Springboard Data Science 6 months program

Capstone Project 1

Predicting the number of visitors in USA neighborhoods

Coline Zeballos

2019

Table of Contents

Proposal *3*
The problem I want to solve **3**
My potential client and why he cares about this problem **3**
The data I will be using, and how I will acquire the data **3**
My deliverables **3**
Data Wrangling *4*
Datasets **4**
Step 1: Importing data and viewing it **4**
Step 2: Identify and group columns for first analysis **4**
Step 3: Split, format, rename columns and deal with NaN values **4**
Step 4: Identify Top ten brands for each column **4**
Conclusions **5**
End of Data Wrangling *5*
Exploratory Data Analysis *6*
Context **6**
Findings **6**
End of Exploratory Data Analysis *9*
Inferential Statistics *10*
Context **10**
Findings **10**
Machine Learning *12*
Context **12**
Findings **12**
End of Machine Learning *13*
Conclusions *14*
Annexe *15*
Column description **15**

Proposal

The problem I want to solve

Predict the number of visitors in each CBG/neighborhood each day and each hour.

Underlying questions: What are the most popular brands in a neighborhood? Are there regional preferences for some brands over others? What times do people visit certain census block groups (CBG) (ex. Manhattan during the day vs. night)

My potential client and why he cares about this problem

Restaurants, shops/store, companies, mayor of a city could be interested to know which area attracts the most people to make decisions such as:

- where to implement a new restaurant, shop, ... strategically (according to the brands also present in the area for example)?
- what type of business to implement in which area (depending on the amount of people during the day)?
- when to organize an event (depending on the amount of people per day): beginning, middle of week, or weekend?
- when and where do we need for public transportation (for a mayor, transportation company)?

The data I will be using, and how I will acquire the data

A Census Block Group (CBG) is the most granular level the US Census Bureau reports data on, and covers ~1500 households. SafeGraph derived the popularity of a CBG or distances traveled to a CBG by analyzing a large panel of GPS movement data. The authors also combined this GPS data with our dataset of 5 million building footprints for Points-of-Interest in the U.S. (SafeGraph Places) to determine visits to places like stores or restaurants. They used these visit counts to derive consumer insights such as top brands (ex. McDonald's) visited in a CBG. The data I will use is already put together in this way: https://www.kaggle.com/safegraph/visit-patterns-by-census-block-group

My deliverables

- This report
- A presentation deck
- Jupyter notebooks containing the code

All deliverables are available on the following Github repository:
https://github.com/Crosita/Capstone-Project-1

Data Wrangling

Datasets

CBG patterns.csv
CBG geographic data.csv

After importing the two datasets and observing their content, we will merge them, split some columns, remove unused columns and rows in order to have a clean dataset on which further analysis will be easier to perform and communication facilitated.

Step 1: Importing data and viewing it

Summary of this step:

- import dataset 1 and 2, check type, shape, delete unnecessary columns, check null rows in key column "census_block_group"
- merge them to create one single dataset "common" containing 220331 rows and 13 columns

Step 2: Identify and group columns for first analysis

Summary of this step:
The goal of this step is to check the consistency of the data. With large datasets like the one we are working on, it is difficult to check the level of sanity of a column knowing the number of rows. I suggest to group the columns by similar format and to perform checks on a group of columns.

- Group 1_column "census_block_group": check that all CBGs are numbers of the same length and unique
- Group 2_columns "visitor_home_cbgs", "visitor_work_cbgs", "popularity_by_day": check that delimiter ":" is always in the same position
- Group 3_columns "raw_visit_count", "raw_visitor_count", "distance_from_home": check that all rows are numbers
- Group 4_columns "related_same_day_brand", "related_same_month_brand", "top_brands": no particular check
- Group 5_column "popularity_by_hour": no particular check

Step 3: Split, format, rename columns and deal with NaN values

Summary of this step:
As we saw in Step 1, some columns require some splitting and information extraction because as such the data is not useful. These concerned columns are: "popularity_by_day", "popularity_by_hour", "visitor_home_cbgs", "visitor_work_cbgs". The goal of this step is therefore to:

- create several columns out of "popularity_by_day" and "popularity_by_hour"
- remove empty rows
- create new columns calculating ratio of "popularit_by_day" and of "popularity_by_hour"
- append all new columns to the main dataset
- extract COUNT, MAX, MIN, AVG from "visitor_home_cbgs" and "visitor_work_cbgs"

