Declaration on Plagiarism

This form must be filled in and completed by the student submitting an assignment

Name/s:	Tanmay Potbhare Atharva Joshi
Student Number/s:	Tanmay – 21262012 Atharva - 20211526
Programme:	MCM (MSc in Computing)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	26th November, 2021
Module Coordinator:	Dr Suzanne Little

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Tanmay Potbhare **Date:** 26th November, 2021

Name: Atharva Joshi **Date:** 26th November, 2021

ASSIGNMENT REPORT

THE AIRLINE CARRIERS OF UNITED STATES AND THEIR DELAY AND CANCELLATIONS

Dublin City University

ABSTRACT:

This is an analysis and visualization of the United States domestic flights from the year 2009 to 2018. There are a lot of airlines which help people in transport for ease and comfort and mostly speedily delivery from one point to another. Thus, in this project the airlines which is focused on is the United States domestic carriers which are there to transport people to and fro from one destination to another. Airlines plays a very important role as it goes through various risk factors and have to deal with many natural issues while travelling.

In this project we focus on the domestic carriers of United States which conclude total of twenty carriers or simply airlines and why they get delayed and cancelled and what are the reasons of this. On the basis of this, which is the risk factors and reasons of being cancelled which will be explained later we try and analyse the best airlines with a smaller number of delays and cancellations and try and visualise the same.

DATA COLLECTION:

In this, there are ten datasets in CSV formats. So, basically the ten datasets which are present they have the data of the United States domestic carriers dated from 2009 to 2018. These ten datasets contain 28 columns and each data set of at least 750 MB. Each dataset has information about the carriers, delay and cancelation and moreover they contain very specific details about reasons of cancellation of each airline. There are cancellation codes given in the dataset which give the reason of why specific airline was cancelled and it is a numerical data where if the cancelled code is 1 for the specific reason, then it is cancelled because of that specific reason.

These CSV files are created by Yuan Yu 'Wendy' Mu, three years ago and was updated two years ago, who was a student of University of South California, Los Angeles. They are available on a platform called Kaggle, linked below. There is no mention about where this data is compiled from.

Dataset Link: https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018?select=2009.csv

The dataset can be considered as Big Data set as it fulfils one of the three V's of Big Data.

Volume: The available dataset takes a total of 8GB of storage which incudes, text, numbers and date series, time. In this assignment, we have no used all 28 columns but the columns which are used are: The data series, time series which is calculated in minutes and all the numerical data. Then there is textual data which contains the flight carrier's information about the departure arrival and the names of carriers. This is dated from 2009 to 2018.

DATA EXPLORATION, PROCESSING, CLEANING, AND/OR INTEGRATION:

DATA CLEANING AND PRE-PROCESSING:

The Exploration of dataset is done by creating data frame and then working on the exploratory operations. We used Pandas, Plotly, matplotlib, and seaborn to manipulate, clean and integrate the data frames and convert given data into considerable sensible visualizations.

In the earlier stages we had to see how can we integrate all the years data and use needed columns for visualizations. We then had to come up with a data frame concatenated_df() which focused on integrating the only mentioned columns from all the year datasets. It was a huge dataset as well as a huge task to take all 28 columns and process and mould the data as per requirement, so, groupby() function was a must.

There is a column named OP_CARRIER which is one of the main data as it contains the names of the Airline Carriers. So, originally, they were just defined as initials for example if AS is mentioned we had to replace it with Alaska Airlines using the data frame object.
Source: https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States

VISUALIZATION:

We have created graphs or visualizations for Delay and Cancellation of Airlines. Then there is one interactive chart for the reason of cancelled flights depending on the cancellation code.
As mentioned in the above lines we wanted to show a short or brief picture of the delay of flights in terms of its arrival and departure. We created two charts for the year 2017 and 2018 for arrival delay of flights and departure delay of flights.
So, for Arrival delay and Departure delay we created two data frames named Arr_Delay_df and Dep_Delay_df respectively. Then we used a groupby() function to group delay attributes with unique valu which is Fl_Date that is flight date. We used a line chart so we needed continuous data. We also used a range slider which is a part of plotly library to display the
delays of arrival and departure.

