Applied Data Science Capstone, Week 2 Modeling Foursquare Restaurant Ratings

Paul Anderson

1. Description of the problem

For the final assignment of the Applied Data Science Capstone, I've decided to explore the problem of finding a suitable neighborhood to open a new restaurant in Toronto.

The solution of this problem is aimed towards a possible investor who has these traits:

- The investor has no particular interest in any neighborhood or type of restaurant.
- The investor is interested in the safest type and location of a restaurant that can be opened in Toronto.

The concept "safe" in this particular context means: finding the type of restaurant and neighborhood that would improve the chances of the restaurant being highly rated according to Foursquare.

To achieve the goal of figuring out the type and location of restaurant; I will try to model Foursquare restaurant ratings in Toronto. I decided to build this model because from previous observations of acquaintances who opened venues similar to a restaurant, I noticed they chose the business and location out of convenience and personal taste, not if the market values that particular combination of location and type of business. They chose the location because it was close to where they lived; the type of business was a long time ambition and/or saw similar businesses nearby.

Another reason I chose to build this particular model is to try to reduce some uncertainty when a person attempts to start a business, in this case a restaurant. Out of all the possible decisions a person can make, what are the decisions that can have a bigger impact on the success of the business? How is the success of the business going to be measured? I hope the approach I will use in this model can reduce the amount of uncertainty involved, by giving an approximation of possible outcomes derived from choosing a neighborhood and type of restaurant. In this case, success is going to be measured by the restaurant rating.

2. Data.

In order to build a model than can extrapolate the restaurant ratings in Toronto, I will use the following groups of data:

1. Average commercial rental cost in Toronto per neighborhood.
2. Population description of each neighborhood.

3. Restaurant traits and ratings of each neighborhood.
4. Geographic data of Toronto's neighborhoods.

2.1. Average commercial rental cost in Toronto.

This information will be obtained by using Selenium to scrape commercial rent webpages. The search will be based in Toronto, and no further analysis will be made to determine how appropriate the venue might be for a restaurant. The goal is to get a baseline of how much it costs to rent a commercial venue in each neighborhood of Toronto. Rental cost; aside from determining the location of the restaurant, might also determine the price tier the restaurant could have.

The code used to scrap the webpages can be seen here and here.

Follow this link to see how the rental information was put together from the scrapped data.

2.2. Population description of each neighborhood.

This information will be obtained from the 2016 Toronto Census, readily available from Toronto's government page. I will focus on information regarding the population of each neighborhood, age segregation, average income, population density. These traits may affect the type of restaurant that's preferred on each neighborhood. I will not take into account any race related statistics.

The way the census database was wrangled and the features I decided to use can be seen here.

2.3. Restaurant traits and ratings.

Foursquare will be used to obtain the restaurant's rating, categories and other relevant traits. The search will be made on a neighborhood basis, tweaking the search radius for each neighborhood to get the most possible amount of data related to restaurants.

I go into great detail at how I queried Foursquare following this link.

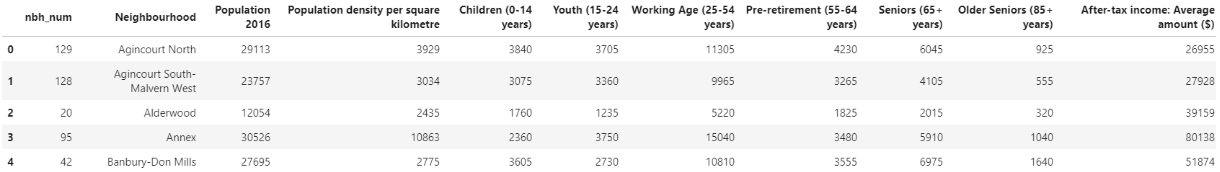
2.4. Geographic data of Toronto's neighborhoods.

This information will be obtained from Toronto's government page. It's a GEOjson file that describes the geometry of each neighborhood in Toronto. This information will be used to fine tune the search radius and to segment the rental venues locations.

3. Data Analysis

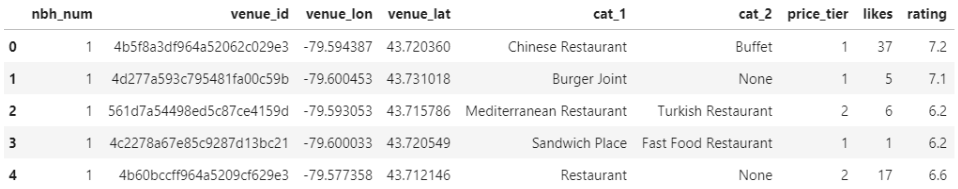
The result of the data wrangling is 3 data frames that contain the following information, the code used can be seen here.

toronto_merged.csv:



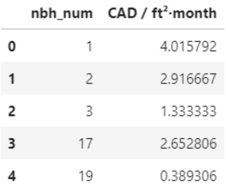
	nbh_num	Neighbourhood	Population 2016	Population density per square kilometre	Children (0-14 years)	Youth (15-24 years)	Working Age (25-54 years)	Pre-retirement (55-64 years)	Seniors (65+ years)	Older Seniors (85+ years)	After-tax income: Average amount ($)
0	129	Agincourt North	29113	3929	3840	3705	11305	4230	6045	925	26955
1	128	Agincourt South-Malvern West	23757	3034	3075	3360	9965	3265	4105	555	27928
2	20	Alderwood	12054	2435	1760	1235	5220	1825	2015	320	39159
3	95	Annex	30526	10863	2360	3750	15040	3480	5910	1040	80138
4	42	Banbury-Don Mills	27695	2775	3605	2730	10810	3555	6975	1640	51874

ratings_raw.csv:



	nbh_num	venue_id	venue_lon	venue_lat	cat_1	cat_2	price_tier	likes	rating
0	1	4b5f8a3df964a52062c029e3	-79.594387	43.720360	Chinese Restaurant	Buffet	1	37	7.2
1	1	4d277a593c795481fa00c59b	-79.600453	43.731018	Burger Joint	None	1	5	7.1
2	1	561d7a54498ed5c87ce4159d	-79.593053	43.715786	Mediterranean Restaurant	Turkish Restaurant	2	6	6.2
3	1	4c2278a67e85c9287d13bc21	-79.600033	43.720549	Sandwich Place	Fast Food Restaurant	1	1	6.2
4	1	4b60bccff964a5209cf629e3	-79.577358	43.712146	Restaurant	None	2	17	6.6

area_rent.csv:



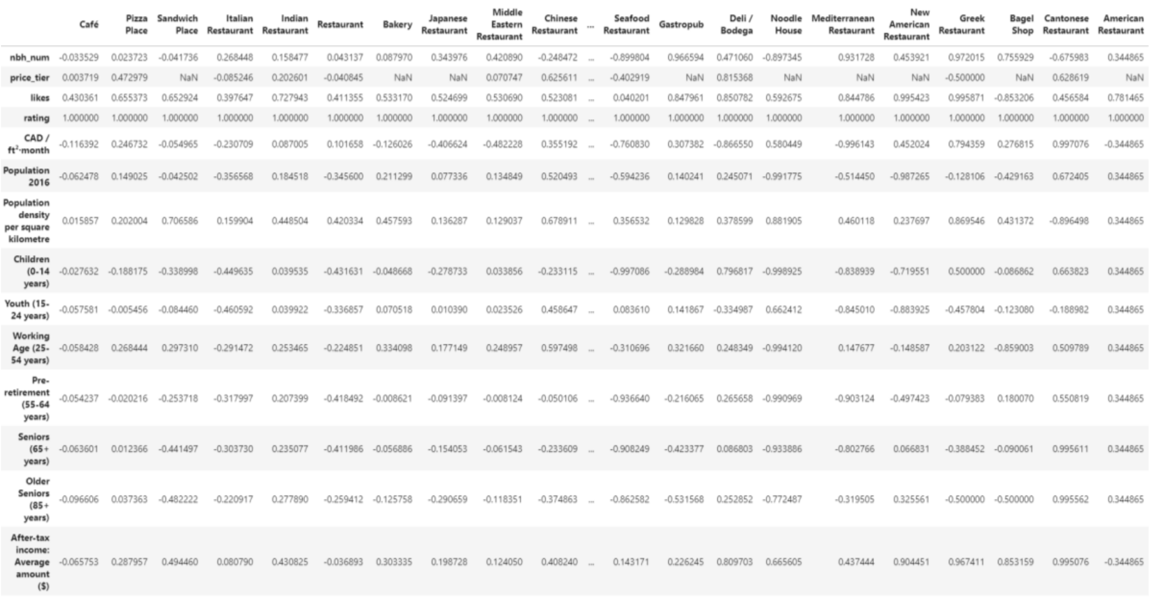
	nbh_num	CAD / ft^2·month
0	1	4.015792
1	2	2.916667
2	3	1.333333
3	17	2.652806
4	19	0.389306

The data scrapping resulted in rental data from roughly 50% of the neighborhoods. Since I want to model the rating using the rental cost, I decided to drop from the toronto_merged.csv data frame those neighborhoods that did not have rental information. This decision was also related with the query limitations I have for the Foursquare API account. Using a personal account allowed me to make up to 500 premium calls per day, which is the type of call needed to get the venue ratings. Getting more than 500 venues among all the neighborhoods would have meant having to query Foursquare over the span of several days to avoid any costs.

This limitation of the premium calls placed an unforeseen constraint on the model I want to build. For the sake of applying the most amount of tools learned from the course, I decided to go forward with the analysis. However, I have to acknowledge the model that can be made from this amount of data is instructional at best, given it lacks the rigor or depth to be called academic. It also lacks the refinement to be productionized for a real application.

