

DS4A FINAL REPORT

Dunkin' vs. Starbucks: Using Foot Traffic Data to Inform Targeted Marketing Strategies

Team 60

Paula Espitia, Kelly Hopkins, Dondré Marable

1. Introduction

1.1 Overview

When the Covid-19 pandemic began to explode across the U.S. in the beginning of 2020, it brought a wave of economic destruction that has impacted many industries and the population as a whole. Behavioral changes as a result of the pandemic have affected the ways in which consumers interact with their surroundings and participate in the economy. For example, individuals may no longer have a daily commute where they usually stop for a cup of coffee on the way to work. Individuals may visit their favorite stores less now as a precautionary measure against exposure to the virus. While it was already important for brands that relied on in store visits to have a keen understanding of foot traffic data for themselves and competitors even before the pandemic, it is imperative now to have a competitive advantage. Foot traffic data can give insights into market penetration, the demographics of visitors, and temporal popularity of store locations, all of which are necessary for effective marketing campaigns. Understanding the demographic profiles of visitors to a specific location and segmenting them by factors such as age, household type, income, and educational attainment can ensure that stores are providing services and an experience tailored toward their customer base.

Since the vast majority of Americans possess a mobile phone and carry these devices with them everywhere, there is a massive amount of foot traffic data that is always being collected. This allows anyone with access to conduct market analysis of competing brands in a predefined geographic location. This type of analysis can be an invaluable tool for franchisees, who may not have the financial resources for an analytics team and lack the technical resources to do the analysis themselves.

1.2 Objective

We decided to apply this type of market analysis to two very popular competing brands, Starbucks and Dunkin', to demonstrate its utility. To narrow our spatial focus, we decided to only look at Starbucks and Dunkin' stores located in the Philadelphia metro area. Our solution is a dashboard that will allow users to easily see the performance of these two competing brands, across days and time. Additionally, we can provide demographic insights for store locations so that operators can better understand their customer base, as well as the customer base of their competitors. Combined, we believe that this information can help enable better and more targeted marketing campaigns that impact the bottom line. To summarize, we have two main objectives for this project. First, we will identify times, both days and specific times of day, that

we believe would be the most effective for a potential marketing campaign, such as targeted mobile offers. Our second objective is to identify a target audience, using demographic profiling, so that the content of the marketing campaigns can be tailored.

2. Data Analysis & Computation

2.1 Datasets and Cleaning

We used SafeGraph's Core Places dataset for Starbucks and Dunkin' location information and foot traffic data. Since the data is licensed and we cannot link directly to it, you can access the Core Places documentation [here](#). This data is updated monthly and includes locations that are both open and closed. Due to limitations around obtaining data older than three months, we were restricted to only looking at data from October, November, and December. For each location, the dataset included the number of visits per day, both daily and hourly popularity counts, and the home census block groups (CBG) of visitors. The CBG data would prove to be valuable for demographic analysis of visitors.

In addition to their paid data products, Safegraph also hosts the free Open Census Portal, which is a standardized and compiled extract of the 7500+ demographic groups that are tracked by the U.S. Census Bureau. We used this census information to get a rough estimate of the demographic breakdowns of the census block groups where the visitors in this dataset live. Once we calculated those percentages, we used them as raw probabilities of the customer segments to which those visitors belong. We recognize that this method won't give us perfectly accurate results, however, it's reasonable to assume that visitors patronizing a store from a census block group that's mostly affluent will be mostly affluent themselves, for example. Even with this optimized census dataset, pulling all of the necessary variables is a tedious process, so we created several functions to help automate some of these repetitive tasks. Below, you can see one of the helper functions we wrote to pull a set of variables from the census data and return a dataframe with the columns renamed.

```
def read_census_data(variable_dict):
    dfs = []

    data_sets = sorted(list(set([x[0:3].lower() for x in variable_dict])))
    print("Data Sets: ", data_sets)

    for data_set in data_sets:
        dfs.append(pd.read_csv(f"{data_path}/raw/census/data/cbg_{data_set}.csv", dtype={'census_block_group': "object"},
                               usecols=["census_block_group"] + [x for x in list(variable_dict.keys()) if x[0:3].lower() == data_set])\
                    .set_index('census_block_group'))

    df = pd.concat(dfs, axis=1)
    df = df.rename(columns=variable_dict)

    return df.reset_index())

read_census_data({
    "B01001e1": "total_population",
    "B02001e2": "pop_white",
    "B02001e3": "pop_black",
    "B02001e5": "pop_asian",
    "B03002e1": "pop_hispanic"
})
```

