Capstone Project - Analyzing King County Crime using Geographic Data

Part I Introduction

The main goal of this project is to use location data from FourSqure to analyze King County Crime. Crime rate has increased significantly for the past few years in King County, WA due to drug abuses and pricey housing. The ideas is to see if we could estimate the crime based on the location characteristics such as venues around the neighborhood. For example, we want to answer question like would crime more or less likely to happen in neighborhood with more venues of certain type? The report could be used for the following purposes for example:

- House buyers or renters who want to settle in relatively safe neighborhood can use this as a rough guidance by examining the surrounding venues.
- Optimize allocation of police patrol resources based on surroundings in the locations.
- Optimize use of County funds to improve public safety.

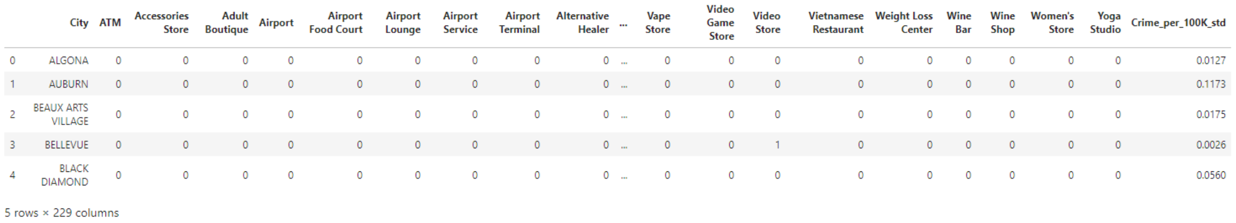
Part II Data Description:

There are four main data sources used in this project.

- **King County Population by City**: http://worldpopulationreview.com/us-counties/wa/king-county-population/
- **King County Sheriff's Office Crime Data**: https://moto.data.socrata.com/dataset/King-County-Sheriff-s-Office/4h35-4mtu
- **Simplemaps Geographic Data**: https://simplemaps.com/data/us-cities
- **FourSqure API Location Data**: http://www.foursquare.com

The population data is mainly used to normalized the crime data. The latest estimated population available by city within the county is 2016. We then extract the the population growth for 2017 and 2018 to project the population by city to 2018. The Crime data is at the incident level. We aggregate it to the city level to calculate the number of crimes and then merge it with the Population data. The Geographic Data has the list of U.S. cities with their latitudes and longitudes. We then attach it to the dataset and finally we bring in the venues data from FourSqure using the lat/long information.

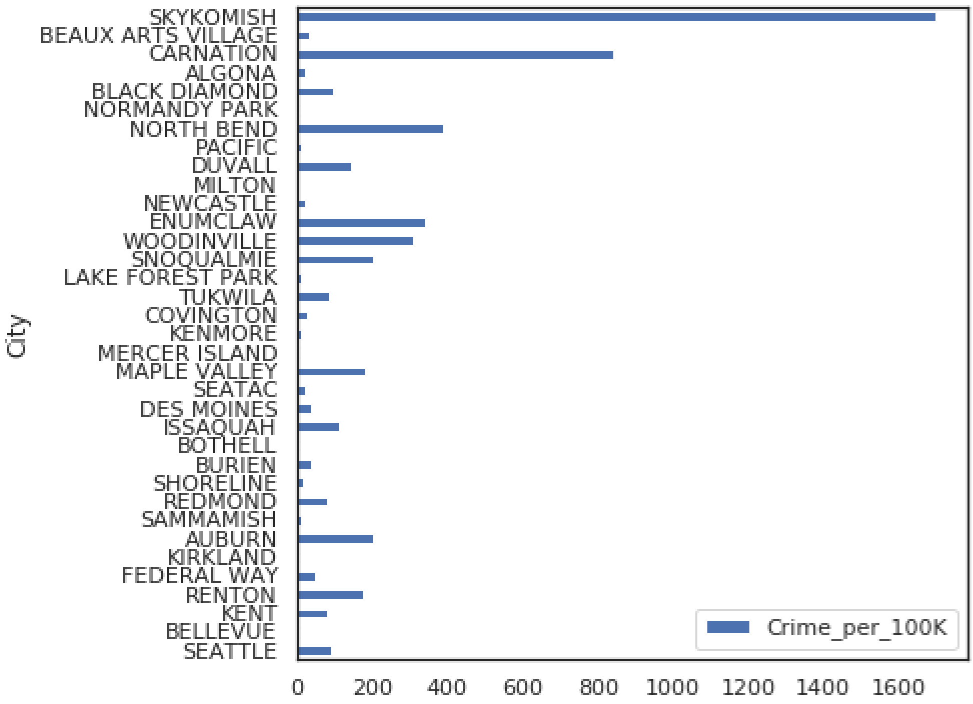
After completing the cleaning and preparations above, our final dataset looks like the following:



