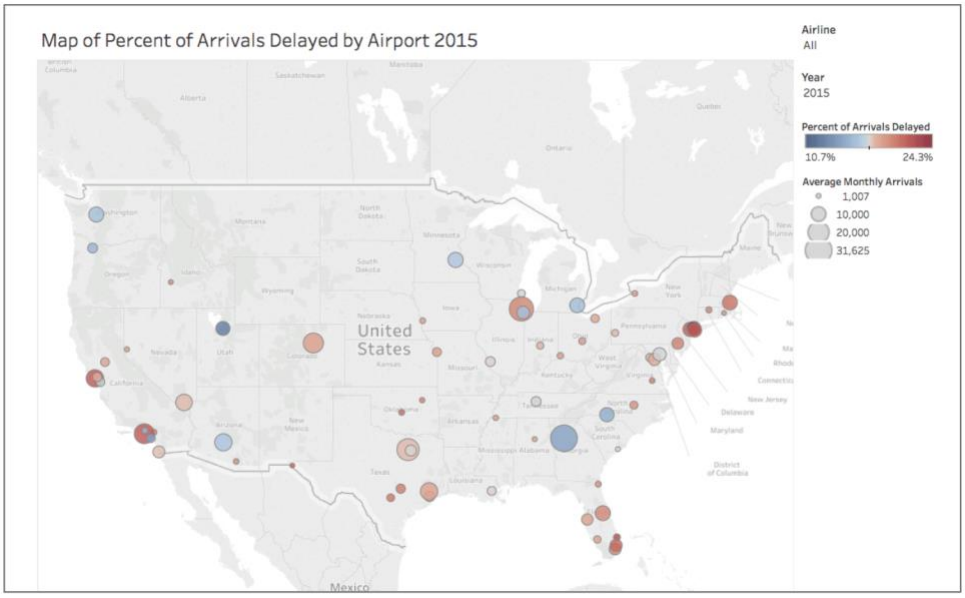
The above visualization shows the overall delay of different airlines and also the visualizations of delay due to different reasons (Airline issue, Weather, Security, Air System) of all airlines. The visualization is useful to easily interpret the analysis with respect to specific reason of delay but for overall delay it fails to represent the information correctly. To eliminate this issue we plan to plot a visualization which is easy to interpret the overall delay minutes and also the delay minutes due to specific reason.

Source of Visualization 3:



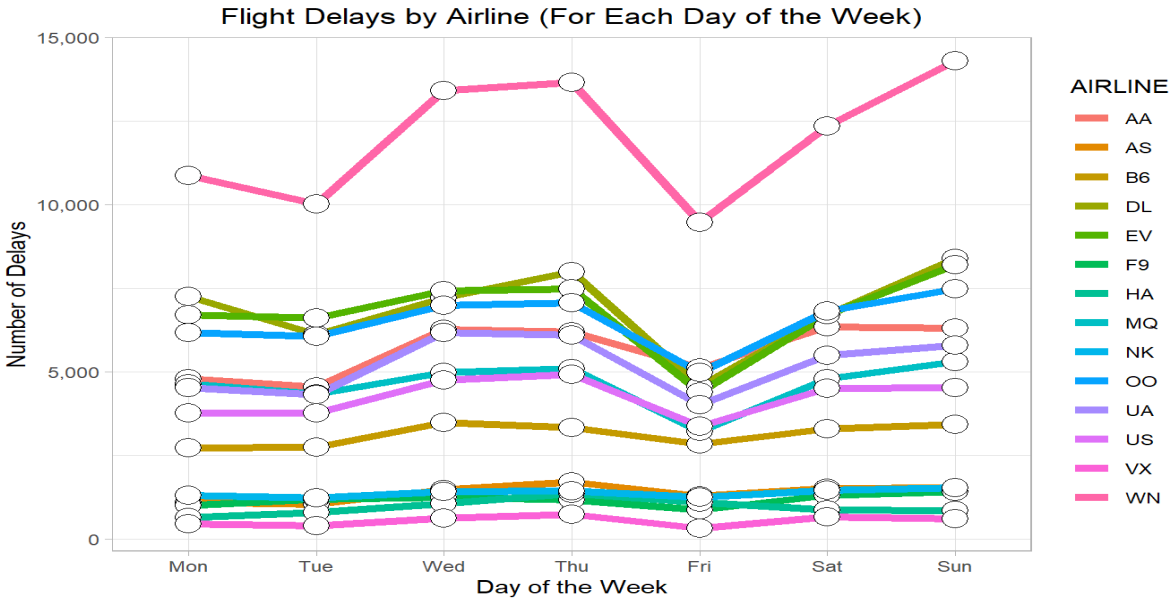
Here in this visualisation, the heat map represents helps in identifying the location with high delay of time in arriving but as we can see from above snapshot that by hovering over the airport we can see the percent of flights on time for that airport. But we can't compare among different airlines from a specific location i,e we wouldn't be able to know which airlines flights has high percentage of on time arrival and which doesn't. So we are planning to plot a visualization which gives delay of arrival information of different airlines along with overall percentage from specific location of airport rather than giving only the overall percentage of delay or percentage of on time arrivals.

Source of Visualization 4:



Here in this plot, the points are labelled with location names for easy identification but we can't identify the difference between the color of the circles which represents the percent of arrival flights which are delayed in that airport. For example consider the Louisiana, Tennessee and Illinois airports, all the three airports have almost similar color so that we can't differentiate these three airports for which airport has highest delay or lowest delay.

Source of Visualization 5:



In this plot we can see that there are more delays in the American airline(AA) when compared to the southwest airlines(WN), but we know that American airlines will operate more

flights when compared to southwest airlines, So obviously there will be more delayed flights in American airlines when compared to southwest airlines, But the probability that American airline will be effective is more when compared to the southwest airlines. Hence the percentage of flights delayed is good indicator than the number of flights delayed for the effective operation indication. Hence the above visualisation might misleading the viewers by interpreting that American airlines(AA) has many number of delays than all the other airlines.

Contribution:

In our project, we have modified the existing visualizations of flight delays and cancellations. The first visualization represents the average departure delay time and arrival delay time for different airlines, the existing version of this visualization represented the total delay time for different airlines which can mislead the viewers because as the total delay time contains all the delay times, the highest delay time occurred on a particular day contributes more to overall delay of the particular airline and this can lead to misinterpretation that the particular airline has more delay time and which may lead to serious loss in business of the particular airline. So, we have taken average departure delay time and average arrival time which will solve the issue discussed above. In the second visualization, which represents the average monthly delay times of different airlines, the existing version of this visualization represented the stacked bar graph where the months are represented by hue colors and the months are not distinguishable in this plot and it is hard to compare two airlines by month.

