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'.

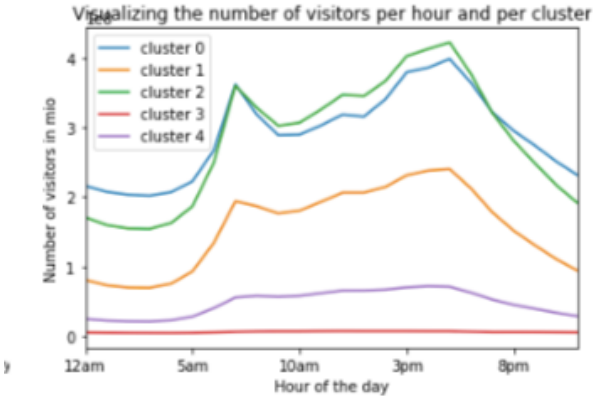
Goal:

Predicting the number of visitors per 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 predicted the number of visitors for each cluster.

Cluster	R^2 LR	Main hour	R^2 DT	Main hour
Cluster 0	0.95	12am	0.91	4pm
Cluster 1	0.77	12am	0.72	5pm
Cluster 2	0.89	12am	0.81	4pm
Cluster 3	0.90	5am	0.59	1am
Cluster 4	0.60	12 am	0.46	8am

LR = linear regression
DT = decision tree



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.

One the left 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. **Not the Linear Regression… why??**
Find the coefficients related to each x variable in X matrix.