Popular neighborhoods:

aka. why do people choose to live in certain neighborhoods?

Two questions:

- Can we predict neighborhood population density from geographical, population, and socio-economic factors?

- Are people moving in or moving out of a neighborhood?

Data sources:

- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M allows us to group Toronto neighborhoods by postal code

- http://cocl.us/Geospatial_data allows us to get the geographical latitude and longitude of the postal code

- https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods allows us to find the population, income, ratio of people renting and commuting: i.e population attributes of Toronto neighborhoods

- Finally, Foursquare gives us socio-economic attributes: i.e venues providing different facilities: shops, professional opportunities etc for different post codes and its associated neighborhoods

Data

- Associated with each postal code (and its associated neighborhood) we compile data related to:

A. Location: Distance to Downtown in Kms

B. Population: density, average income, ratio of people commuting and renting apartments, change of population from 2001 to 2006

C. Socio-economic: area averaged number of venues in each postal code in broad categories

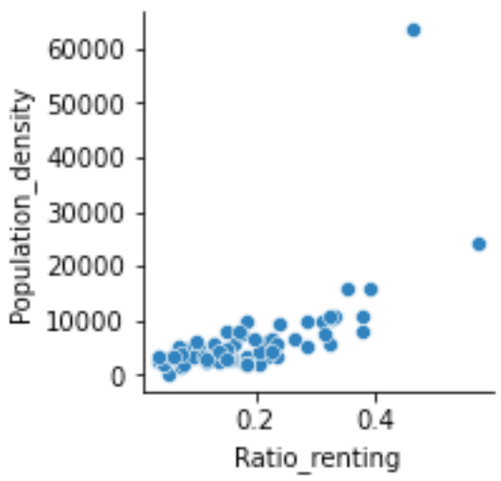
Question 1

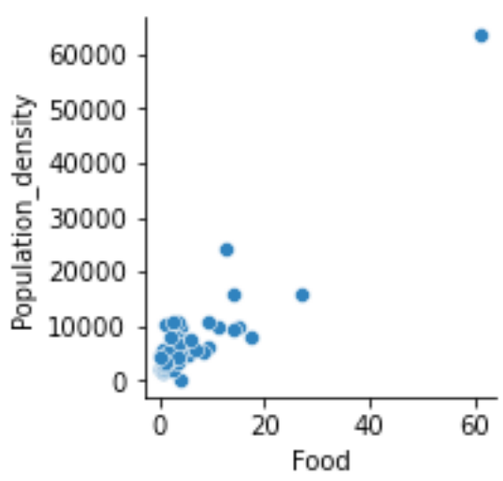
- Can we predict neighborhood population density from geographical, population, and socio-economic factors?

- Approach: Using carefully designed multiple linear regression to estimate pollution density

Exploratory analysis: part 1

- Population density shows clear linear (more generally, monotonic) relationship with some of the feature variables, eg. Ratio of people renting, or number of Food places per sq.km in that postal code.





Exploratory analysis: part2

- Correlation (although dominated by outliers) can still give a good idea that population density is moderately to strongly correlated (>0.6) to quite a few of the features.

```
Travel                          0.813121
Outdoors                        0.474614
Food                            0.909420
Arts                            0.188136
Shops                           0.654445
Residence                       0.852728
Professional                    0.905675
Distance_to_Downtown [Km]      -0.441215
Ratio_commuting                 0.581641
Ratio_renting                   0.642328
Average_income                 -0.149398
Population_density              1.000000
Population_change_ratio        -0.058450
```

Exploratory analysis: Part3

- Fitting a linear model quantifies these relationship further.

OLS Regression Results

Dep. Variable:	Population_density	**R-squared (uncentered):**	0.972
Model:	OLS	**Adj. R-squared (uncentered):**	0.966
Method:	Least Squares	**F-statistic:**	174.6
Date:	Mon, 15 Feb 2021	**Prob (F-statistic):**	1.83e-42
Time:	18:27:57	**Log-Likelihood:**	-645.35
No. Observations:	73	**AIC:**	1315.
Df Residuals:	61	**BIC:**	1342.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Travel	224.6775	354.001	0.635	0.528	-483.191	932.546
Outdoors	-86.0299	142.991	-0.602	0.550	-371.958	199.898
Food	450.6776	64.067	7.034	0.000	322.567	578.788
Arts	-1356.8667	394.281	-3.441	0.001	-2145.280	-568.454
Shops	40.3763	45.220	0.893	0.375	-50.046	130.798
Residence	585.8983	271.974	2.154	0.035	42.052	1129.745
Professional	234.4353	111.482	2.103	0.040	11.514	457.357
Distance_to_Downtown [Km]	1.9575	37.394	0.052	0.958	-72.816	76.731
Ratio_commuting	1190.7941	5532.669	0.215	0.830	-9872.462	1.23e+04
Ratio_renting	1.588e+04	3203.404	4.956	0.000	9470.953	2.23e+04
Average_income	-0.0017	0.005	-0.328	0.744	-0.012	0.009
Population_change_ratio	3210.5517	1640.275	1.957	0.055	-69.380	6490.483

Summary of the regression:

- A very high fraction of variance (R-squared) in the population density can be “explained” by the model

- The coefficients of some of the features are significantly different from zero (P>|t| being less than 0.05)

- Some of the features are correlated with each other (high condition number)

- We cannot discount the presence of additional features that are causal to some of the other features

We want to *predict* and not find the best model

- Since we want to *predict* and not find the best model we use scimitar learn to create this imperfect but workable model.
- Using this model we find consistent explained variance ***higher than 0.8 for the test set***

Question 2

- Are people moving in or moving out of a neighborhood?