Step 4: Identify Top ten brands for each column

Summary of this step:

The goal of this step is to focus on the three columns regarding brands which are: "related_same_day_brand", "related_same_month_brand", "top_brands" and extract information of the top 10 brands. In the end, we will have the following dataframes:

- "common3" containing the main data,
- "brand_day_df" containing the top 10 brands per day
- "brand_month_df" containing the top 10 brands per month
- "top_brands_df" containing the top 10 brands overall

Conclusions

- Our data is now available in 4 csv files: "common3", "brand_day_df", "brand_month_df", "brand_top_df"
- Our data doesn't contain NA rows
- Our columns have the right type (float, object,...)
- Unused columns have been deleted
- Two datasets have been merged together
- Our data is ready to be analyzed

End of Data Wrangling

Exploratory Data Analysis

Context

The goal of the exploratory part of the project is to explore the data, identify some interesting areas of analysis. Each step will answer a specific research question and derive conclusions. This report will summarize the findings of the exploratory data analysis of my dataset. This analysis follows the data wrangling step which aimed at preparing the data for further analysis. As a reminder, the goal of my project is to predict the number of visitors in each CBG. Questions that will be answered around the main goal are *What time do people visit certain census block groups (ex. Manhattan during the day vs. night)? What are the most popular brands in a neighborhood? Are there regional preferences for some brands over others?*

Findings

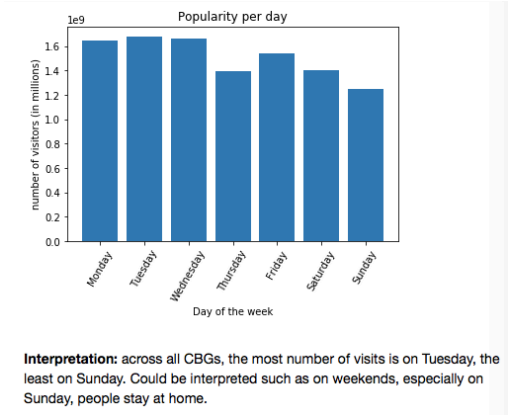
1. What are the top visited brands?
    1. United States Postal Service
    2. Subway
    3. Dollar General
    4. McDonalds
    5. Shell Oil
    6. Cricket Wireless
    7. Starbucks
    8. Family Dollar Stores
    9. The American Legion
    10. Walgreens

2. Where are the most visited CBGs located?

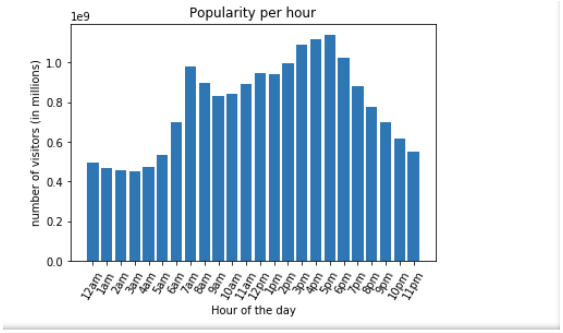
By plotting the top CBGs on the map, we were able to identify the top 10 most visited CBGs in the USA:

1. NY
2. NY
3. Los Angeles
4. Houston
5. NY
6. Atlanta
7. Washington D.C.
8. San Francisco
9. Washington D.C (Maryland)
10. Orlando

3. When do people "visit" or move between CBGs? Which day and at which time?

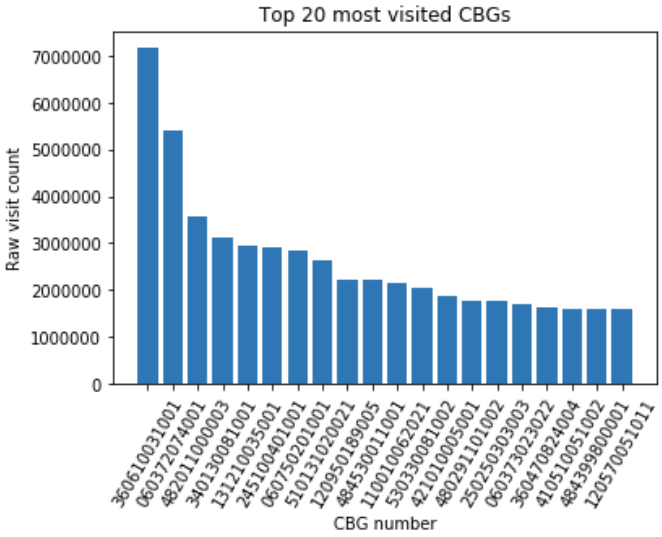


Interpretation: across all CBGs, the most number of visits is on Tuesday, the least on Sunday. Could be interpreted such as on weekends, especially on Sunday, people stay at home.