Data Sets: ['b01', 'b02', 'b03']

	census_block_group	total_population	pop_white	pop_black	pop_asian	pop_hispanic
0	010010201001	745	585	160	0	745
1	010010201002	1265	1083	104	9	1265
2	010010202001	960	361	568	0	960
3	010010202002	1236	615	571	24	1236
4	010010203001	2364	1481	515	27	2364
...
220328	721537506011	921	752	0	0	921
220329	721537506012	2703	2230	207	0	2703
220330	721537506013	1195	1085	25	0	1195
220331	721537506021	2005	1662	68	0	2005
220332	721537506022	736	588	0	0	736

Figure 1. Helper function to pull a set of variables from the census data

2.2 Exploratory Data Analysis

Our entire data ingestion and processing pipeline was written in Python, utilizing common data analysis packages like numpy, pandas, and seaborn. The data itself was fairly easy to download and import into our project, which is to be expected from a paid dataset. Each month of our three-month data sample was delivered as its own csv file along with additional supplemental data sets about Safegraph's sampling metrics. A basic EDA experiment on the Safegraph revealed that it was relatively clean, with minimal missing values. There were a lot of extremely useful features available for us to answer a number of different business problems, but to answer our specific problem statement, the features we ultimately used the most include longitude, latitude, raw_visits, raw_visitors, visits_by_day, visitor_home_cbgs, and poi_cbg. The visitor_home_cbg column was one of the features that come as a json string, so some preprocessing and feature engineering was needed to get that information into a workable

format for analysis. Below is an example of some of the iteration needed to handle each instance of this feature.

```
df = sg[sg['poi_cbg'].notnull()].copy()

dfs = []
for i,row in df.iterrows():
    d = eval(row.visitor_home_cbgs)
    df = pd.DataFrame(data={
        "cbg": list(d.keys()),
        f"visits_{row.brands.lower()}": list(d.values())
    })
    dfs.append(df)

visits = pd.concat(dfs)
visits
```

	cbg	visits_starbucks	visits_wawa
0	420912032043	20.0	NaN
1	420912012013	9.0	NaN
2	420912032044	9.0	NaN
3	420912032082	9.0	NaN
4	420912068011	8.0	NaN
...
120	340010112024	NaN	4.0
121	340076109001	NaN	4.0
122	340076088004	NaN	4.0
123	340076074021	NaN	4.0
124	421010313004	NaN	4.0

Figure 2. Processing of visitor_home_cbg column

Initially, we focused our EDA on a single store and made basic charts illustrating popularity by day and popularity by hour. Later, we realized a better approach would be to look at the CBGs that contain both a Dunkin' and Starbucks store. We created a new dataframe showing the number of visits to Dunkin' and Starbucks in each CBG that contained both locations (Figure 3). Columns providing the top brand and the top brand's percentage of total visits were also added.

```
t[t.top_brand_pct < 1]
```

cbg	visits_starbucks	visits_dunkin	top_brand	top_brand_pct
090138815004	4.0	4.0	tie	0.500000
100010401001	4.0	12.0	dunkin	0.750000
100010402011	18.0	11.0	starbucks	0.620690
100010402021	28.0	16.0	starbucks	0.636364
100010402022	38.0	44.0	dunkin	0.536585
...
421330212202	4.0	4.0	tie	0.500000
421330240012	4.0	4.0	tie	0.500000
450510604051	4.0	12.0	dunkin	0.750000
511076110251	4.0	4.0	tie	0.500000
511539012321	4.0	5.0	dunkin	0.555556

Figure 3. Resulting data frame for top brands in shared CBGs

Using the resulting dataframe, we were able to create a map comparing foot traffic between Dunkin' and Starbucks in the heart of Philadelphia (Figure 4). Green represents the CBGs where more visitors went to Starbucks and orange represents CBGs where more visitors went to Dunkin'. An opacity attribute was added to color the CBGs more strongly based on the brand's foot traffic percentage share. This analysis gave us the popularity of these brands by location. While Starbucks wins the market penetration battle in downtown Philadelphia, it is overtaken by Dunkin' in the regions surrounding this area.