So, we have drawn a line plot which easily allows us to compare different airlines by month and as the months are not stacked here, it is easy to distinguish and see the delay time for different months of airlines. In our visualization, we have highlighted the airlines which have the maximum average delay time and minimum average delay time by representing them with dark bold colors. We can also select only particular airline by just clicking on it which makes it easier for viewers who are interested in only particular airline. The third visualization represents the fraction of delayed flights by size of the circle and average delay time by range of color for the airports of US state on a map, the existing version of this visualization represented the average delay time for airports of US state on a map and the color range used here is very limited and the size of the circles does not represent anything here. So, we choose a wide color range so that the circles are distinguishable and easy to interpret. We have also taken size of the circle to represent the fraction of delayed flights. The higher fraction of delayed flights is represented by larger circle and smaller fraction of delayed flights by small circle and if viewer hover to a particular circle, it gives data of delayed flights and average delay time in particular airport. The fourth visualization represents the average flights cancelled with respect to different airlines. We have used bar graph to represent the average flights cancelled with respect to different airlines in USA.

This visualization represents the number of delayed flights for different reasons for each airline in the US. This can help the user to plan when to arrive at the airport and choose the airline that better suits their interest and comfort. For example, if the security delay is high for the airline user want to select, he/she can plan at what time they want to arrive at the airport based on the security delay of the airline reasons for each airline in the US.it

Objectives:

After studying the existing visualizations and understanding positives and negative factors of the visualizations, the key issues we want to solve is data misinterpretation,

complexity of the existing visualizations by create visualisations in such a way that are less complex and not misleading the viewers. Viewer can analyse and identify the airline for which he/she can travel with less probability of delay or cancellation of flight.

Visualization 1:

We plot a standard bar plot visualization of average arrival and department delay representing using different bars(bins) and the x-axis represent the different airlines and y-axis representing average delay in minutes.

Visualization 2:

So we would like to plot a graph between the month and percentage delay in flights with the airline as a legend using scatter plot and the line plot. By seeing this plot we can get the information such as in a particular month which airline is good to choose (which has less percentage of delay in the chosen month). This visualization is easy to compare and come to conclusion on choosing a specific airline to travel.

Visualization 3:

The another visualisation is to identify the particular delay of flight of a particular airline at specific location. In this visualization we will be representing each airport with a point and corresponding label(name of city) using latitude and longitude coordinates of the airport and represent the delay in flights of every airline from specific airport and use the heatmap representation to indicate the overall delay at that airport.

Visualization 4:

For knowing the delay reason distribution for the cancellation of flights, we will be plotting the bar plot with number of flights cancelled by that reason vs airline, so one can understand the probability of each reason for cancellation easily.

Visualization 5:

The last visualisation is regarding the reason for the delay of flight and percentage of flights delayed due to that reason. And here we can also indicate the average delay time of the flight due to that reason So by seeing the data and the visualisations obtained it will be easier for the viewer to choose the airline in which the delay is low.

One can also easily match the conditions such as the weather, month, airline and obtain the delay in the flight and can estimate it, so that he/she can make convenient travel plan.

Analysis of Data:

We have obtained the dataset from Kaggle which is referred from US Bureau of Statistics website. It is a tabular dataset. The dataset contains three tabular forms (CSV files) named “airlines.csv”, “airports.csv”, and “flights.csv”. Nominal, quantitative and ordinal datatypes are present in the data set. The categorization of each variable in the dataset into different data types is as follows:

Column Name	Data Type	Description
Month	Ordinal	Month of flight scheduled

Year	Ordinal	Year of Flight Trip
Day	Ordinal	Day of flight scheduled
Day of Week	Ordinal	Day in a Week(Mon,Tue,Wed..)
AIRLINE	Nominal	Airline Identifier
FLIGHT_NUMBER	Nominal	Unique Number of the flight
TAIL_NUMBER	Nominal	Identifier of aircraft
ORIGIN_AIRPORT	Nominal	Starting Airport
DESTINATION_AIRPORT	Nominal	Destination Airport
SCHEDULED_DEPARTURE	Nominal	Planned time of departure
DEPARTURE_TIME	Quantitative	Difference of wheel off and taxi out
DEPARTURE_DELAY	Quantitative	Total delay on the departure
TAXI_OUT	Quantitative	The time duration between the departure from the origin airport gate and wheels off
WHEELS_OFF	Quantitative	The time point that the aircraft's wheels leave the ground
SCHEDULED_TIME	Quantitative	Planned time needed for the flight trip
ELAPSED_TIME	Quantitative	Sum of AIR_TIME,TAXI_IN,TAXI_OUT
DISTANCE	Quantitative	Distance between the two airports
AIR_TIME	Quantitative	The time duration between wheels_off and wheels_on time
WHEELS_ON	Quantitative	The time point that the aircraft's wheels touch on the ground
TAXI_IN	Quantitative	The time duration between wheels-on and gate arrival at the destination airport
SCHEDULED_ARRIVAL	Quantitative	Planned arrival time
ARRIVAL_TIME	Quantitative	Sum of WHEELS_ON,TAXI_IN
ARRIVAL_DELAY	Quantitative	Difference of ARRIVAL_TIME and SCHEDULED_ARRIVAL
DIVERTED	Quantitative	Aircraft landed on airport that out of schedule
CANCELLED	Quantitative	Flight Cancelled
CANCELLATION_REASON	Nominal	Reason for Cancellation of flight
AIR_SYSTEM_DELAY	Nominal	Delay caused by air system
SECURITY_DELAY	Nominal	Delay caused by security
AIRLINE_DELAY	Nominal	Delay caused by the airline
LATE_AIRCRAFT_DELAY	Nominal	Delay caused by aircraft
WEATHER_DELAY	Nominal	Delay caused by weather
STATE	Nominal	state code
IATA_CODE		Location identifier
AIRPORT	Nominal	Name of the airport
COUNTRY	Nominal	country’s name
LATITUDE	Quantitative	coordinate of the location
LONGITUDE	Quantitative	coordinate of the location

Candidate Visualization methods:

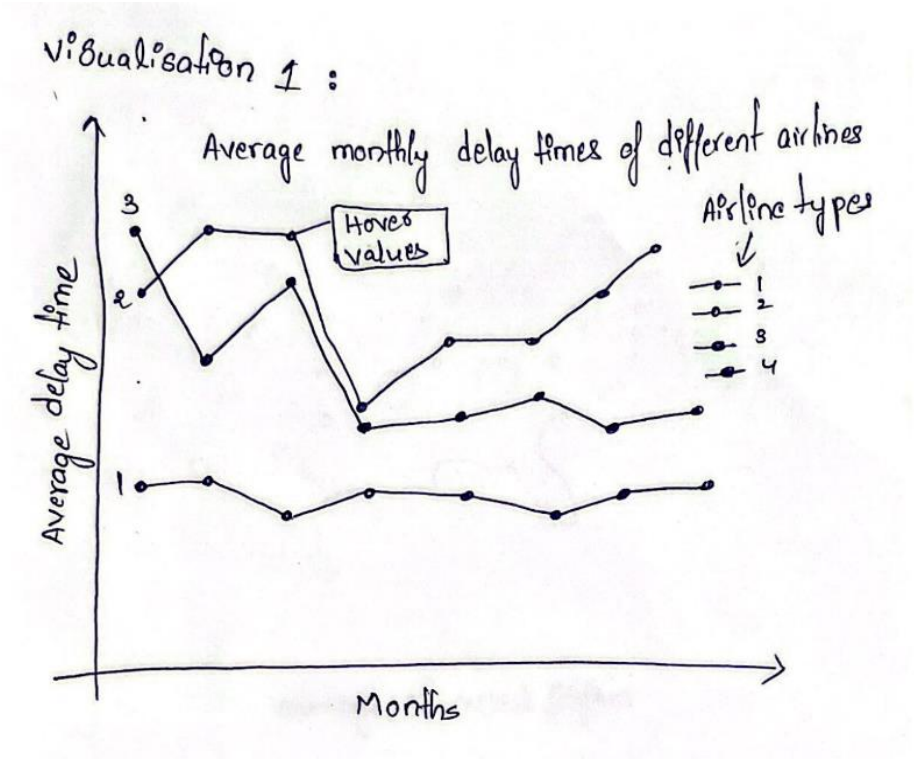
We apply scatter plot, line graph, bar graph, and heatmap visualization methods to create the visualizations. We use bar graph to create a visualization of total departure delays and arrival delays with respect to airlines as bar graph conveys relational information quickly. We want to use the scatter plot and line graph to create a visualization of percentage of delays with respect to each month as scatter plot can represent the data point and line graph can be used to connect the data points which makes it easy to compare percentage of delays in different airlines. We want to use the heat map to create a visualization of delay of flights of a specific airline with respect to a particular airport as using the heat map makes it easy to visualize in which airport most delays occur through the colour intensity. We want to use the bar graph to visualize flight cancellation by reason as bar graph is simple and understandable to represent different reasons and percentage of flights. We want to use the bar graph to visualize flight delays with respect to different reasons as bar graph is easy to understand and analyse.

As mentioned, there are three csv files and flights.csv has 31 columns in total of different variables mentioned above in the table, airports.csv has 7 columns which are variables used to locate the airport and airlines.csv has 2 columns to define the short code for all the airlines available in the dataset.

For visualizing the percentage of delays with respect to each month we need columns like month, airline identifier, departure_delay, arrival_delay which are present in the flights.csv and airlines.csv dataset. For visualizing the delay of flights of a specific airline with respect to a particular airport, we need columns like departure_delay, arrival_delay, airline, destination_airport which are present in flights.csv,airports.csv. For visualizing flight cancellation by reason we need columns like cancelled reasons which are present in flights.csv. For visualizing flight delays by reason we need columns like air_system_delay, security_delay, airline_delay, late_aircraft_delay, weather_delay which are present in flights.csv.

Sketches and prototypes: : Below are the rough sketches of the visualisations that we want to plot in this project

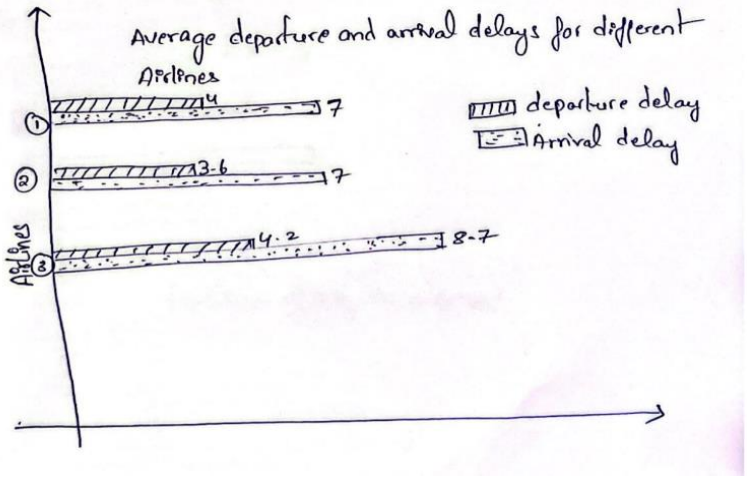
1)



We want plot the above visualisation were each line graph represents the delay of the specific airline in a particular month. Here we can also compare the delays of the airlines for every month and we want to make it more interactive by adding the hover values.

The line plot will be better for this visualisation because for every month we can easy compare the delay time between the airlines which helps to choose the airline and the cons of this plot is the plot will become very messy when there are more airlines. To overcome this we used interactive techniques such as top 4 and last 4 delay time airlines will be bold in color and remaining are pale. Even when we hover over the plot we will get the airline name and delay time . We can also select the airline so that the particular airline will have a bold line and rest will be pale.

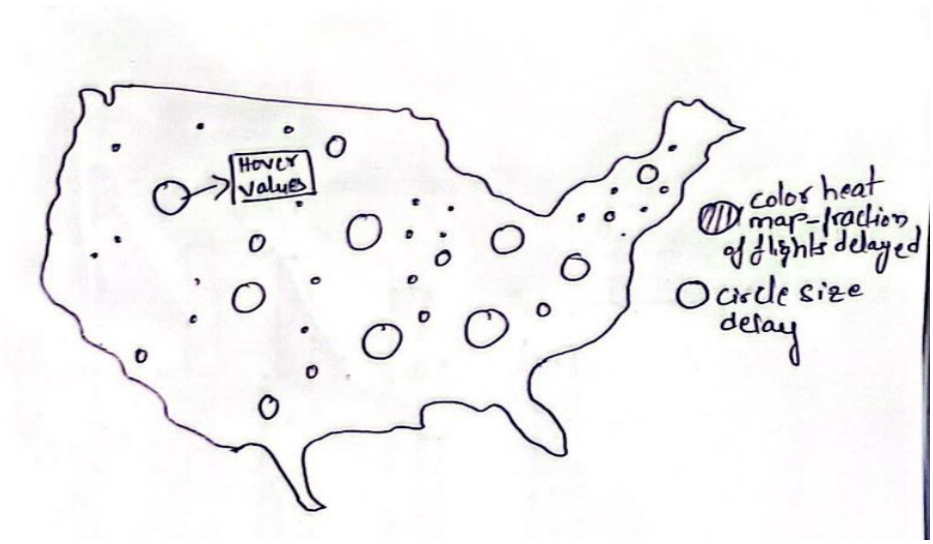
2)



The next plot we want to make is a bar plot which indicates the average and arrival delays for different airlines and the plot will also have a number on the edge of the bar so that it can be accurately represented.

The bar plot will be more useful when we are comparing two different classes. As we are comparing the average departure and arrival delay it will be the better choice of drawing this plot.

