Data Visualization Final Project

Byron Han

Dataset introduction

What is this dataset?

The dataset is about the information about all flights enplaned or deplaned in SFO between July 2005 and Dec 2018, the columns of the data include the year and month of which the data point is taken, the name and code of operating airline, the geo region of the airline, the terminal name, the price category of the ticket and total number of passengers this flight transported in that month, etc.

Where did you get it from?

I got this dataset from the SF governmental website. It is a .csv file downloaded from data.sfgov.gov. It is a perfect analogy of the time series financial data which I am super interested in to analyze trend and anomalies. There is nothing more perfect for us to dig into this monthly data and find interesting patterns and detect outliers from it.

Why did you choose this dataset?

I have always wanted to analyze financial data, which due to the security reason and various non-disclosure competitive constraints, none of the financial institutions I have contacted with allows me to use their labelled data. I tried to download some financial data from open API but either they only allow me to acquire a small fraction of the stocks or equity I am interested in or they just charge too much for me to be able to afford. Therefore, I found this exciting SFO data which is both interesting enough for me to do an analysis and a great analogy for the time series data. Furthermore, this dataset has all the requirements and geo-information which I need for this final project.

What types of questions were you hoping to explore with this data?

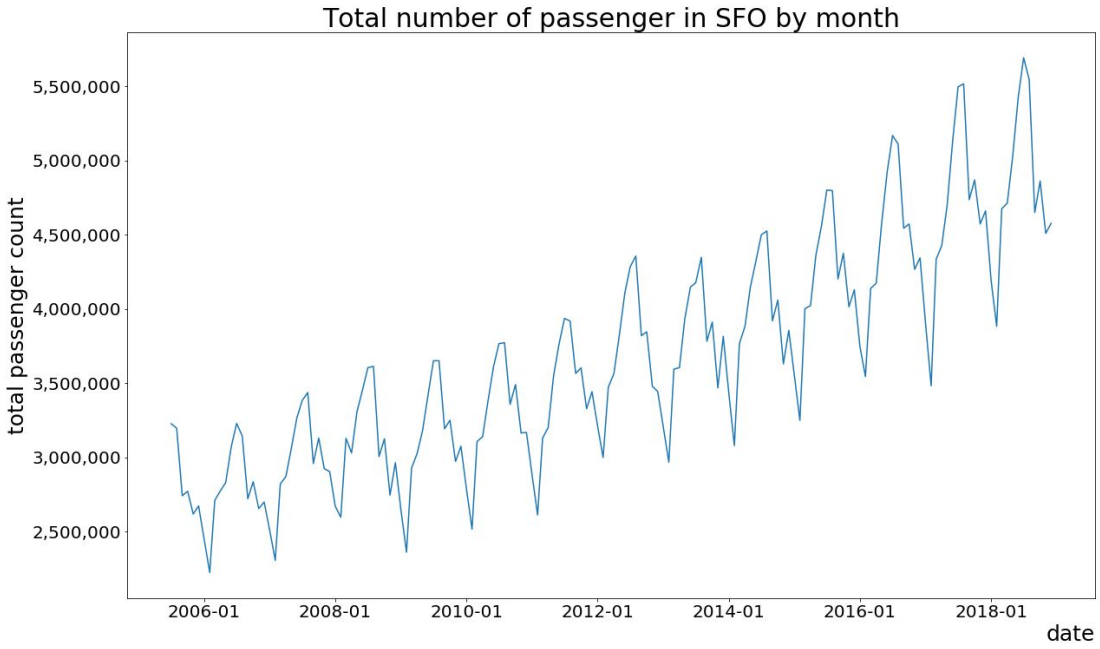
Overall, I want to find out how the total number of passengers varies with respect to different covariates, i.e. how the number of passengers varies with airline companies, the geo-regions etc. as the year goes forward.
I also have at least one question to answer using each of the plots. I will illustrate them in the next section and append the textual description in the next section.

Summary of the data

I want to answer a particular question with respect to each plot. I will illustrate each plot and each question separately.