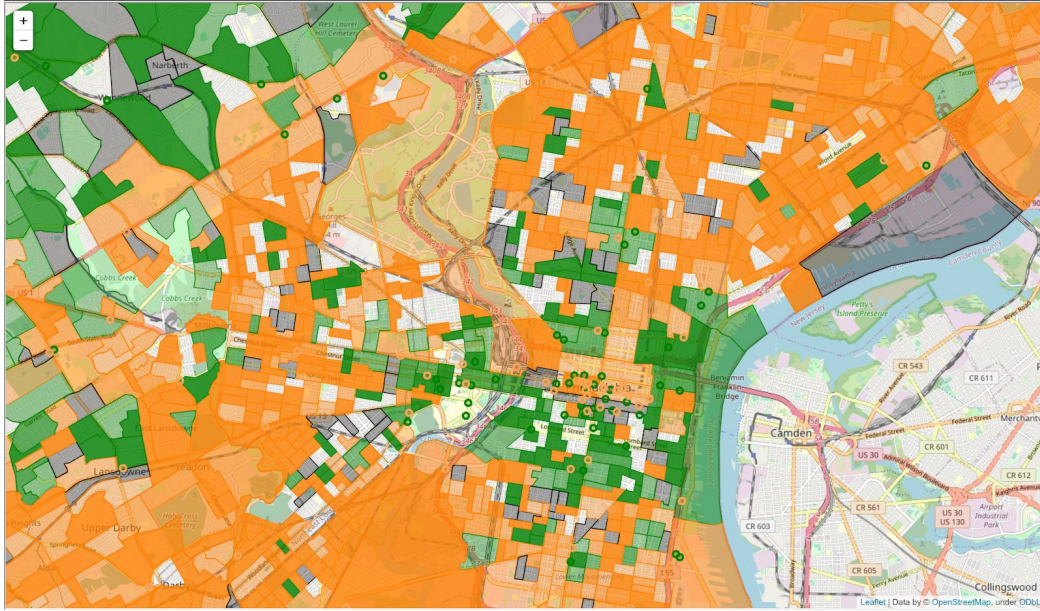


Figure 4. Leaflet showing brand foot traffic in shared CBGs

Since we now had a better understanding of geographic popularity, we turned our focus to temporal popularity. For this, we needed to create similar maps, but at specific days and times. We generated a map for each hour of the day and day of the week that shows foot traffic to both brands. Figure 5 is a sample map that shows foot traffic to the two brands at midnight, while Figure 6 is a sample map showing foot traffic on Monday. There seems to be a Starbucks location in downtown Philadelphia that receives a good amount of foot traffic in the late night hours.