	City	ATM	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	Alternative Healer	...	Vape Store	Video Game Store	Video Store	Vietnamese Restaurant	Weight Loss Center	Wine Bar	Wine Shop	Women's Store	Yoga Studio	Crime_per_100K_std
0	ALGONA	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0.0127
1	AUBURN	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0.1173
2	BEAUX ARTS VILLAGE	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0.0175
3	BELLEVUE	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0.0026
4	BLACK DIAMOND	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0.0560

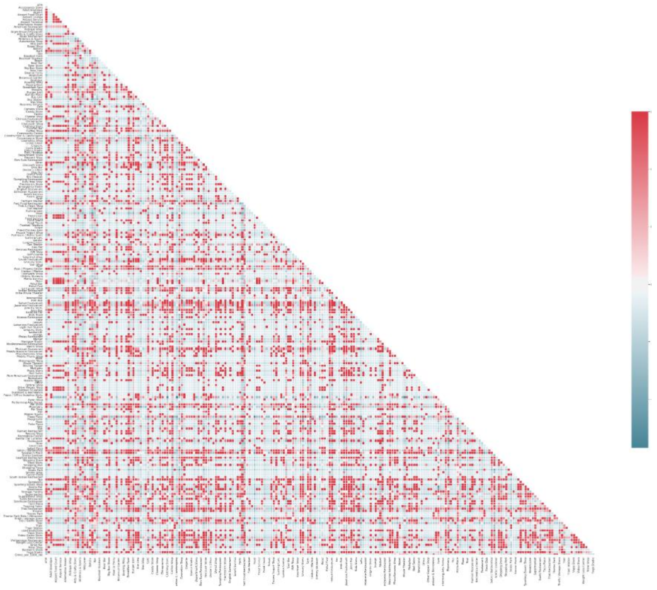
5 rows × 229 columns

The dataset contains 34 samples and 229 features. From here we could see which city has the most crime per 100K population, which is a pretty standard way to measure crime rate.