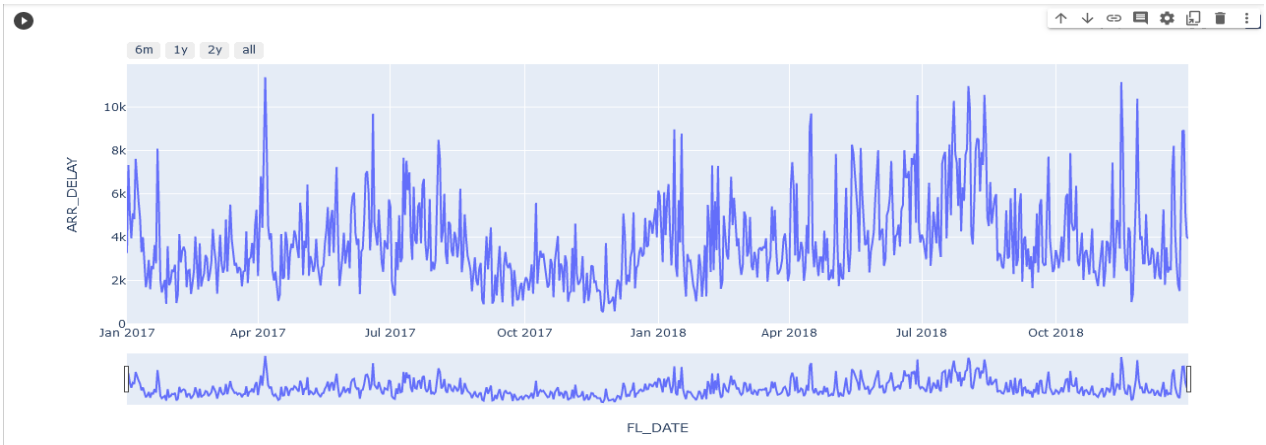

Fig 1: Arrival Delay of Airlines

As the Arrival delay same is with the departure delay a line chart and a range slider which gives a clear picture if the range on the slider is moved from left to right or vice a versa. On the X-axis is shown the years and on the Y-axis the departure delay.

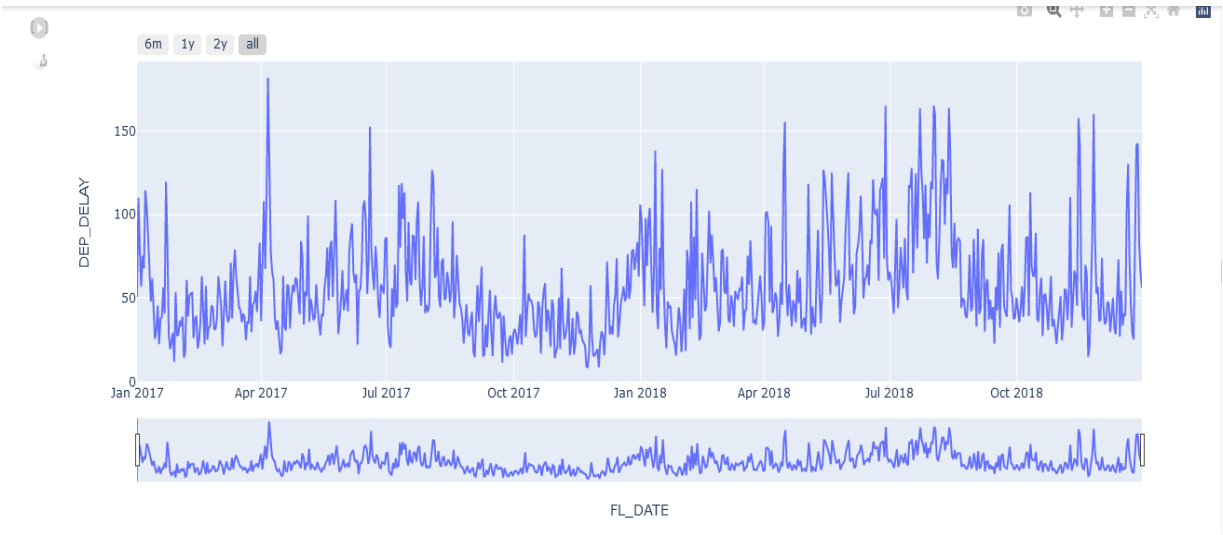

Fig 2: Departure Delay

Then we planned to show the Cancellation of flights. The choice of this dataset was a bar graph as it was the most suitable option compare to scatter plots or line graphs or other available charts. This is done by using plotly library and pandas library as well. We chose the bar graph as it is only option which can explain cancellation for all the years for 2009 to 2018 in a comparatively good fashioned way. As this shows the carriers on the X-axis with Year on the Y-axis and colours are given for the airlines as well. So, each bar will show the cancellation of the airline for all the years.

