Predicting the crime rate of Chicago

Jingwei Hu

August 1st 2019

1.Introduction

Assume that we want to run our businesses in the city of Chicago, we would concern about the crime rate of the possible site.

In this project, I want to figure out the factors that would have an impact on the crime rate in the city of Chicago. And use those factors to access the safety level.

As we know, many factors might have an impact on the crime rate whether positive or negative. For example, the poverty rate or household income will affect the crime rate. Murders or kidnappings are more likely to occur in outdoor places like a park or forest. Thefts would prefer appearing in a transportation center rather than a garden.

I will use the power of data to explore the relationships between those factors. So we can pay more attention to the place where would have less criminal events.

2. Data acquisition and cleaning

2.1 Data sources

Based on the definition of our problem, factors that will influence the crime rate in Chicago are:

* Household density, unemployed rate, average incomes in the neighborhood.

* All the venues and its category in the neighborhood.

I decided to use the 5-digit zipcode and its boundary to define the neighborhoods. Following data sources will be needed to extract/generate the required information:

* zip codes and its boundary from geospatial datas.

* census tracts' boundary and the identifier of the city. This would be used to align data from different sources.

* venues and their type and location in every neighborhood will be obtained using Foursquare API

* the category hierarchy of venues will also be obtained from Foursquare API

* the demographic data includes civilian, labor force, unemployment, household, and income will be obtained from US government official website.

* the crime data occurs in the city of Chicago will be collected from this source.

2.2 Data cleaning

First, parse zip codes and its boundary from geospatial file which was in the form of JSON, then parse census tracts and its centroid and boundary from KML files, evaluate the area of each tract. Finally, align them use the boundary and location, merging census tracts in the same zip code area. Now we get the census tract and zip code information.

We obtained venues information around each census tract's centroid, within three kilometers radius. Then, we dropped out the duplicated venues and assign the parent

category and zip code to each venue. We finally collected almost four thousands venues in Chicago. Then we group those venues by its zip code and count them into ten super categories. Now we obtained venue categorical information in each zip code area.

The crime data in 2018 was filtered, and remove the rows without a location. And then assign a zip code according to its latitude and longitude. At last, we group them by zip code and count the number of crimes occurred in each zip code area as well as the number of street crime that occurred in those areas.

Because it’s hard to find demographic data in 2018, So I chose the data in the year 2017. Multiplying the mean household incomes and the number of households to get the total incomes in a census tract. Grouping those data by zip code, and calculate the mean household incomes in each zip code area. The same as the calculation of the unemployed rate, using the civilian labor force and unemployed labor force data.

3. Methodology

In this project I will use crime count, venue proportion, unemployed rate and income of each zip code area to predict the crime rate in that area.

In first step above, I have collected the required data.

Second step in my analysis will be calculation and exploration of 'relationship' between different factors and the crime rate. I would use the different graph to show their relation.

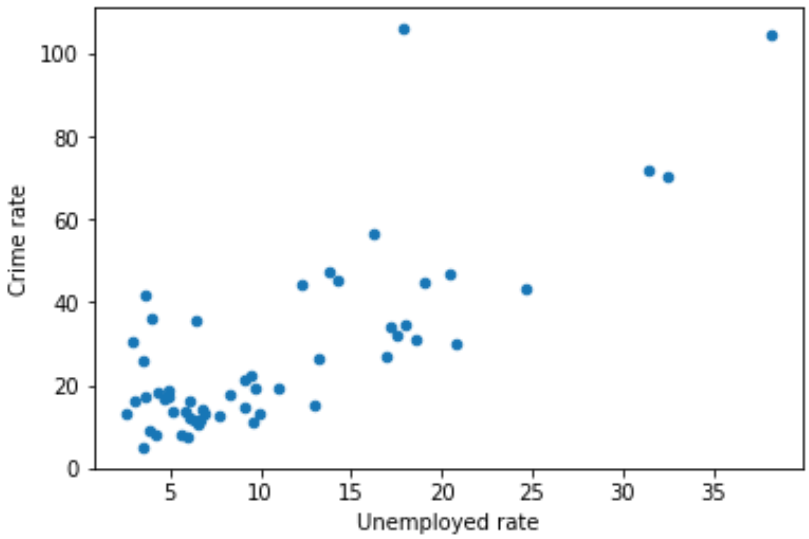
In third and final step we will focus on using model to predict the crime rate and discuss that how different factors influence the crime rate.

Are their correlations linear? To what extend they influent each other, and how to choose and transform the right feature.

4.Analysis

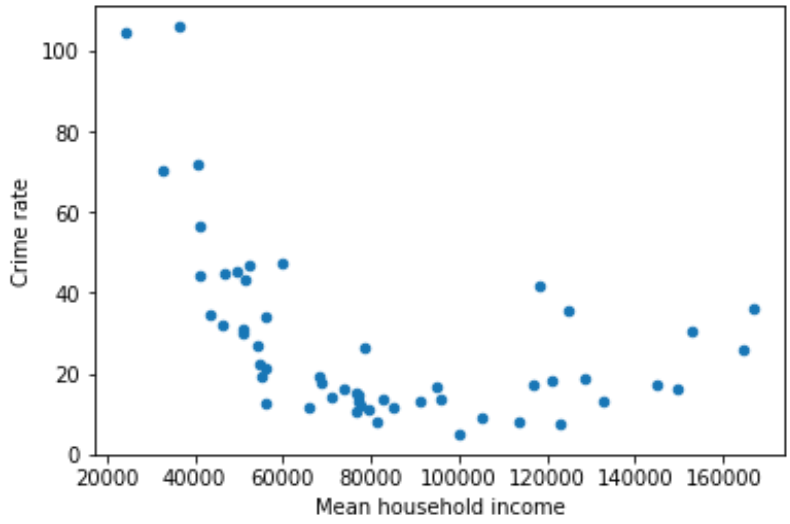
4.1 To prepare the required data, merging data frames into a single one using the zip code.