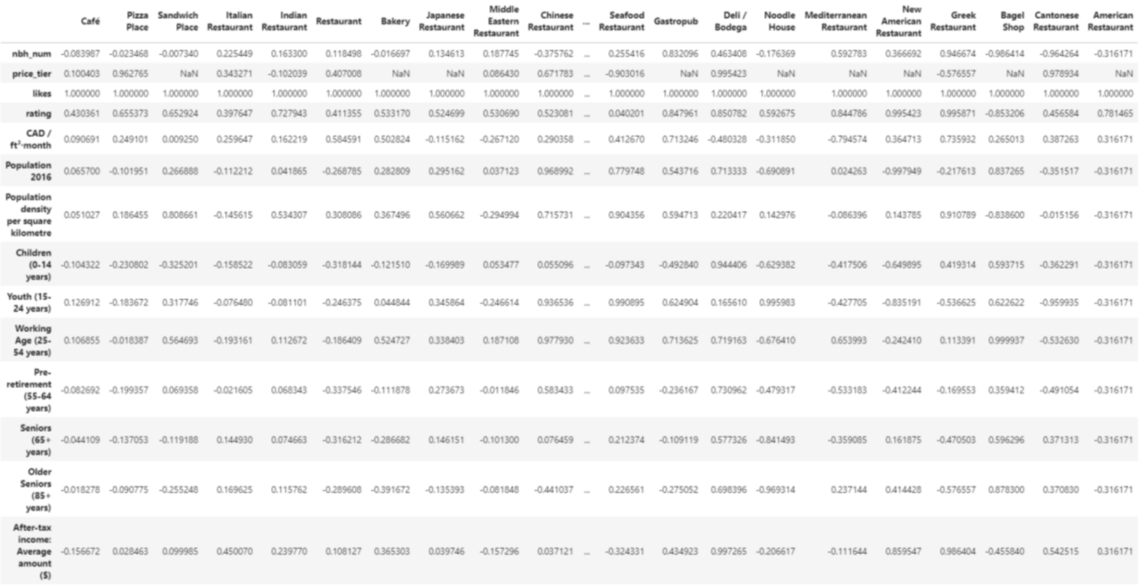
When all the data frames were merged I get a data frame with 17 features. I decided to explore how the population data of each neighborhood correlates with the ratings and likes of each type of restaurant got on a given neighborhood. To achieve that I got the amount of restaurant categories, filtered the master data frame on each of those categories, got the correlation matrix of the filtered data and finally, I extracted the rows pertaining to the correlations I was interested in.

The correlation matrixes I got are shown next:



	Café	Pizza Place	Sandwich Place	Italian Restaurant	Indian Restaurant	Restaurant	Bakery	Japanese Restaurant	Middle Eastern Restaurant	Chinese Restaurant	...	Seafood Restaurant	Gastropub	Deli / Bodega	Noodle House	Mediterranean Restaurant	New American Restaurant	Greek Restaurant	Bagel Shop	Cantonese Restaurant	American Restaurant
nbh_num	-0.033529	0.023723	-0.041736	0.268448	0.158477	0.043137	0.087970	0.343976	0.420890	-0.248472	...	-0.899804	0.966594	0.471060	-0.897345	0.931728	0.453921	0.972015	0.755929	-0.675983	0.344865
price_tier	0.003719	0.472979	NaN	-0.085246	0.202601	-0.040845	NaN	NaN	0.070747	0.625611	...	-0.402919	NaN	0.815368	NaN	NaN	NaN	-0.500000	NaN	0.628619	NaN
likes	0.430361	0.655373	0.652924	0.397647	0.727943	0.411355	0.533170	0.524699	0.530690	0.523081	...	0.040201	0.847961	0.850782	0.592675	0.844786	0.995423	0.995871	-0.853206	0.456584	0.781465
rating	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
CAD / ft^2·month	-0.116392	0.246732	-0.054965	-0.230709	0.087005	0.101658	-0.126026	-0.406624	-0.482228	0.355192	...	-0.760830	0.307382	-0.866550	0.580449	-0.996143	0.452024	0.794359	0.276815	0.997076	-0.344865
Population 2016	-0.062478	0.149025	-0.042502	-0.356568	0.184518	-0.345600	0.211299	0.077336	0.134849	0.520493	...	-0.594236	0.140241	0.245071	-0.991775	-0.514450	-0.987265	-0.128106	-0.429163	0.672405	0.344865
Population density per square kilometre	0.015857	0.202004	0.706586	0.159904	0.448504	0.420334	0.457593	0.136287	0.129037	0.678911	...	0.356532	0.129828	0.378599	0.881905	0.460118	0.237697	0.869546	0.431372	-0.896498	0.344865
Children (0-14 years)	-0.027632	-0.188175	-0.338998	-0.449635	0.039535	-0.431631	-0.048668	-0.278733	0.033856	-0.233115	...	-0.997086	-0.288984	0.796817	-0.998925	-0.838939	-0.719551	0.500000	-0.086862	0.663823	0.344865
Youth (15-24 years)	-0.057581	-0.005456	-0.084460	-0.460592	0.039922	-0.336857	0.070518	0.010390	0.023526	0.458647	...	0.083610	0.141867	-0.334987	0.662412	-0.845010	-0.883925	-0.457804	-0.123080	-0.188982	0.344865
Working Age (25-54 years)	-0.058428	0.268444	0.297310	-0.291472	0.253465	-0.224851	0.334098	0.177149	0.248957	0.597498	...	-0.310696	0.321660	0.248349	-0.994120	0.147677	-0.148587	0.203122	-0.859003	0.509789	0.344865
Pre-retirement (55-64 years)	-0.054237	-0.020216	-0.253718	-0.317997	0.207399	-0.418492	-0.008621	-0.091397	-0.008124	-0.050106	...	-0.936640	-0.216065	0.265658	-0.990969	-0.903124	-0.497423	-0.079383	0.180070	0.550819	0.344865
Seniors (65+ years)	-0.063601	0.012366	-0.441497	-0.303730	0.235077	-0.411986	-0.056886	-0.154053	-0.061543	-0.233609	...	-0.908249	-0.423377	0.086803	-0.933886	-0.802766	0.066831	-0.388452	-0.090061	0.995611	0.344865
Older Seniors (85+ years)	-0.096606	0.037363	-0.482222	-0.220917	0.277890	-0.259412	-0.125758	-0.290659	-0.118351	-0.374863	...	-0.862582	-0.531568	0.252852	-0.772487	-0.319505	0.325561	-0.500000	-0.500000	0.995562	0.344865
After-tax income: Average amount ($)	-0.065753	0.287957	0.494460	0.080790	0.430825	-0.036893	0.303335	0.198728	0.124050	0.408240	...	0.143171	0.226245	0.809703	0.665605	0.437444	0.904451	0.967411	0.853159	0.995076	-0.344865