Interpretation: across all CBGs, during the day the most number of visits is at 17h, with another peak at 7am. This could be interpreted such as people go to work and come back home during these two peak hours.

4. Is there a big difference in popularity/number of visits between the top 10, 50, 100 CBGs?



The most visited CBGs has twice the number of visits than the third most visited CBG. After the top 200, all CBGs have the same amount of visits approx.

Interpretation: the difference between the 2 or 3 first CBGs and the rest identified above is confirmed.

5. What are the most popular brands during the week (Monday to Thursday)? during the weekend (Friday to Sunday)?

To answer these questions, I divided days into categories: week days and week-ends, and calculated the most popular brands for each category.

The most visited brands during the week are:

1. United States Postal Service
2. Subway
3. McDonalds
4. Dollar General
5. Cricket Wireless
6. Shell Oil
7. Starbucks
8. Family Dollar Stores
9. Walgreens
10. CVS

The most visited brands during the weekend are:

1. United States Postal Service
2. Dollar General
3. Subway
4. McDonalds
5. Shell Oil
6. The American Legion
7. Exxon Mobil
8. BP
9. Family Dollar Stores
10. Chevron

6. What are the most popular brands in the morning? in the afternoon? in the evening?

To answer these questions, I divided hours into categories: early morning, morning, afternoon, evening, and calculated the most popular brands for each category.

During the early morning, the most visited brands are:

1. United States Postal Service
2. Dollar General
3. Aflac (American Family Life Assurance)
4. The American Legion
5. Shell Oil
6. Vfw (Veterans Of Foreign Wars)
7. Cricket Wireless
8. Family Dollar Stores
9. Subway
10. Marathon Petroleum

During the morning, the most visited brands are:

1. United States Postal Service
2. Subway
3. Dollar General
4. McDonalds
5. Shell Oil
6. The American Legion
7. Cricket Wireless
8. Family Dollar Stores
9. Starbucks
10. Aflac (American Family Life Assurance)

During the afternoon, the most visited brands are:

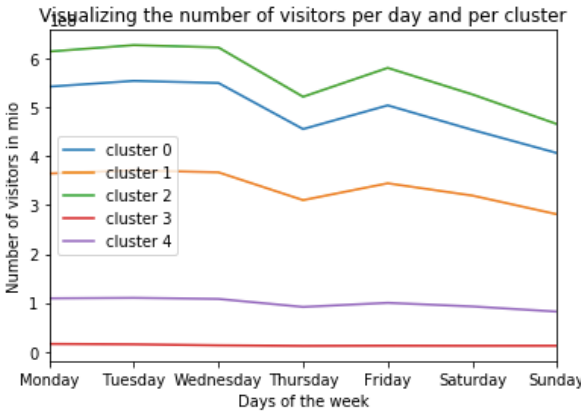
1. United States Postal Service
2. Subway
3. Dollar General
4. McDonalds
5. Shell Oil
6. Cricket Wireless
7. Starbucks
8. Family Dollar Stores
9. Walgreens
10. CVS

During the night/evening, the most visited brands are:

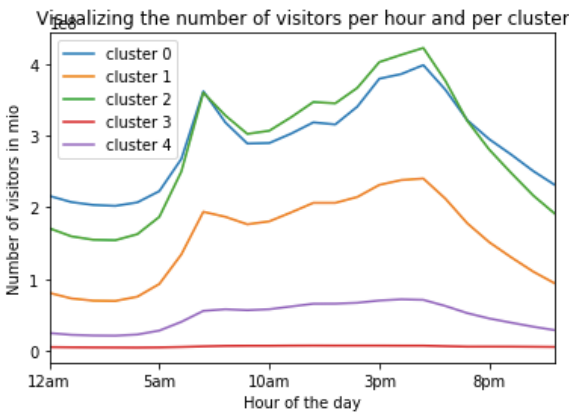
1. United States Postal Service
2. Cricket Wireless
3. The American Legion
4. Dollar General
5. Aflac (American Family Life Assurance)
6. Subway
7. Family Dollar Stores
8. National Association For The Education Of Young Children
9. VFW (Veterans Of Foreign Wars)
10. Boys & Girls Clubs Of America

7. Results of data clustering

I performed a cluster of the dataset based on the columns linked to visits: “raw_visit_count”, “12am” – “11pm”, “Monday”-“Sunday”. After normalizing the data, I calculated the measure of inertia and used the elbow method to determine the ideal number of cluster. Once I obtained the “labels” or cluster id, I added this information to the full dataset and analyzed the characteristics of each cluster.



