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.

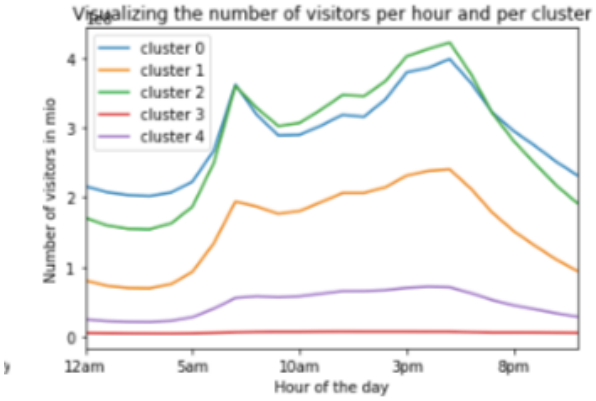
Goal:

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.

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.**)

Figure 1.

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.

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)?