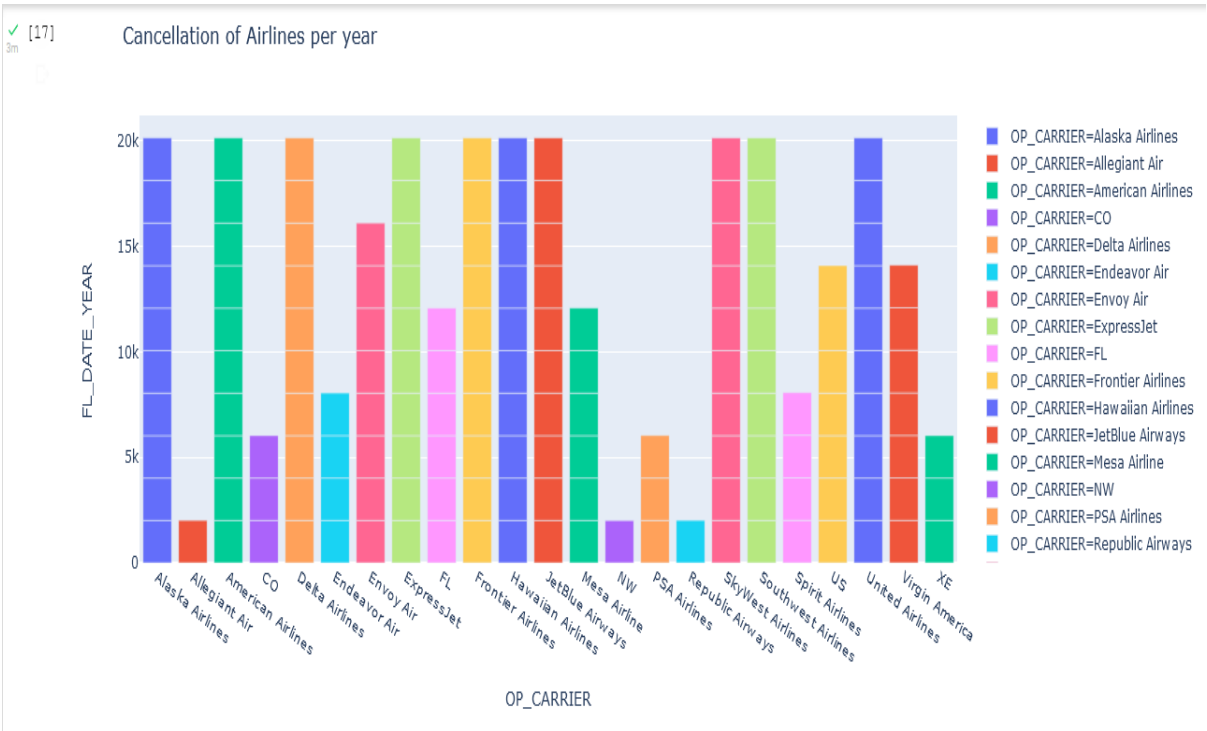

Fig 3: Cancellation of Flights according to years (2009 to 2018)

We chose to show the reasons behind the cancellation of airlines using one attribute of dataset which is Cancellation_Code. The airlines are then categorised in different categories where they will be defined as per their codes.

So, There are three categories A - Airline Cancellation, B - NAS (Air) Cancellation, C – Security Cancellation, D - Weather Cancellation.
When the airlines are matched with one of these codes it is understood that why was the reason for that particular airline to get cancelled on that date or year.

This chart is also a Bar graph which uses drop down menu to select the airlines and then it shows the cancellation count according to the canellation code or cancellation reasons.