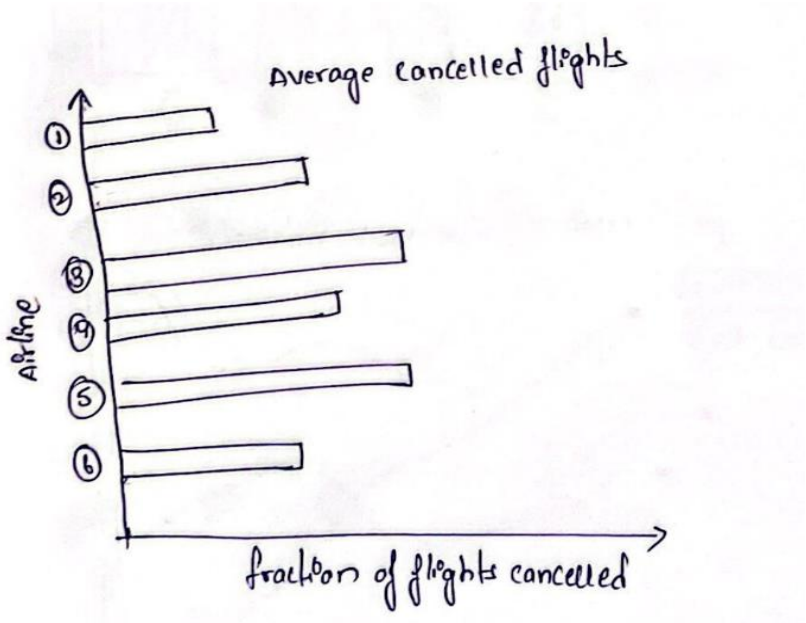
3)



Here in this visualisation we want to make a plot in such a way that each point represents a airport and size of the point indicates delay and the color represents the fraction of flights delayed. Even we want to add the hover values for each point.

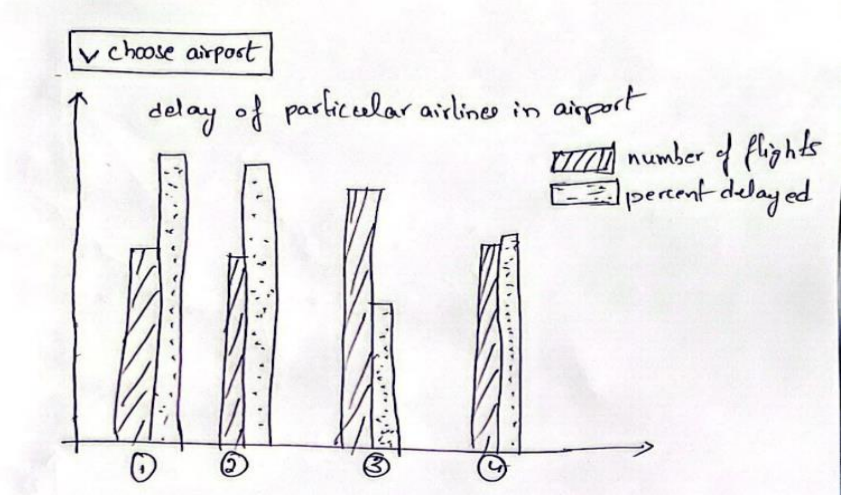
So here we used different techniques in a single plot i.e heat map and the scatter plot with size of point matching the density. As we cannot differentiate area easily i.e. stevens power law, we added a hover value for this to avoid the misconception. We thought of using the density equalising cartogram for this but this cartogram will entirely distort the graph and may be misleaded.

4)



Now we want to plot a bar chart which will gives the information about the average number of flights cancelled for a particular airline. Here also it is better to use the bar plot because we are comparing the values between the airlines.

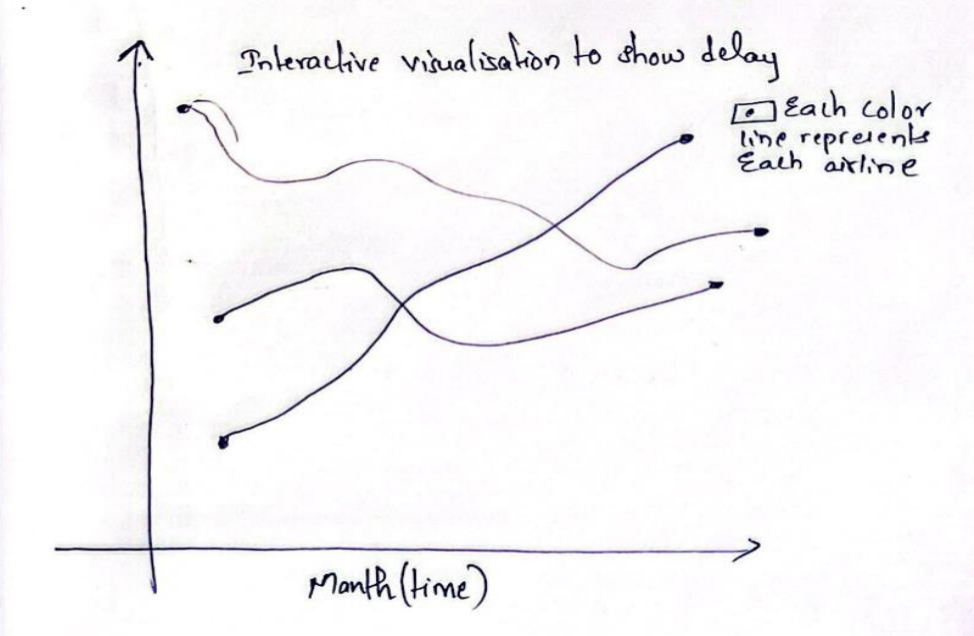
5)



This is a interactive visualisation , if we give the input airport then it will plot a bar plot which gives the number of flights of an airline are there in the airport and the percent delayed will also be plotted.

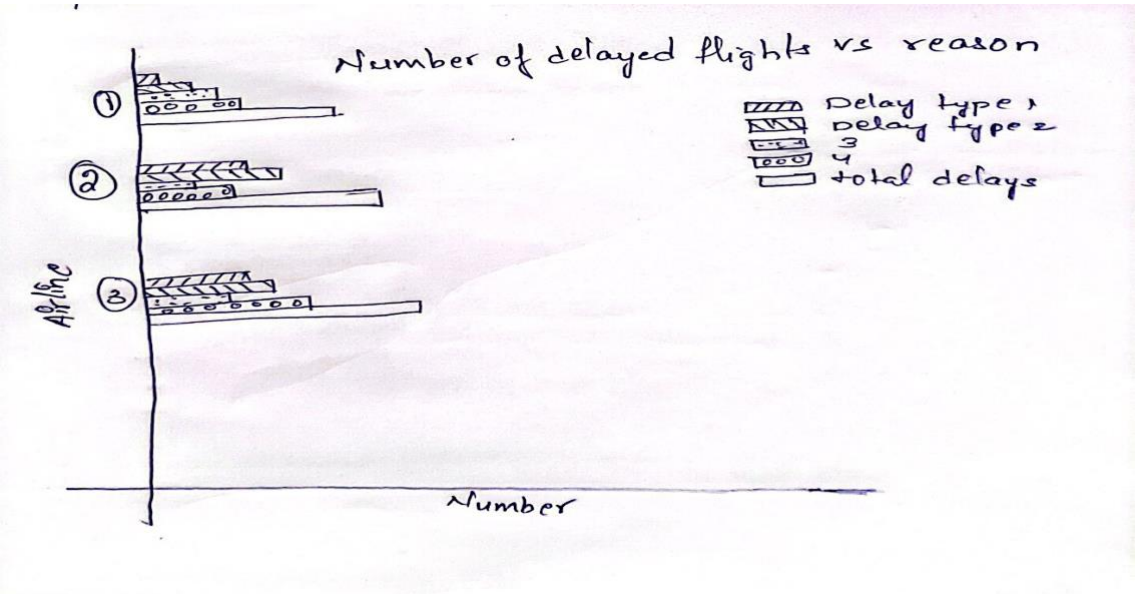
Here we can also use the pie chart for this visualisation but if we use the pie chart it will hard to compare the delay of the airlines because there is only slight difference between the delay of each airline.

