**Course Project  - Various Sampling Methods Report**

Shaun Pritchard

Rasmussen College

QMB4000

Benjamin Tasker

December 15, 2020

**Course Project  - Various Sampling Methods**

Initially I implemented several methods to reduce the dimensionality of the flight data through implementation of sampling methods. I started by importing a flight dataset into my localhost MySQL workbench.

**Selecting Sample Data:**

I then created a subset of data with a limit of 10,000 rows of the corresponding data. Which is 15.5% of the dataset population of 65000. I determined the initial sample size by using calculations and a sampling distribution method based on a confidence level of 95% and margin of error of 1% based on a population parameter and sample size of **8367**. Therefore 10,000 samples is more significant in returning non-sampling errors. The sampling methods are standardized practice of (NIH) National Institute of Health  and The Advanced Analytics Institute Using Cochran’s Formula

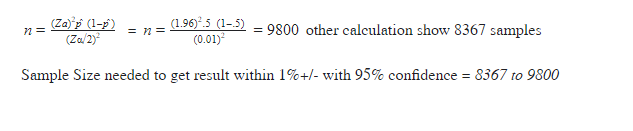
**Calculations Variables**:

n = sample Size

CI = Confidence Interval = σ = x̄ ± z\* σ / (√n)  = N(**μ**,σ2) = 0.01

CL = Confidence Level of 95%, Zα is 0.01

Z Score = 1.96 = Zα/2

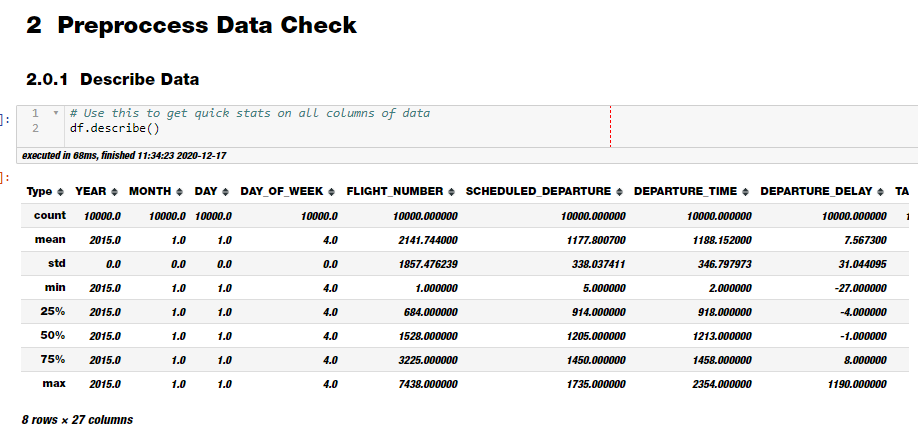
P = population = 65,000  


n=(Za)2p̂  (1-p̂ )(Zα/2)2  = n=(1.96)2.5  (1-.5) (0.01)2=9800 other calculation shows 8367 samples

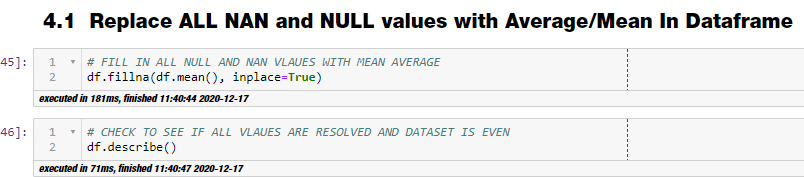
Sample Size selected is calculated to get accurate result within 1%+/- with 95% confidence = *between 8367 – 9800 features*. With a 10,000-row subset  we are above the mark on sample size being appetite for our data.

**Cleaning and Preprocessing Data:**

Initially I accessed the dataset using the Python framework and ran several tests to check the integrity of the data and to find any missing values or inconsistencies after I was connected and imported. I found that several columns of data from the set of 10,000 samples had many missing and NULL values.

