2. Data acquisition and cleaning

2.1. Data sources

Crime data:

We downloaded the crime dataset of most recent year of San Francisco and Chicago from DataSF (here) and DATA.GOV (here) respectively. The two websites provide constantly updated crime data recorded by local police departments. These data include the coordinates of police station, date of incident, crime description, etc.

Neighborhoods json file:

For San Francisco data, we used the Analysis Neighborhood defined for the purpose of providing consistency in the analysis and reporting of socio-economic, demographic, and environmental data, and data on City-funded programs and services, whose json file can be downloaded from here.

For Chicago data, we downloaded the json file from here, which is a repository on Github that records the method to transform coordinates into according neighborhood name in Chicago.

Venues data:

The venues data were generated from the Foursquare API. The data include the popular spots and their categories in each neighborhood.

2.2. Data wrangling

According to our goal, we only need the categories of crime, description of crime and coordinates of the police station where the incident was reported to. These data correspond to the column 'Category', 'Descript', 'Y' and 'X' in San Francisco's crime data, and 'PRIMARY DESCRIPTION', 'SECONDARY DESCRIPTION', 'LATITUDE', and 'LONGITUDE' in Chicago crime data. Therefore, we extract these columns and drop the rest.

The only clues to check if a crime belongs to credit card fraud are its category and description. We found that credit card fraud should belong to 'FRAUD' and 'DECEPTIVE PRACTICE' in San Francisco and Chicago crime data respectively. Afterwards, we look further into their criminal descriptions and found the most related descriptions are 'CREDIT CARD, THEFT BY USE OF' and 'CREDIT CARD, THEFT OF' for San Francisco data and 'CREDIT CARD FRAUD' for Chicago data. After dropping the data with invalid coordinates (i.e., incomplete or not located in San Francisco or Chicago), there are 23068 and 4537 credit card fraud crimes in San Francisco and Chicago respectively.

To find out the neighborhoods with exclusively high risk of credit card fraud, we calculated the credit card fraud rates of each neighborhood as number of credit card

fraud of a neighborhood divided by the total number of credit card fraud in San Francisco or Chicago. We then drew a choropleth of San Francisco data and found that some neighborhoods did have exclusively higher credit card fraud rates than others (Fig. 1). We labeled the 'outliers' of credit card fraud rates of neighborhoods as outliers and the other as 'average'. There are 4 (out of 41) and 6 (out of 78) neighborhoods labeled as outliers for San Francisco and Chicago respectively (Fig. 2). The descriptive statistics of credit card fraud rates of neighborhoods are shown in Table 1.

We used the explore function of Foursquare API to get the top 100 popular venues of each neighborhood within the radius of 500 m from the centroid of the neighborhood. Because the categories returned by the explore function could be too specific to serve as the feature of a model of broad use. Therefore, we looked up the primary category of each category, and replace the original categories with their primary categories. According to the Foursquare website, there are 10 primary categories in total, including 'Arts & Entertainment', 'College & University', 'Event', 'Food', 'Nightlife Spot', 'Outdoors & Recreation', 'Professional & Other Places', 'Residence', 'Shop & Service', 'Travel & Transport'. Lastly, we used the venues data of neighborhoods to calculate the percentage of each primary category of each neighborhood. The metric of San Francisco is then used for training model, and the metric of Chicago is used for evaluating the performance of the model.

Table 1. Descriptive statistics of credit card fraud rates of neighborhoods in San Francisco and Chicago (unit: %).

	San Francisco	Chicago
Mean	2.44	1.28
Std	2.88	2,36
Min	0.01	0.00
25% quantile	0.76	0.40
50% quantile	1.54	0.77
75% quantile	2.62	1.27
Max	13.03	18.17

Fig 1. Boxplot of credit card fraud rates of neighborhoods in (a.) San Francisco and (b.) Chicago.

