REPORT ON HYPOTHESIS TESTING

## Problem Statement

**In an effort to do this, we need to identify some areas and periods of interest via sampling stating the reason for the choice of method, then perform hypothesis testing with regards to the claim that we will have made. An example of claim to test would be "Is the number of Blue Cars taken in area X different than in area Y? Is it greater in area X than in area Z? Etc''. The selected periods of interest are either weekdays or weekends but not a mix of both. You can also consider postal codes 75015 vs 75017 to some of the areas of interest.**

The data we will be working with is from a car sharing company containing details such as geospatial data,postcodes,ID,dates and times of pick up and drop off times.

Our null hypothesis is that the number of Blue Cars taken in area postal codes is different.

Our alternative hypothesis is that the number of Blue Cars taken in area postal codes is not different.

The premise of our hypothesis is to conclude if blue cars are the most used to and from different postal codes and locations.

## Data Description

The data contains the following fields:

* Address
* Cars
* Bluecar counter
* Utilib counter
* Utilib 1.4 counter
* Charge Slots
* Charging Status
* City
* Displayed comment
* ID
* Kind
* Geo point
* Postal code
* Public name
* Rental status
* Scheduled at
* Slots
* Station type
* Status
* Subscription status
* Year
* Month
* Day
* Hour
* minute

The dataset and glossary to use for this project can be found here [[http://bit.ly/DSCoreAutolibDataset (Links to an external site.)](http://bit.ly/DSCoreAutolibDataset)] and here [[Link]](http://bit.ly/DSCoreAutolibDatasetGlossary) respectively.

* The provided dataset is a daily aggregation, by date and postal code, of the number of events on the Autolib network (car-sharing and recharging).

## Hypothesis Testing Procedure

Before performing the hypothesis testing on our dataset we will first have to perform an Exploratory Data Analysis, where we will deal with outliers, anomalies, and missing data within the dataset.

We will then Plot appropriate univariate and bivariate visualizations summaries recording our observations.

We will then after take a sample from the dataset to do our hypothesis testing, picking a sample of about 40 from the dataset to do our test on,which will hopefully satisfy our hypothesis assumption.

We will use an alpha level of 0.05

## Hypothesis Testing Results

We have failed to reject our Null hypothesis :-(

Our T-test ==> 0.0 and P Value ==> 1.0

Since our p value is greater than our alpha level we will fail to reject a null hypothesis.

## 

## Summary and Conclusions

From our analysis we can conclude that :

1. Paris is the most frequented city followed by Gennevilliers and Rueil-Malmaison
2. Our data is normally distributed
3. We have few outliers in our data
4. From our scatter plot our data seems to show positive correlation

From our first analysis we see that we have

two columns with the most number of missing values

* Scheduled at with 99% missing values
* Displayed comment with 97% missing values