**2017 Starbucks Dataset - Executive Summary**

Team Members: Cara Jiayi You, Ho Tai Peter Law, Crystal Jiawen Li

Introduction

Starbucks had quickly established itself as a premium coffee brand after initially launching in the 1970s. It’s exponential growth and expansion during the 1990s and the early 2000s had made the brand an instant classic for business school case studies globally and had remained as the market leader in multiple industries and market sectors even as of today.

As business school students from a top school, despite not having direct access to Starbucks's internal data, our team is still interested in understanding key factors that influences Starbucks’ strategic decision with its business. We have found a location dataset of Starbucks stores last updated in 2017 and have decided to gather more associated data in order to build a larger dataset for understanding, modeling and forecasting Starbuck’s current and future business strategy within the United States.

Dataset expansion and Data Cleaning

The Starbucks locational dataset originally consists of 13 columns, including the brand of the Starbucks store, the unique number and the name for each store, ownership of store, address of the store, contact information of the store, time zone and the longitudinal location data for each store. Since our group’s focus is on domestic business strategy, we have removed all entries outside of the United States and reduced the number of entries from approximately 24,000 entries to 13608 entries. Then we have standardized all the inconsistent postal codes to standard 5-digit zip codes and checked for potentially missing data.

Upon finish cleaning the original dataset, we have continued to seek for related data to expand our current dataset while using the locational data as the matching source and Store Number as the primary key. We have then implemented data from 5 more dataset to provide a wholistic view of the demographic data of each county, including the per capita personal income from 2015-2017 (PCPI), the population each county by the end of 2016, the education level of residents within the county (2013-2017), the race of residents, the number of qualify voters, the category of occupation of residents in percentage, the number of employed residents, the daily commuting method in percentage, crime rate per 100000, the distribution of major crime types and the total population of each county.

All additional data incorporated into the original dataset are mapped based on the county name we previously have identified for each store. For any grouping and naming inconsistency found during the implementation process of additional datasets, we have standardized them with the naming system from the original dataset and recalculated any percentage according to the county population. Other than naming and mapping, very few entries from the additional datasets had missing data and required additional cleaning.

Dataset insight

Since our team would like to have a wholistic understanding Starbucks’ current and future business strategy and we do not have direct access to Starbucks’ sales and consumption data in the United States, we believe the best way to produce a snapshot of the customer profile and explanation for store distribution strategy is to combinate demographic data with its current store location data. To ensure data quality, we have focused our search on data collected and provided by the accredited sites such as government resources, academic databases Kaggle datasets that had been checked by numerous data enthusiasts that used location data as their primary key.

On the website of U.S Department of Commerce, we have successfully found a dataset from 2017 that consist of the per capita personal income (PCPI) at the county level, a demographic variable and potentially good indicator of consumption power. In addition, we have also included PCPI data from the year of 2015 and 2016 to potentially check if there is any correlation between trend of income and the number of stores in county. By comparing combination of PCPI along with the education level, race, gender, and occupation type of residents with the number of stores within each county, we would have a better understanding on which type of neighborhoods and customers do they potentially prefer and how do they accommodate the deployment of that strategy. We could also understand how many customers each store is potentially serving in the county, state, and national level and evaluate which states are Starbucks more popular in, as well as what states do Starbucks still have more potential to develop in. We would also be able to compare counties within the same state to predict whether Starbucks would expand its business into those regions and even expand to counties in the adjacent states.

Besides base demographics above, we have also included the crime rate and poverty percentage to check if the instability and the disparity of wealth within the counties would have an impact on store opening strategies. We are also curious whether the repositioning of Starbucks from a high-end, exclusive premium brand to an accessible brand that serves quality coffee, according to Starbucks’ 10K report, has increased its popularity among those who would potentially buy coffee through the grab-and-go while commuting to work. Hence, we have also incorporated the commuting method of residents and the mean commuting time of each county into the dataset to explore whether the commuting method and average time spent to potentially explore whether they could be correlated with store location selection.

Limitations of the dataset

Despite our best effort, the demographics data applied to the dataset is based on county level instead of neighborhood level, of which limits its application in metropolitan area and potentially be less accurate and/or predictive for counties that are larger in size or counties that are combined within the original dataset. In addition, since all datasets used are found from online sources instead of scraping directly from a website source, we are unable to check whether there are missing entries or if there’s any data reported inaccurately. From a macro perspective, although all datasets are adopted from credible sources, we are unable to further verify if there are any discrepancy among variable definitions across different dataset or if data collectors had miscounted or miscalculated any demographical variables.