Conclusion: Cluster 2's characteristics is the most populated cluster. Then closely comes Cluster 0. Cluster 3 is too small to be representative. A similar observation across all clusters is that days from Monday to Wednesday are the highest of the week over all clusters. Thursday undergoes a clear decrease in the number of visitors overall. Friday goes up again, but lower than the begining of the week. Then, the number of visitors goes down until Sunday, which is clearly the day of the week with the lowest number of visitors.



Conclusion:
As expected, the ranking of the clusters is the same as in the previous graph. Also as previously, there are common trends between the clusters. There is a peak at 7am, then down progressively until 9am, then up progressively until a second peak at 4pm. The number of visitors then decreases very quickly until 5am.

The next step is to create a predictive model to predict the number of visitors per cluster (0 to 4). This will be the machine learning part.

End of Exploratory Data Analysis

Inferential Statistics

Context

The goal of the machine learning part of the project is to build a model to predict the number of visitors in a cluster of CBGs, clusters that we have created in the Exploratory Data Analysis section previously. Intuitively, I would like to use the variables visitors per day of the week and those per hour of the day. Before building the predictive model, we will check correlation between the variables mentioned above - in case of a perfect correlation between the independent variable (number of visitors a.k.a. 'raw_visit_count') and the variables to predict it, we will not be able to use these last variables.

Findings

1. Checking correlation between variables

In particular, I will be focusing on the relationship between the variable 'raw_visit_count' and the other set of variables that I plan to use to predict the number of visitors.

```
x_cols1_df.corr()
```

	Unnamed: 0	raw_visit_count
Unnamed: 0	1.000000	-0.031252
raw_visit_count	-0.031252	1.000000
raw_visitor_count	-0.016909	0.815855
distance_from_home	-0.002220	0.021636
latitude	0.008857	-0.137880
longitude	-0.000999	-0.013560
12am	-0.018774	0.801545
1am	-0.017330	0.771748
2am	-0.016828	0.758236
3am	-0.016858	0.757771

Correlation between 'raw_visit_count' and variable representing popularity by hour

There is a positive and quite important correlation between the variables ['12am',...,'11pm'] and 'raw_visit_count'. This allows the possibility to use these variables in the model.

```
x_cols1_df.corr()
```

Monday	-0.029872	0.994422
Tuesday	-0.030834	0.996822
Wednesday	-0.032047	0.993425
Thursday	-0.031268	0.997531
Friday	-0.032046	0.996169
Saturday	-0.031653	0.990627
Sunday	-0.029654	0.984290
Monday(ratio)	0.001670	-0.023853
Tuesday(ratio)	0.000420	-0.024037
Wednesday(ratio)	0.000629	-0.053401
Thursday(ratio)	-0.002432	0.044798
Friday(ratio)	-0.001024	0.061575

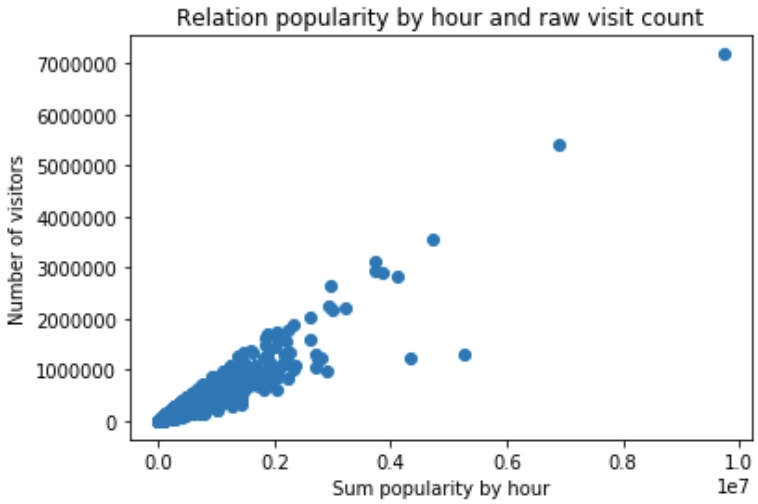
Correlation between 'raw_visit_count' and variable representing popularity by day

The variables ['Monday',...,'Sunday'] are highly, almost fully, correlated to 'raw_visit_count'. This is surprising and unexpected. There is a risk that I will not use these variables to predict the number of visitors in the model since they will not bring any additional information. We will explore this below.