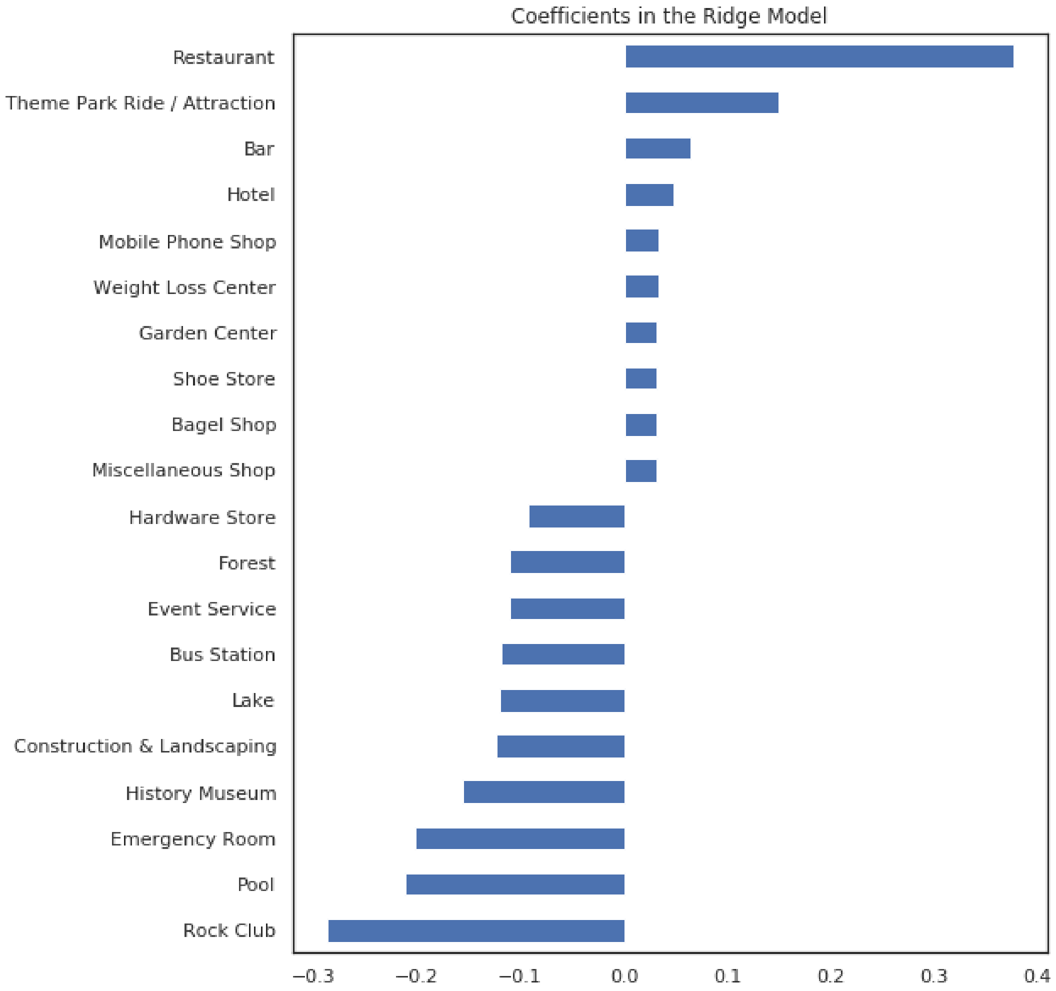
Part III - Models

Before we start building any models, we want to visualize the correlation matrix on the heat map. The low correlation between the features and the target variable is mostly due to sparsity in the dataset.



Because of the fact that the number of features are way more than the number of observations, we here explore two method Ridge regression and Principle Component regression to analyze the crime data. These are known to work well for dimensionality reduction.

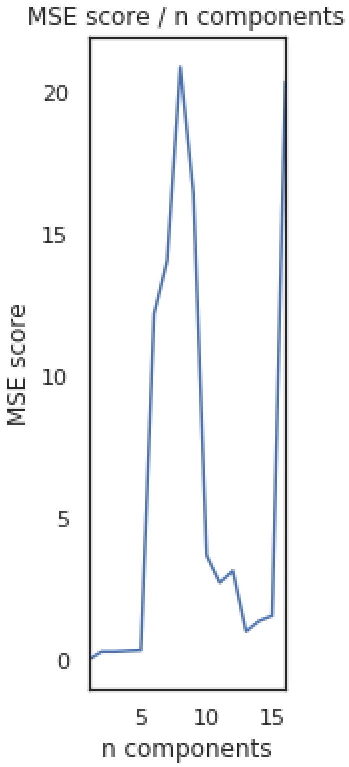
For Ridge regression, we apply a cross-validation to choose the tuning parameter alpha. The code below yields the alpha value that results in the smallest CV error. In this case, the alpha selected is 0, which is equivalent to least squares method. We then take the optimal alpha and refit the model to obtain the coefficients. Below are the top 10 most positive/negative coefficients.



The coefficients are quite intuitive. We could see crime happen more frequently in neighborhoods that has more venues that people visit on daily basis.

We use the mean squared error (MSE) for model comparison. The MSE for Ridge regression is 0.0547.

We next apply the PCA to the dataset to obtain the principle components. Similarly, we also apply a CV procedure to determine the best n components to include in the secondary linear regression model.



The MSE of the optimal PCR model is 0.0569, which does not seem to be better over the Ridge regression above.

Part IV - Conclusion

Crime rate seems to be correlated with the venues data. We have tested Ridge regression and Principle Component regression on the dataset. Ridge regression seems to be better in terms of prediction. In addition, Ridge regression also yields interpretable results where we could see the most predictive features associated with crime rate. From the list of variables by Ridge regression, it's not surprising that venues that people visit on daily basis will have more crimes, whereas places that people don't visit regularly (bus station) have less crimes.