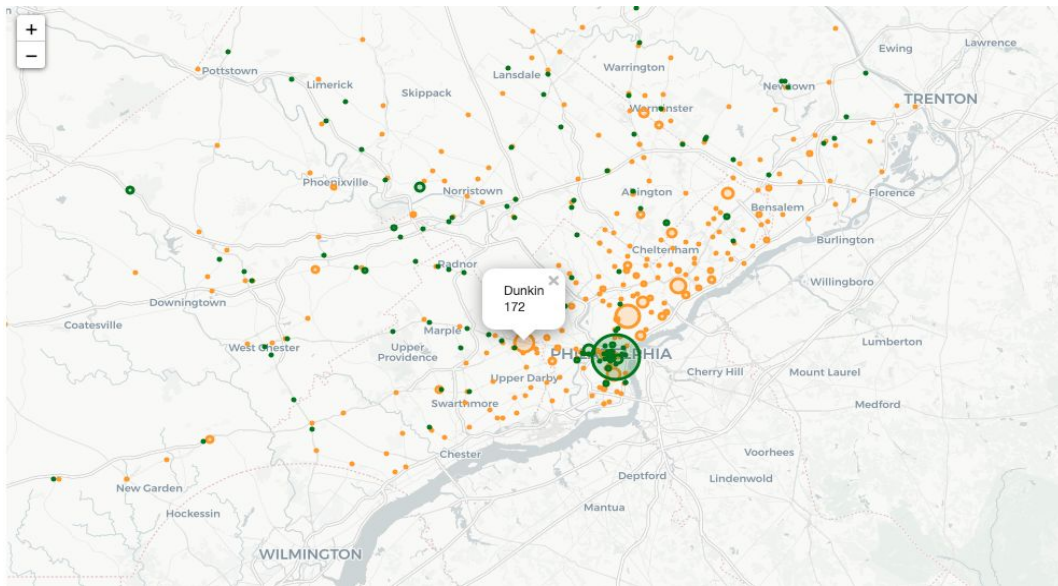


Figure 5. Sample map of foot traffic at midnight

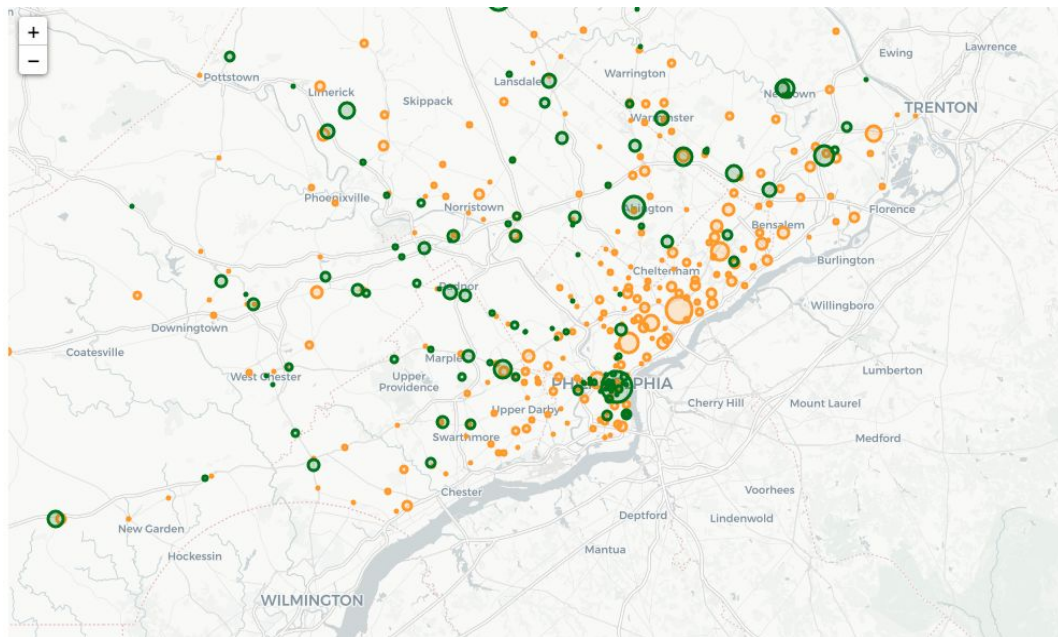


Figure 6. Sample map of foot traffic on Monday

2.3 Statistical Analysis

We decided to focus our analysis on just a few demographic factors, including sex by age, household income, and sex by occupation. We observed that Dunkin' received a larger market share of individuals up to 17 years old and between the ages of 22 and 40. Furthermore, Dunkin' also received more visits from individuals with annual household incomes lower than \$55k.

