Exploratory Data Analysis

Context:

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?*

Link to Github repository: https://github.com/Crosita/Capstone-Project-1

Findings:

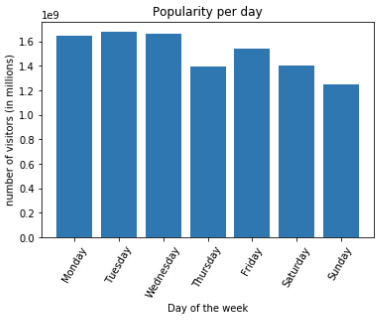
1. **What are the top visited brands?**
    1. United States Postal Service
    2. Subway
    3. Dollar General
    4. McDonalds
    5. Shell Oil
    6. Cricket Wireless
    7. Starbucks
    8. Family Dollar Stores
    9. The American Legion
    10. Walgreens

2. **Where are the most visited CBGs located?**

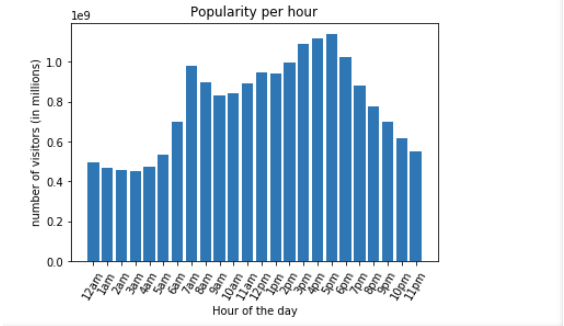
By plotting the top CBGs on the map, we were able to identify the top 10 most visited CBGs in the USA:

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?**

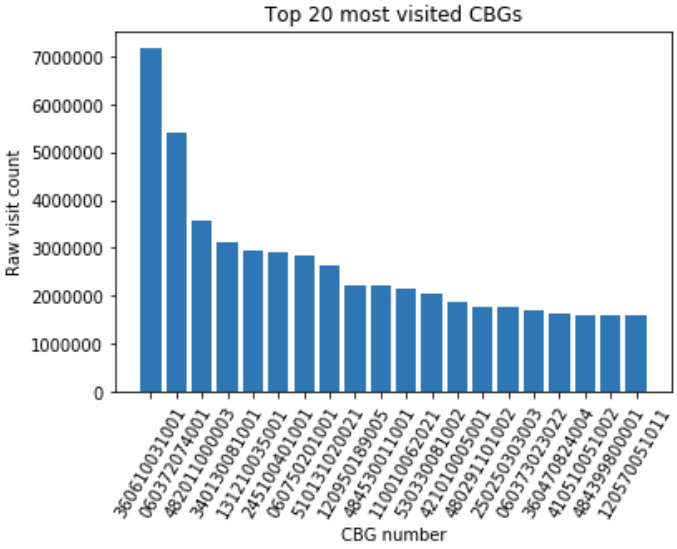


Interpretation: across all CBGs, the most number of visits is on Tuesday, the least on Sunday. Could be interpreted such as on weekends, especially on Sunday, people stay at home.



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?



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

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 vists approx.

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)

During the morning, the most visited brands are:

1. United States Postal Service
2. Subway
3. Dollar General
4. McDonalds

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 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

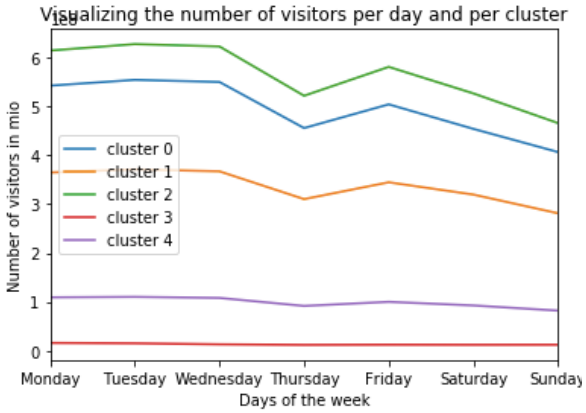
5. Shell Oil
6. The American Legion
7. Cricket Wireless
8. Family Dollar Stores
9. Starbucks
10. Aflac (American Family Life Assurance)

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

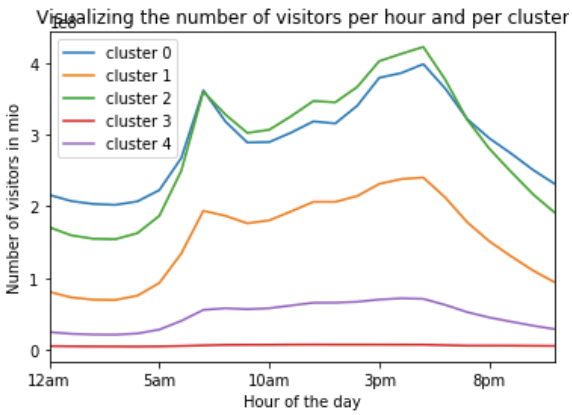
1. United States Postal Service
2. Cricket Wireless
3. The American Legion
4. Dollar General
5. Aflac (American Family Life Assurance)
6. Subway
7. Family Dollar Stores
8. National Association For The Education Of Young Children
9. VFW (Veterans Of Foreign Wars)
10. Boys & Girls Clubs Of America

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 predicting model to predict the number of visitors per cluster (0 to 4). This will be the machine learning part.