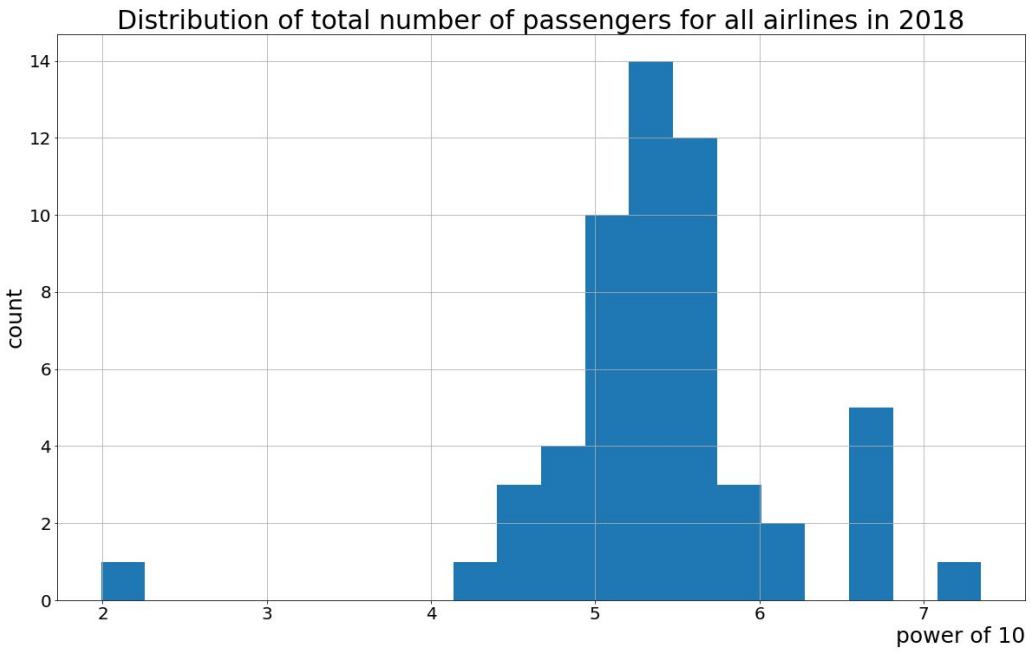
Line plot

This plot shows how the number of passengers varies with respect to the year number. It allows us to see the general varying trend of the time series variance in the dataset.



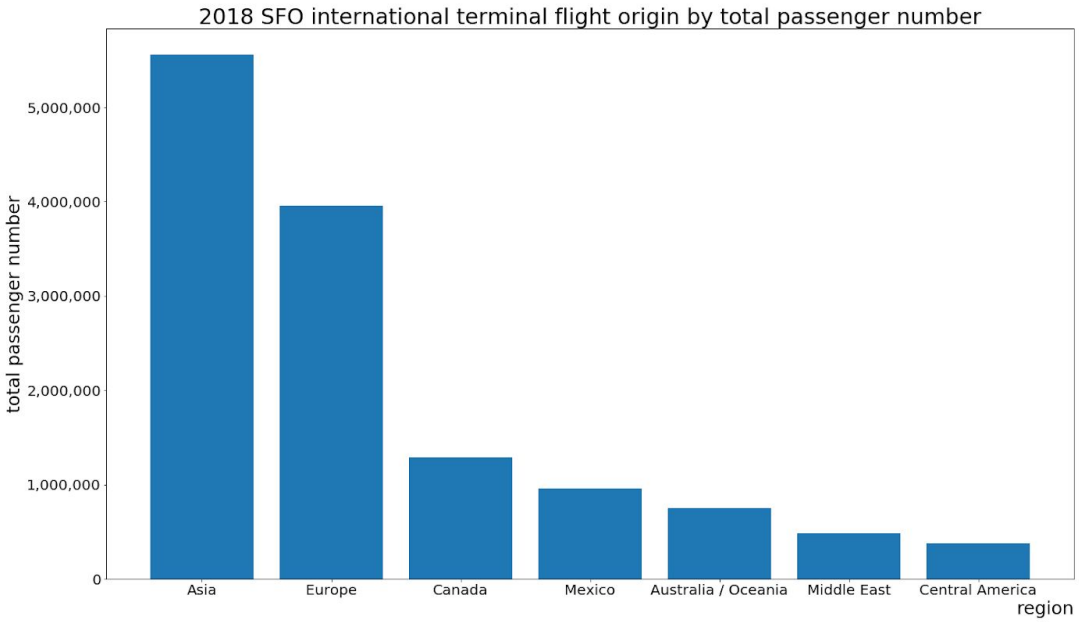
Histogram

This plot shows the distribution of passengers within 2018 for all airlines with respect to the number of passengers. We can see most airlines, 14 of them, transported 1 million passengers in 2018.