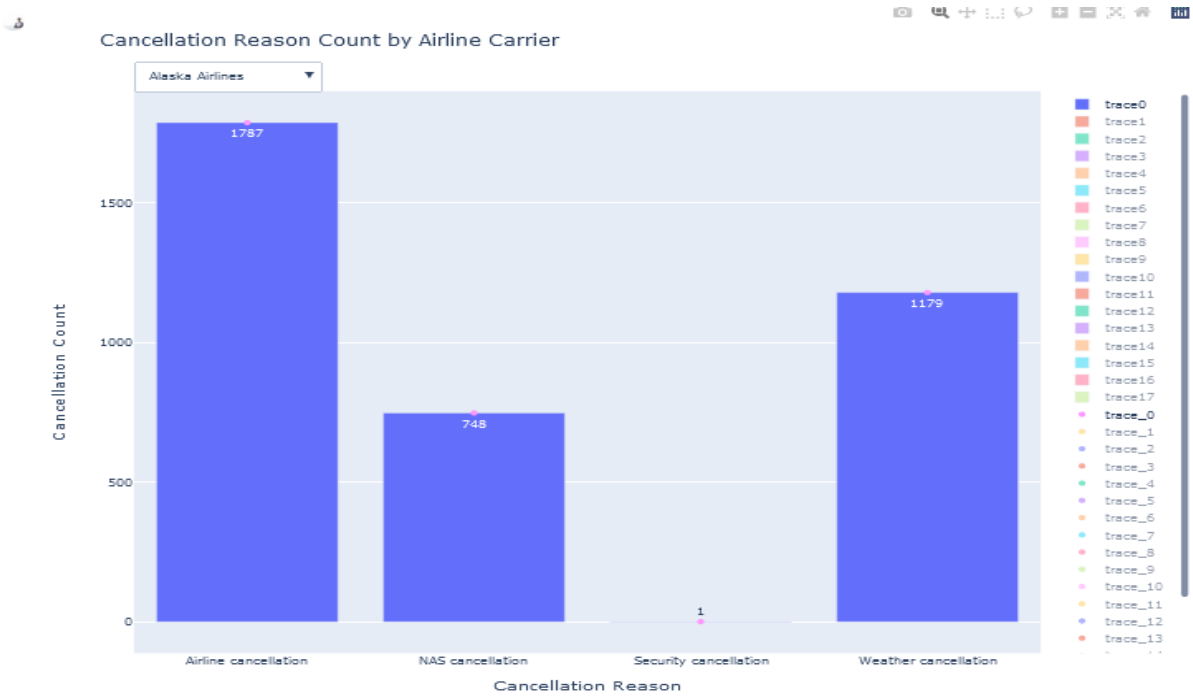

Fig 4: Reason of Cancellation and it's Count of Airline Carrier

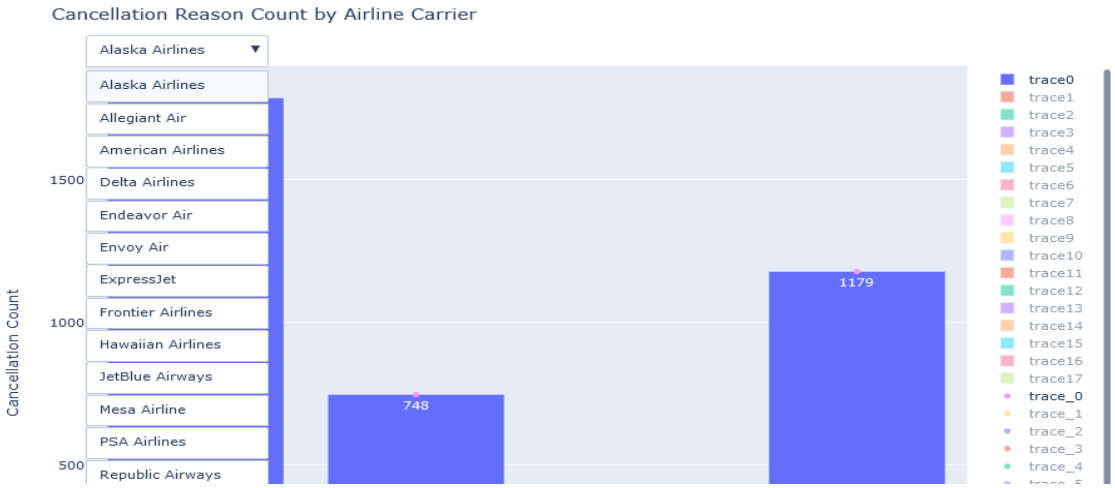

Fig 5: Cancellation reason with Drop Down of Airlines

CONCLUSION

In this project, we learned to use the Plotly library, the Seaborn library and Pandas library. As we have never worked on these libraries on my previous projects. Numpy library was the one we wanted to use but did not get the chance of using it. In this visualizations, all four presented are interactive and thus, with the help of plotly and pandas library we were able to get through the interactive bar graphs and the interactive line graphs with range sliders. By using the data to visualize bar graphs and line graphs we can conclude that the bar graphs and line graphs have portrayed a better way of showing the airline delays and cancellation, but yes not the best.

The line graphs which are created for the arrival and departure delays of the airlines can be improved and can be used different type of chart to display the same information more effectively. The main feature of bar graphs for cancellation reason was a drop down as it was a huge data and successfully showing each and airlines mapped with their reasons was difficult. The user will have to select which flights cancellation reason he wants to see and then proceed selecting the airline.

After, successfully creating and visualizing the graphs for the first time, we can surely conclude that we did learn the better side of python using the libraries and got a hands on experience about data cleaning, exploring, pre-processing an management, and data visualization.

REFERENCES

1. Link to dataset: https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018?select=2009.csv

2. Link to list of United States Airline: https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States

3. Link to plotly Documentation: https://plotly.com/python/

4. For creation of Line graphs: https://plotly.com/python/line-charts/

5. For creation of Bar graphs: https://plotly.com/python/bar-charts/

6. Link to Seaborn Documentation: https://seaborn.pydata.org/