2. Digging the unexpected high correlation between 'raw_visit_count' and popularity by day variables

We see that the sum of visitors on Monday, Tuesday (N=14'935) until Sunday equals exactly the amount of visitors 'raw_visit_count'. This explains the perfect correlation identified above. To conclude, we will not be able to use the popularity by day to predict the number of visitors.

3. Relation between popularity_by_hour and raw_visit_count



This graph confirms the result of the correlation analysis above: it seems there is a good correlation, but not perfect. We will use the variables representing the popularity by hour to predict the number of visitors.

Machine Learning

Context

The goal of the machine learning part of the project is to build a model to predict the number of visitors in a cluster of CBGs, clusters that we have created in the Exploratory Data Analysis section previously. Intuitively, I would like to use the variables "visitors per day of the week" and those "per hour of the day". Before building the predictive model, we checked correlation between the variables mentioned above – we found a perfect correlation between popularity by day and 'raw_visit_count'. Therefore we decided not to use popularity by day variables in the predictive model. However, we will use the variables relative to popularity by hour to predict the number of visitors, represented by independent variable 'raw_visit_count'. Finally, we refined the goal of the project after realizing that our dataset does not allow to build a predictive model to predict the number of visitors per hour of the day: this is because we only have one observation of number of visitors ("raw_visit_count") per CBG; this observation is an average number of visitors during a month. This is a limit of the dataset.

Goal

Predicting the number of visitors using a representative hour of the day

Findings

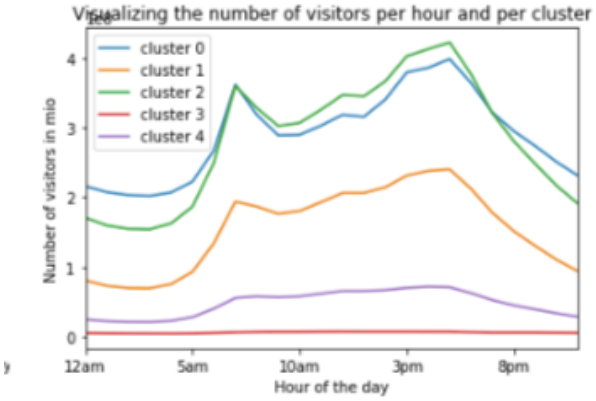
Our dataset contains 220'240 lines, each representing a CBG. In the exploratory data analysis section, we created 5 clusters (0 to 4) grouping CBGs together according to their similarities in visitors per day and hour. We therefore created a predictive model for the number of visitors for each cluster. The predictive model used is a decision tree.

Results:

Cluster	R^2	Representative Hour
Cluster 0	0.91	4pm
Cluster 1	0.72	5pm
Cluster 2	0.81	4pm
Cluster 3	0.59	1am
Cluster 4	0.46	8am

Interpretation:

These results should be interpreted as follows: to estimate the average number of visitors per month in each cluster, measuring the number of visitors during the unique representative hour identified for each cluster is representative.



Conclusion:
As expected, the ranking of the clusters is the same as in the previous graph. Also as previously, there are common trends between the clusters. There is a peak at 7am, then down progressively until 9am, then up progressively until a second peak at 4pm. The number of visitors then decreases very quickly until 5am.

Figure 1.

On the left, in **Figure 1.**, were the results of the clustering analysis at the end of the exploratory data analysis.
All clusters had a peak hour at 7am then a higher one at 4pm.

These results coincidence with the predictive model's using the decision tree method. Cluster 3 seems to be a bit different which can be explained by the small amount of CBGs present in this cluster. (see **Figure 2.**)

11.9 Count the number of lines/CBGs per cluster

```
In [154]: #groupby to see the content of x_cols1
          x_cols1_df.groupby(82).count().iloc[:, 1]

Out[154]: 82
          0    150075
          1     13861
          2     54386
          3        56
          4      1862
          Name: 1, dtype: int64
```

Comment: we have 5 clusters each containing respectively 150'075, 13'861, 54'386, 56, 1'862 CBGs. We are going to analyze what characteristics each contain.

