Research Proposal

Coffee drinking is essential in many people's lives. The second wave coffee came with an increase in the quality, not just exponent growth in quantity like the first wave, of coffee readily available. Big companies like Starbucks started running coffee shops as profitable businesses as coffee started to become a luxury product rather a necessity. Buyers started to be more conscious of other elements of a coffee brand, such as the environment. This study aims to explore the relationship between the discrepancies in the overall rating for stores of a standardized coffee chain and the demographic characteristics of the nearby neighborhood.

Why is this research question important? Starbucks cares about surrounding neighborhoods as it makes efforts to reinvent the store experience to speak to the heart and soul of local communities. The majority of previous studies of Starbucks focus on the effect of its strategies to boost the consumers' demand. This project will expend the research by measuring the market reaction using the rating and reviews on Starbucks stores on Yelp to explore more reasons behind consumers' thoughts and feedbacks for Starbucks stores.

This study aims to help analyze the discrepancies in overall preferences of surrounding neighborhoods with different demographics. People associate a corporate image with consumers' rating, which reflects customer satisfaction and thus influences consumers' loyalty. Positive corporate and high consumer rating is critical to help companies achieving higher sales and generating more revenues. As a result, it is significant to analyze the relevant key influencers of consumer ratings for Starbucks as the rating will provide valuable business insights for its current and future market.

The rating for each Starbucks store in New York City can be obtained by interacting with the Yelp Fusion application programming interface (API) with a private API user key that

authenticated by Yelp for users to access all endpoints after registering for the application. The Yelp Fusion API contains endpoints for search experiences, such as business search, business details, and reviews. Developers can use location parameters and any search related keywords that can be categorized into search groups to call the search endpoint. The response provides businesses with information like name, address, yelp ratings, price levels, coordinates that made up by latitude and longitude, and review counts.

A function with R scripts can be written to acquire the rating (from 0 to 5), the review number, and the coordinates of all Starbuck stores by interacting with Yelp API. Later with these data, this research can calculate the number, average rating, and average review number of Starbucks stores. This study further needs the coordinates of each Starbucks store to match the store with the demographic and economic conditions of the nearby neighborhood.

Moreover, the New York City census from the American Community Survey in 2015, along with additional raw data from the US Census Bureau website, provides 2,167 demographic and economic data of surrounding neighborhoods for this study to analyze. The joint dataset contains coordinates, county code, county names, borough names, total population, number of men, number of women, percentage of the population by different races, number of citizens, median house income and error, and income per capita and error. Data for percentage under the poverty level, percentage of children under the poverty level, percentage of various professions, percentage of various ways of commutes, and average minutes of commute time are also available in this dataset. There are data in the percentage of the employed, the percentage of people who work at home, the percentage of the employed in private and public industry, the percentage of the self-employed, the percentage of unpaid family work, and the unemployment rate available.

The demographic and economic information is collected by “census tracts”, areas that are roughly equivalent to neighborhoods established by the Bureau of Census for analyzing populations. A census tract generally encompasses a population between 2,500 to 8,000 people. This research finds an online Federal Communications Commission (FCC) census block lookup tool to retrieve the census block code for a 200 x 200 grid containing New York City and some surrounding areas by interacting with FCC API. This dataset contains the coordinates and associated census block codes, along with the state and county names. Although one classmate suggests using the data downloaded from StreetEasy regarding the income in the neighborhood, this study will not use the data from StreetEasy. This study needs more variables than only income to have a more comprehensive picture of how other demographic characters influence people’s rating on Starbucks.

With information on coordinates of both the Starbucks stores and the census tracts, this study can match them together to conduct statistical analysis. Accordingly, this research project intends to use R scripts to explore the relationship between ratings and the number of Starbucks and the demographic and economic factors in the areas divided by census tracts. This study aims to employ the Ordinary Least Squares regression to examine the relationship. Statistically significant influencers in the previous study will be used as parameters to predict the rating for a Starbucks store in the area by machine learning techniques. Parameters can be adjusted based on the accuracy of the model. The success of the model will be measured by the accuracy of the prediction.

Granted, coffee in different Starbucks stores might taste different as they are made by different people, but the recipe and materials for the same type of coffee should be standardized to be roughly the same. Therefore, it is likely to be the case that the demographic and economic

characteristics of the surrounding neighborhood become critical influencers of the rating for stores in one area. For areas that are crowded by office buildings, lots of reviewers of that shop might not live in the area. White collars are likely to buy coffees near their offices. Similarly, potential consumers of coffee shops near shopping malls are not necessarily from those areas. Yet, people from the surrounding neighborhood still make up a large number of consumers in those shops.

As this study intends to examine consumers' overall feedbacks to Starbucks store in New York City, Yelp is a suitable sample pool due to its comprehensive and diverse demographic components. Nevertheless, Yelp rating is only one of the measurements for the overall feedback from the consumers to the Starbucks stores. Therefore, it is also worth conducting sentiment analysis on the reviews to reveal more reasons behind the ratings. However, Yelp API was restrictive when it came to reviews as only three reviews per business are available. I instead found a curated dataset by Stonybrook University that collected reviews. The main issue with that is that it's a little dated and a lot of the restaurants no longer exist. This created some headaches merging on the data, so I might have to keep the sentiment analysis on reviews for the business separate. Meanwhile, there are data about health code violations from the Department of Health and Mental Hygiene, which might provide some insights into the difference in reviews for different Starbucks.

As the US Census Bureau website and the American Community Survey in 2015 are authoritative data sources, this study can rely on the data from the New York City census. Nevertheless, this study should still consider human biases in the process of collecting data about demographic and economic characteristics.

Overall, this study will retrieve a dataset contains information about total population, gender and race constructions, income, age, and occupation and the overall rating and consumers' reviews for Starbucks stores in New York City from Yelp.