Name:	Rohit Toshniwal
Student Number:	*******
Programme:	MCM1
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	13 Dec 2019
Module Coordinator:	Dr Suzanne Little

Abstract

Flight delay is one of the most common but an unpleasant experience that people dread to have. There could be some reasons which are inevitable such as weather conditions, air trafficking or any unforeseen event; but there also could be some reasons which can be dealt with by improving the process.
After looking at the data, I had a couple of questions in mind:
What factors contributes the most to flight delay?
Which Airline has the most delay in 2015?
Which state in America has the most flight cancellation?
What are the conditions that led to the cancellation in Airports?
This data visualization analyses the variety of factors responsible for and associated with flight delays for different airlines. I have tried to answer all the above questions with the help of data visualization. I am analysing 2015 Flight Data to make suggestions about traveling around them world. I compared airlines based on Delays and Cancellations. I build a dashboard to learn which airlines have the highest cancellations and delays.

1. Dataset

The flight delay and cancellation data were collected and published by the DOT's Bureau of Transportation Statistics. Once the project topic is decided, I went through various websites, viz. Kaggle.com, AirlineBureau.com, and qundal.com, but I got the data of 2015 Flight Delays and Cancellations from kaggle.com.
There are three csv files in the dataset: 1. Flights 2. Airports 3. Airlines. The dataset is of 565 MB and seems to be structured with flights.csv having 1,048,576 number of rows and 31 number of columns, airlines.csv file having 14 number of rows and 2 number of columns and airports.csv file having 322 number of rows and 7 number of columns. The different data types present are:

1. temporal: YEAR, MONTH, DAY, DAY_OF_WEEK
2. Spatial: AIRLINE, ORIGIN_AIRPORT, DESTINATION_AIRPORT, CITY, STATE, COUNTRY
3. Ordinal: SCHEDULED_DEPARTURE

4. Nominal: AIRPORT
5. Interval: DEPARTURE_DELAY
6. Ratio: LATITUDE, LONGITUDE
7. Boolean: DIVERTED, CANCELLED

There are three aspects of big data in this dataset,
first volume, as there are 1,048,576 records available and the
second one is velocity, because for January 2015 there were only 3325 rows, for February 2015 it became 5789, for March 2015 it was 17834 and for April 2015 it increased to 132,818 rows which is almost 110 times the January 2015.
For variety, as the data is related to flight delay and cancellations, the details given in the dataset is about each and every aspect. There are variety of columns which gives detailed information and we can easily predict data using the dataset.

2. Data Exploration, Processing, Cleaning and/or Integration

The dataset contains 1,048,576 records. The exploratory analysis is done using Python PANDAS and is shown below:

```
df.describe()
```

	DEPARTURE_DELAY	ARRIVAL_DELAY	SCHEDULED_TIME	ELAPSED_TIME	DELAY_LEVEL
count	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06	5.714008e+06
mean	9.294842e+00	4.407057e+00	1.418940e+02	1.370062e+02	3.487349e-01
std	3.688972e+01	3.927130e+01	7.531400e+01	7.421107e+01	6.172869e-01
min	-8.200000e+01	-8.700000e+01	1.800000e+01	1.400000e+01	0.000000e+00
25%	-5.000000e+00	-1.300000e+01	8.500000e+01	8.200000e+01	0.000000e+00
50%	-2.000000e+00	-5.000000e+00	1.230000e+02	1.180000e+02	0.000000e+00
75%	7.000000e+00	8.000000e+00	1.740000e+02	1.680000e+02	1.000000e+00
max	1.988000e+03	1.971000e+03	7.180000e+02	7.660000e+02	2.000000e+00

Explanation:

- count is the number of non-empty rows
- min is minimum value from each column
- max is maximum value from each column
- Mean is the average value of the variables
- Std is the standard deviation each column
- Q1 (25%) and Q3 (75%) are first and third quartile, Median is the middle value and Skewness is the measure of asymmetry.

