**Exploratory Data Analysis**

**By Lavanya Telapudi**

**Introduction:**

This project includes 9 datasets with which we needed to work upon and carry out detailed analysis of each dataset. The datasets includes the airlines, planes, flights and weather information from which we need to see deep and understand the pattern of data and carve out conclusions like cancellation rates, delay rates and much more. Multiple graphs have been used to depict the pattern of data so that we can comprehend the trend with the graph and understand the intensity of data. Here we have used Jupyter Notebook to carry out EDA and the code. The methodology used in our EDA process is we first read all the files provided and then understood the size of the data and the columns present. Using .columns(), .size, valuecounts() and other methods we understood on which dataset we need to work upon for a particular question.

The datasets are large, with thousands of rows and columns, and we can't manually apply filters  in Excel to detect the pattern and check various observations or outcomes, therefore we need to use the Python programming language to make the process easier, more efficient, and faster. Python includes various built-in tools and modules for persforming analysis, and it is one of the easiest languages to learn. The graphs, heatmaps, and other python features make it extremely interesting and dynamic. Python is useful for solving linear and logistic equations, as well as extracting the precision and accuracy of mathematical functions. We will also show how we analyzed the data and what our conclusions are.

The present section of this study is considered to illustrate the utilization of flight 10 flight dataset for the analytical dataset. Considering the above aspect, the dataset consists of data about arrival and departure time.

**1.a: Calculate the total number of seats for all the planned flights for each destination separately?**

Performing Inner Join with flights and planes data-set to calculate the total seats for the planned flights w.r.t each destination and find total number of seats with each destination.

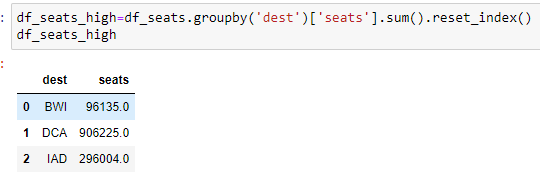


Fig 1.1 : total number of seats for all the planned flights

**#b.What is the day of the year with the highest number of flights?**

Calculating the highest number of flight as per date and sorting according to flight and Date column with highest to lowest

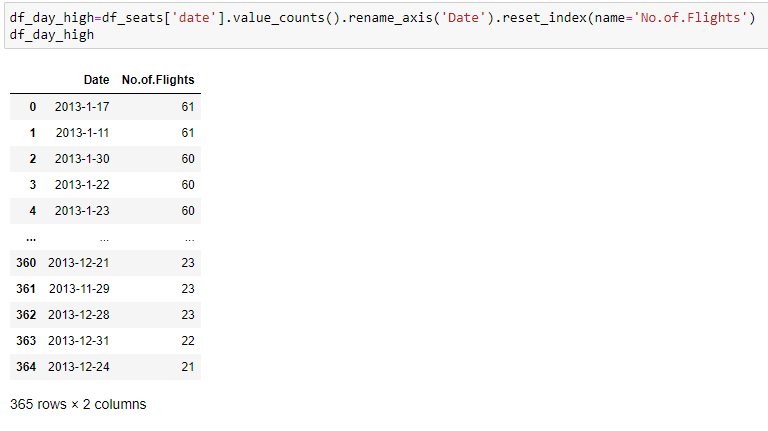


Fig 1.2: day of the year with the highest number of flights

**#c. What is the day of the year with the highest number of seats available on that day?**

Calculating the highest number of seats as per date. 28th February, in 2013 has the highest number of seats with count of 5379 seats

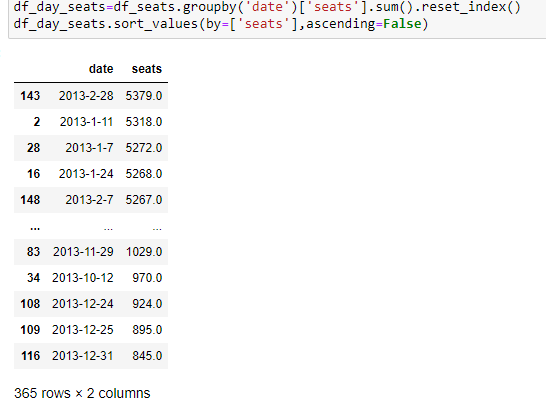


Fig 1.3 :day of the year with the highest number of seats

**2a. What day of the year most cancellations happened?**

The condition for cancelled flights is that the departure time and arrival time both is empty. This is for the reason that, if the departure time is present, then the flight flew. The same with arrival time, the flight has landed.

