**New Business Location Selection in Seattle**

**Dongdi Zhou**

**August 4, 2020**

**Introduction**

**Background**

Seattle has been one of the fastest growing large American cities for nearly a decade. According to a Washington Office of Financial Management (OFM) survey, Seattle's population grew by 22.78% between 2010 and 2019. There is no doubt that more people are attracted to this booming city to establish their own businesses.

**Problem**

As the 15th largest city in the United States, Seattle is made up of over 30 neighborhoods. Each neighborhood has its own unique characteristics and the demographics of each neighborhood can also be very different. Therefore, it is difficult to determine where is the ideal location for a new business without an understanding of this fast-growing market.

**Interest**

The main goal of this analysis is to provide essential data and assist a new business's location selection process. For illustration purposes, a new bakery is used as an example. This new bakery features top quality bread and pastries that are made with trendy ingredients. Its store environment is Instagram-worthy. Its target market is composed of young and educated individuals or families under the age of 35 who have medium to high income level. This analysis goes through the data of each Seattle neighborhood from the following aspects:

1. median age
2. annual household income
3. crime rate
4. current market saturation

**Data Acquisition and Cleaning**

1. Median age by Seattle neighborhoods – [link](https://www.seattle.gov/Documents/Departments/OPCD/Demographics/AboutSeattle/Census%202010%20Community%20Reporting%20Areas%20Subject%20Report.xlsx)

When I was looking for demographic related information like median age, I first thought about the U.S. Census data that is available to the public. The City of Seattle organizes the U.S. Census data by neighborhood. However, different city departments and non-city entities define neighborhoods differently based on many factors. Some districts and neighborhoods are informal with varying boundaries and names. Some neighborhoods may overlap and be referred to by different names by community members. For this analysis, I chose the 2010 U.S. Census full report by Seattle Community Reporting Area (“CRAs”). The CRAs were established for the U.S. Census purpose specifically. The city is divided into 53 CRAs whose name and boundaries are the closest to what are commonly referred to as Seattle neighborhoods. Also, compared to the other definitions, CRAs provide the proper level of detail for this analysis.

The full report is in Excel format, but it includes a lot of formatting such as title rows, spacing rows/columns, merged cells, etc. The data for each CRA is shown in columns and different demographic data such as ethnicity, age and sex are shown in rows. The data cleaning process mostly involves removing the formatting rows and columns and extracting the one row of data that represents the median age for both sexes using the Python Pandas module. The dataframe is then transposed to show the CRAs as rows and median age as the only feature.

A screenshot of a cell phone

Description automatically generated

Figure 1: first 5 row of median age by neighborhood dataframe

1. Annual Household Income - [link](https://www.irs.gov/pub/irs-soi/17zp48wa.xlsx)

I could not find a dataset that includes annual household income by Seattle neighborhood. The closest data I could find was the Individual Income Tax Statistics by zip code in the State of Washington disclosed by the Internal Revenue Service. The latest data available was from 2017. One of the features included in this dataset is Adjusted Gross Income (“AGI”) which is defined as gross income minus adjustments to income. Gross income includes wages, dividends, capital gains, business income, retirement distributions as well as other income. Adjustments to income include educator expenses, student loan interest, alimony payments or contributions to a retirement account. Depends on how the returns were filed, the AGI can be a fair representation of an individual or a married couple’s annual income level, i.e. annual household income.

The rows of the dataset are organized by zip code and each zip code is further divided into 6 buckets based on the AGI range in 1 total row. The original dataset also includes 151 columns, but only the zip code, size of adjusted gross income and number of returns columns are relevant to this analysis. To clean up the data, the following steps are performed:

Step 1: the total row for each zip code is removed and columns other than the relevant columns are dropped.

Step 2: a U.S. zip code database is downloaded which includes all U.S. zip codes and their corresponding primary cities. The zip code database is merged with the AGI dataframe so that the zip codes in AGI dataframe can be associated with a city. Any city other than Seattle is then dropped.

Step 3: Pandas.pivot\_tab is then used to convert the size of adjusted gross income from rows to features. The Seattle zip codes stay as the rows.

Step 4: The number of returns for each AGI range is converted to percentage.

Step 5: Certain features are further grouped into 3 features.

A screenshot of a cell phone

Description automatically generated

Figure 2: first 5 rows of dataframe for percentage of returns filed in six AGI buckets by Seattle zip code

A screenshot of a cell phone

Description automatically generated

Figure 3: first 5 rows of dataframe for percentage of returns filed in three AGI buckets by Seattle zip code

1. Crime Data - [link](https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5)

Seattle Police Department shares Seattle crime data from 2008 through the present. The dataset includes 17 features such as incident date and time, crime type, neighborhood, longitude and latitude information. The neighborhoods in this dataset are referred to as Micro-Community Policing Plans (MCPP) neighborhoods which are very similar to the neighborhoods defined for the U.S. Census. For this analysis, I am interested in the number of incidents by Offense Parent Group in each MCPP because the Offense Parent Group provides enough detail on the types of crimes. Since this feature contains categorical data, I used Pandas One Hot Encoding to convert the Offense Parent Group into dummy numerical values for each incident and its corresponding MCPP where the incident happened.