Table 1. Correlation matrix of the ratings for each type of restaurant.



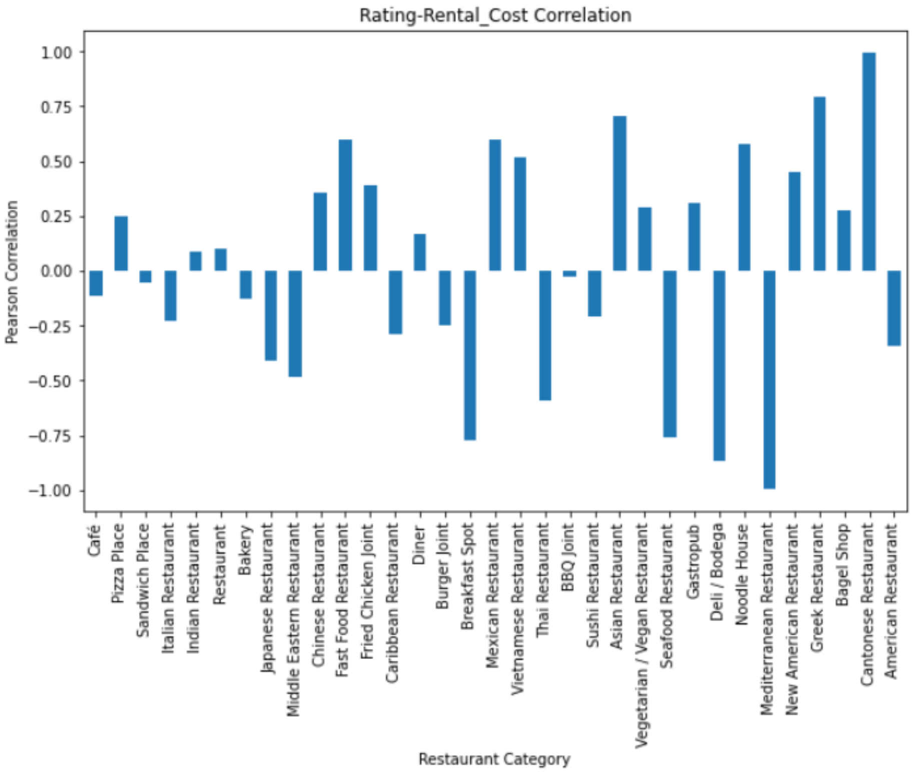
	Café	Pizza Place	Sandwich Place	Italian Restaurant	Indian Restaurant	Restaurant	Bakery	Japanese Restaurant	Middle Eastern Restaurant	Chinese Restaurant	...	Seafood Restaurant	Gastropub	Deli / Bodega	Noodle House	Mediterranean Restaurant	New American Restaurant	Greek Restaurant	Bagel Shop	Cantonese Restaurant	American Restaurant
nbh_num	-0.083987	-0.023468	-0.007340	0.225449	0.163300	0.118498	-0.016697	0.134613	0.187745	-0.375762	...	0.255416	0.832096	0.463408	-0.176369	0.592783	0.366692	0.946674	-0.986414	-0.964264	-0.316171
price_tier	0.100403	0.962765	NaN	0.343271	-0.102039	0.407008	NaN	NaN	0.086430	0.671783	...	-0.903016	NaN	0.995423	NaN	NaN	NaN	-0.576557	NaN	0.978934	NaN
likes	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
rating	0.430361	0.655373	0.652924	0.397647	0.727943	0.411355	0.533170	0.524699	0.530690	0.523081	...	0.040201	0.847961	0.850782	0.592675	0.844786	0.995423	0.995871	-0.853206	0.456584	0.781465
CAD / ft^2·month	0.090691	0.249101	0.009250	0.259647	0.162219	0.584591	0.502824	-0.115162	-0.267120	0.290358	...	0.412670	0.713246	-0.480328	-0.311850	-0.794574	0.364713	0.735932	0.265013	0.387263	0.316171
Population 2016	0.065700	-0.101951	0.266888	-0.112212	0.041865	-0.268785	0.282809	0.295162	0.037123	0.968992	...	0.779748	0.543716	0.713333	-0.690891	0.024263	-0.997949	-0.217613	0.837265	-0.351517	-0.316171
Population density per square kilometre	0.051027	0.186455	0.808661	-0.145615	0.534307	0.308086	0.367496	0.560662	-0.294994	0.715731	...	0.904356	0.594713	0.220417	0.142976	-0.086396	0.143785	0.910789	-0.838600	-0.015156	-0.316171
Children (0-14 years)	-0.104322	-0.230802	-0.325201	-0.158522	-0.083059	-0.318144	-0.121510	-0.169989	0.053477	0.055096	...	-0.097343	-0.492840	0.944406	-0.629382	-0.417506	-0.649895	0.419314	0.593715	-0.362291	-0.316171
Youth (15-24 years)	0.126912	-0.183672	0.317746	-0.076480	-0.081101	-0.246375	0.044844	0.345864	-0.246614	0.936536	...	0.990895	0.624904	0.165610	0.995983	-0.427705	-0.835191	-0.536625	0.622622	-0.959935	-0.316171
Working Age (25-54 years)	0.106855	-0.018387	0.564693	-0.193161	0.112672	-0.186409	0.524727	0.338403	0.187108	0.977930	...	0.923633	0.713625	0.719163	-0.676410	0.653993	-0.242410	0.113391	0.999937	-0.532630	-0.316171
Pre-retirement (55-64 years)	-0.082692	-0.199357	0.069358	-0.021605	0.068343	-0.337546	-0.111878	0.273673	-0.011846	0.583433	...	0.097535	-0.236167	0.730962	-0.479317	-0.533183	-0.412244	-0.169553	0.359412	-0.491054	-0.316171
Seniors (65+ years)	-0.044109	-0.137053	-0.119188	0.144930	0.074663	-0.316212	-0.286682	0.146151	-0.101300	0.076459	...	0.212374	-0.109119	0.577326	-0.841493	-0.359085	0.161875	-0.470503	0.596296	0.371313	-0.316171
Older Seniors (85+ years)	-0.018278	-0.090775	-0.255248	0.169625	0.115762	-0.289608	-0.391672	-0.135393	-0.081848	-0.441037	...	0.226561	-0.275052	0.698396	-0.969314	0.237144	0.414428	-0.576557	0.878300	0.370830	-0.316171
After-tax income: Average amount ($)	-0.156672	0.028463	0.099985	0.450070	0.239770	0.108127	0.365303	0.039746	-0.157296	0.037121	...	-0.324331	0.434923	0.997265	-0.206617	-0.111644	0.859547	0.986404	-0.455840	0.542515	0.316171