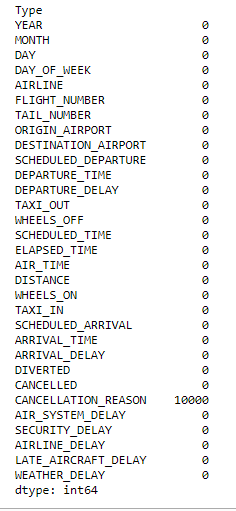


I used python with the code below to calculate all mean variances of data for each column and replace the missing values with the new mean features to even out the data.



**Results**

The results below show the full data set has NO missing NAN values in features. The column of data remains with 10000, NAN values based on cancelation reasons. This column of data is categorical arbitrary data. I proceed to drop this column from the dataset because it holds no significance to the questions that need answered.



Code used to drop CANCELLATION\_REASON data column



From this point I was able to start grouping data and running inferences on specific columns and features to answer questions and produce business logic and analytics visualization. I have included a deck which highlights the main features and answers the main business questions separately.

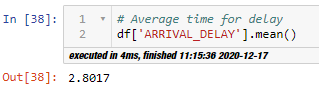
**Data Dictionary**

For ease of understanding the following data to answer the business questions I created a useful dictionary for the CEO below based on the important variables used for the analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Variable name** | **Unit** | **Description** |
| Airline | AIRLINE | VARCHAR | Airline that is responsible for flight. |
| Airport | AIRPORT | VARCHAR | Airport hub for origins and departing flights. |
| Arrival delay time | ARRIVAL\_DELAY | FLOAT64 | Time delay in which flights arrived. |
| Origin | ORIGIN\_AIRPORT | VARCHAR | Destination flight started from. |
| Destination | DESTINATION\_AIRPORT | VARCHAR | Destination flight went too. |
| Delay in flight Arrival | ARRIVAL\_DELAY | FLOAT64 | Delayed from when the flight left it arrived at the destination. |
| Time of flight arrival | ARRIVAL\_TIME | DATETIME | Time the flight arrived at the destination. |
| Delay in flight departure time | DEPARTURE\_DELAY | FLOAT64 | Time delay from the time the airplane departed to start is flight. |

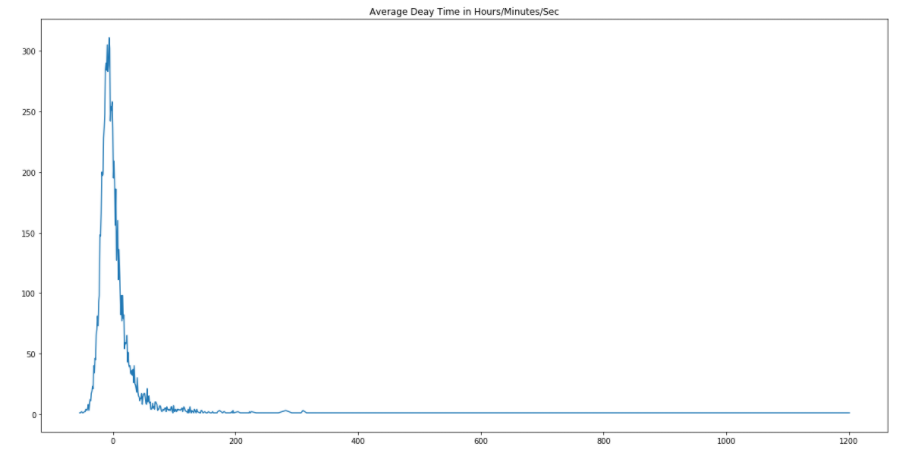
**What is the Average delayed arrival flight time?**

Essentially the average delay in flight times based on units provided in the data set is mean time of 2.80 a median time value of -4.0.