Conclusion and Future Application

In conclusion, even though there are a few limitations with the current dataset we have cleaned and built, we believe this dataset could be useful for Starbucks’ competitors from the coffees and drinks sector, fast food sector and even fast moving consumer goods sector as they would be able to build forecasting and predictive models after combining with their own first party data and be able to move along with, or even ahead of Starbucks in terms of business expansion and identify the best counties with most fitting demographics where Starbucks had found success in. Similarly, students and institutions could also leverage the foundation we have built and incorporate information from their academic database to potentially build a more well-rounded business case profile background when studying for the success and business strategy of Starbucks.

References

https://investor.starbucks.com/financial-data/annual-reports/default.aspx (Starbucks 2017 10k)

https://www.kaggle.com/starbucks/store-locations (Original Starbucks Dataset)

https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas (PCPI 2015-2017)

<https://data.ers.usda.gov/reports.aspx?ID=17829> (Demographics: Education data)

<https://www.kaggle.com/tsnowak/us-geographic-codes> (County Conversion Code)

<https://www.kaggle.com/muonneutrino/us-census-demographic-data> (Demographic data)

<https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county> (Demographic Data: Crime rate)

|  |  |  |
| --- | --- | --- |
| index | variable | description |
| 1 | Brand | Starbucks or associated brands |
| 2 | Store Number |  |
| 3 | Store Name |  |
| 4 | Ownership Type | Franchise or Company owned |
| 5 | Street Address |  |
| 6 | City |  |
| 7 | State/Province |  |
| 8 | Country |  |
| 9 | Updated\_PostCode | Cleaned, 5-digit postal code |
| 10 | Timezone |  |
| 11 | Longitude |  |
| 12 | Latitude |  |
| 13 | County Code |  |
| 14 | County Name |  |
| 15 | PCPI\_2015 | Per capita income of 2015 |
| 16 | PCPI\_2016 | Per capita income of 2016 |
| 17 | PCPI\_2017 | Per capita income of 2017 |
| 18 | Less than a high school diploma, 2013-17 | Percentage of population whose educational level is Less than a high school |
| 19 | High school diploma only, 2013-17 | Percentage of population whose educational level is high school |
| 20 | Some college or associate degree, 2013-17 | Percentage of population whose educational level is college |
| 21 | Bachelor's degree or higher, 2013-17 | Percentage of population whose educational level is Bachelor or higher |
| 22 | Poverty% | % under poverty level |
| 23 | TotalPop | Total population |
| 24 | Male\_18+ | Number of men above 18 |
| 25 | Female\_18+ | Number of women above 18 |
| 26 | Hispanic | % of population that is Hispanic/Latino |
| 27 | White | % of population that is white |
| 28 | Black | % of population that is black |
| 29 | Native | % of population that is Native American or Native Alaskan |
| 30 | Asian | % of population that is Asian |
| 31 | Pacific | % of population that is Native Hawaiian or Pacific Islander |
| 32 | VotingAgeCitizen | Number of citizens |
| 33 | Professional | % employed in management, business, science, and arts |
| 34 | Service | % employed in service jobs |
| 35 | Office | % employed in sales and office jobs |
| 36 | Construction | % employed in natural resources, construction, and maintenance |
| 37 | Production | % employed in production, transportation, and material movement |
| 38 | Drive | % commuting alone in a car, van, or truck |
| 39 | Carpool | % carpooling in a car, van, or truck |
| 40 | Transit | % commuting on public transportation |
| 41 | Walk | % walking to work |
| 42 | OtherTransp | % commuting via other means |
| 43 | WorkAtHome | % working at home |
| 44 | MeanCommute | Mean commute time (minutes) |
| 45 | Employed | Number of employed (16+) |
| 46 | PrivateWork | % employed in private industry |
| 47 | PublicWork | % employed in public jobs |
| 48 | SelfEmployed | % self-employed |
| 49 | FamilyWork | % in unpaid family work |
| 50 | Unemployment | Unemployment rate (%) |
| 51 | crime\_rate\_per\_100000 | aggravated assaults |
| 52 | MURDER |  |
| 53 | RAPE |  |
| 54 | ROBBERY |  |
| 55 | AGASSLT |  |
| 56 | BURGLRY |  |
| 57 | LARCENY |  |
| 58 | MVTHEFT | motor vehicle thefts |
| 59 | ARSON |  |