Table 2. Correlation matrix of the likes for each type of restaurant.

3.1. Data findings.

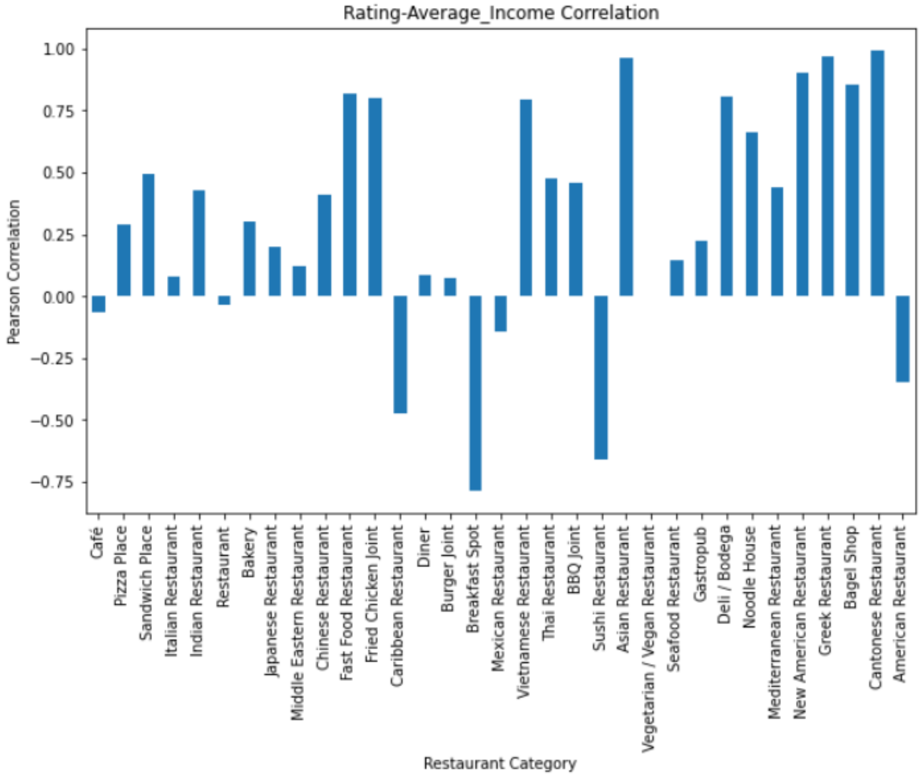
In the particular case of the Café category, there doesn't seem to be any correlation with any of the features of the neighborhood with the rating of the Café. Since all its correlation values are close to zero, there is however some correlation in the rating of the Café with the amount of likes it has: 0.43. It seems how well rated a Café is depends solely on the likes, and not in any apparent fashion by the population of the neighborhood. The price tier of the Café (how much it costs to buy something at the restaurant) doesn't seem to affect the rating of the Café at all; it has a correlation value of 0.0003. One interesting thing to note with the Café category is that its rating is negatively correlated with the rental cost per square feet of the neighborhood. This would lead to believe the more a Café spends on rental, the lower its rating becomes. Exploring this hypothesis might prove interesting with more data analysis tools.

