# Applied Data Science Capstone

Modeling Foursquare Restaurant Ratings

Paul Anderson, 2020

# Modeling Foursquare Restaurant Ratings

- Some people open a business like a restaurant based on a hunch, feelings or convenience.

- Having a model that can predict to some extent the rating a restaurant based on neighborhood and type of restaurant, can reduce some of the guesswork.

# Data

1. Average commercial rental costs in Toronto per neighborhood (Web scrapping).
2. Population description of each neighborhood (Census).
3. Restaurant ratings and traits (Foursquare API).
4. Geographic data of Toronto's neighborhoods (Toronto.org).

# Data Analysis, input data.

| | 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 |


| | 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 |


| | nbh_num | CAD / ft²·month |
|---|---|---|
| 0 | 1 | 4.015792 |
| 1 | 2 | 2.916667 |
| 2 | 3 | 1.333333 |
| 3 | 17 | 2.652806 |
| 4 | 19 | 0.389306 |

# Data Analysis, correlations.

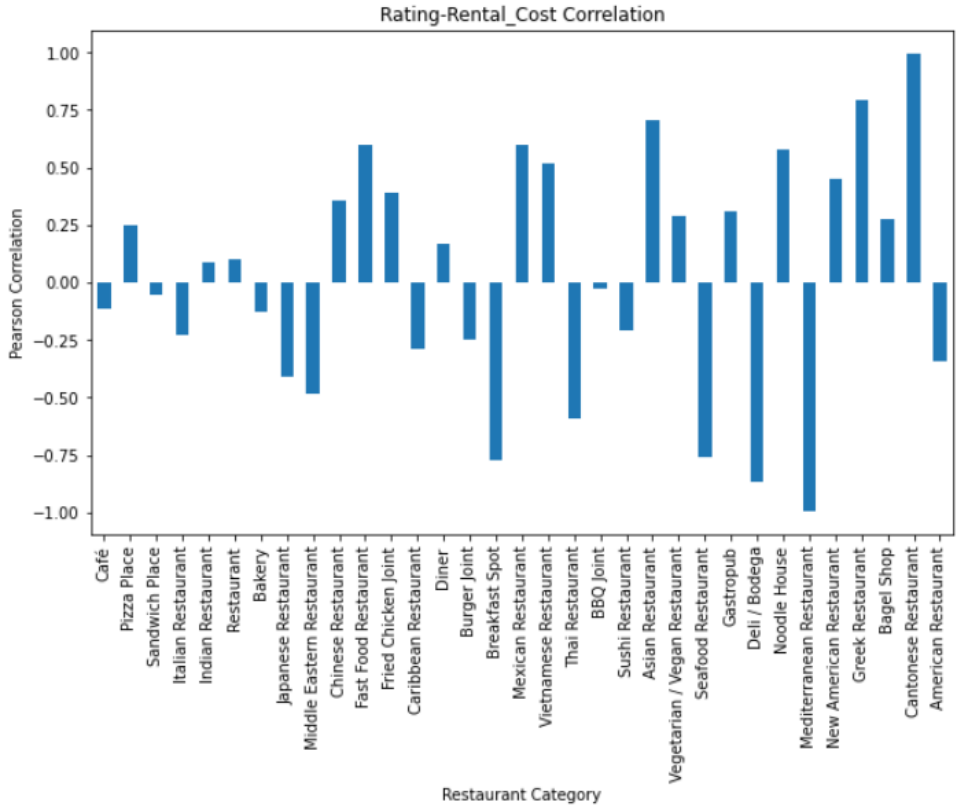
| | Café | Pizza Place | Sandwich Place | Italian Restaurant | Indian Restaurant | Restaurant | Bakery | Japanese Restaurant | Middle Eastern Restaurant | Chinese Restaurant | ... | Seafood Restaurant | Gastropub | Deli / Bodega | Noodle House | Mediterranean Restaurant | New American Restaurant | Greek 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 |