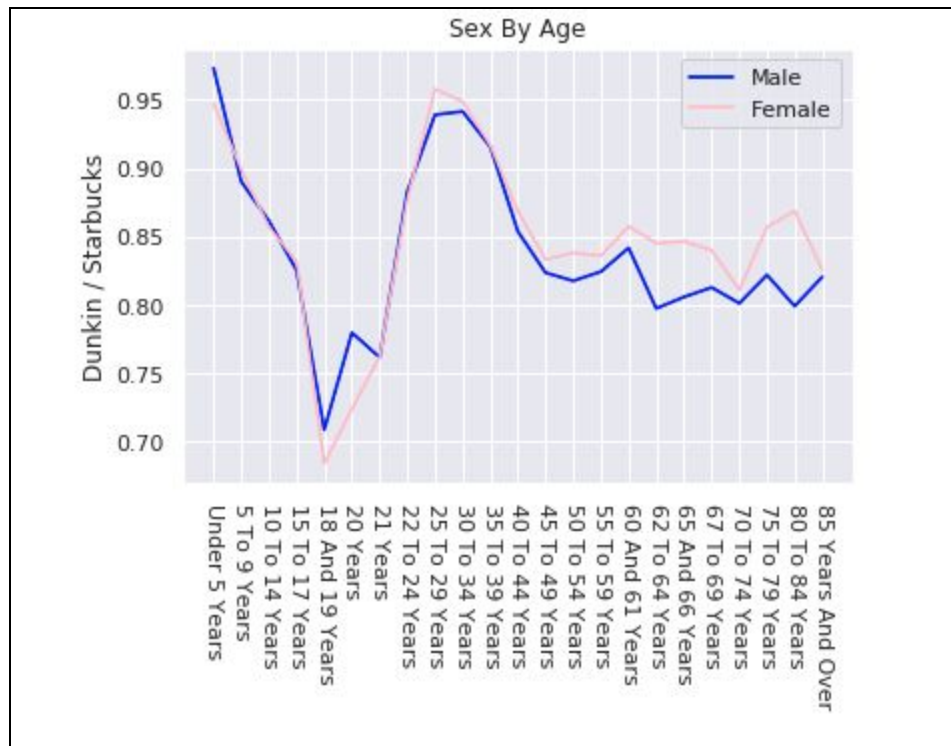


Figure 7. Dunkin' vs. Starbucks demographic analysis of sex by age

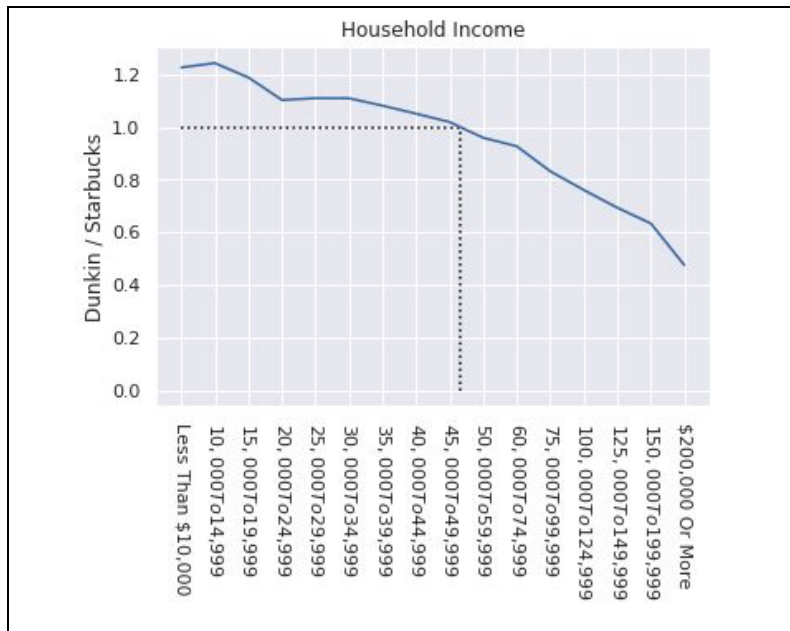


Figure 8. Dunkin' vs. Starbucks demographic analysis of household income

After calculating and aggregating the number of visitors per each chosen demographic group for each CBG in our study, we applied KMeans clustering to these customer segments to determine similar shopping behavior. Using the elbow method, we discovered that 4 was the optimal k value.