To get a better view of the correlation values of the ratings with some features of interest, I created the following plots:



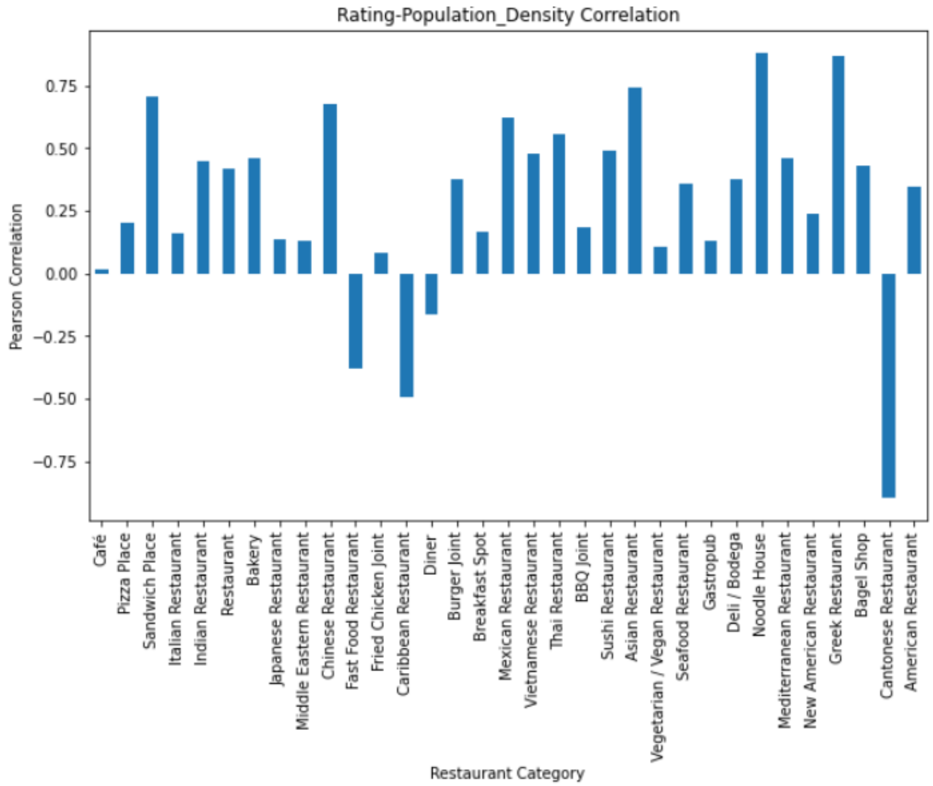
Extending the rental cost correlation with the restaurant rating analysis. It can be seen there is an almost equal amount of positive and negative correlations depending on the type of the restaurant and the rental cost of the neighborhood. For some reason, it seems that a Cantonese Restaurant, the more it spends on rental, the better its rating, it’s the most positive correlation. Personally, I’ve never been on a Cantonese Restaurant so I can’t muster an explanation about it. But I can attribute this to the small pool of data gathered; this correlation may be due to some outliers. As I mentioned earlier, no serious inference can be made from the data.

Seeing how the rating correlates with the rental cost, it seems rental cost is an adequate feature to predict the rating of a restaurant in Toronto.



Another interesting feature to explore is the average income of the neighborhood and the restaurant ratings. The plot shows that; overall, the better the average income of the neighborhood, the better ratings a restaurant will get. This observation holds true when it comes to ethnic restaurants. Maybe people with better income tend to visit this particular type of restaurant for leisure and tend to enjoy the experience, giving overall higher ratings.

