# 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 mentionned 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 |

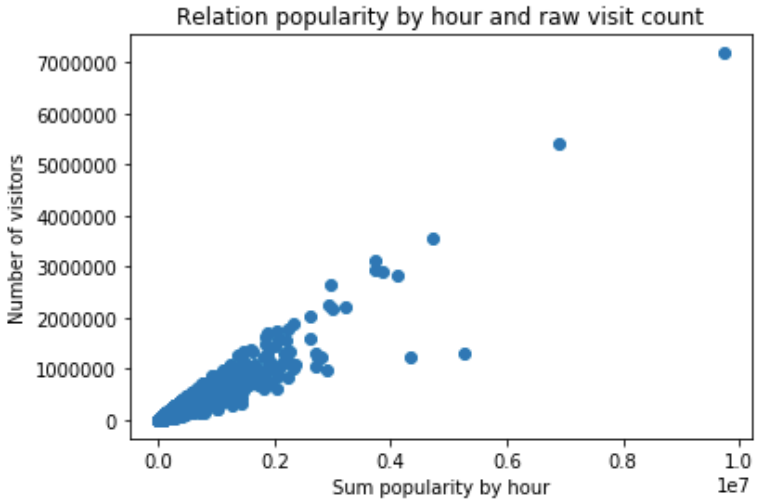
Correlation between 'raw_visit_count' and variable representing popularity by day

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.