Processing and Cleaning:

In my dataset, there is lots of temporal data. So, I fixed and cleaned all the temporal columns. Moreover, in the **SCHEDULED_DEPARTURE** variable, the hour of the take-off is coded as a float where the two first digits indicate the hour and the two last, the minutes. This format is not convenient, and I thus convert it. Finally, I merge the take-off hour with the flight date. To proceed with these transformations, I define a few functions.

The content of the **DEPARTURE_TIME** and **ARRIVAL_TIME** variables can be a bit misleading since they don't contain the dates. For example, in the first entry of the data frame, the scheduled departure is at 0h05 the 1st of January. The **DEPARTURE_TIME** variable indicates 23h54 and we thus don't know if the flight leaved before time or if there was a large delay. Hence, the **DEPARTURE_DELAY** and

ARRIVAL_DELAY variables proves more useful since they directly provide the delays in minutes. Hence, in what follows, I will not use the **DEPARTURE_TIME** and **ARRIVAL_TIME** variables.
Finally, I clean the data frame throwing the variables I won't use and re-organize the columns to ease its reading. At this stage, I examine how complete the dataset is: I observed that the variables filling factor is quite good (> 97%). Since the scope of this work is not to establish the state-of-the-art in predicting flight delays, I decide to proceed without trying to impute what's missing and I simply remove the entries that contain missing values.

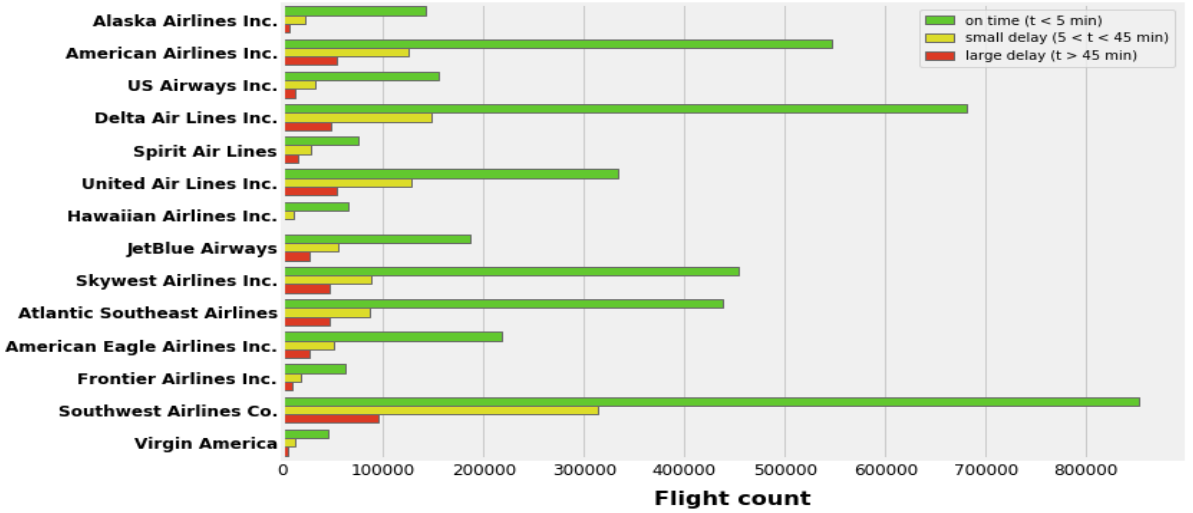
How did you choose the attributes to visualise?
For the first graph, I wanted to plot graph for airlines with their flight count based on the delay level, so I chose DEPARTURE_DELAY and AIRLINE as my attributes.
For the second visualization, I wanted to gain insights of all aircraft carriers based on the delay/cancellation reasons. For this we created a dashboard that gives details about the carrier specific information.

3. Visualisation

1. Count Plot graph gives a count of flights for three different time delay intervals

Choice of Graph:

The below Count Plot graph gives a count of the delays of less than 5 minutes, those in the range 5 < t < 45 min and finally, the delays greater than 45 minutes. **Count Plot** Show the counts of observations in each airline using bars. A **count plot** can be thought of as a histogram across a categorical, instead of quantitative variable.