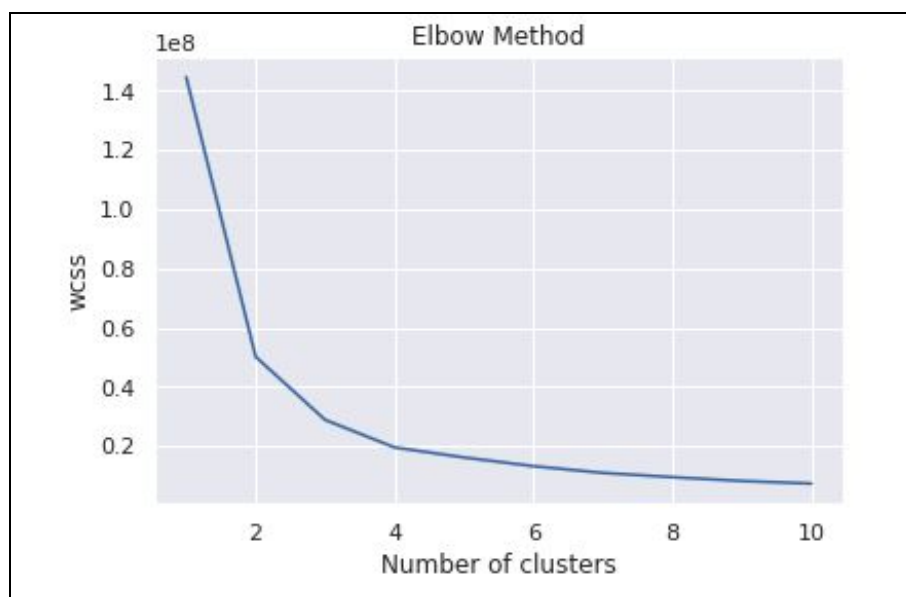


Figure 9. Elbow method to inform number of clusters

Since we had aggregate values for each demographic across our 4 clusters, we needed to normalize the data so that the clusters could be compared more accurately. This was accomplished by dividing each value by the number of rows represented in that cluster. PCA dimension reduction was used to visualize our four clusters.

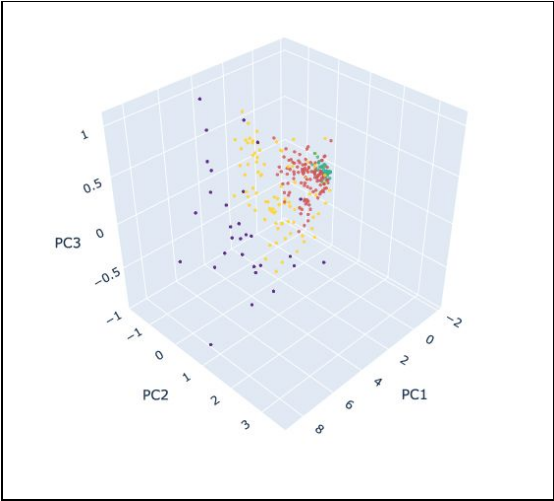


Figure 10. Cluster visualization

As a next step in our analysis, we wanted to create heat maps that show the proportion of each of our chosen demographics in the cluster, for Dunkin’ stores. Though our clusters were not as different as we would have hoped, we did find that Dunkin’ stores in Philadelphia have two significant clusters. One cluster consists of higher income individuals in the 50-60 age range while the other cluster consists of lower income individuals in the 22-34 age range. Larger numbers on the heat map represent higher engagement with customers of that demographic.