6)



This is also a interactive visualisation where there will a point for each airline and this will move along a line which shows how the delay varies with respect to date.

7)



The above visualization is a comparison of different airlines delay reasons w.r.t overall delayed fines of those airlines. This visualization is helpful in choosing an airline for travel plan based on above data of delay reason. For example if there is more delays of particular airline flights due to weather conditions, one can avoid choosing such airline to travel during bad weather conditions.

Failed Experiments :

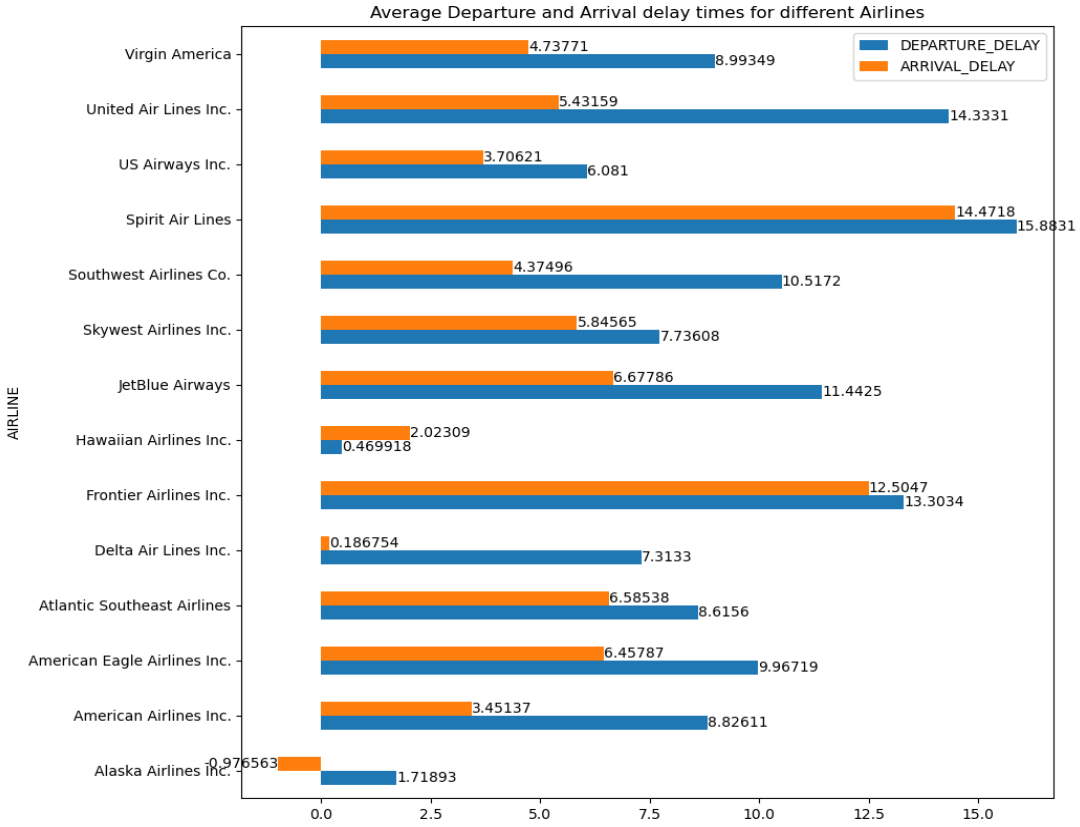
We tried to plot the interactive visualisation as in such a way that we will get a drop down of the airports, so that we have to choose the airport(Shown in visualisation 5 above). Then we

will get a bar chart that gives the delay of the each flight of particular airline. This one failed because there will be lots of airports(328) so it is very difficult to search for the airport in that drop down list.

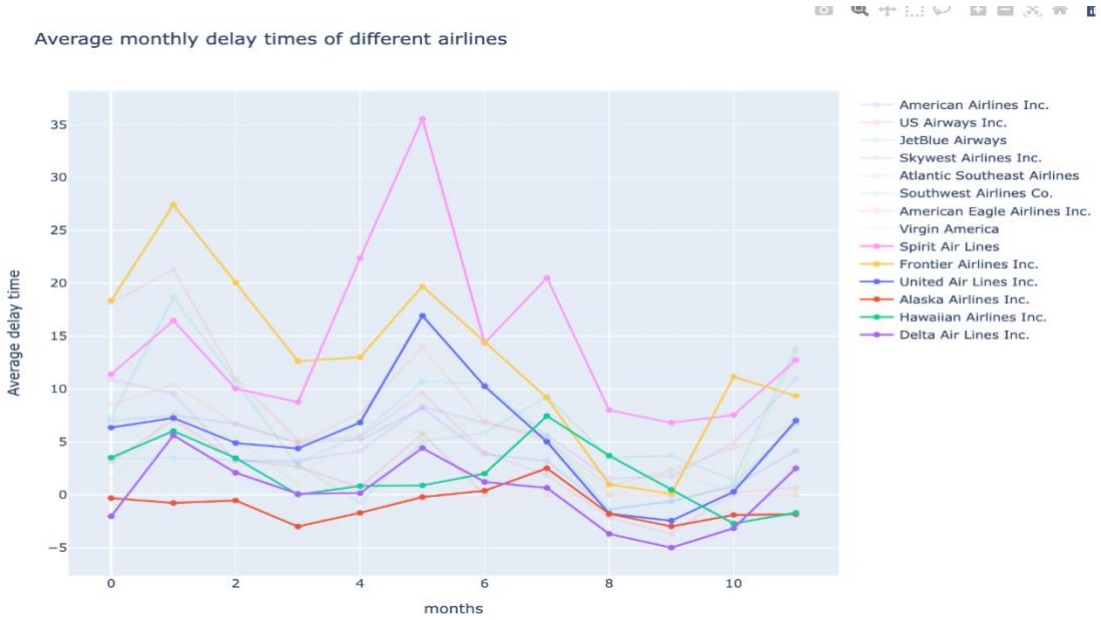
Tried to make an interactive visualisation where the delay in the airport is visualised. Here we tried plot the delay vs time graph in which the node moves as the delay changes for every day(Shown in visualisation (6) above). We can give the input speed of how fast the graph varies. We tried to plot this graph but we used our full efforts but not able to draw it. We can't figure it how to show the visualisation for all airports and for all airlines.