Design Choices:

Color- I used green color to show on time flights, Yellow color for small delay flights and Red color to indicate large delay flights.
Shape- I have used bars to show the counts of each Airline.
Layout and Structure- The airlines names have been placed on Y-axis and flight count placed on X-axis. There are legends that show the different time delay intervals.
Font and Labels- The font used here is Source Sans Pro. The labelling is done for the Airlines, Flight Count and different time delay intervals.

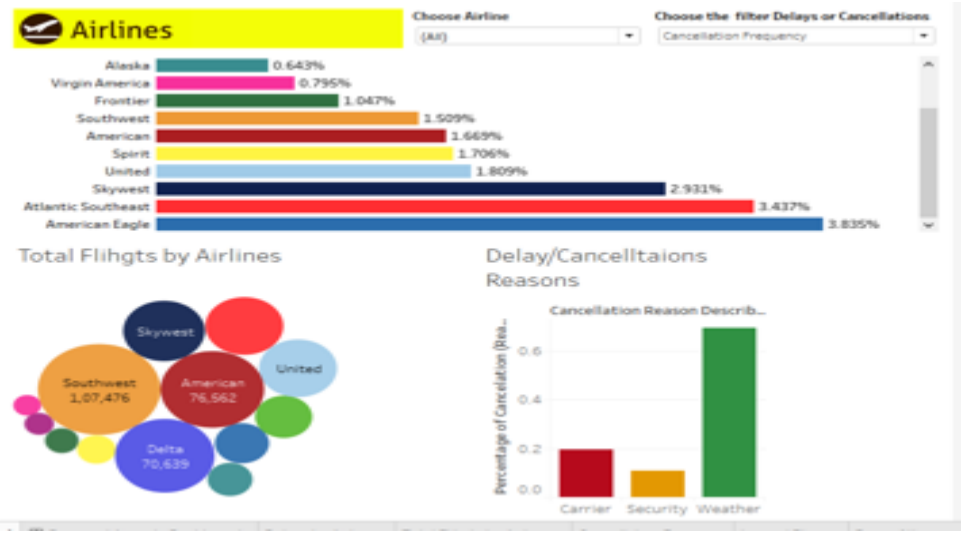
2. Airlines Delays and Cancellations Dashboard in Tableau

Choice of graph:

I wanted to gain the insights based on the Delay and cancellations reasons. For this I created a dashboard that gives details about the carrier specific information. I have used three different worksheets to create the dashboards. For the first worksheet I have used bar graph, for the second and the third worksheet, I used bubble graph to show information about number of flights and percentage of delay reasons respectively.

Note: Refer attached screencast video for visualization of Tableau Dashboard

Animations and Interactions: I have added two dropdown filter in the dashboard, one is to select Airline and other is to filter delays or cancellations which shows interactivity between different airlines and their delay reasons and flight cancellation reasons.



List of Tools or Libraries Used:

1. Python: Pandas (a Python library) has been used to clean the file to remove the string and null values from the original file and derived file.
2. Tableau: It is a visualization tool used to develop an Airline Dashboard and also to create multiple worksheets to visualize the data.
3. Matplotlib, Seaborn and Plotly libraries in Python used for creating count plot graph.

4. Conclusion

Critical Analysis:

The Dashboard has the following issues:

1. If the percentage and number of flights are small in amount, then the bubbles shrink in size and hence doesn't show the details in the smaller bubble.
2. Animations in dashboards are not that interactive in tableau.

Were there aspects that you think could be improved upon?

1. There should be data of flight prices to study correlation with delays
2. The delay reasons could have been shortened for the longer ones by abbreviating the Airline names and/or the delay reasons.
3. For the first graph, I could have implemented scatter plot instead of count plot, but I created a function which was only suitable for count plot.

Were there effects or functionality that you were technically unable to achieve?

1. Tableau: When I tried to add some animation in my dashboard in Tableau, so I couldn't due to lack of features for animation.
2. Plotly: Plotly didn't allow me to create count plot using Plotly functions, so I used seaborn to create count plot.

References

[1]: https://www.kaggle.com/usdot/flight-delays
[2]: https://www.tableau.com/learn/training
[3]: https://seaborn.pydata.org/tutorial.html