Process book for Final Project

by

Yining Huang, Zhaoning Wang

**Motivation and Background**

“The COVID-19 pandemic is a serious global health threat, and CDC is committed to stopping its spread”(CDC). Our lives have changed dramatically since the COVID-19 started. We can never imagine staying home for several months, all work and school became online. Every year for thanksgiving and Christmas, most people booked flights to return to their families. Nonetheless, it is scary for people to get on the plane with others this year because of the risk of getting infected. Therefore, we hope to analyze this dataset and understand how the airport traffic is affected by the COVID-19.

**Research Question**

To better understand the impact of COVID-19 on airport traffic, we came up with some research questions are as follows:

1. How daily trip numbers(March 15th 2020 to October 15th 2020) have been decreased compared to Feb 1st to March 15th 2020?

2. Is it still decreasing or increasing back after lock down for 6 months?

3. Which city has been affected the most due to COVID-19? Which city has been affected the least?

4. How is COVID-19’s impact on the four different countries (the U.S., Canada, Chile and Australia)?

**Data Description**

The COVID-19's impact on airport traffic dataset is publicly available from Geotab (geotab.com). It shows the airport traffic from March 15th 2020 to October 15th 2020 and variable “percentage of baseline” compared to the number of trips on specific days (N1) with the same day of week in the baseline period which is Feb 1st to March 15th 2020 (N2).

Percentage of baseline = X100

Therefore, the larger number percentage of baseline is, the less flights it has on that day which means COVID-19 made an enormous bad impact on the airline.

This dataset was collected into a csv file which has 5936 rows and 11 columns. The columns mainly include AggregationMethod, Date, Version, AirportName, PercentOfBaseline, Centroid, City, State, ISO\_3166\_2, Country and Geography. Variable “Centroid” includes two parts which are Latitude and longitude. The URL link is as follows:

<https://www.kaggle.com/terenceshin/covid19s-impact-on-airport-traffic>

**ETL Process**

After setting working dictionary to my directory which contains “covid\_impact\_on\_airport\_traffic.csv”, the read.csv function was used to read this csv file with the stringsAsFactors parameter equal to TRUE and save it on a variable named data.

Dim function is used to print the dimension of data which is 5936 observations of 11 variables. Get the statistical summary of columns in the dataset using the summary function. To better understand the data, use str to check the structure of data. From the str function, we realize that the Date column was recognized by R as a set of Factors. Use as.Date(data$date) in R which can convert factor to Date. Centroid was also recognized as Factor, we use as.character to convert characters. Use gsub from Library stringi to remove POINT[] from each centroid and divide the column to longitude and latitude. Use View(data) to check the two new columns. Since variable “AggregationMethod”, “Version” is not related to the research question and variable “Geography” can be represented by new variable longitude and latitude. We removed column 1, 3, 11 and saved the new dataset as data2.

Use write.csv to download data2 as airportdata.csv.

**Design of visualization**

For this part, Tableau is used to perform data visualization as the main tool. After importing the processed data into Tableau, we made the following data visualization for our research questions:

1. For question one, a symbol map is used to show the location of the airport which this dataset collected geographically. The size of the circle represents the sum of percent of baseline, which means how serious each city and state were affected by COVID-19 from March 15th 2020 to October 15th 2020. The bigger the circle, the more severe the damage.

2. In sheet 02, the line chart shows the relationship between date and sum of percentage of baseline which can clearly reflect whether the airlines increased or decreased as the date changes from March to October 2020.

3. Sheet 03 is a horizontal bars chart showing the relationship between cities and sum of percentage of baseline. “Month” is added with different colors and it shows the degree of damage in different cities and different months.

4. As for question 4, we build a box plot chart to show how different countries have been affected based on the percentage of baseline to compare the different degree of damage.

The link of our data visualization is as follows:

<https://public.tableau.com/views/1_16068739922320/Dashboard1?:language=en&:display_count=y&publish=yes&:origin=viz_share_link>

**Implementation**

In sheet 01, longitude and latitude which has been generated by Tableau are used as column and row to build the symbol map. Country, state and city are shown as marked by different colors. Different states of the same country are also displayed in the same color for easy distinction. A size mark which appears as a circle is also added to show the degree of damage based on the sum of percentage of baseline from March to October 2020. We also add a map layer and some filters here which can easily help find the specific country, state and city if the audience has interest.

Moreover, for the line chart in sheet 02, we build the visualization with the sum of percentage of baseline as rows and date as columns. The classification standard for column is month. Country is shown as marked by different colors. From this we get four lines, which represent the trend based on the sum of baselines in different countries from March to October.

In addition, when the sum of the percentages of the baseline and the city were used as columns and rows, we drew a horizontal bar graph to show the extent of damage caused by COVID-19 in different cities. We also added the dates that are specifically displayed as months and marked them as colors and labels to further show how affected the different months are. The results are sorted in descending order according to the value.

For the COVID-19’s impact on the four different countries, we use a box plot chart to show the relationship between column “country” and row “percentage of baseline”, which can reflect the average level of damage in different countries.

**Results and conclusions**

In conclusion, airlines between countries have been drastically reduced due to COVID-19 generally. From the symbol map, four countries on three continents(28 airports total) were collected into this data and the states and cities of America and Canada are generally more severely affected by COVID-19 compared with Australia and Chile. Taking the United States as an example, the collected cities are mainly concentrated in states with higher population density. Generally speaking, there will be more airlines with higher population density, which enhances the contrast of data.

For America and Canada, the airlines have been reduced significantly since March, and the decline began to slow down in April but is still decreasing. It reached the maximum reduction in September and basically returned to the level of March in October. This means airlines between different countries are recovering after six months of lockdown. However, compared with America and Canada, the decline in Australia and Chile is relatively flat, but they are also recovering from September.

In addition, it can be seen from sheet 03 that New York is the most severely affected city by COVID-19, and every month since March, it has suffered more damage than other cities. Santiago is the least affected city based on this dataset.

In sheet 04, due to the characteristics of boxplot chart, here we focus on the average level of COVID-19’s impact in different countries. We already know for airports in the US and Canada total count of percent of baseline are significantly higher than Australia Chile, but we are still wondering if it is just due to the number of airports, since the dataset concluded more airports in the US and Canada. However, from the box plot chart, the average airlines of America are still the most affected by COVID-19, followed by Canada. Australia and Chile are relatively less affected by it. Although there is an outlier in Chile’s airlines, Chile is the least affected country.