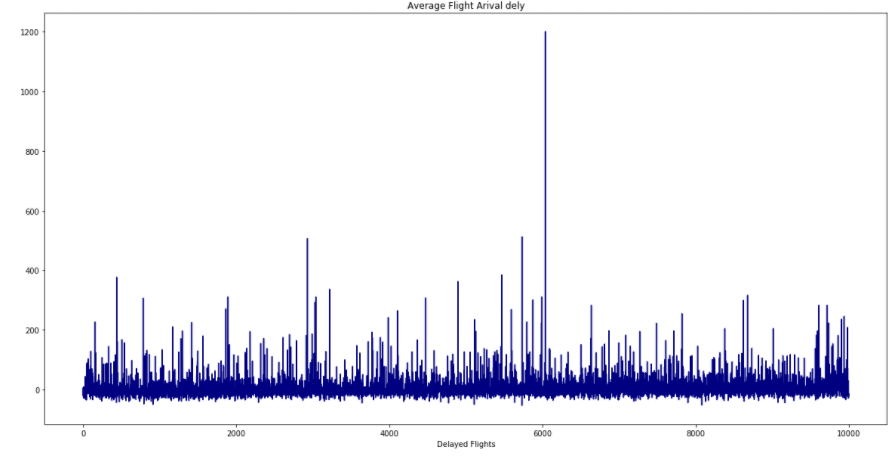


To calculate and visualize these average delays in flight times I ran 2 inferences and plots. The first line chart shows the peak of the frequency in. This gives a good look visually of the frequency data. The 2nd plot chart I used was a histogram. As you can see below from the histogram, I first grouped the feature data of the ARRIVAL\_DELAY column while using descriptive statistics techniques to average all the mean values of each feature. I selected the histogram because histograms  are used to show distributions of variables while plotting quantitative data with ranges of the data grouped in the mean intervals.

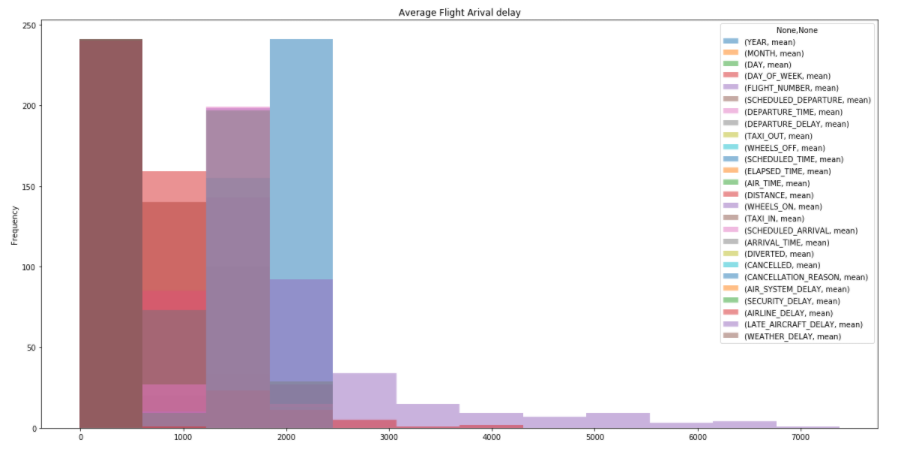
**Frequency Plot of Average Delay Time:**



**Frequency Plot-2 of Average Delay Time:**



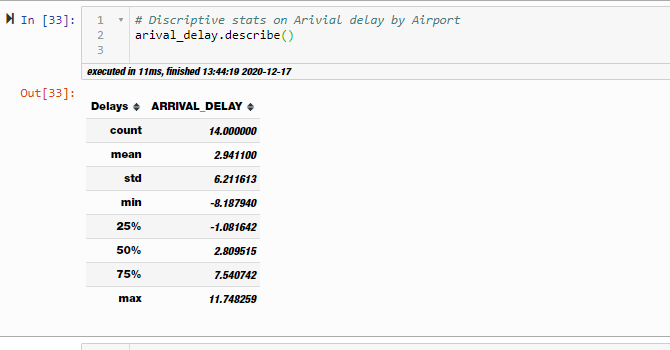
**Histogram Mean of Average Delay Time:**