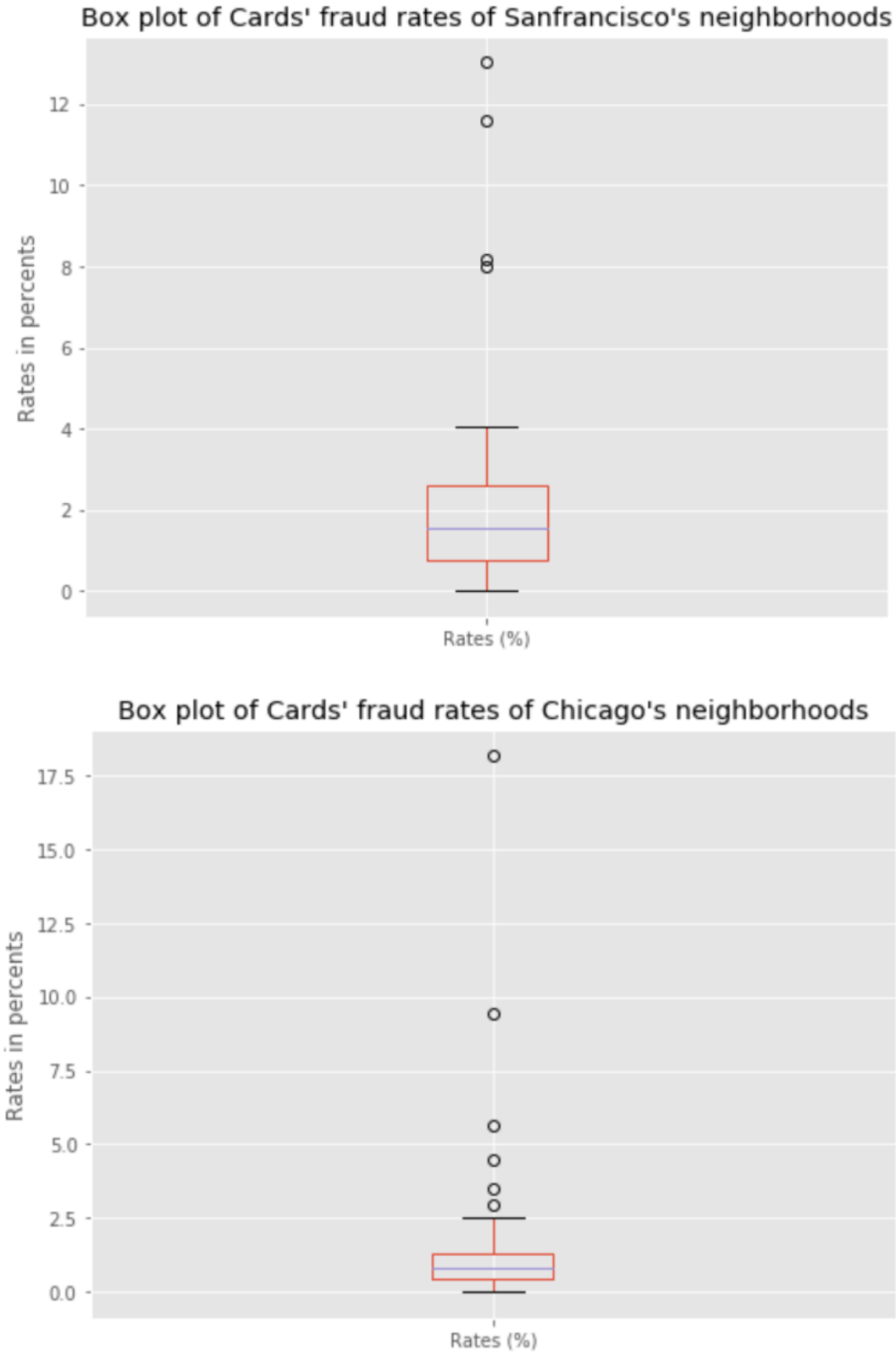

Fig 2. Choropleth of credit card fraud rates of neighborhoods in San Francisco.