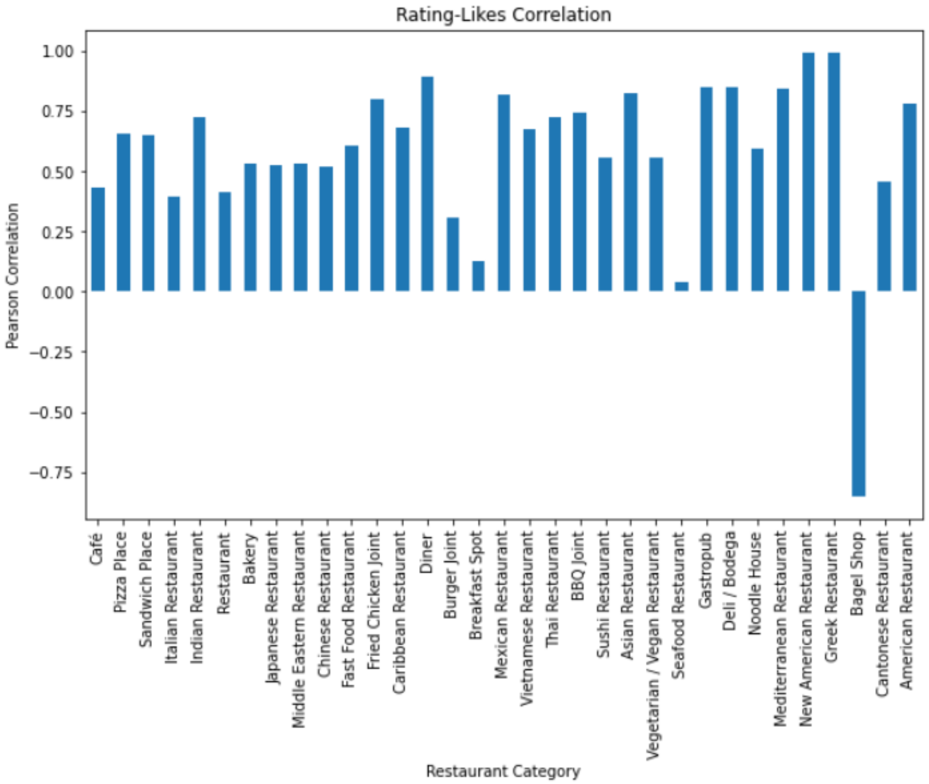
When it comes to restaurants that don’t serve any particular type of food; or more common at least, the average income of the population doesn’t seem to affect the rating much. With the exception of the restaurants labeled as “Breakfast Spot” which seems to exhibit a negative correlation with the average income.



Evaluating the correlation of the population density with the restaurant ratings, we see the same trend the Café category exhibits; no correlation at all with the population density. However, the Cantonese Restaurant category shows another curious feature; the more densely populated the neighborhood is, the lower its rating is. Since this data analysis is not taking into account geographical data, it would be interesting to see where and how many Cantonese restaurants are found on Toronto and further explore what lies on their surroundings that would explain this behavior.

Only exploring these previous features, it seems the population and rental costs of a given neighborhood can give some insight on the ratings a restaurant might get. Some correlations are tenuous at best; as such any regression model built with this data might not be too accurate to predict the ratings.

Another feature that seems to have a particularly interesting correlation on the ratings is the children population. Looking back at Table 1. The children population seems to have an overall negative correlation on the rating of all types of restaurant, with some positive outliers close to 1. The children population correlation is not only mostly negative; it’s also significant in many categories, being the strongest in some. This trend could be further explored getting the tips of the venues along with the reviews and analyze if any words relating to families with children visiting might affect the restaurant rating.



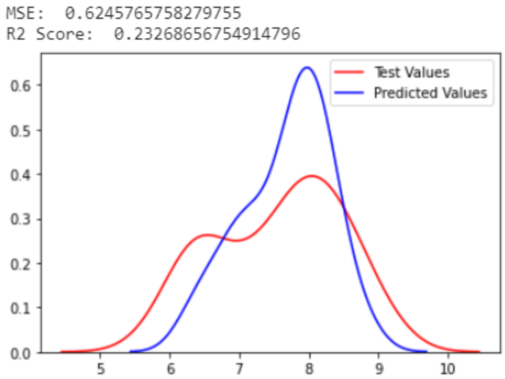
One last feature that I would like to explore is the correlation of the likes of a restaurant and its rating. Overall the rating of a restaurant is positively correlated with the amount of likes, with some outliers that could be explained to the data itself and not a particular trend. One category that proves to show an interesting correlation is the Breakfast Spot, it’s not affected by likes much, but it’s the restaurant category that holds the most negative correlation when it comes to the average income of the neighborhood.

3.2. Regression Models

In an attempt to predict a restaurant rating based on the neighborhood and type of restaurant, I will try to fit the data with a multiple linear regression model, multiple polynomial regression model and Ridge regression.

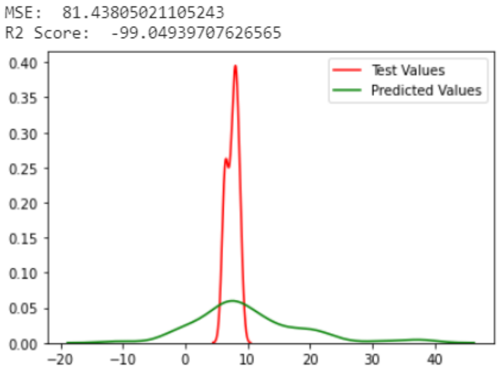
The model will have two categorical data columns, neighborhood and restaurant type. I will use two approaches to these categorical features. In one approach I will use label encoding for neighborhood names, while for the other I will use one-hot encoding. In all approaches I will use one-hot encoding for the restaurant category. I will not use the column called “cat_2” that was obtained, since this feature is not present in all the restaurants.