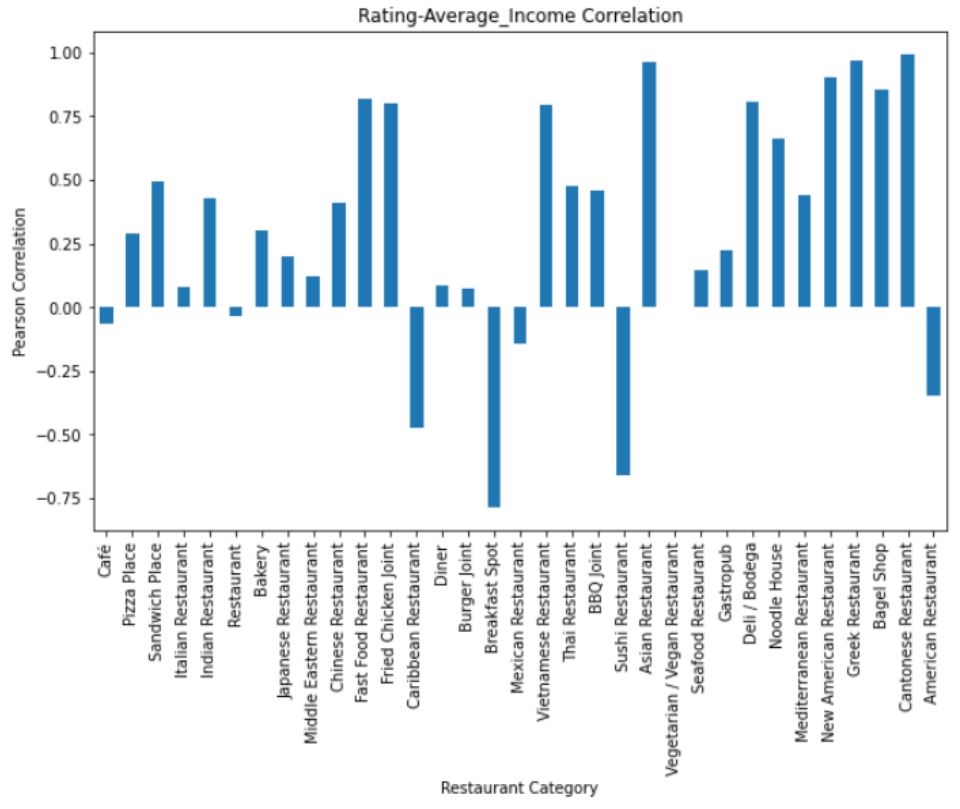
Rating correlations.

Likes correlations.