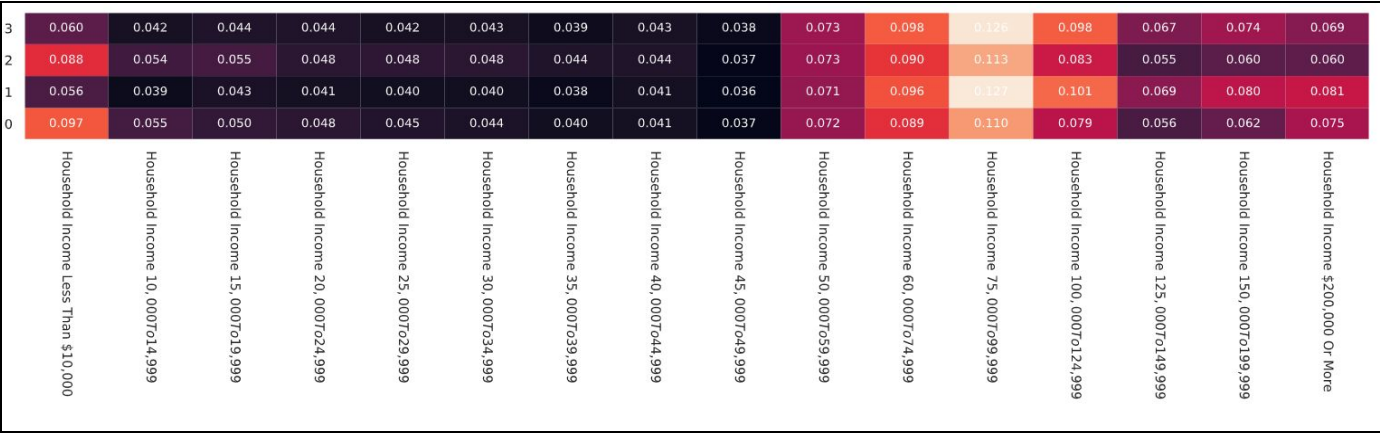


Figure 11. Heat map showing demographic clustering for household income

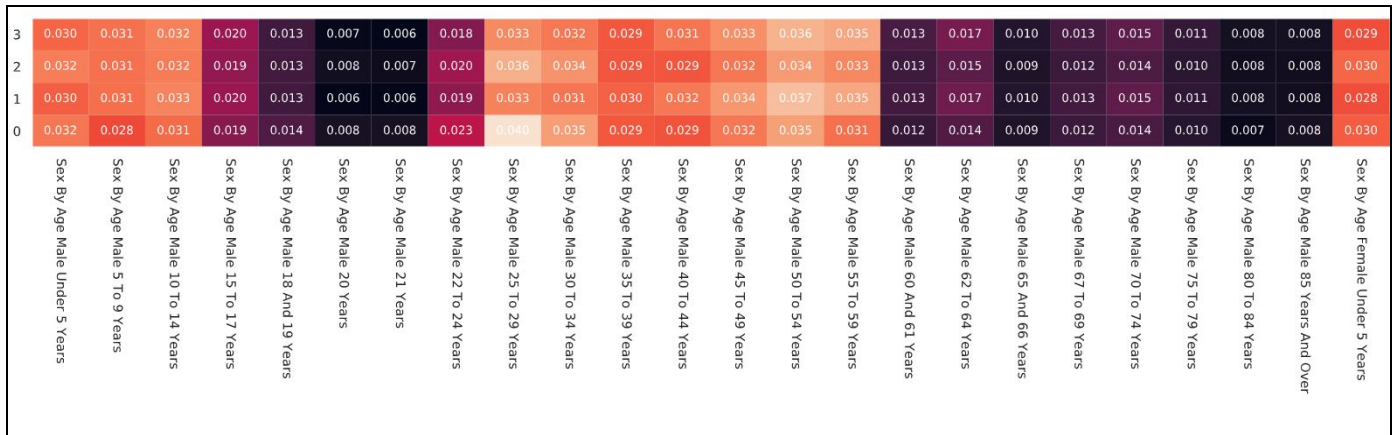


Figure 12. Heat map showing demographic clustering for sex by age

3. Description of Application/Dashboard

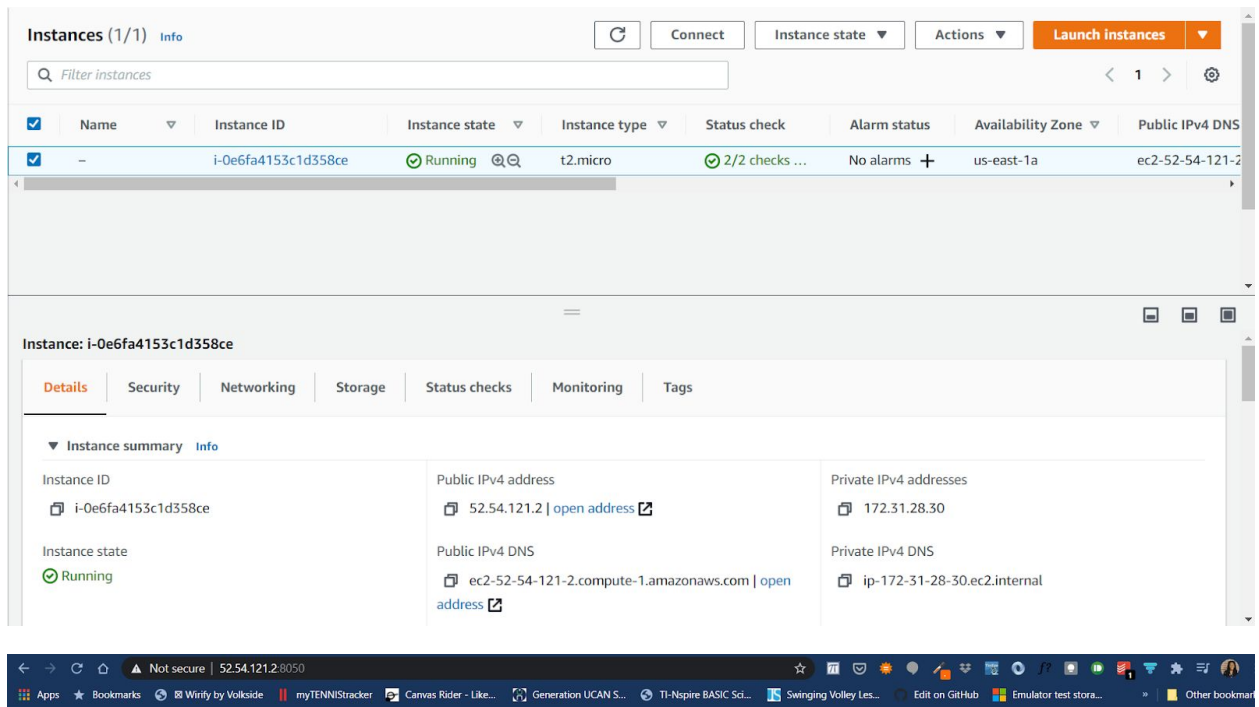
3.1 Use Case

Our dashboard allows users to easily see foot traffic to Dunkin' and Starbucks stores in the Philadelphia metro area. We've included a choropleth map that gives an overview of each brand's foot traffic in the heart of Philadelphia, over the three month period. Additionally, we have charts displaying the brand popularity by day and hour. However we thought it would be beneficial to give users the ability to see the performance of individual locations across days and time. Our interactive map widget gives users the ability to filter by day or time. After selecting one of these options from the dropdown menu, there is a slider that allows the user to pick a day of the week or hour of the day. The size of the markers are proportional to the foot traffic volume to that location.

3.2 Engineering

We created individual maps for each day and hour. On the dash app, we are using a callback function to load the cached html file of the map into an iframe based on the day and time being sent from the slider. This helps load times and ensure that the maps retain their full functionality. To make this application as portable as possible, we decided to recreate the conda environment inside a Docker container so that it can easily be deployed to a variety of hosting locations. Currently, we have the docker container for this application running on a single EC2

instance. Future plans include hosting the application in an environment where it can be easily replicated for greater performance and accessibility, like Elastic Beanstalk on AWS or Heroku.



The top screenshot shows the AWS Management Console interface for an EC2 instance. The instance is named "i-0e6fa4153c1d358ce", is in the "Running" state, and is of type "t2.micro". It is located in the "us-east-1a" availability zone. The instance has a public IPv4 address of "ec2-52-54-121-2" and a public IPv4 DNS address of "ec2-52-54-121-2.compute-1.amazonaws.com".