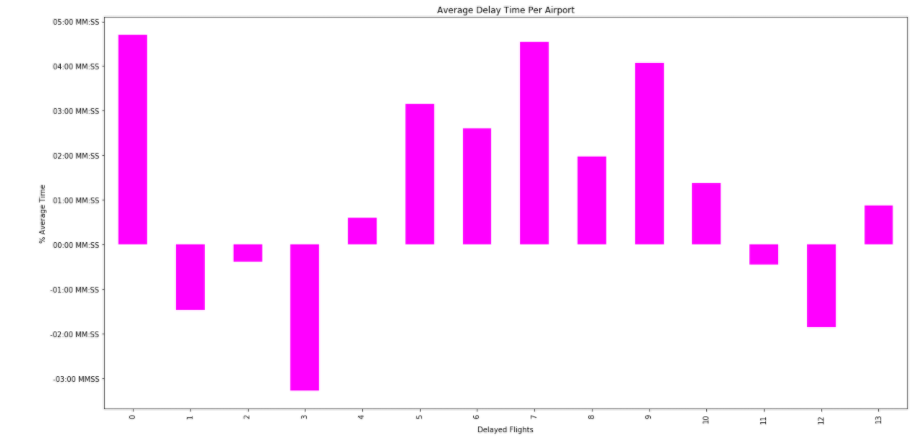
**What is Average delayed arrival flight time  based on each airport location**

Using techniques with grouping the AIRPORT variable with AVERAGE\_DELAY variable I was able to determine the  average delay time per airport was *2.94 (HH:MM:SS) .*  I calculated descriptive statistics followed by plotting the mean variance distribution on a time sequence chart and trend line as follows to get accurate depiction of average delayed arrivals per airport:

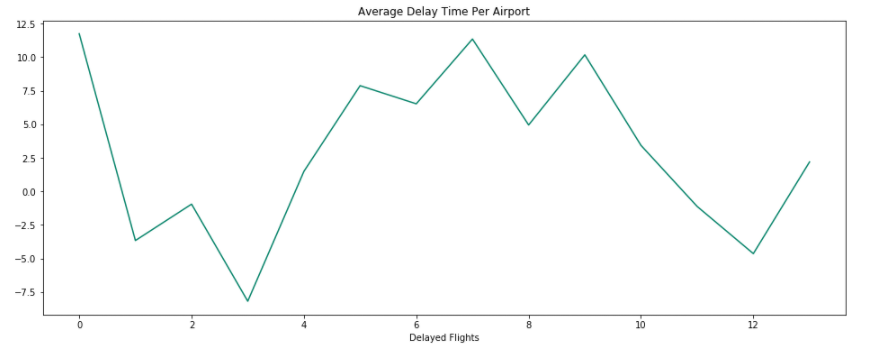
**Descriptive Statistics:**



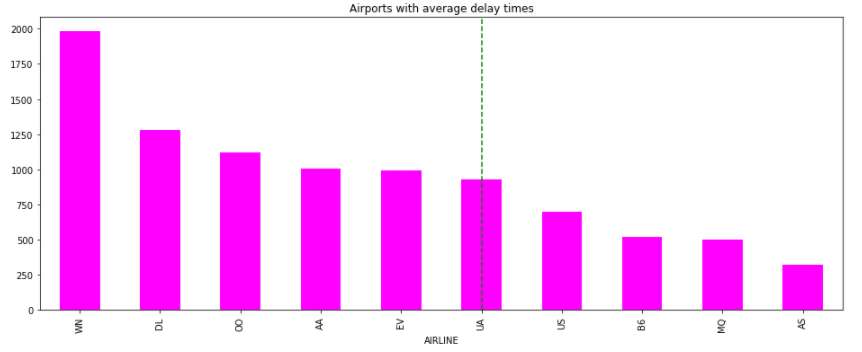
**Plot Distribution of Average Arrival Delays Per Airport:**