Results and insights :

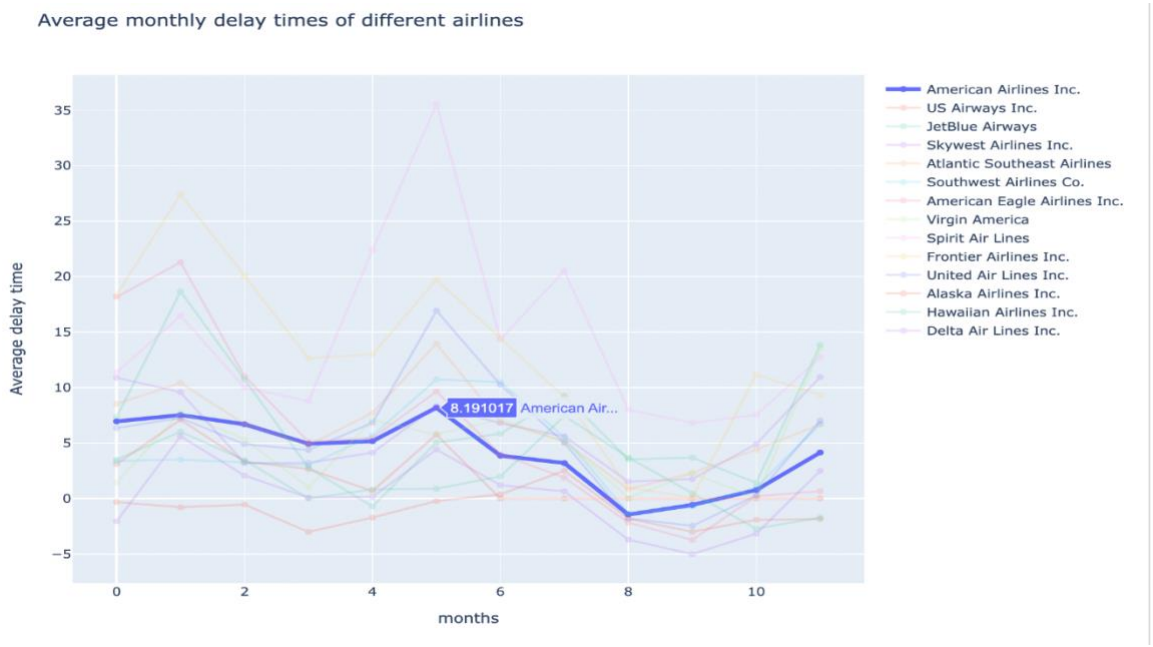
Visualization 1:



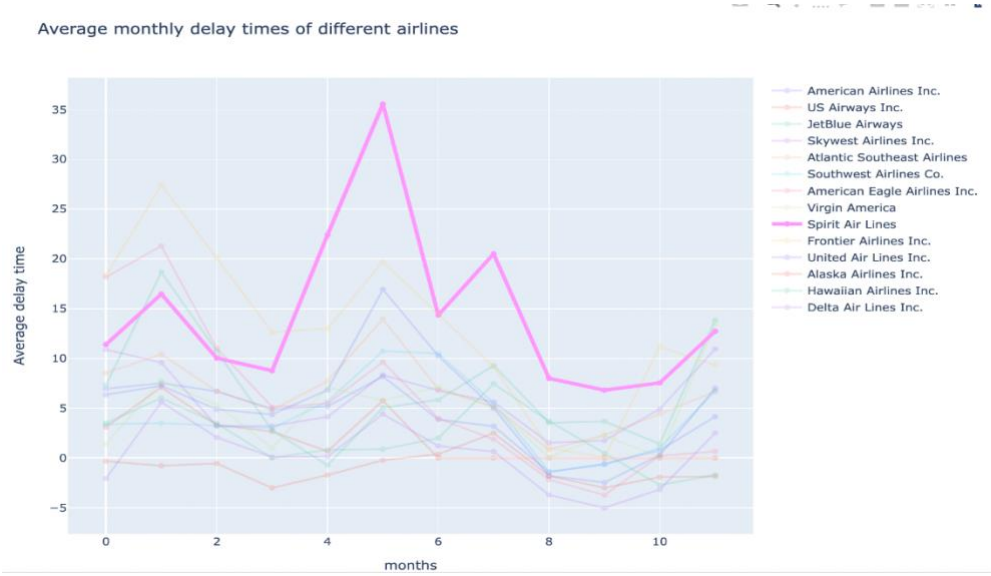
Visualization 2:



This is the first visualisation we got the line graph where top three and last three airlines are bold and rest are pale.



When we hover on to each line, a value will be displayed i.e. average delay time and the airline name in which the line plot belongs. We can also select the airline from the list on the right so that it will be highlighted.

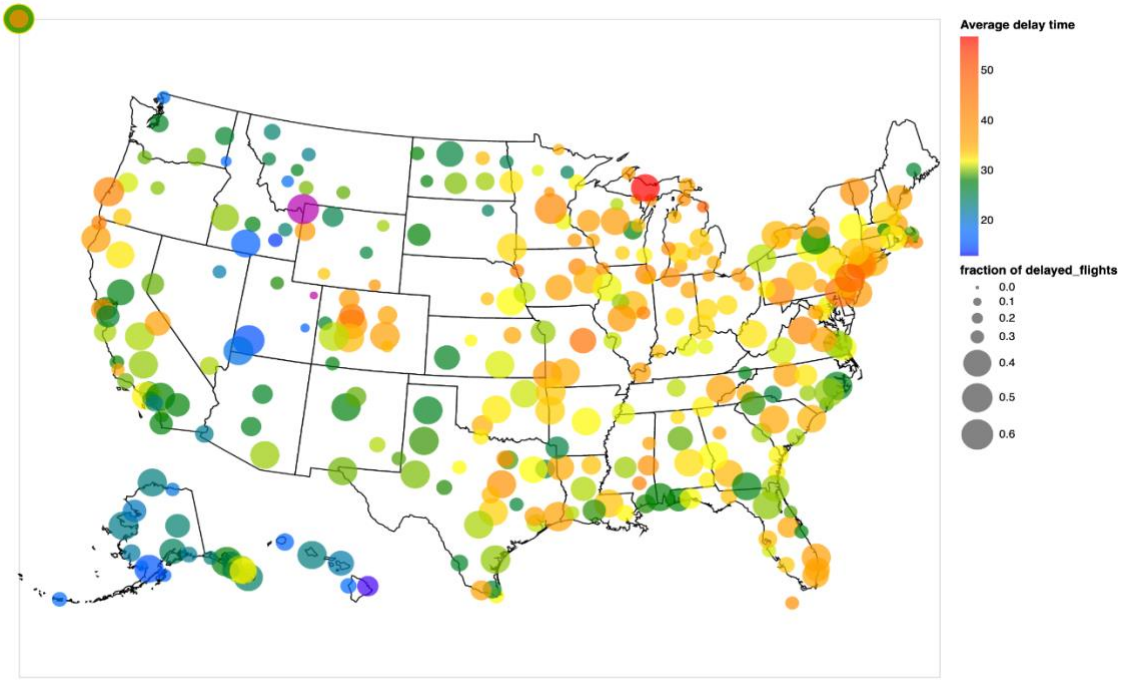


We can also select the line plot that we want to analyse so that the line will be highlighted and we can see which line is selected on the right