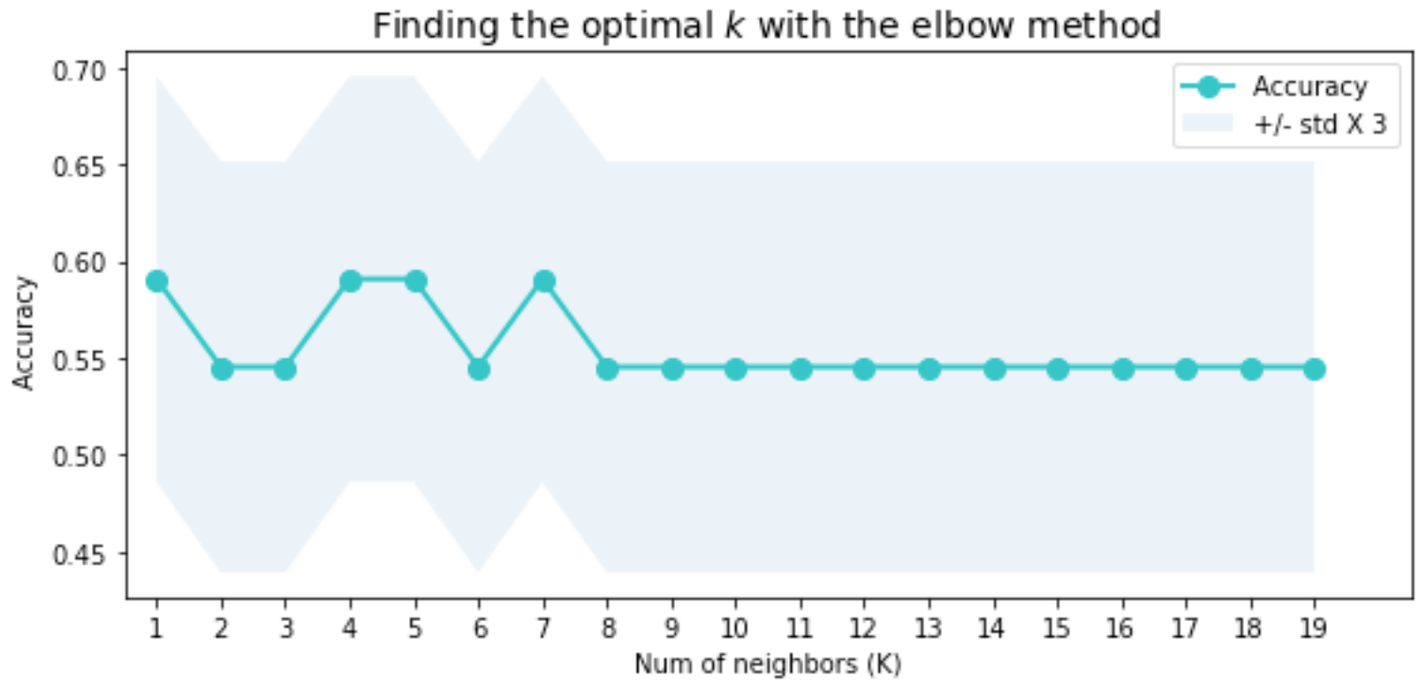
- Approach: Testing various approaches to classify whether people are moving in or out of neighborhoods: a binary classification problem

- Positive change from 2001 to 2006 is labeled 1 and negative change as 0

- Major problem is that positive change has somewhat smaller number of occurrence than negative change

Approaches

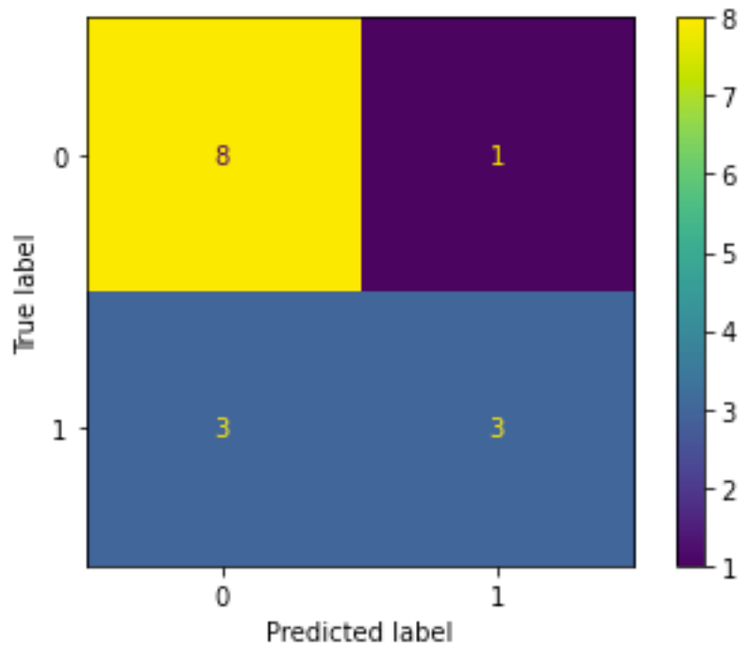
- KNN (after standardization of features) is quite ineffective at prediction in the test set



Approaches

- Decision trees and Logistic regression fare better with accuracy of between 0.7-0.8 after a best model (for the train set) has been found by grid search. Across samples, the false negatives causes issues with F1 score.

Decision Tree confusion matrix



Conclusion

- Neighborhood population density can be quite well predicted from geographical, population, and socio-economic factors
- Whether people are moving in or out of areas could not be very satisfactorily determined with this sample size