6th March, in 2013 has the highest number of flights cancelled with count of 46 flights˙

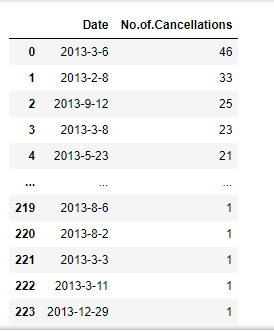
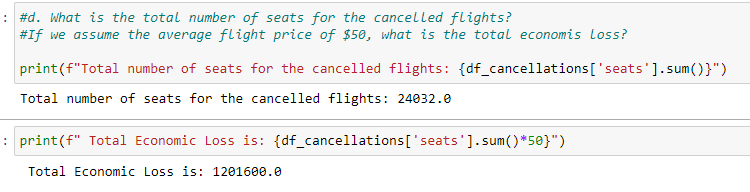


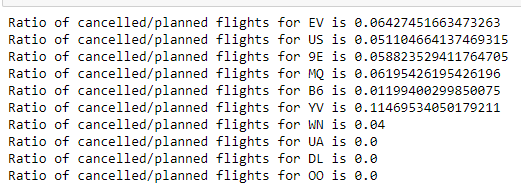
Fig 2.1 : most cancellations as per date

**#d. What is the total number of seats for the cancelled flights? If we assume the average flight price of $50, what is the total economis loss?**



**#e. Determine the ratio of cancelled flights/planned flights for each airline company, list it, and determine the most and least reliable airline company (most reliable = the one that has the smallest ratio of cancelled/planned)**

Here the most reliable airline company is UA,DL,OO as they have 0 cancellations and the least reliable airline company is YV as it has the cancellation rate of 11.4%



**3a. Calculate the average arrival delay for all the flights that took place in the same day and plot it (x = 1:365, y = daily average delay). On the same plot, please mark the Federal Holidays from the federal-holidays-2013.xlsx dataset.**

Initiate the two variables as x axis and y axis which we will use to plot the graph for average arrival delay of all flights on the same date.

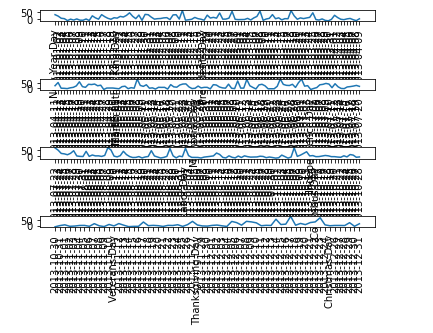


Fig 3.1: The average arrival delay for all the flights

**#d. Calculate the average arrival delay for all the flights for each arrival airport (e.g. IAD, DCA, and BWI) and determine most and least reliables (most reliable = the one that has the shortest average delay)**

To calculate the avg arrival delay for all flights for each airlines we need to do groupby for destination and arr\_delay.

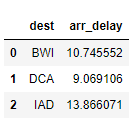


Fig 3.4: average arrival delay for all the flights for each arrival airport

**#e. Calculate the average arrival delay for all the airlines and determine most and least reliables most reliable = the one that has the shortest average delay)**

To calculate the avg arrival delay for all airlines for each airlines we have done the groupby for arr\_delay and carrier.

****

Fig 3.5 : average arrival delay for all the airlines

**#f. What day of the week we had the highest average delay?**

To calculate the highest avg arrival delay we have done the groupby for arr\_delay and day of week and then finding the mean of it.

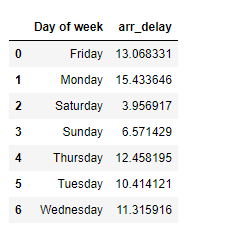
****

Fig 3.6: day of the week with the highest average delay

**#g. Which one had a higher average delay: flights that took off in the morning (6 am to 10 am), noon (11 am to 2 pm), afternoon (3 pm to 5pm), or evening (6 pm to 10 pm)?**

For solving this problem, to calculate the highest arrival delay for all flight we have used two counters again which are # Morning 6amto10am Noon 11amto2pm Afternoon 3pmto5pm or Evening 6pmto10pm.