Figure 2.

End of Machine Learning

Conclusions

The initial goals of this project were the following:

- predict the number of visitors in each CBG/neighborhood each day and each hour
- what are the most popular brands in a neighborhood
- are there regional preferences for some brands over others
- what time do people visit certain CBG

How was each goal answered or not, and why?

Goals	Findings
What are the most popular brands in a neighborhood?	We created a function to calculate the top x brands in a dataframe (df). The example given in the jupyter notebook is using a df equal to the overall dataset and x set to 10, which gives the overall top 10 brands across all CBGs. However, if the df is filtered on specific CBGs forming a neighborhood, the function will return the top 10 brands in this neighborhood.
Are there regional preferences for some brands over others?	This question was not developed in the project due to a lack of time and focus regarding the main goal.
What time do people visit certain CBG?	We calculated the total number of visitors per week across all CBGs and represented these amounts on a histogram. We see a distribution of visitors during the week. To go further, the reader/manager can filter on one particular CBG to know its distribution of the number of visitors during the week. The same was done for the number of visitors during the day, per hour.
Predict the number of visitors in each CBG/neighborhood each day and each hour	To simplify the analysis, we created clusters of CBGs based on their similarity in terms of visits per day and hour, and the number of visitors each attracted. Before building a predictive model, we realized that the granularity of the data would not allow to predict the number of visitors per hour and day of the week for the following reason: the data available per CBG is an average of visitors in a month. We decided to use a decision tree to find the most significant hour of the day to predict the monthly visitors per cluster.

Additional questions that were answered:

- Where are the most visited CBGs located?
- What are the most popular brands during the week (Monday to Thursday), and during the weekend (Friday to Sunday)?

To conclude, the lack of granularity in the data did not allow us to perform a predictive model as in depth as initially stated. However, we answered several interesting and useful questions that can be used by restaurants and shops, for event organizers and many other types of organizations. On a more personal touch, this project was my first end to end data science project, including data collection, wrangling, exploring, using machine learning techniques to derive insights and finally communicating results in a simple and understandable way for managers. Thank you to my mentors, Lukas and Rafael for their support!

Annexe

Column description

Column 1: census_block_group
The unique 12-digit FIPS code for the Census Block Group. Please note that some CBGs have leading zeroes.

Column 2: date_range_start
Start time for measurement period as a timestamp in UTC seconds.

Column 3: date_range_end
End time for measurement period as a timestamp in UTC seconds.

Column 4: raw_visit_count
Number of visits seen by our panel to this CBG during the date range.

Column 5: raw_visitor_count
Number of unique visitors seen by our panel to this POI during the date range.

Column 6: visitor_home_cbgs
This column lists all the origin home CBGs for devices that visited a destination in the CBG listed in the column census_block_group (the destination CBG). The number mapped to each home CBG indicates the number of visitors observed from this home CBG that visited census_block_group during this time period. Home CBGs with less than 50 visitors to census_block_group are not included.

Column 7: visitor_work_cbgs
This column lists all the work-location CBGs for devices that visited a destination in the CBG listed in the column census_block_group (the destination CBG). The number mapped to each work CBG indicates the number of visitors observed with this work CBG that visited census_block_group during this time period. Work CBGs with less than 50 visitors to census_block_group are not included.

Column 8: distance_from_home
Median distance from home traveled to CBG by visitors (of visitors whose home we have identified) in meters.

Column 9: related_same_day_brand
Brands that the visitors to this CBG visited on the same day as their visit to the CBG where customer overlap differs by at least 5% from the SafeGraph national average to these brands. Order by strength of difference and limited to top ten brands.

Column 10: related_same_month_brand
Brands that the visitors to this CBG visited on the same month as their visit to the CBG where customer overlap differs by at least 5% from the SafeGraph national average. Order by strength of difference and limited to top ten brands.

Column 11: top_brands
A list of the top brands visited in the CBG during the time period. Limited to top 10 brands.

Column 12: popularity_by_hour
A mapping of hour of the day to the number of visits in each hour over the course of the date range in local time.

Column 13: popularity_by_day
A mapping of day of week to the number of visits on each day (local time) in the course of the date range.