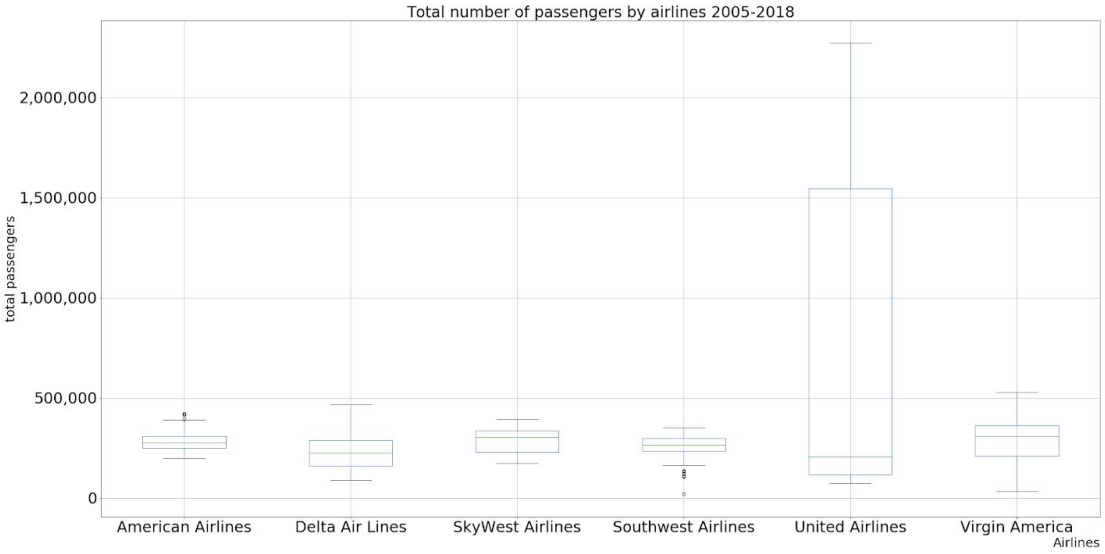
Barplot

This plot shows the relative comparision of the place of international passengers come from in 2018. we can see most travellers come from Aisa and Europe. And Aisa is almost as much as all other places excluding Europe combined



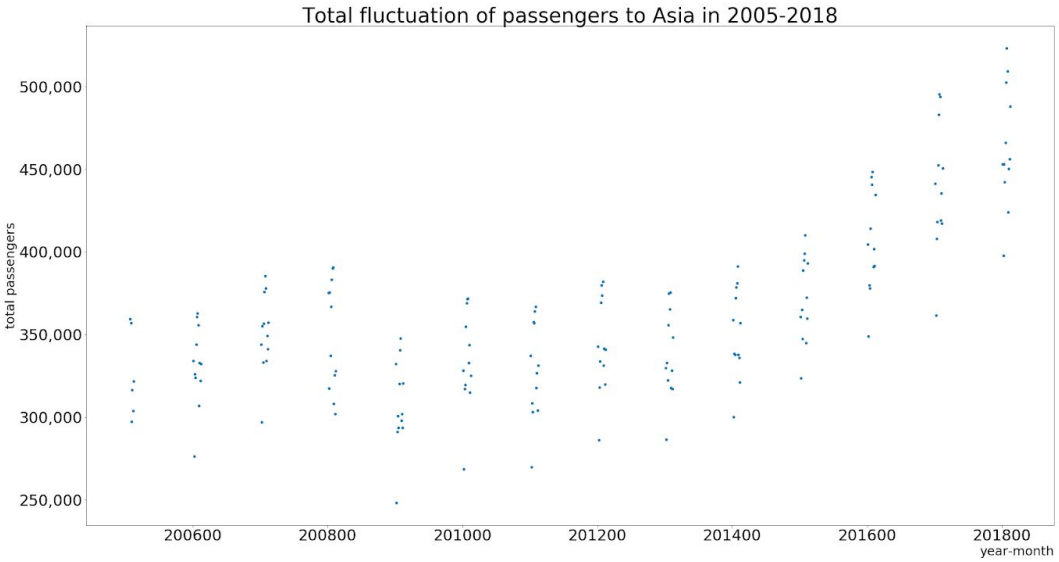
Boxplot

This boxplot shows for all those international passengers which major airlines transported the most passengers in 2005-2018, and the distribution of them. We can see in some of the years United Airline send over 2 million passengers, but mostly it sends as much as the other dominating airlines.

