INTRODUCTION

The purpose of this project is to gain and share a better understanding of coffee shops in Los Angeles county. This project assumes a business need for business intelligence relative to this subject-matter. Stakeholders of this business intelligence can be assumed to be investors, coffee shop proprietors or, more practically, Los Angeles local coffee consumers.

The ultimate question that this project seeks to answer is:
"What are the measurable factors that contribute to a coffee shop being highly rated by customers?"

DATA

Data was initially sourced from Foursquare using its native RESTful API to create a basic geographic framework. Yelp! was mined using BeautifulSoup to obtain relevant customer feedback (rating / review_count) as well as individual coffee shop website.

METHODOLOGY

In keeping with the Data Science Methodology, a Business Understanding was first explored before a descriptive model was chosen as an Analytical Approach. Data requirements were partly uncertain so discovery was conducted to see what sources might allow in support of this project- ultimately data requirements were made to include Foursquare and Yelp!. Data was collected as described above and understood by using Foursquare as a backbone, with newly acquired Yelp! data appended in. The resulting dataset was carefully tidied up into the CSV file found here:

https://github.com/jjbodell/Coursera_Capstone/blob/master/Final/LA_Tidied.csv

Data was then modeled into meaningful complimentary components as follows:

Tuples were defined for neighborhood and coffee shop name. A tuple for price was also added after employing one hot encoding to count the number of dollar currency symbols captured to create a price measure of 1, 2 or 3.

Booleans were defined for category and chain. Category was simplified to describe whether a coffee shop exclusively sold coffee or whether that shop had additional offerings (such as baked goods, juice, specialty teas, etc.). Chain was validated by count- where a count of one indicates FALSE and a count of greater than one indicates TRUE.

Ratings were grouped by neighborhood as an average and review_count were binned by first quartile, mean and third quartile- both scaled to neighborhood and for individual coffee shops.

Much of the remaining attributal data was left as referential lists.

The ten highest rated neighborhoods for each review_count bin were first isolated to create a digestible slice of the initial data that could be explored in greater depth. Those neighborhoods were identified as individual clusters containing any number of assigned coffee shops- so the next step was to define the contents of each cluster.

RESULT

Initial analysis found rating as an independent variable with neighborhood, prominence (density) and affordability as dependent variables. Given this, the Downtown neighborhood stood out- though this may have been an outlier as Downtown had a significantly higher concentration of coffee shops as compared to any other neighborhood. It is undeniable that coffee shops are abundant in the Downtown neighborhood- but quantity does not necessarily correlate to quality. An interesting correlation to note- is that because of the saturation of coffee shops in the Downtown neighborhood, coffee shops were generally more affordable in this area- likely due to natural competition.

With the model being developed to a functional state, modeling was revisited to explore more definitive metrics. Three neighborhoods stood out with statistical significance across all three of the above mentioned metrics- Los Feliz, Studio City and Beverly Grove. Coffee shops of these three neighborhoods were explored primarily on the basis of rating and review_count. Applying this model solely on the basis of rating yielded favorable weighting to neighborhoods with a higher concentration of coffee shops-similar to the wall found with Downtown.

Overall, Beverly Grove was found to be the most favorable neighborhood for coffee patrons. Beverly Grove found an optimally positive correlation between ratings, prominence and affordability. A closer look at Beverly Grove also found that these metrics may have related directly to a stronger focus on coffee- with 80% of the coffee shops in Beverly Grove emphasizing coffee and with 20% advertising non-coffee beverages and food items.

DISCUSSION

Numerous assumptions were made in this course of this project to supplement greater time and energy devoted to the analysis, which seemed to have a point of diminishing return. I'm now aware that my hometown of Santa Monica is among the top 20 Los Angeles neighborhoods to grab a cup of coffee.

CONCLUSION

Downtown is such a strong outlier in all areas that it not only acted as an outlier but forced me to revisit my approach and modeling of the majority of this project. Neighborhoods can be quantified in terms of customer satisfaction, however, this exercise may certainly be improved upon by taking into consideration a great number of factors and by refining the weighting relationships between those factors. I enjoyed this project.