**Trend Line of Average Arrival Delays Per Airport:**



I also went ahead and plotted the average delay time based on the airport to get an inverse view of delay time while marking the average mean section. We can see that airports UA hold the means standard for average delay times:

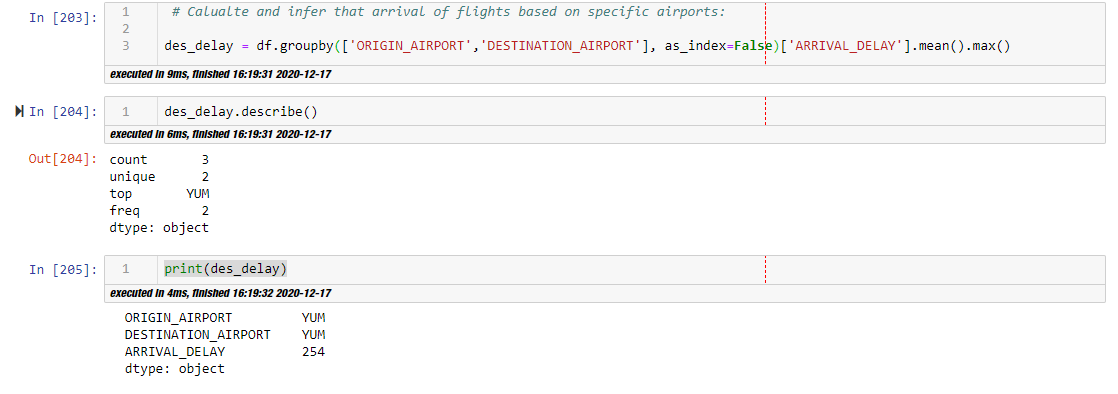


**What Origin and destination airports have the longest delayed flight times on average**

For these questions some more advanced calculations and inferences were taken to achieve the results. It was specifically asked to determine the average between to destination and origin airports. This needs no consideration to the airline itself just the locations therefore I used the ORIGIN\_AIRPORT and the DESTINATION\_AIRPORT as my independent variables. I then grouped those variables while using that variable ARRIVAL\_DELAY. The ARRIVAL\_DELAY variable took is the total amount of time in delay resulting from each airport origin and destination. I contemplated dividing the difference of the DEPARTURE\_DELAY time with the actual ARRIVAL\_TIME and it gave me the same calculation that the ARRIVAL\_DELAY  column already had inferred.

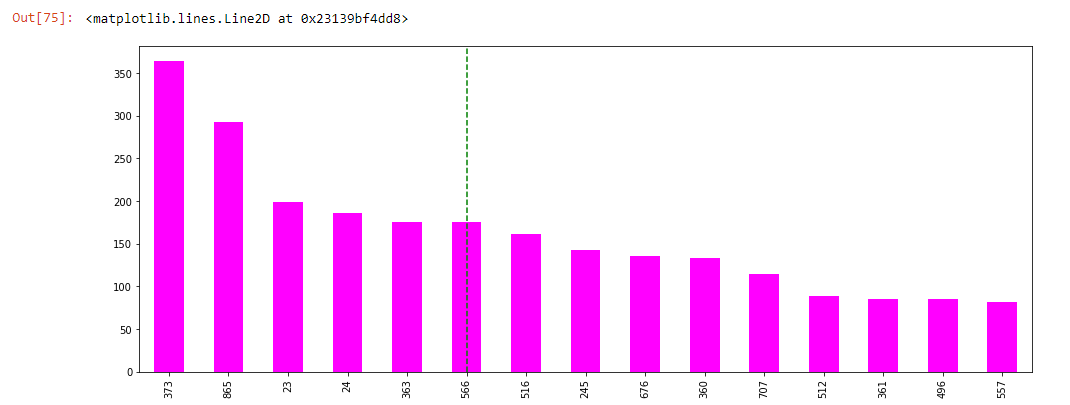
Essentially,  we can set 2 columns in a vector to determine the variance for each set of data and apply the average times in delays to and from each destination to get the results of this report. I started with simply joining the data and running descriptive statistics to determine variance and then used the data to plot frequency and bar charts. Bar charts are best used nominal features. Here are my results.