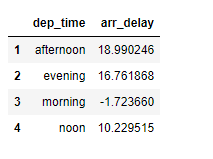
****

Fig 3.7: Highest average delays

**#h. Determine the number of airplanes used in these flights manufactured by BOEING, EMBRAER, and AURBUS separately.**

For solving this problem, we need the airplanes details for each flight manufacturer.

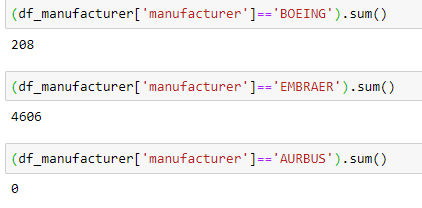
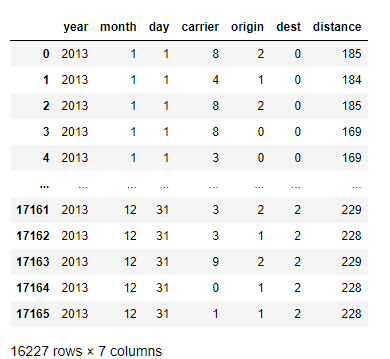
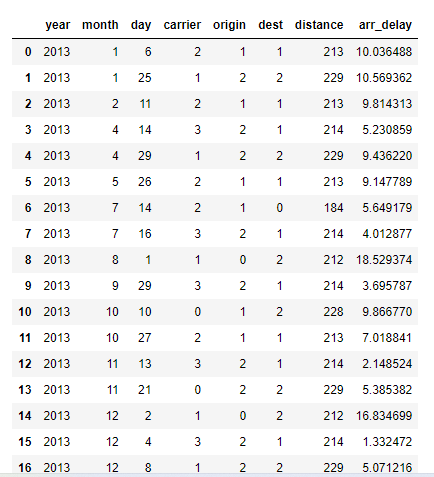
****

Fig 3.8: Manufactured flights

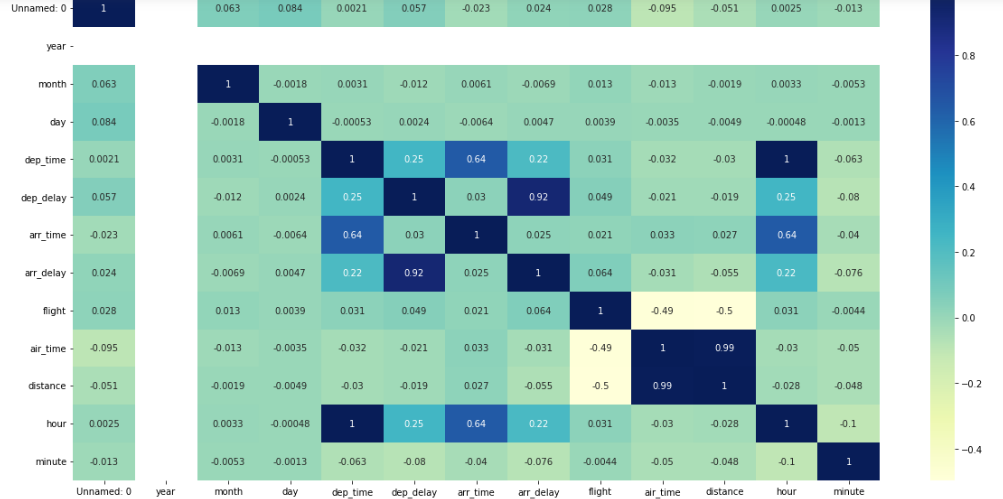
**Linear Regression Model**

We performed was to split the dataset into 70% (train) and 30% (test) and build the Linear Regression Model with the following output: We plot the scatter plot with training and testing data.

****

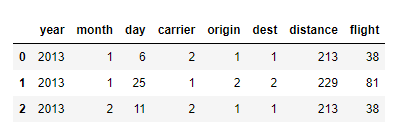
****

we plot the heat-map to understand the correlation:

****

**Logistic Regression Model**

Our approach to solve the query was to take the columns arrival delay and departure delay. To build the Logistic Regression we needed the data in binary format and hence the logic created was if arr delay is positive then it is 1 or else 1. And also,need to find accuracy.

****

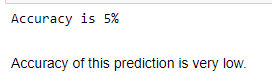
****

Fig 5 : Accuracy