| | 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 |

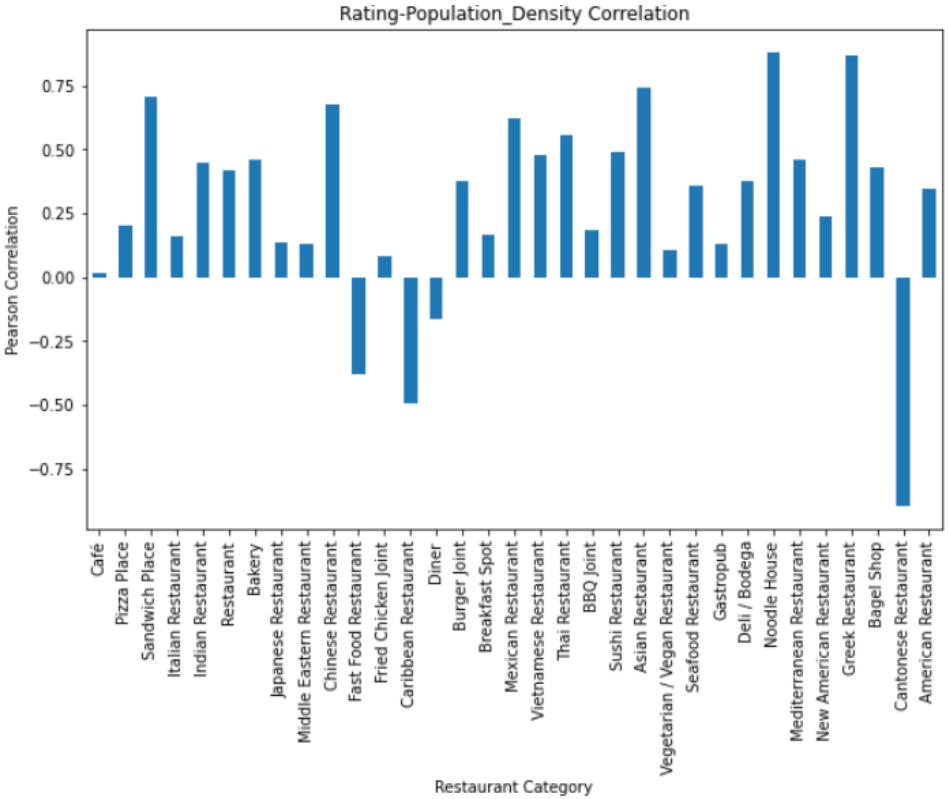
# Data results

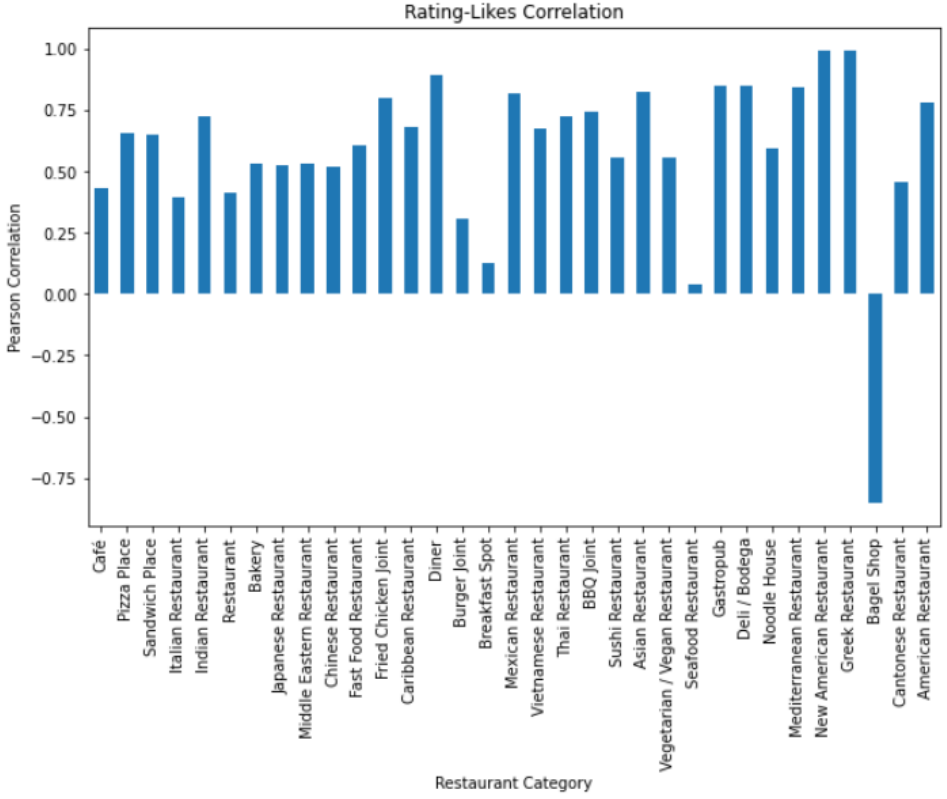



Restaurant categories are more affected by some neigborhood features than others.

# Data results

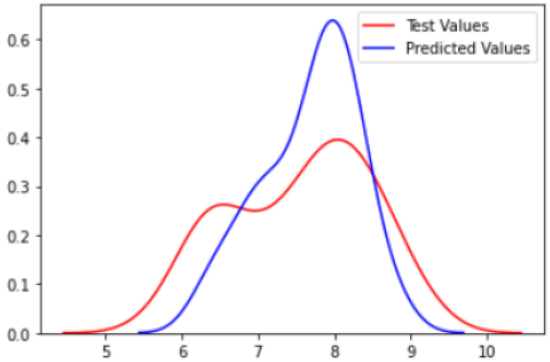



Amount of likes seems to be the best predictor for amount of likes. Likes won't be taken into account, because these come after the restuarant exists, not before.

