**Report on Hypothesis testing**

**Problem Statement**

We have been tasked to understand electric car usage by solving another research question therefore we need to perform hypothesis testing with regards to the claim that we will have made.

We had two filed provided:

* <http://bit.ly/DSCoreAutolibDataset> was the dataset for analysis. It contained data from two regions California and Paris for the course of six months.
* <http://bit.ly/DSCoreAutolibDatasetGlossary> contained the descriptions for the dataset for analysis

Steps to follow when solving the research question

* Find and deal with outliers, anomalies, and missing data within the dataset.
* Plot appropriate univariate and bivariate summaries recording our observations.
* Implement the solution by performing hypothesis testing.

I formulated the hypothesis:

* Ho: There is no difference in the number of blue cars taken from Paris and California.
* Hi: There is a difference in the number of blue cars taken from Paris and California.

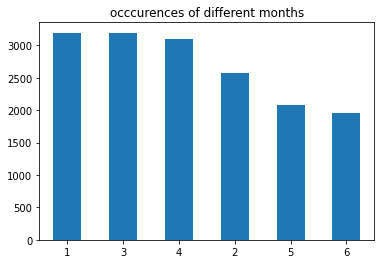
With a significance level of 0.05

This is a hypothetical hypothesis for an investor looking to invest in the electric car industry and wants to pick a region to invest in. The hypothesis aims to check if there is a need to be specific about a region to invest in or randomly pick the region.

**Data Description**

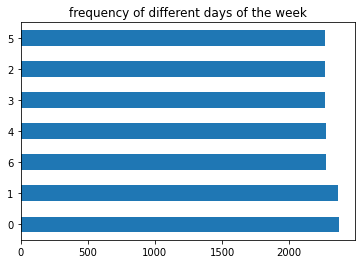
There were no null or duplicated entries from either dataset.

However, There were a lot of outliers out of 16085 entries, 7269 were flagged as outliers. Because getting rid of the outliers would negatively impact our results I did not remove the outliers.

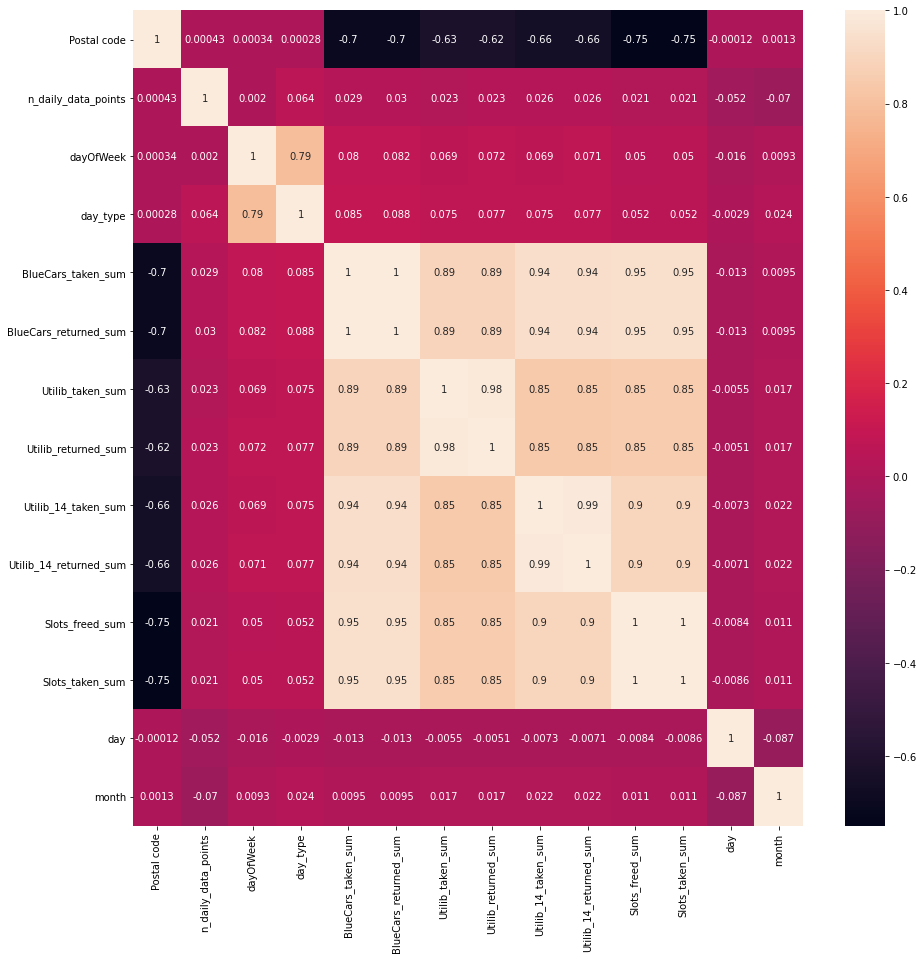
The dataset was pretty elaborate, I performed some univariate and bivariate analysis on it.

This was a representation of how active the months were.

January and February were the most active months



This was a representation of how active the days of the week were.

Monday and Tuesday were the most active days.

A correlation of the Factors in the dataset shows Cars being taken and returned, and slots freed and taken have a high positive correlation.

**Hypothesis Testing Procedure**

The dataset provided contained a pretty large number of postal addresses. However, There were no cities or countries associated with any entry.

From external research from <https://worldpostalcode.com/france/ile-de-france/paris> and <https://www.zip-codes.com/state/ca.asp>, I discovered the postal addresses belonged to two different Regions: Paris and California.

This is where my hypothesis stems from.

I wanted to see if the two had the same level of activity and thus if both were a suitable option for one looking to invest in the electric blue car or were significantly different and one had to pick one region to invest in after a bit more research.

I used two sampling methods to come up with my data:

* Simple random sampling to narrow down to a month for the hypothesis testing.
* Stratified sampling because I realized California had more entries compared to Paris

I set a confidence level of 95%

Once I sampled the data I realized it was not normally distributed so I transformed it using a box-cox transformation.

I wanted to determine if there is a significant difference between the means of two groups, which may be related to blue cars taken. Therefore I used a two-sample t-test for my hypothesis test.

**Hypothesis Testing Results**

* The Test was productive since the results show there is a difference in the number of blue cars picked up in California and Paris.
* The t statistic=2.15 is higher than the critical value 1.64 at a 95% confidence level for this test we reject the null hypothesis and accept the alternative hypothesis.
* The p-value was 0.0321 which was also less than the significance level of 0.05
* The point estimate of n1-n2 is -164
* The confidence level range is from -3.778671665752602e+41 to 2.4746163135172328e+42

**Discussion of Test Sensitivity**

I calculated the power of the test using the statsmodels module.

This estimated the power of the test as 0.1842 with alpha as 0.05 level of significance. Therefore the probability of accepting the null hypothesis when it is false is 18.42 %.

**Summary and Conclusions**

Based on the hypothesis test the activity in California and Paris is not the same meaning one has higher activity, with a likelihood of creating more profit for the investor.

Therefore before investing in any region you should make an investigation to find out which region has a higher activity rate so as to choose a region yielding the highest profit.

This estimated the power of the test as 0.1842 with alpha as 0.05 level of significance. Therefore the probability of accepting the null hypothesis when it is false i.e the probability of avoiding a type II error is 18.42 %.

There were a lot of outliers that may have impacted the hypothesis therefore more research in the matter should be done.

For more detailed work here, [GitHub link](https://github.com/draglar/car-sharing-and-recharging/blob/main/Ian_Muigai_Dsc6_Core_week_4_.ipynb)