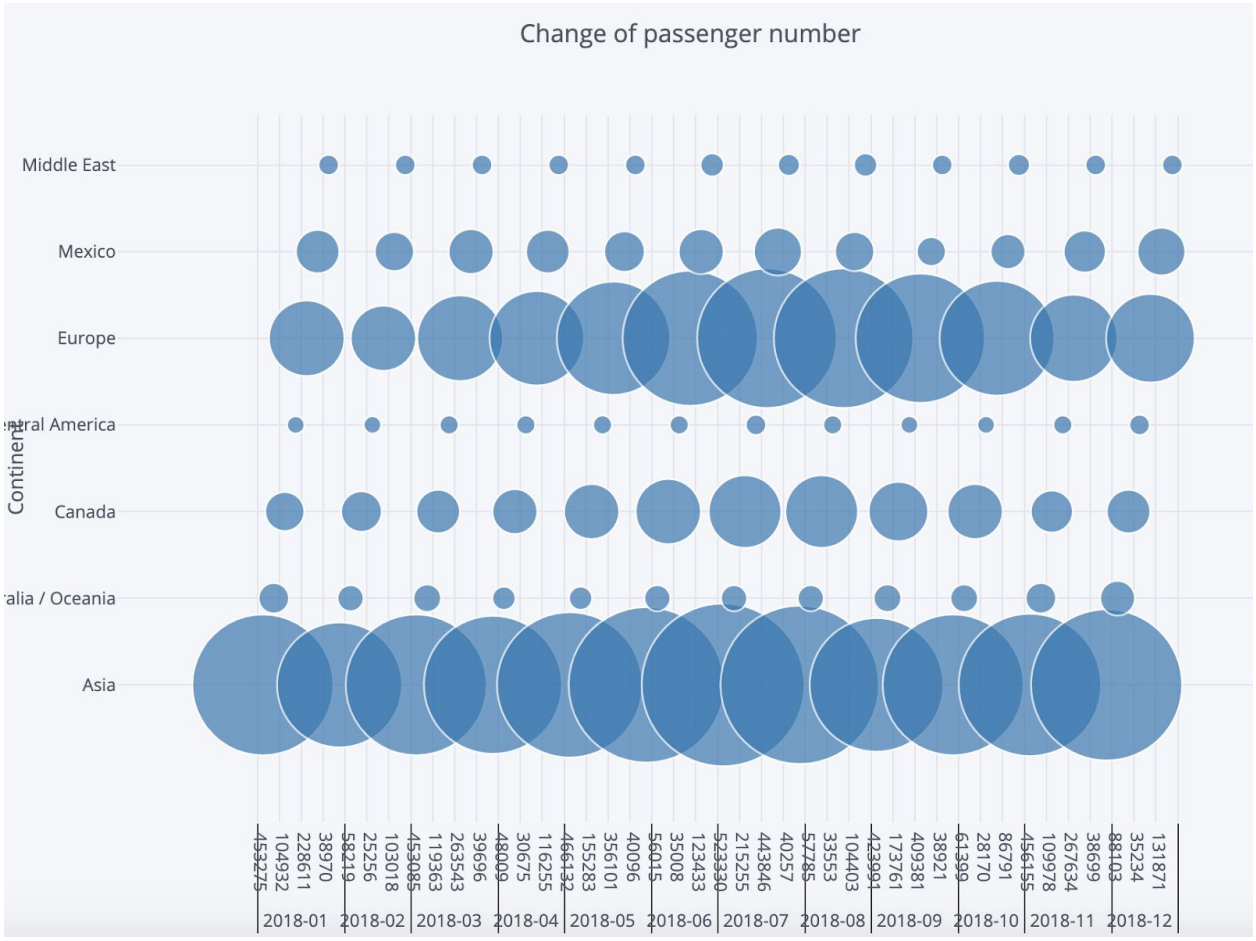


Scatterplot

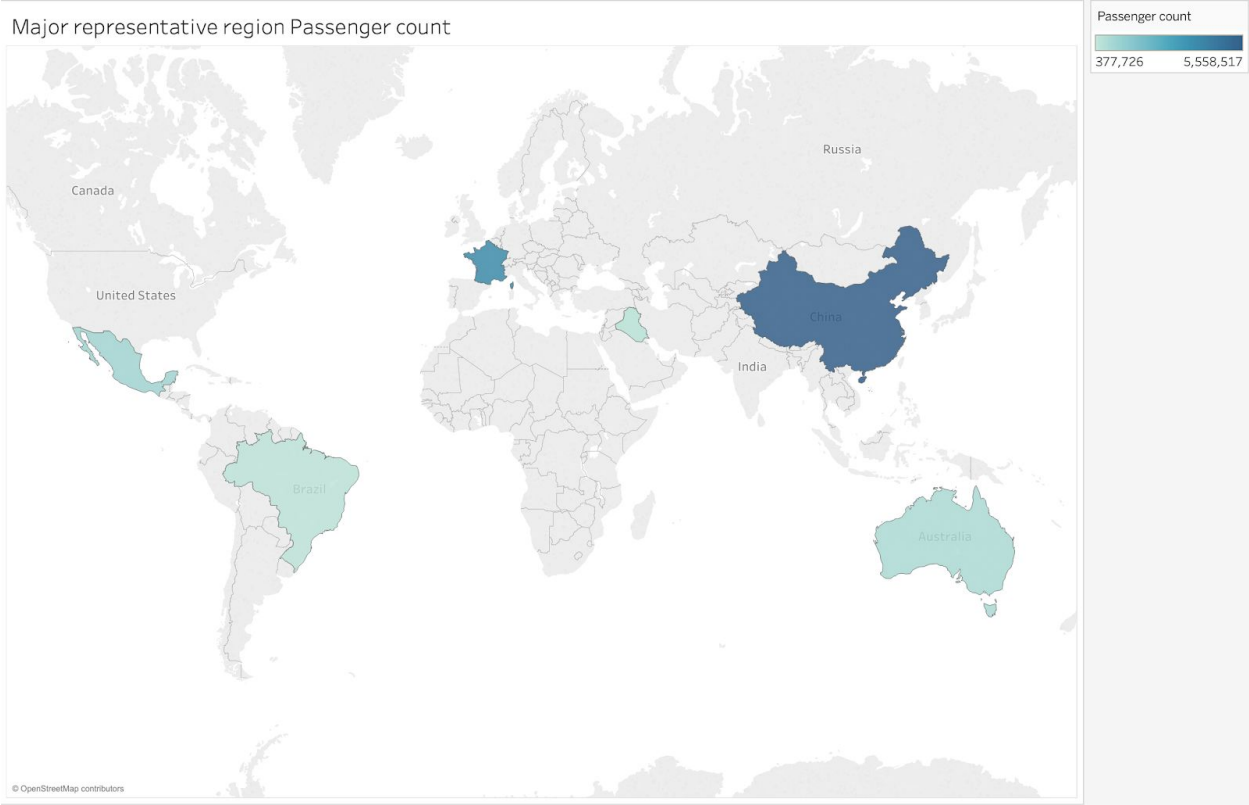
This plot shows for the major destination Asia, how much does the number of passengers vary within 2005-2018. And how much the number of passenger varies within each year.