The code used to generate the following graphs can be seen here.



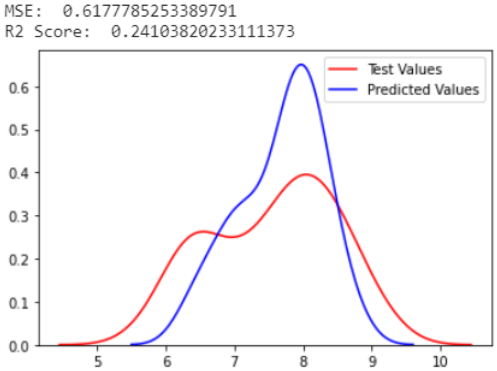
```
MSE:  0.6245765758279755
R2 Score:  0.23268656754914796
```

Model 1. Multiple linear regression, using label encoding for the neighborhood name.



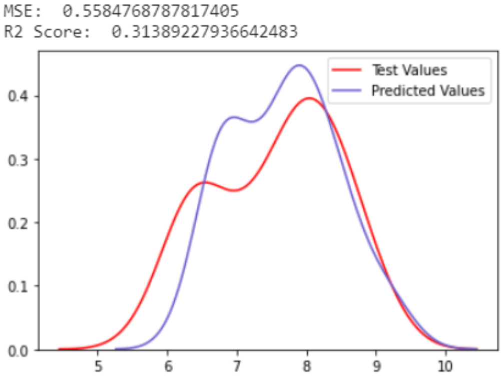
```
MSE:  81.43805021105243
R2 Score:  -99.04939707626565
```

Model 2. Polynomial regression, using label encoding for the neighborhood name and a second order degree polynomial feature on the population data frame.



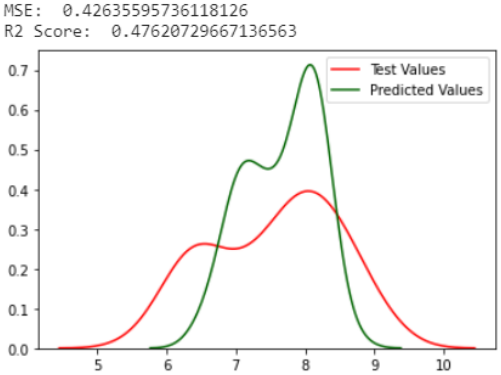
```
MSE:  0.6177785253389791
R2 Score:  0.24103820233111373
```

Model 3. Ridge regression, alpha = 0.1. Label encoding for neighborhoods.



```
MSE:  0.5584768787817405
R2 Score:  0.31389227936642483
```

Model 4. Ridge regression, alpha = 0.1. One-hot encoding for neighborhoods.



```
MSE:  0.42635595736118126
R2 Score:  0.47620729667136563
```

Model 5. Ridge regression after using grid search, alpha = 1.5. One-hot encoding for neighborhoods.

It can be seen on the Jupyter Notebook that complements this part of the report; several other linear regression models were used. I decided to focus the on Ridge regression since it was the model that worked best when using one-hot enconding, the other models just didn't work.

Considering how little correlation there was between some features and the small amount of data, I consider this model to be adequate enough with the current tools I have at the moment.

In hindsight, some sort of data normalization might have been necessary before training the models. I would like to create a function that would allow me to input the neighborhood and restaurant type and have the model predict a possible rating.

4. Conclusions and closing words.

Without using exact numbers, I spent close to 90% of the time spent on this project getting and cleaning data, learning how to scrape data and learning some more Python along the way.

The goal of reducing the amount of uncertainty when choosing a restaurant was partially solved. I managed to train a model that can output ratings. But it still needs work, the best model isn't particularly accurate or reliable and there's still the matter of interacting with the model to get ratings from neighborhoods and restaurants on command. I consider with further study of more complex machine learning models, better predictions can be made.

One constraint that proved difficult to overcome was getting more data to feed the model. Scraping rental values for Toronto from webpages was tedious; and frankly at times, above my skill level when it comes to Python and programming, but I wanted to scrape the data myself.

There is a plethora of rental API that give access to much more granular rental data, but after 3 days of trying to find and API that would allow me to use it without paying, or with a model similar to that of Foursquare’s API, I decided to settle with the data scrapped and move forward with the project.

There were some interesting insights from the data; some of these might be obvious for people who work with real state, but for me it was enlightening to see how a venue like a Café seemed to be impervious to the population data of the neighborhood. It would partly explain why there are so many Cafés. They can generally get good ratings if the coffee is good enough. I consider with more data scrapping tools, it can be further explored why the Cafés rating was mostly independent from the neighborhood data.

Long story short: open a Café shop anywhere in Toronto, make sure the coffee is good, people know where it is, and it should be a safe bet.