Final Capstone Writeup

Project Title

Optimal Coffee Shop Location NYC vs Toronto

Introduction/Business Problem

I am an avid Coffee drinker, and was curious how neighborhood demographics impacted the number of coffee shops present in various neighborhoods. I approached the problem as if I had a hypothetical client who was interested in opening a coffee shop in Houston. Houston is an upcoming metropolitan area that can be compared in size more like Toronto, but in potential size to New York City. Therefore I wanted to use data from both of these cities to analyze if there are relationships between demographic factors of neighborhoods and coffee shop locations, to determine if there is an optimal neighborhood to choose in Houston.

Data

To gather the necessary data, I will be using a mixture of Wikipedia webpages, New York City Data from the census bureau, and of course the Foursquare server. From foursquare I will be pulling information on the number of coffee shops that are located in each of the suburbs identified on the Wikipedia articles. The census data will be used as independent variables when building out the models..

Methodology Section

I have outlined my methodology in my Jupyter notebook, with annotations to make it easy to follow.

Results Section

My results indicate that surprisingly, both cities show a statistically significant negative correlation between population and number of coffee shops. For the New York city data, the coefficient was -.00008076. This might seem really small, but keep in mind the population number is very large, and is being regressed against a much smaller coffee shop number. This value had a t score of -8.853, making it statistically significant even at the 99.99% level. It’s also interesting that in the New York City data that squaring the population is also a statistically significant measure. This had a t score of 6.803, which was also statistically significant at the 99.99% level. Due to limitations in available datasets that compromise all of the New York City Neighborhoods, these are the only variables that I was able to generate regressions on.

For the Toronto data we had more in depth demographic information, which led to interesting regression results. Once again we found a negative correlation between population and coffee shops, with a coefficient of -0.0003, and an associated t value of -1.7. This was statistically significant at the 90% level, indicating a slightly weaker significance level than the New York City data. For Toronto, the population squared was not a statistically significant factor. Other independent variables that I thought could have an impact on how many coffee shops were present included average income, density and the percentage of renters. Looking at the t values for all three of those factors, however, indicates that none of them were statistically significant.

Discussion Section

Armed with this data, I looked into Houston neighborhoods to see which neighborhood had the highest and lowest population values. If my client shares the viewpoint that fewer coffee shops means less competition and a bigger ability to get new clients, then they would be interested in the neighborhoods of Bear Creek, Copperfield Area, Katy-North and Sugar Land South just to name a few.

Ultimately the data could be made a lot more reliable if more demographic information was available. Dealing with severe limitations on the New York City dataset limited the variables that could be explored. Further research could also include analyzing the ratings of the various coffee shops, seeing how many are franchises vs mom and pop stores, how education levels impact the number of coffee shops, and many other factors.

Conclusion

In conclusion, I feel like I learned so much throughout all of the Coursera courses on how to gather data from numerous sources (Foursquare, Wikipedia, csv files), successfully clean all the data to get working information, and then be able to process the information to reach meaningful conclusions for my client. I hope you enjoyed my project!