**Descriptive Statistics Average Delay of Origin and Destination Flights:**

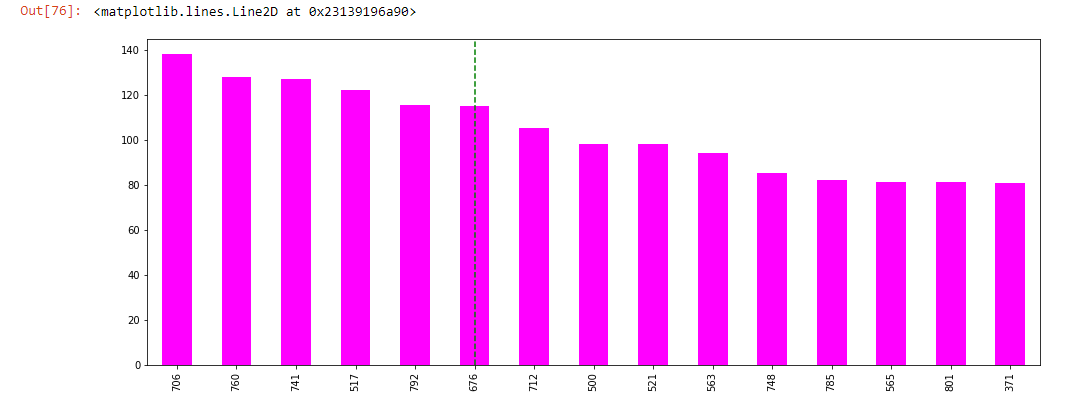
****

**Frequency Plot of Average Delay of Origin and Destination Flights:**

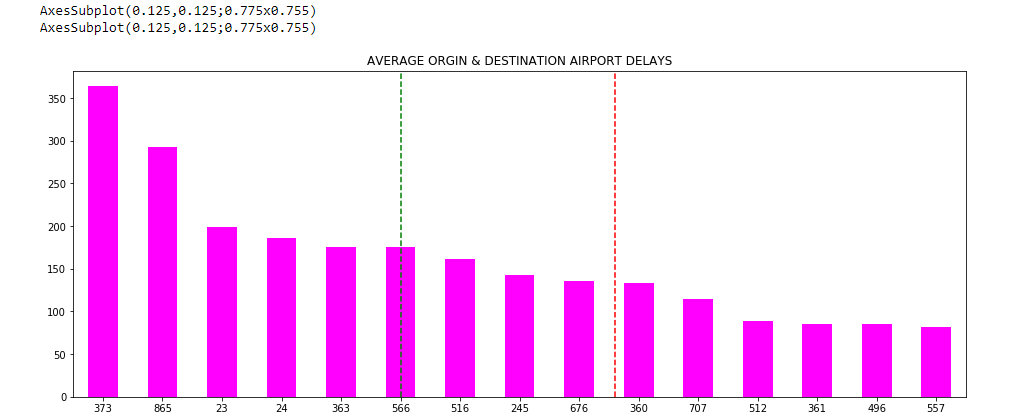
**ORIGIN\_DELAYS:**

****

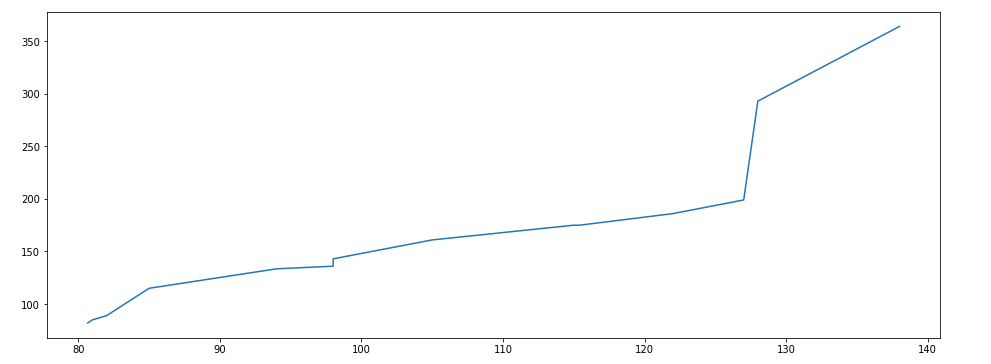
**DESTINATION\_DELAYS:**

****

**Bivariate Plot of Average Delay of Origin and Destination Flights:**

****

**Bivariate Frequency of both columns**

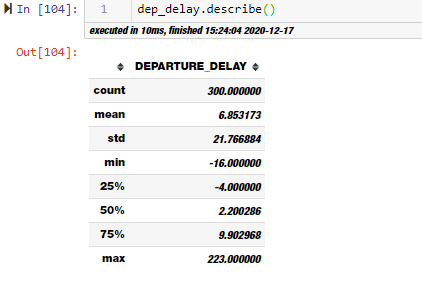
****

**What is the Average delay in departure time per airport by starting destination?**

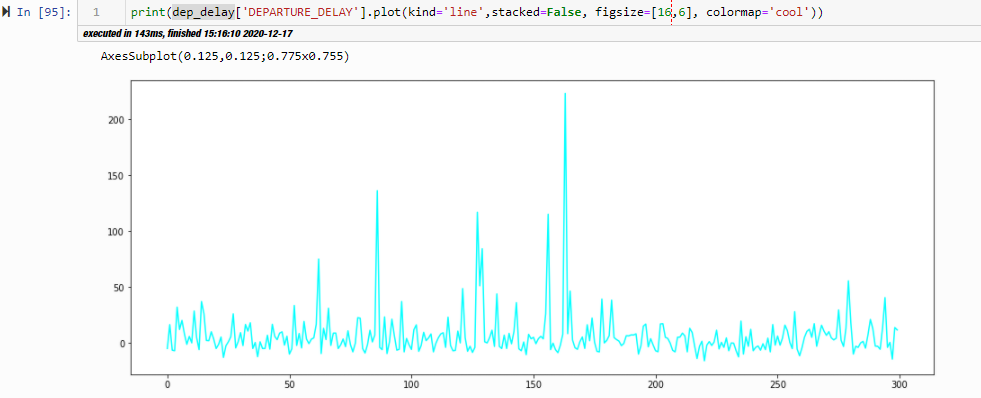
For this final question, I simply grouped the ORGIN\_AIRPORT variable to the time in DEPARTURE\_DELAY variable with a calculated mean to average the departure delay time for each ORGIN\_AIRPORT.

I simply ran descriptive statistics to get needed data to answer the question and plotted frequency line graph and mean average bar chart because the data is quantitative but nominal as shown below.

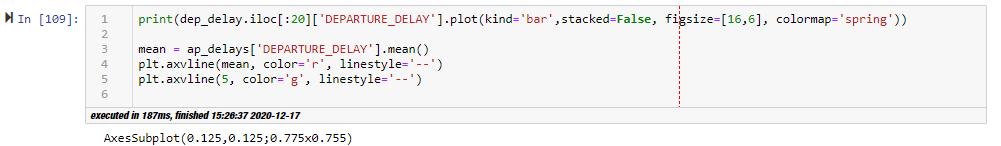
**Descriptive Statistics Average Delay of Origin Airports:**