4.2 Unemployed rate positive related to the crime rates.



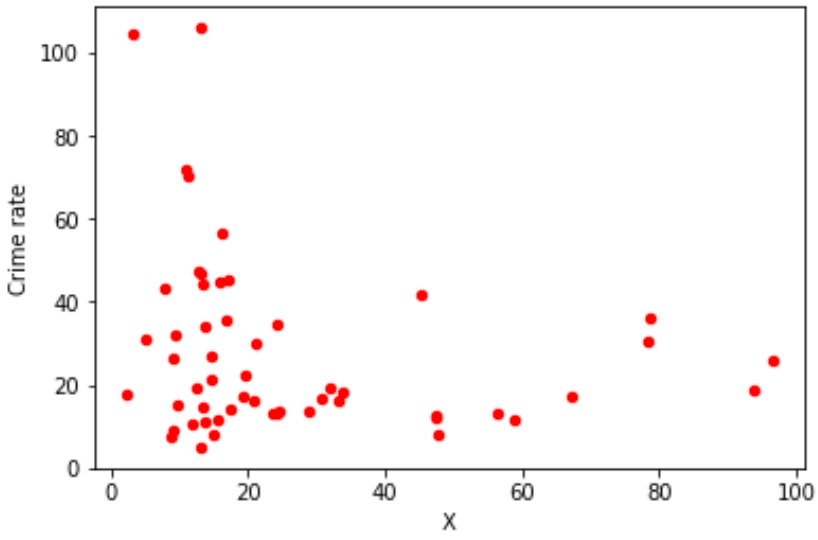
The Pearson correlation coefficient between crime rate and unemployment was 0.772. So there is a linear-like positive correlation between them.

4.2 Income negative related to the crime rates.



We can tell from the graph above that there is a strong relation between the mean household income and the crime rate, but we also found that this correlation is a non-linear relation, like a bowl.

4.3 Household density negative related to crime rate



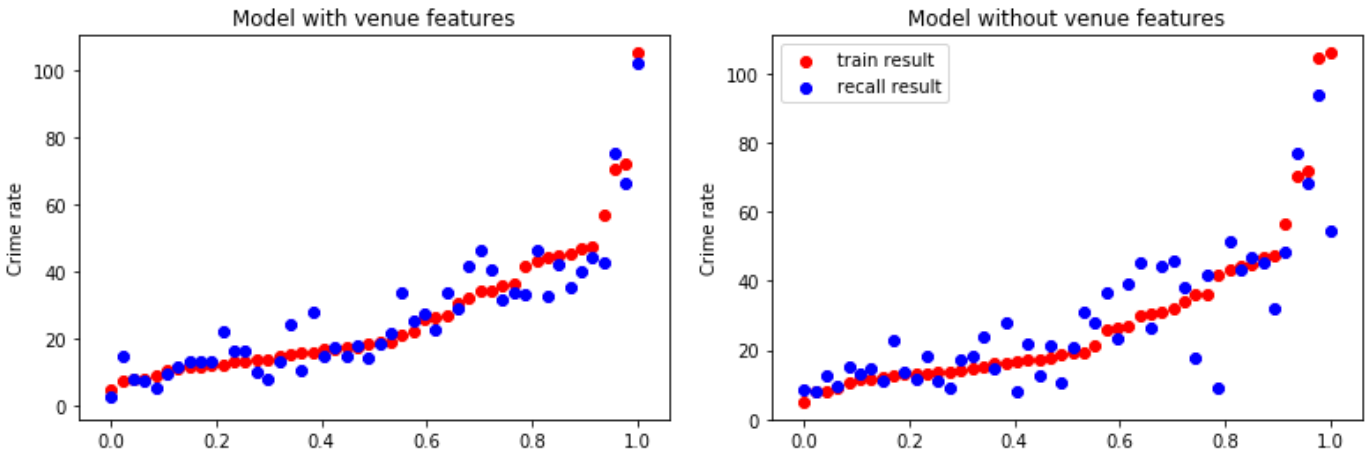
Compare with the unemployment rate or income, the relation between household density and crime rate seems weaker than those above. But in the area with high household density, the crime rate is significantly lower than other places.

5. Predictive Modeling

Since not all the relations are linear, and the number of data points are less than 100, so I choose curve_fit method to fit my data.

At first ,I picked only unemployed rate, household density, and mean household income as the feature. Then I added the venue category proportions and venue density to the model.

R2 Score without venue features	0.74
R2 Score with venue features	0.92



According to the graph we can see that venue features help us to refine the model.

5. Results and Discussion

The result shows that unemployed rate have the strongest negative linear impact on the crime rate, and the influence from household income is a little bit different. It seems that when an area is poorer it would have higher crime rate, meanwhile when an area is richer it is also more likely to have criminal events. But the slope of the two directions are not the same.

Looking at graphs above in more detail, I initially thought venue datas did not effect the crime rate. But when I tried use different features to predict the model, I found that the venue data could help to refine the model.

6.Conclusion

Purpose of this project is to predict the crime rate. Since we are more likely to run our own business or investment in an area with lower crime rate, we would be interested in the factors that influence the crime rate. By refine and visualize the data that we collect from the Foursquare API and government census website, we find the relationship between different factors, and get a simple and rough model to predict the crime rate.

If we want to compare different possible sites to run our own businesses, we can evaluate their crime rate and compare them.