Data Story Telling Project with Tableau

This project is part of Udacity Data Analyst Nanodegree where I have created an **explanatory** data visualization from a data set that communicates a clear finding or that highlights relationships or patterns in a data set.

Summary:

Here I have created visualizations of the Flight Delays during the time period from 2008 to 2018 and the dataset has been downloaded from the United States Bureau of Transportation Statistics website ((https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1) as mentioned in the Udacity's Data Resource Option section.

The visualizations I have created from the dataset will address the following questionnaire:

1. How is the variation of Number of Flights and the Flight Delayed Time over the years 2008 to 2018
2. How the Top 10 busiest airports have been contributing in the Percentage of Delays of the different carriers
3. How the airlines have been affected by different reasons of delays.
4. How much time has been delayed for each type of reason and how the different average delay types have been changing across the time period
5. How is the No of arrival flights, diverted and cancelled flights across the months of the different year.
6. Airports with more flights delays and if the no of arrival flights has any relation with the delay time
7. How is the average Late Aircraft type delay across the seven important cities of United States.

Data Assessing and Data Wrangling:

Before creating visualizations in tableau, to explore the dataset, I have saved the downloaded csv file into a pandas dataframe and observed the details of different columns available in the data. Here I have renamed some column names like 'arr_flights' -> *arrived,* 'arr_del15' -> delayed, 'arr_delay' -> *minutes_delayed* . Also 'airport_name' has the city, state and airport name all together, so I split them.

Finally, I have saved the dataset to an excel file 'airline_delay_data_final.xlsx' to be used for the visualization.

Design:

Initial Draft -> For the first visualization, I have plotted line charts of the Total number of Flights and total Delay time over the years. The two plots have been drawn to each other for comparing their trends across time.

Next, I have selected bar charts to display the percentage of the delay of the top 10 frequently used airports by the airlines. The selection of small set of airports is because of the ease interpretations. The color coding of the bar charts easily depicts how the top 10 frequent airports are contributing to the delays of the airlines over the other airports.

Next, I have created a bar chart to show the different type of total delays of all the carriers in the dataset. The different color coding and selecting of bar chart visualization helps to distinguish the various types of delays easily across the carriers and it also gives a comparison view as well.

In the next two plots, I have plotted bar chart to display the total time delayed for each different cause and a line chart to show the variation of the delay time due to different causes over the years. Later I have put these two separate plots into a dashboard to display the different delay time spreads across the years. The color coding for different types of delays in the line chart is easy to interpret with the help of legend and tooltip. Also, I have added the Year as filter for selecting single and multiple years.

In the following plot, I have created line charts to display the average number of arrival flights, cancelled and diverted one through the months. The different colors of the individual lines are easily detectable.

Next, its time to try out a map plot. So, at first, I have changed the geographic role of dimension 'Airport' to airport one for plotting the map. This plot basically displays the number of flights and the total delayed time for all the airports. To make the visualization more meaningful, I have encoded the total number of flights in the size and the total delayed time in the color. Also, the airport name has been added as the detail marks so that anyone can get to know the names of the airport while moving the mouse over the circles. It helps to easily display which airport has the max/min number of flights and the small/large total delay time as well.

Finally, I have created the tableau story with all the previous plots and added captions to each of them.

Visualization link on Tableau Public:

https://public.tableau.com/profile/swaranika#!/vizhome/FlightsDelayProject-Version1/FlightDelay2008-2018Story

Second Draft **->**

After getting the feedback on my initial draft of visualization from friend and colleague, first I have changed the total delay time to average delay time in the first line plot as the average measure is more meaningful for interpretation here and then used the aggregated value of them as aggregate functions allow to summarize or change the granularity of data. For that, I have created the calculated fields for all the average value of different causes of delays and used them for visualizations.

For the other plots like the different types of delays and the different causes of delays across the year also, I have changed the total value to the average value. In the dashboard of the first draft, there was no provision for highlighting a particular cause of delay. I have added Highlight as action so that if we choose the average delay of a particular cause, it will display the same cause over the years in the bottom figure of the dashboard greying out the others line charts.

From the feedback, I have also added the year filter to the plot where the average no of arrived flights, cancelled and diverted flights have been displayed. Also, I have added captions for the plot showing the percentage of the delayed count of the airlines for the top 10 most frequently used airports as the legends of IN/OUT was not that descriptive. I have added a text in the visualization of the story section of the same plot as well because it will be easy to understand just viewing the visualization itself without going through the story caption.

In the second draft, I have added one more map plot showing the average late aircraft delay time of the seven most important cities in United States as the late aircraft delay time was the most contributing cause of the total delay types.

For both the map plots in second draft, I have used the average delay time and then the used the aggregated value of them instead of the total value.

Also, the test part of the story was not properly displaying because of the font size, so I have edited it and then uploaded it with the same name in tableau public.

Link for second draft:

https://public.tableau.com/profile/swaranika#!/vizhome/FlightsDelayProject-Version2/FlightDelay2008-2018Story

Feedback:

I have showed my visualization to one of my friends (who is familiar with tableau and analytics) and a colleague of mine (who is not used to with the data analytics and visualization – just to get feedback from neutral point).

While doing my first two plots of the first draft, I was struggling with the Year filter as initially I had forgotten totally to change it to discrete. At first, I was getting the Year filter as slider as the dimension was continuous. My friend has pointed this one and I had changed the Year as discrete to select a particular year.

Other Feedbacks for Draft 1:

- What about the Year in the plot for arriving, diverting and cancelled flights? You can use the same Year filter across different worksheets in tableau so that instead of individually adding it every worksheet, you can add at a single go.
- You have used the Total number of delay time for all the causes in almost all the visualizations, in case of comparison purpose, it will be more useful if you can use mean or median value – it will display more meaningful results.
- I am not quite sure of the legend IN/OUT (10 Most Frequently Used Airports) in the plot, from the visualization only in the final story, I am not able to able to get the meaning of it, can you change the legend title (Added the test in the story and caption in the individual worksheet)
- The color coding is easy to look at in the plots and the tooltip is helping.
- For the dashboard can I choose one delay cause and look at the changes over the year only for that? (I have added the Highlight as Action in the dashboard)

References:

1. Udacity video and Text lessons
2. Tableau Official Training Videos
3. Adding Text in Story: https://www.tutorialgateway.org/create-story-tableau/