Bubble Map (interactive in plot.ly)
This plot uses a visual and interactive way to see who much the number of passengers varies within each year for different regions. The size of the circles represents the amount of variance. We can see European travels to the US mostly during the summer. Whereas Asian travel to US in December.

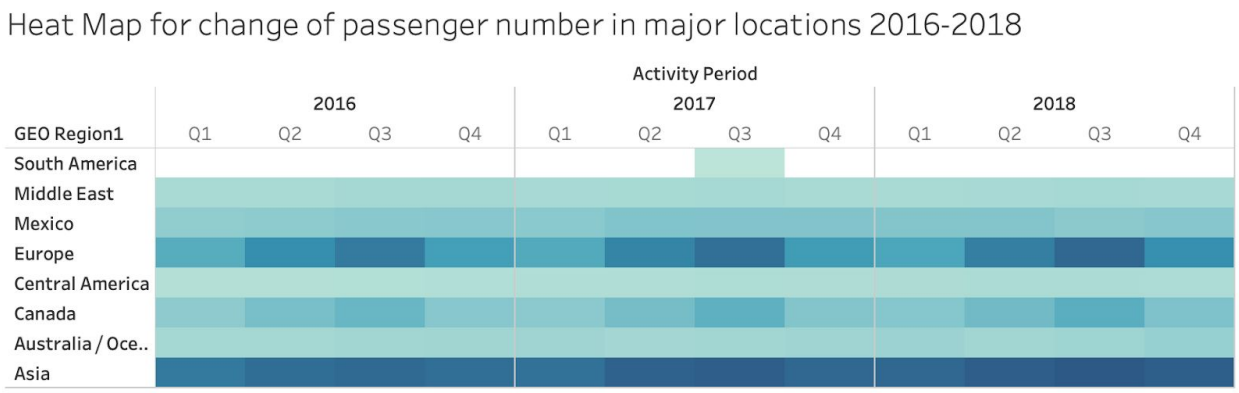


Choropleth Map (interactive in Tableau)
This map shows the for Major destination within top 6 regions, the relative number of passengers of each region travels to the US in 2018. It shows China is definitely the most popular destination for SFO.



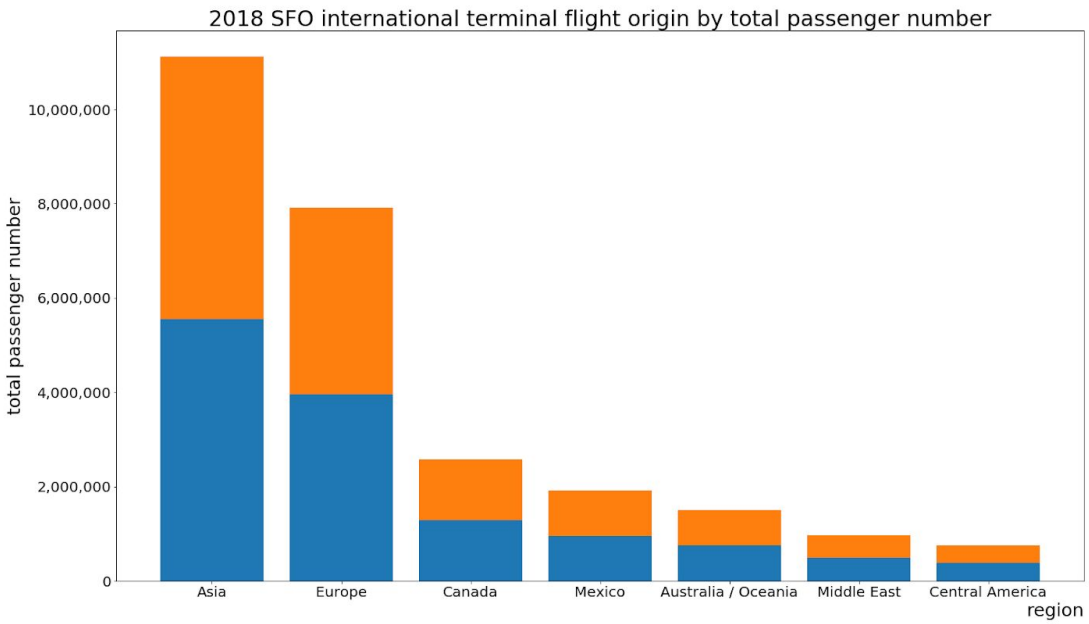
Connection Map (interactive plot in Tableau)
This plot shows the most popular destination in each popular country. And draw a connection line between SFO and the destination.

Heat map (interactive in Tableau)
This plot shows the change in the number of passengers within the top 7 most popular regions for each quarter from 2016 to 2018 for SFO. Q3 is definitely the busiest quarter and Asia is the most popular region.