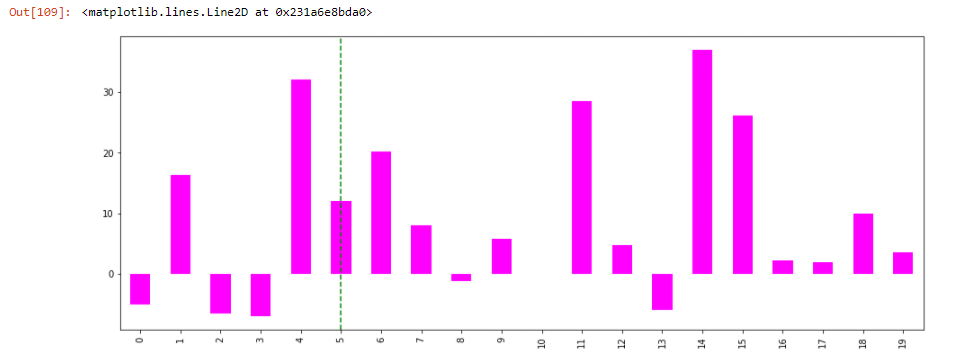


**Frequency Plot of  Average Delay of Origin Airports:**

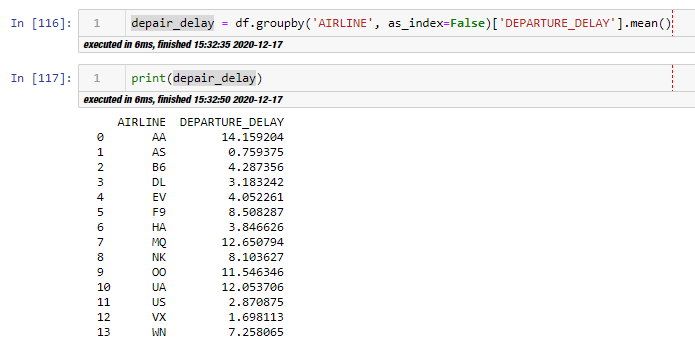
****

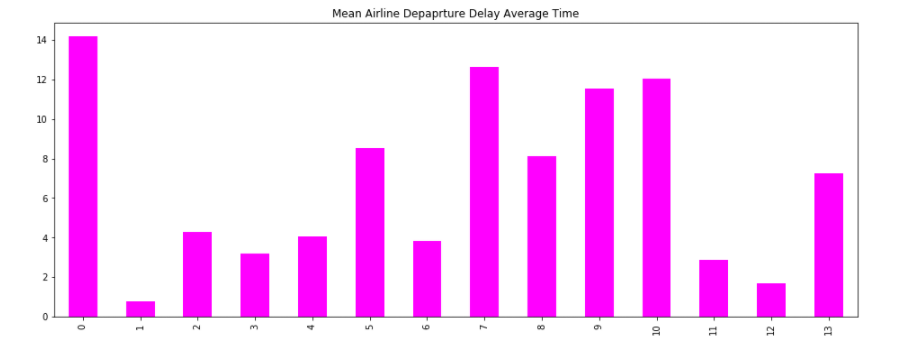
**Plot of  Average Delay of Origin Airports:**





I Also based calculations off the AIRLINE variable to see which AIRLINE had average delay in departure times. Below are the descriptive statistics and plots to visualize the analysis.





**This concludes my following report as follows:**

* **The average delayed arrival flight time is**
* **The average delayed arrival flight time  based on each airport location**
* **The origin and destination airports that have the longest delayed flight times on average**
* **The average delay in departure time per airport by starting destination**

**Overview:**

1. The average delayed arrival flight time is  = 2.8 (hh:mm:ss) means delay time.
2. The average delayed arrival flight time  based on each airport location = 2.9 (hh:mm:ss) means delay time.
3. The origin and destination airports that have the longest delayed flight times on average
4. The average delay in departure time per airport by starting destination = 6.9  (hh:mm:ss) means delay time.