ANALYZING KING COUNTY CRIME USING GEOGRAPHIC DATA

IBM DATA SCIENCE CAPSTONE PROJECT

Analyzing Crime Rate Using Location Data is Valuable for

- 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.
- … and more

Data Source

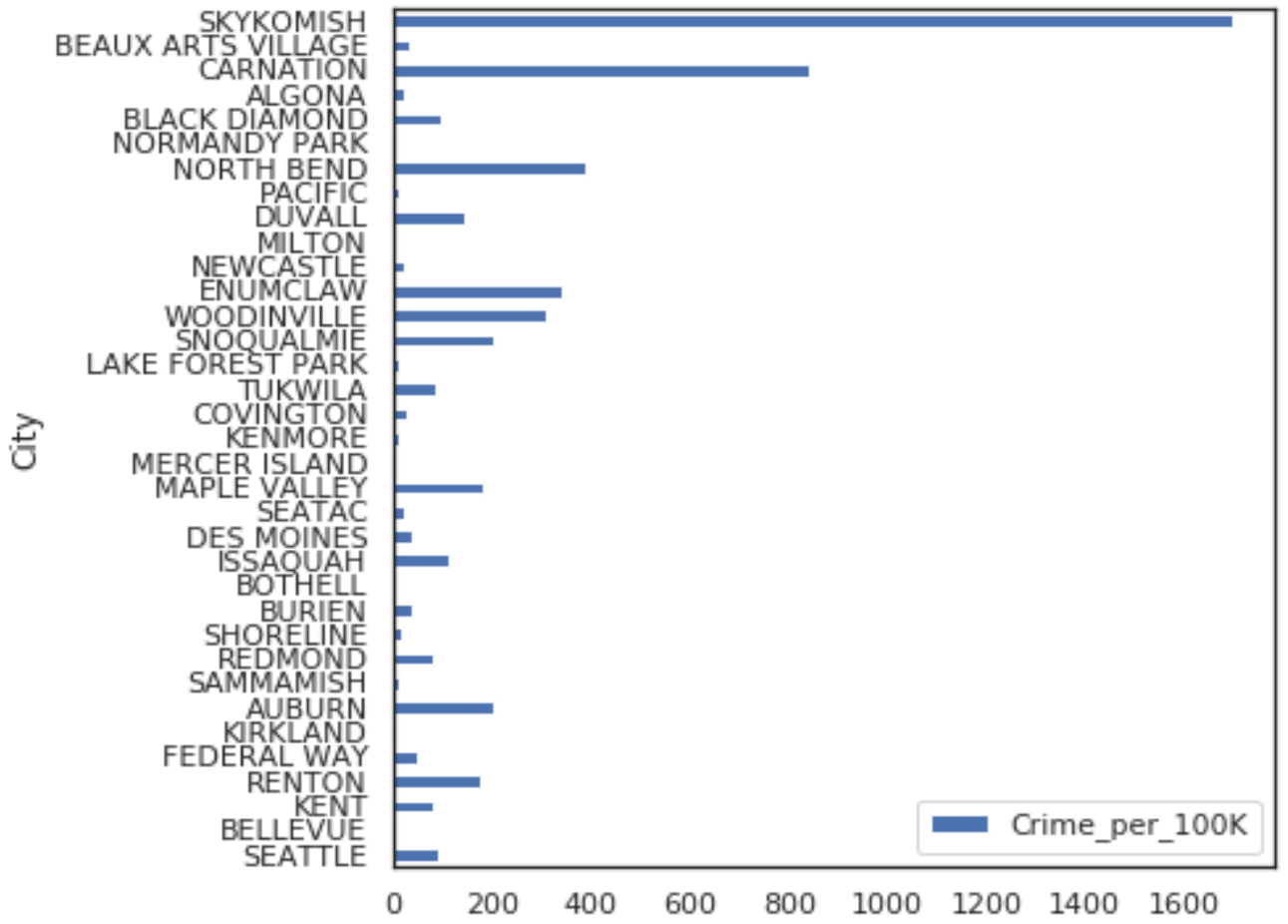
- **King County Population by City**: http://worldpopulationreview.com/us-counties/wa/king-county-population/
- **King County Sheriff's Office Crime Data**: https://moto.data.socrata.com/dataset/King-County-Sheriff-s-Office/4h35-4mtu
- **Simplemaps Geographic Data**: https://simplemaps.com/data/us-cities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