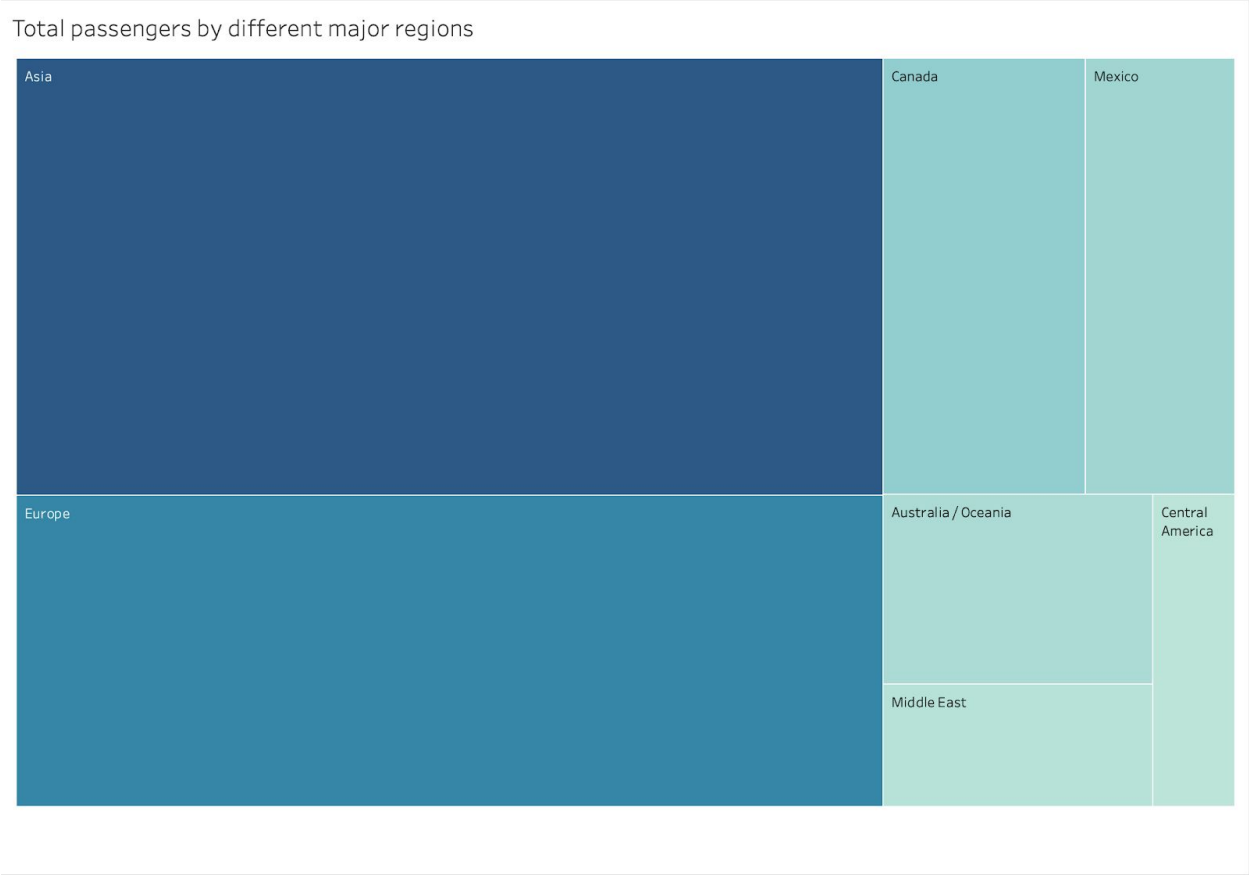


Stacked area

This plot shows the number of passengers leaving from SFO and arriving at SFO for 6 most popular regions, in 2018. It shows each region has a roughly similar arrival to departure proportion.

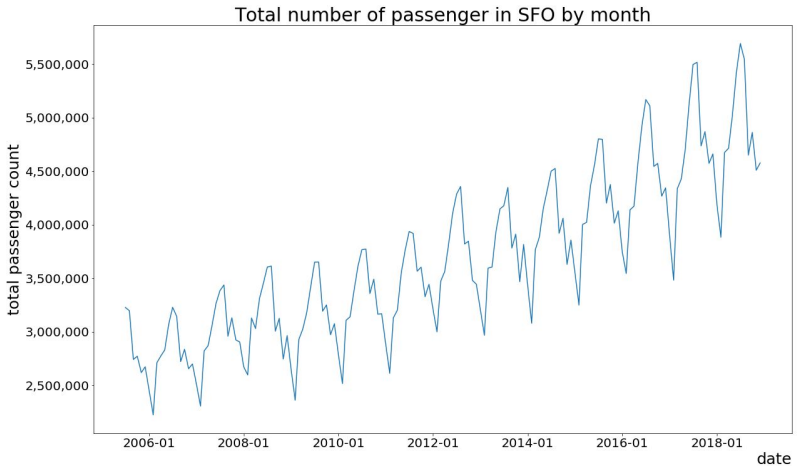


Treemapping (interactive in Tableau)
This shows the relative proportion of the total passengers from each of the region travelling from SFO. We can the size of Asia is almost as the rest(excluding Europe) combined.