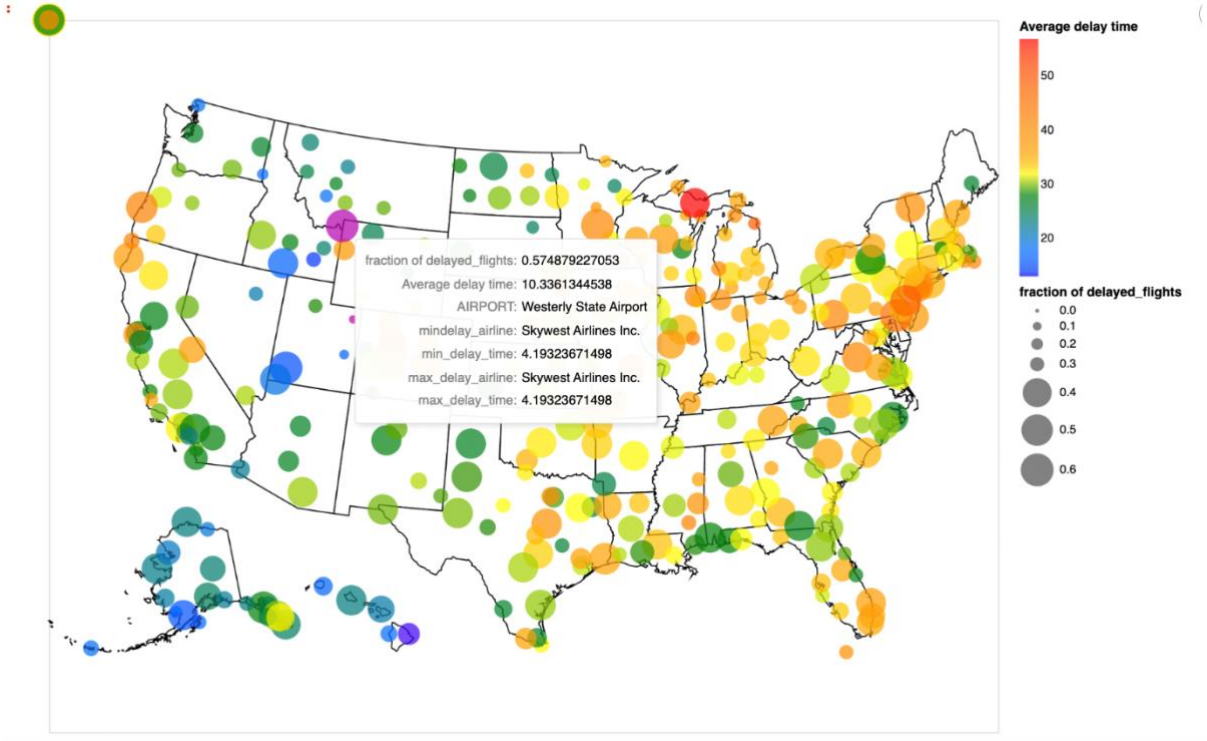
So this visualisation is very useful for comparing the average delay time between the airlines in a particular month.

The above is the bar plot of the average delay times of each airline. Here we also plotted the number at the edge of the bar graph so that we can know easily which airline is delayed most and how much it is delayed.

3)

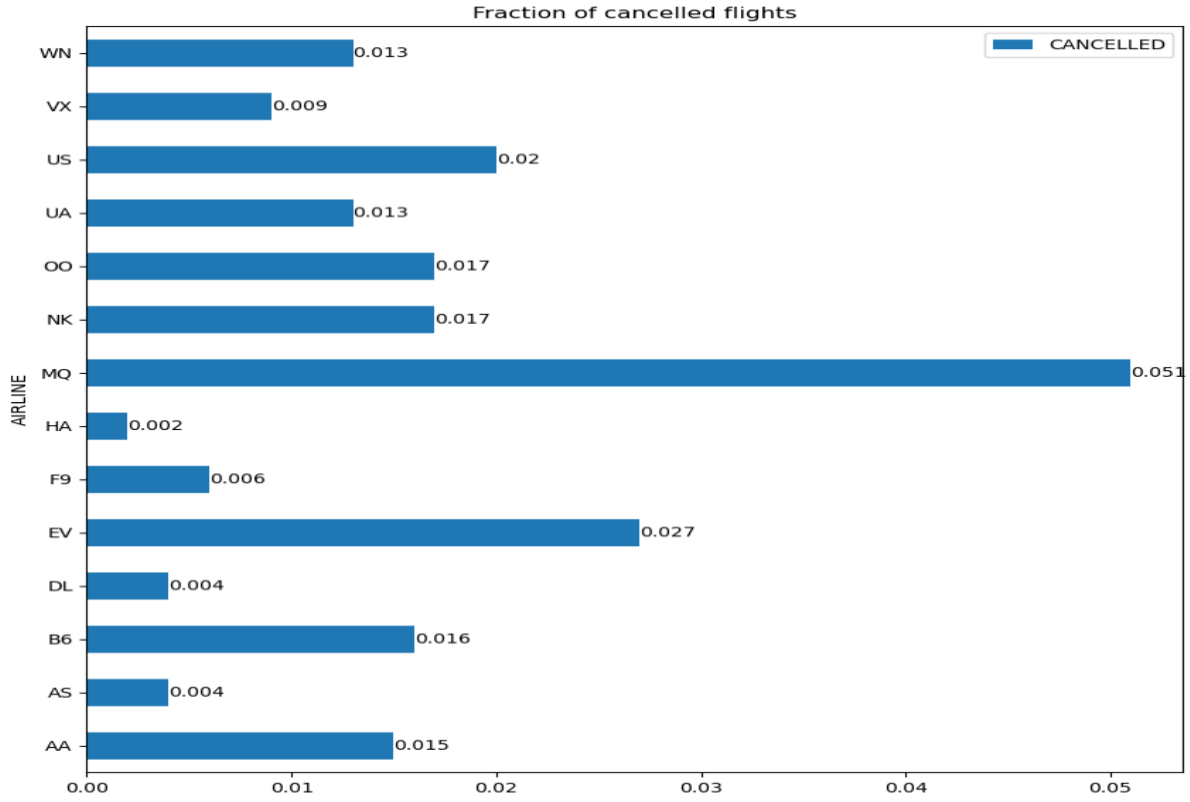


Here we plotted the positions of the airports on the map and used heat map and size representation along with it. Here each mark represents the location of each airport on the USA map and the size of each mark represents the fraction of flights delayed at that airport and the color(heat map) representation shows the average delay time of all the delayed flights at tthat particular airport. So if the color is read in RGB series it represents high average delay time and blue color represents less average delay time.