Final Dataset after Cleaning

	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

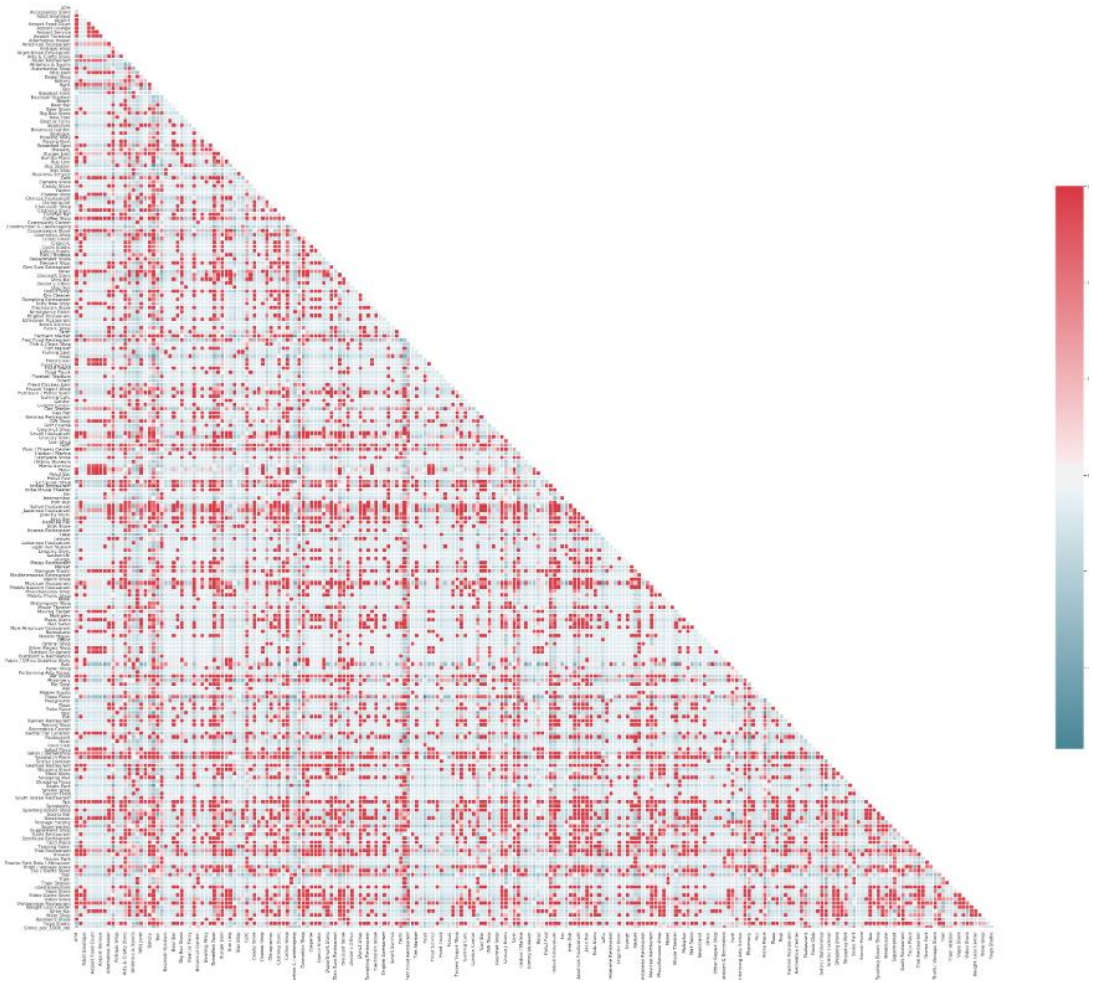
Crime per 100K Population by City

- Incidents: 2019 Jan – August
- Population: growth projected to 2018



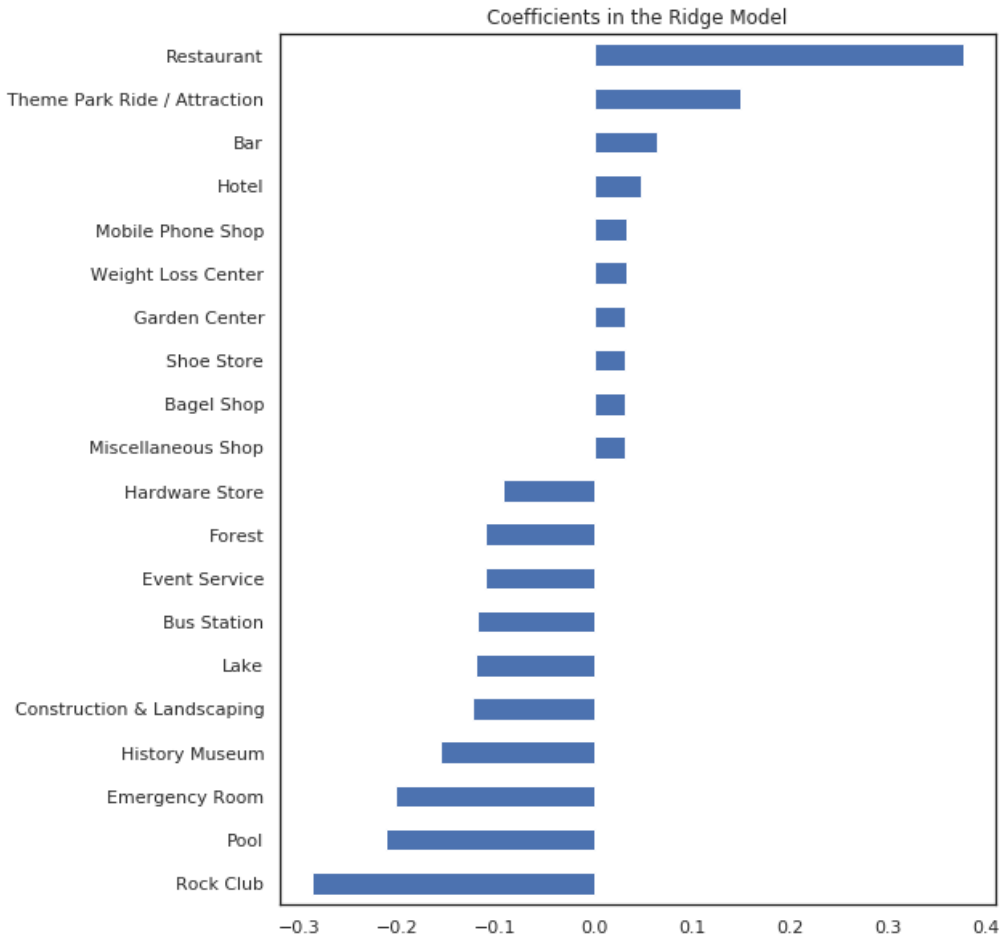
Correlation Matrix Heatmap

- Low correlation between features and target due to sparsity
- Some features are strongly correlated



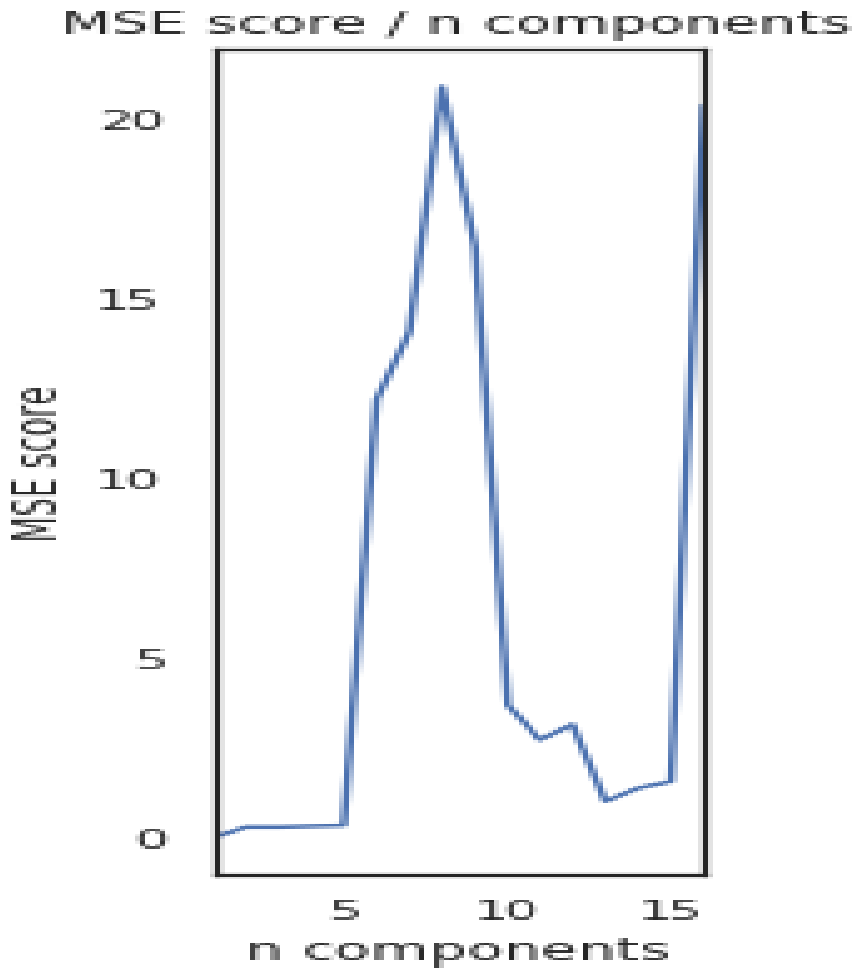
Ridge Regression

- Cross Validation applied to determine the optimal tuning parameter Alpha
- Most positive/negative coefficients: crime happen more frequently in neighborhoods that has more venues that people visit on daily basis
- Model MSE: 0.0547



Principle Component Regression

- Cross Validation applied to determine best n-components
- Model MSE: 0.0569



Conclusions

- Crime rate seems to be correlated with the venues data
- Ridge regression seems to be better in terms of prediction
- Ridge regression also yields interpretable results where we could see the most predictive features associated with crime rate
- Venues that people visit on daily basis will have more crimes, whereas places that people don't visit regularly have less crimes