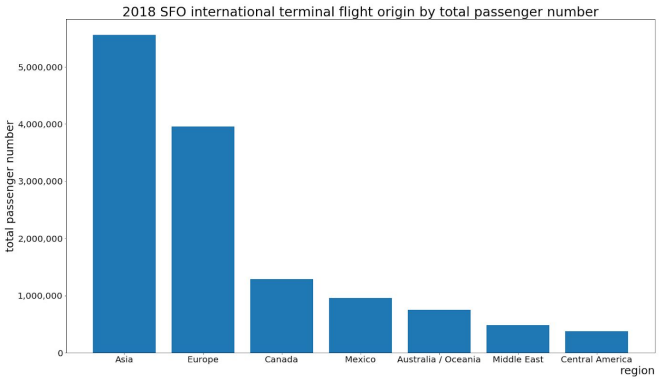


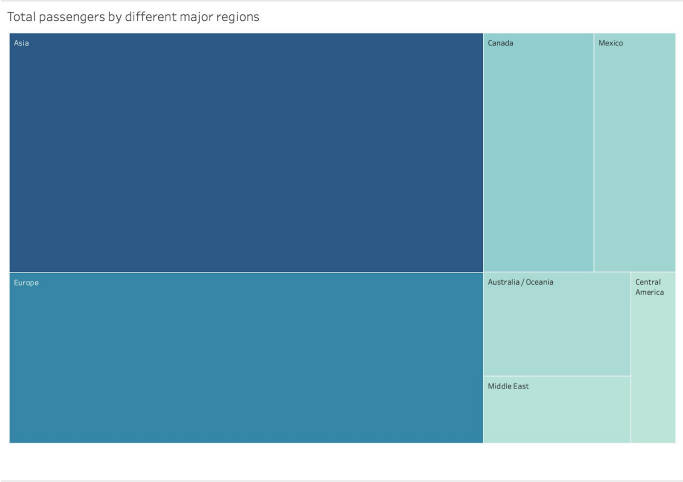
Storyline

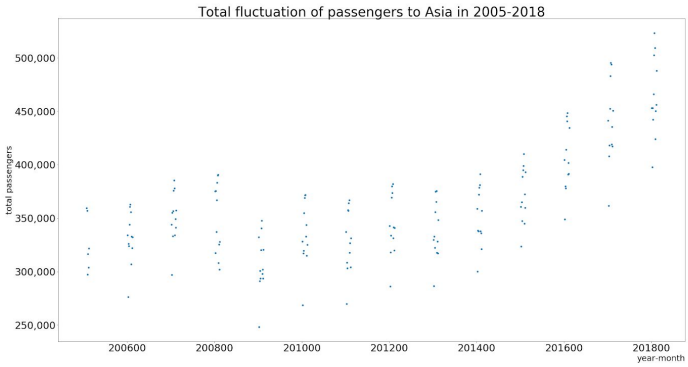
Firstly, we can see from the line plot, there is clearly a trend of increasing and an oscillating number of total passengers for each region. And it makes us wonder, is this true for all regions and for all airlines?

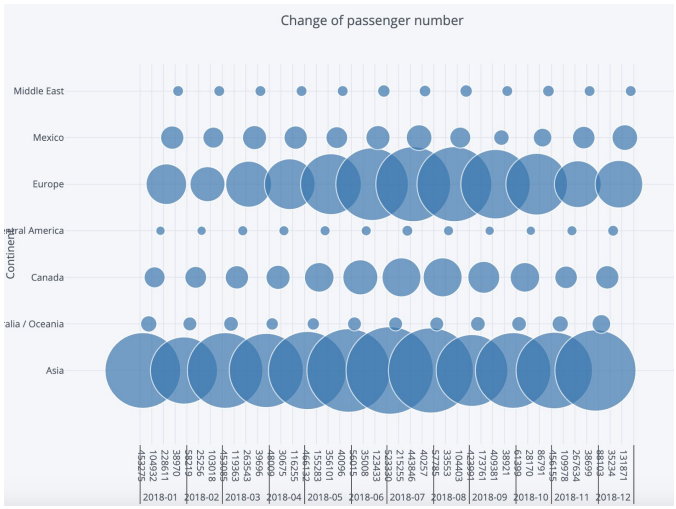


Then we look at the barplot, the treemap, Bubble map we found out. Yes, but not quite.
Asia dominates the flight market for SFO, which is almost all else combined(exclude Europe).
But the trend is not monotonically increasing for all regions. Like Europe reaches its peak around June-July and then decreases.
Whereas, Asia reaches its peak at the November-December period.

