BME/ECE 6850 Pattern Recognition

The progress of the final projects

Zeyu Liu

In the previous final project's proposal, I chose a public dataset named "New York City Airbnb Open Data" from Kaggle. However, with more analysis of this dataset, finally decided not to use the dataset anymore. Because this dataset has been used by too many people, the people in network has very deep analysis of every dates and parameters in this dataset, and repeated the same analysis is meaningless and boring.

So, in an occasional chance, by attending TDI challenge (the data incubator challenge in November 4), I got a new dataset named "Arrest Data from 2010 to Present", This dataset provided by Los Angeles Police Department and has been continuously updated. I chose the dataset which update date is 10/31/2019, this dataset contains 1.3 million of arrest dates, each data contains 17 parameters. The detailed description can be below Found in the URL.

The dataset can be download from the websites: https://catalog.data.gov/dataset/arrest-data-from-2010-to-present

The information of the dataset: https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-Present/yru6-6re4

Now let me explain the progress of my work. I preprocessed the data at first. For example, use the na.omit() function to delete the default value. Because the data set is very large, it will be so slow to analyze the overall data. To make R works faster, I first chose a sample with a subset of the "2018" data. And then successfully solved the following problems.

1. How many bookings of arrestees were made in 2018?
2. How many bookings of arrestees were made in the area with the most arrests in 2018?
3. What is the 95% quantile of the age of the arrestee in 2018? Only consider the following charge groups.
4. What is the Z-score of the average age for each charge group. Report the largest absolute value among the calculated Z-scores.
5. What is the projected number of felony arrests in 2019?
6. How many arrest incidents occurred within 2 km from the Bradbury Building in 2018?
7. How many arrest incidents were made per kilometer on Pico Boulevard during 2018?
8. Remove outliers by filtering out locations where either the latitude or longitude is 2 standard deviations beyond the mean of the subset of identified points.
9. Report the number of arrest incidents per kilometer on Pico Boulevard in 2018.
10. Report the average of the top 5 of the calculated ratios. Consider all records prior to January 1, 2019.

The goal for next step is to further visualize the data and present the data in a more understandable way. For example, display Arrest's coordinate data directly in Google Maps. And observe whether the crime location is regular by showing the Arrest location data of each year.