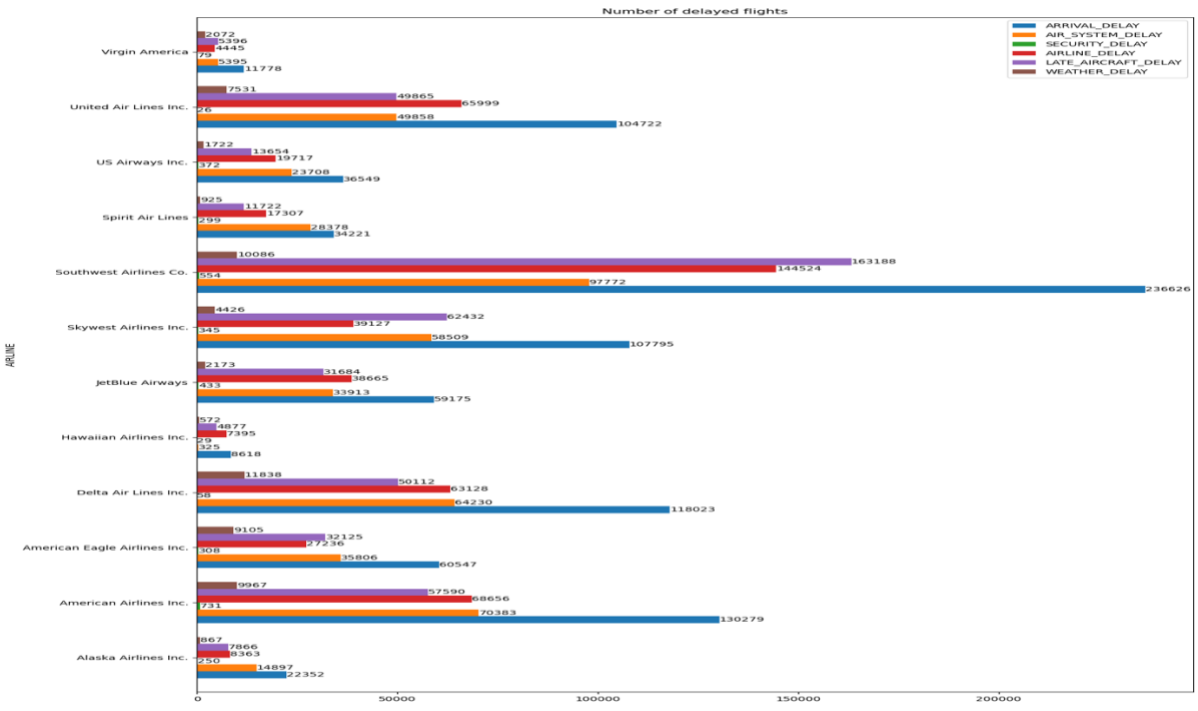
Along with size and heat map representation of fraction of delayed flights and average delay time of airports, when we hover over any mark, the details of that aiport like average delay time of flights, fraction of delayed flights, minimum delay airline and its average delay time and finally maximum delayed airline at that airport and its average delay time at that airport will be displayed.

Visualization 4:



The above bar plot above is a plot of fraction of flights cancelled by different airlines.This bar plot is useful in comparing the different airlines cancellation rates.

Visualization 5:



The above bar plots represents the delay reason for different airlines delayed flights. The data contains five delay reasons for delay in the flights which are delay due to security reasons, weather conditions, air system issue, airline issue, aircraft problem. And each bar has its number of delayed flights value displayed on its edge to make it easy for the visualizer to know the number of delayed flights due to the particular reason and also compare between different airlines.

Discussion and Conclusion :

From the above visualizations we learned how efficient different airlines are operating there flights at different locations of USA. From the visualization 1 we can know which airline is better to travel to reach the destination in time. Visualization 2 tells the efficiency of each airline in different months of the year i.e average delay time of each airline in each month. So from that visualization we can choose a specific airline that have minimum delay time during our month of travel. And Visualization 5 is the bar plots of delay reason and the delay time of different airlines which also helps in making a choice of airline based on safety or urgency of travel plan.

The visualization 3 tells the airport specific average delay time, fraction of delayed flights and minimum delay time airline and maximum delay time airline. From this visualization we can know how the airways perform at different locations of USA from different airports and also tells the maximum delayed airline and minimum delayed airline at the specific airport to make a better choice of travel.

Finally visualization 4 is a bar plot of fraction of cancelled flights of different airlines, which also shows the efficiency of different airlines.

Limitations and Future work :

The limitations of all the above visualizations are they give the analysis based on only 2015 year data of flights which would be biased if there was any airline issue in that specific year and also there might be new airlines present whose information is not provided in any of the above visualizations. The visualization 3 has the limitation of color which is difficult to compare if the contrast of color is less between different marks and the location of the city or airport should be known to check the details of any airport.

These project can be future extended by giving a generalized comparision of flights data by taking data of about atleast 5 latest year. And also visualizations can be made further interactive for example when any airport is selected in the visualization 3, a bar plot of different airlines efficiency at that airport can be given which would be more easier for comparison of different airlines delays at a particular airport.

References:

Primary Source of dataset:

- https://www.transtats.bts.gov/homedrillchart.asp

Dataset:

Author: GAOFENG HUANG

Year: 2019

Link: https://www.kaggle.com/code/together/visualization-flight-delays/data

Background Study or existing Visualizations:

- Author: GAOFENG HUANG
  Year: 2019
  Link: https://www.kaggle.com/code/together/visualization-flight-delays/notebook
- Author: Siva Ranjani Prabasankar
  Year: April -2020
  Link: https://www.linkedin.com/pulse/tableau-data-visualization-flight-delays-us-sivaranjani-prabasankar
- Author: lrakai
  Year: Feb-2018
  Link: https://lrakai.github.io/us-flight-delays/
- Authors: Diane Lee, Paul Roberts, Samuel Shen, Kyle Witt
  Link: https://kylewitt.com/projects/dawdle/Dawdle-Full-Report.pdf
- Author: Cal Henderson
  Year: October-2021
  Link: https://rpubs.com/cal_henderson/829635
- https://stackoverflow.com/questions/53327572/how-do-i-highlight-an-entire-trace-upon-hover-in-plotly-for-python
- https://plotly.com/python/hover-text-and-formatting/