The bottom screenshot shows a web browser displaying a report titled "Using Customer Demographics in Brand Analysis". The report is analyzing foot traffic data from SafeGraph. The report text states: "What factors draw customers to visit their favorite stores, even in a pandemic? In an new era of reduced travel and extra safety precautions, which brands have prevailed so far? These are some of the questions our team sought to answer through our capstone project as considered the economic impacts of the COVID-19 pandemic. By analyzing consumer shopping behavior, we hope to find trends about how people spend money in their communities." A button labeled "Read the Report" is visible.

Dunkin vs. Starbucks: Foot Traffic Share Oct-Dec 2020



4. Conclusions and Future Work

Regarding the temporal popularity of Dunkin' and Starbucks stores from October - December, we found that stores reached peak popularity midweek and during the early afternoon hours. From one perspective, it may be beneficial to deploy offers during these times that influence customers to spend more money during their visits. On the other hand, focusing on driving foot traffic via targeted offers during non-peak hours may be the better play.

Our second project objective focused on understanding the demographics of visitors so that the content of these marketing campaigns can be tailored. Though we didn't find significant differences in our initial clustering, we do believe that there may be demographic differences in visitors to these two brands as it relates to income and occupation. Our analysis subtly suggests that individuals who have more white collar jobs and higher incomes visit Starbucks at a higher rate while the visitors to Dunkin have lower incomes, comparatively, and tend to have blue collar professions. These insights can be valuable for marketing purposes. For example, it could be beneficial for Starbucks stores to push bulk offers to incentivize visitors to purchase food/drinks for the office.

Next steps for this project would include conducting demographic analysis for individual stores. Additionally, we should look at additional demographic factors, such as education, ethnicity, etc. These data points could be added to our Dash app, thus allowing users to understand the demographics of customers visiting a specific store on a specific day and time. This could be an extremely valuable and powerful tool when you consider that Dunkin' franchises 100% of its locations. Allowing franchisees to view and use this data to inform their marketing campaigns could be the difference between a store remaining open or being closed. Regardless, if used correctly, it can have a definite impact on the bottom line.