A screenshot of a cell phone

Description automatically generated

Figure 4: sample data of the crime rate dataframe after one hot encoding

The resulting dataframe is then grouped by MCPP so that the number of incidents occurred in each MCPP can be counted by the type of offense. Last but not least, certain financial crimes such as embezzlement and fraud offenses are dropped from the dataframe.

A screenshot of a cell phone

Description automatically generated

Figure 5: sample data of the final crime rate dataframe

1. Foursquare REST API

The last data source used for this analysis is from the Foursquare REST API. To measure the current market saturation of bakeries in each neighborhood, location data is pulled for bakeries within a 600 km radius of Seattle. The resulting data is transformed into a dataframe and summarized into the number of bakeries by Seattle zip code.

A picture containing screenshot, large, boat, display

Description automatically generated

Figure 6: first five rows of bakery count by zip code dataframe

**Methodology**

The goal of this analysis is to find the Seattle neighborhoods where the defined target market resides judging from four criteria. In the end, neighborhoods that meet at least three of these criteria if not all four can be recommended to a new business. However, as I have discussed in the previous sections, the datasets used for each criteria are from different sources and the neighborhood boundaries are drawn differently by various city agencies. In addition, I could not find a dataset that includes all the neighborhoods with corresponding zip codes. Luckily, I found geojson files for Seattle zip code and MCPP, so I utilize bar charts and the Python Folium module to visualize the four datasets so that the equivalent neighborhood areas can be compared.

Data Visualization

1. Median Age

The bar chart below shows the median age for each CRA in ascending order. The University District is the youngest neighborhood with a median age of 21.3 since most of its residents are students at the University of Washington. The chart also shows Seattle neighborhoods can be divided into 4 groups: median age under 30, median age between 30 and 34, median age between 35 and 40, and median age above 40. The new bakery’s target market are young individuals and families under the age of 35, so I will focus on the neighborhoods within the first two groups.

A picture containing comb

Description automatically generated

1. Annual Household Income

As previously discussed, I use Adjusted Gross Income (AGI) as the measurement of an individual or married couple’s annual household income. For this analysis, I set the target market of the bakery as individuals and married couples with AGI between $50,000 and $100,000. In the bar chart, the percentage of returns with AGI between $50,000 and $100,000 are shown by blue colored bars and the zip codes are organized in descending order of this feature. I also included the percentage of returns in the other two AGI categories in the bar chart as well. This is to help further distinguish neighborhoods in the case of similar percentage of returns with AGI between $50,000 and $100,000. For example, zip code 98103 and 98106 have very similar percentage of returns between $50,000 and $100,000 AGI, 27.43% vs. 27.24%. However, zip code 98106 has 14.13% more returns with AGI above $100,000. Depending on the new bakery’s specific target market, zip code 98106 might be more favorable.

Next, the percentage of returns with AGI between $50,000 and $100,000 is visualized over an interactive Seattle map using Folium so that each zip code’s data can be seen at a glance. The neighborhoods I will consider based on annual household income are Queen Anne, Fremont, Wallingford, South Ballard, West Seattle, Delridge, to name a few.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

1. Crime Rate

Similarly, the number of crime offenses by MCPP is also visualized on the Seattle map for easy interpretation. The safer neighborhoods include Eastlake, Madison Park, South and North Delridge, South Beacon Hill and Alki to name a few.

A picture containing text, map

Description automatically generated

1. Current Market Saturation

Last but not the least, the bakery count by zip code based on Foursquare’s location data is also visualized on a Seattle map. The neighborhoods with few or no bakeries based on Foursquare data are Ballard, Magnolia, Queen Anne, University District, and Delridge.

A close up of a map

Description automatically generated

**Conclusion and future directions**

After evaluating the median age, annual income (AGI), crime rate, and current market saturation of each Seattle neighborhood using Pandas dataframes, bar charts and choropleth maps, the general recommendation for a new bakery will be in Ballard, Queen Anne, Alki, Green Lake and Delridge because they meet three or more target market criteria of the bakery.

However, this conclusion is drawn based on the available data I could find on the Internet and it is just a general overview of the Seattle market. With more and better data, I would like to improve the accuracy of the analysis and further narrow down the location selection by doing the following:

* Consider the population size of each neighborhood and look at crime rate per capita and number of bakeries per capita.
* Factor in average/median rental cost of each neighborhood since it is usually the biggest cost item of a new business.
* Research on zoning and permit requirements of each neighborhood. The low market saturation in some of the neighborhoods could be caused by zoning rules. Then such limitation can be considered by the new business.
* Update the median age and AGI analysis using more recent U.S. Census and IRS data.
* Improve the data availability of current market saturation by using a different API such as Yelp.