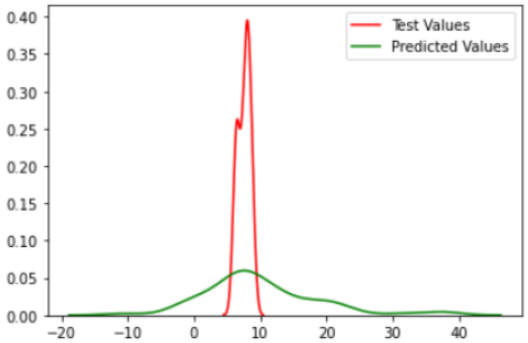
# Regression Models

MSE: 0.6245765758279755
R2 Score: 0.23268656754914796




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

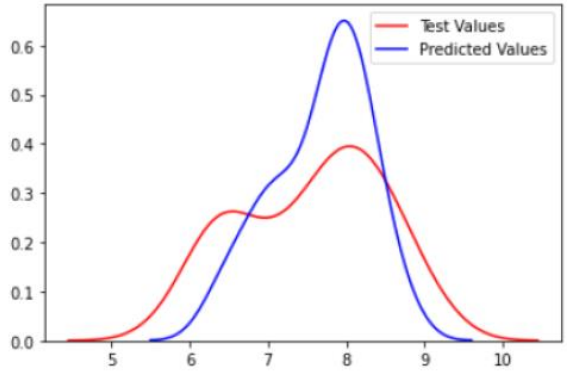
MSE: 81.43805021105243
R2 Score: -99.04939707626565




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

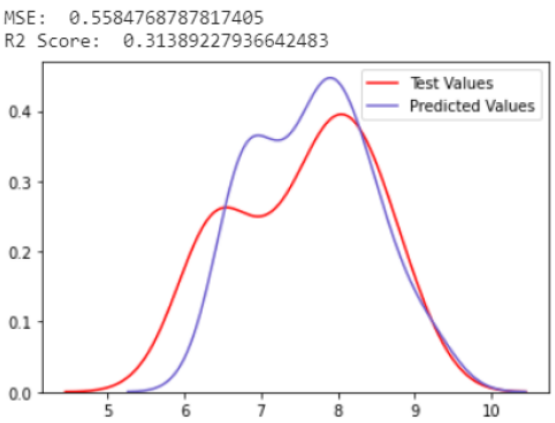
# Regression Models

MSE: 0.6177785253389791
R2 Score: 0.24103820233111373




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

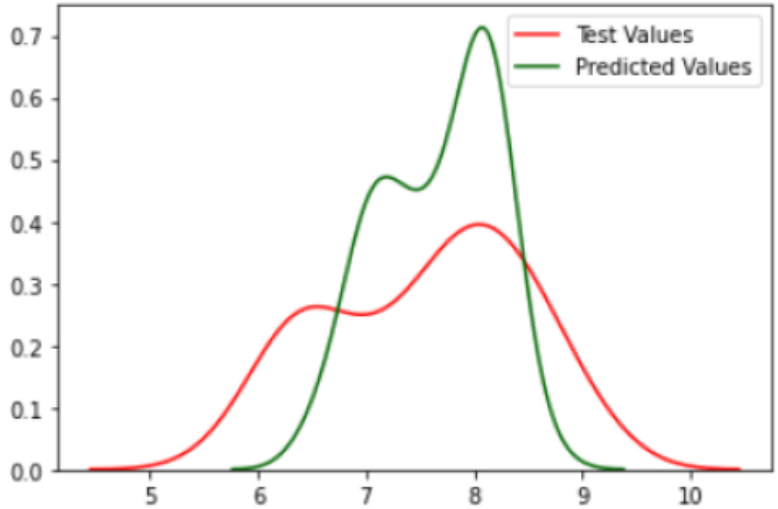
MSE: 0.5584768787817405
R2 Score: 0.31389227936642483




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

# Regression Models

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




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

# Conclusions

- Getting enough data was the main factor limiting  the quality of the model.
- Proper handling of the categorical data improves model quality.
- Based on this sample of data, the safest option is a Café.