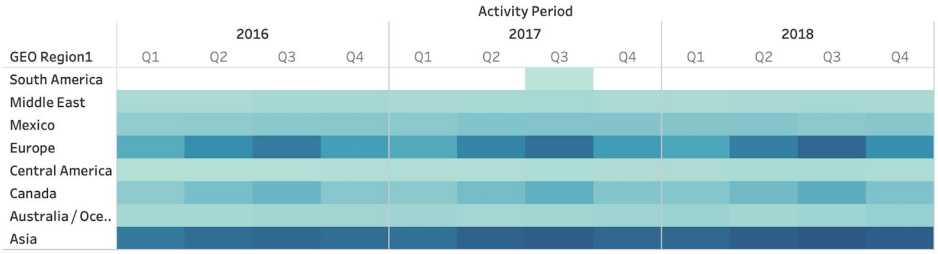


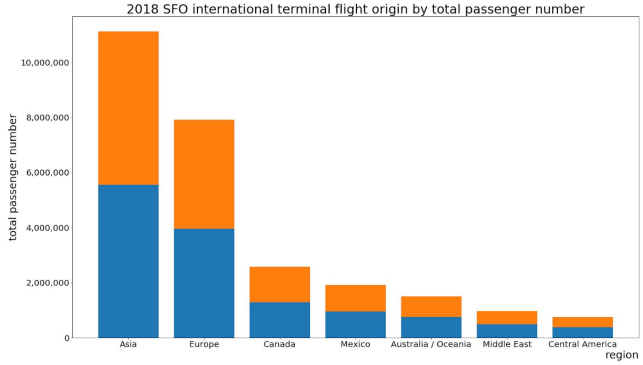




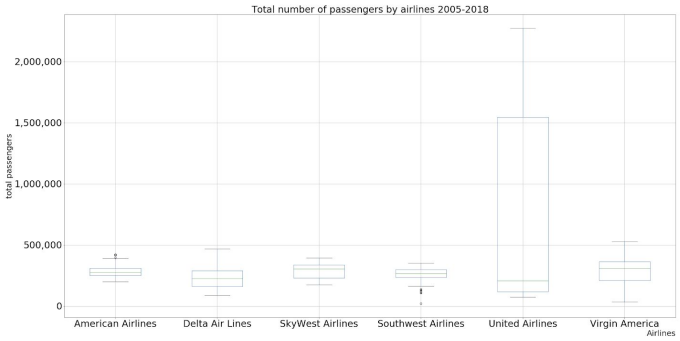


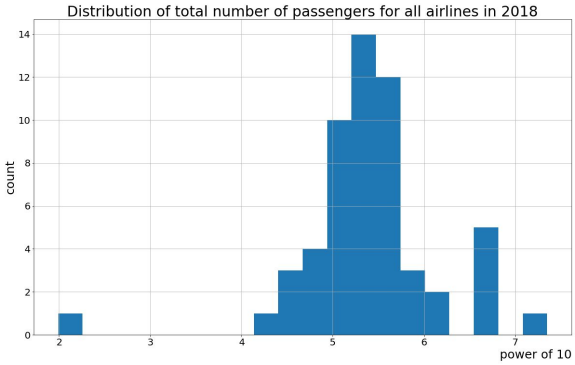
Heat Map for change of passenger number in major locations 2016-2018



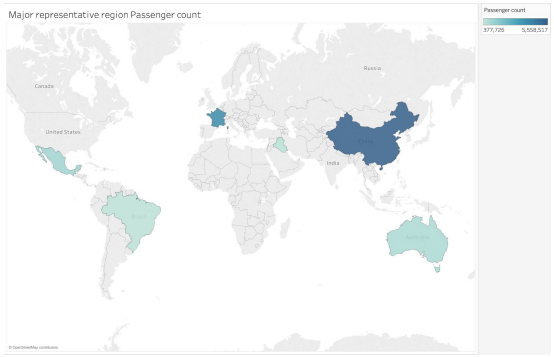


Furthermore, we look at the number of passengers travelling for each airline in histogram and boxplot. And we can discover that indeed some of the airlines are much more dominating than the other airlines.
But also, the difference is only extreme to some certain values because there are some outliers in the dataset which makes the total looks dramatic. If we take a look at the individual ones, it is clear that the distribution almost always has their median around the same values.





So finally we can take a look at the choropleth maps, connection maps and enjoy the beauty behind it. These plots are artworks that show us the dominating destination for each regions travelling from SFO



Conclusion

SFO data is a very exciting dataset containing all information from 2005-2018. I found out about the interesting patterns buried in the seemly noisy number. And saw the clearly time series trend-oscillation pattern.

I always know there is a significant number of Asians living in the Bay area. But there it is relatively small, about 10%. Only after reading and digging into this dataset. I found out that The largest proportion of the flights are directed to Asia.

Another interesting point to notice is that the majority of Asian travellers and European travellers exhibit different habits. European travels the most to SFO in the June-July period whereas most Asian travellers come to SFO during the November-December period of time.

The last interesting point we found is that Although if we look at the number of passengers by Airlines, United seems to dominate all other airlines. But if we dig deeper, we can that the majority of the flights from United are similar to the other major airlines, Only a few times of the year they transported much more than the other airlines.

That being said, Asia is a huge potential for SFO and I believe SFO should pay more attention to its Asian customers.

Appendix for code

Please see the code in this GitHub
https://github.com/ByronHan333/dataviz_final

Link to GitHub

https://github.com/ByronHan333/dataviz_final

Citations

https://plot.ly/python/
https://catalog.data.gov/dataset/air-traffic-landings-statistics
https://pandas.pydata.org/
https://scikit-learn.org/