HYPOTHESIS TESTING REPORT

# INTRODUCTION

Autolib was an electric car sharing service based in France. It maintained a fleet of all-electric Bolloré Bluecars for public use. The service employed a citywide network of parking and charging stations. As of 3rd July 2016, the service had 3, 980 Bluecars and more that 126, 900 registered subscribers.1, 084 car stations were offered in Paris, with 5, 935 charging points. The business expanded to the cities of Lyon and Bordeaux.

To use the cars, users paid for a membership, with an additional cost depending on how much they drove. Memberships could last a day (10€), a week (15€), a month (30€), or a year (144€). After signing up, the user could drive as often as they liked after finding an available car at a nearby station. The cars were dropped off whenever an open spot was found, and plugged to charge. Users paid extra for each half hour a car was used. 7€ for day members, 6€ for week and month-long members, and 5€ for annual users.

# PROBLEM STATEMENT

## Data

The Autolib’ Car-Sharing Service Company dataset contains a daily aggregation, by date, of the postal code and the number of events on the company’s network (car-sharing and recharging). The data was collected during the first half of 2018. Stations in the dataset were based in the departments of the Île-de-France region in France.

## Random Variable

From the analysis of the dataset, Bluecars were the most popular electric cars used by customers. Therefore, the random variable under investigation was the sum of Bluecars taken from various parking and charging stations in the Île-de-France region. The study focused on determining if the mean sum of Bluecars taken were similar for two departments selected during the process of cluster sampling.

## Assumptions

1. The **null hypothesis (H0)** was that the mean sum of Bluecars taken was equal for two departments in the Île-de-France region.
2. The **alternate hypothesis (H1)** was that the mean sum of Bluecars taken was not equal for two departments in the Île-de-France region.

All postal codes within the data were from the same region, therefore, this hypothesis is based on the assumption that departments within the same region should not have a significant difference in the mean sum of daily Bluecars taken from various postal codes. France is a small country, therefore, rejection of failure to reject the null hypothesis will provide insights on the popularity of the car-sharing service in different departments of the same region. Moreover, Île-de-France is the most populated region and contains France’s major attraction. Therefore, this hypothesis will be used to understand the performance of the company.

# DATA DESCRIPTION

## Overview

Data was collected and availed by the car-sharing service company. The dataset used for this project was [[Link]](http://bit.ly/DSCoreAutolibDataset), and this was the link to its description [[Link]](http://bit.ly/DSCoreAutolibDatasetGlossary). As mentioned earlier, the data was collected in the first half of 2018. It contains information about the use of the three types of electric cars (Bluecar, Utilib and Utilib-14) offered by the company. It also provides information about the availability of slots on a daily basis.

## Description

The original dataset had 16, 085 records, and 14 columns. The categorical columns included the postal code, the day of the week and the type of day (weekday/weekend). The rest of the columns included the date and the quantitative columns that contained information on the usage of cars and the availability of slots.

## Sampling

A department column was added to the original data in order to group the postal codes according to the departments they belong to [[Link]](https://www.worldpostalcodes.org/en/france/region/list-of-postal-codes-in-ile-de-france). A list of the numbered departments was generated (sampling frame - present in the Colab notebook). Three out of 7 clusters were randomly selected. They included Essonne, Yvelines and Val-de-Marine. These clusters were then selected from the population. The resulting sample had 3585 records and 14 columns.

## Data Quality

The data lacked erroneous data as well as missing values.

## Data Preparation

The data was clean. Only the column names were changed to make their format uniform. The only anomalies found were outliers. Removing outliers only slightly reduced the positive skew present in the quantitative variables. Therefore, they were retained as they seemed to be significant to the analysis of the service’s performance.

## Descriptive Statistics

### Univariate Analysis

1. **Categorical Variables**

* Val-de-Marne (population = 1.395 million) had the highest number of records, therefore popularity, while Yvelines (1.437 million) and Essone (1.315 million) had roughly the same number of records. As observed, its popularity is independent of the population. [[Val-de-Marne]](https://en.wikipedia.org/wiki/Val-de-Marne) [[Essonne]](https://en.wikipedia.org/wiki/Essonne) [[Yvelines]](https://en.wikipedia.org/wiki/Yvelines)
* The weekday had the highest frequency in terms of records, especially Monday and Tuesday.
* The month of March had the highest number of records. The popularity of the service peaked in March, but dropped in April and May. There was a slight increase in June.

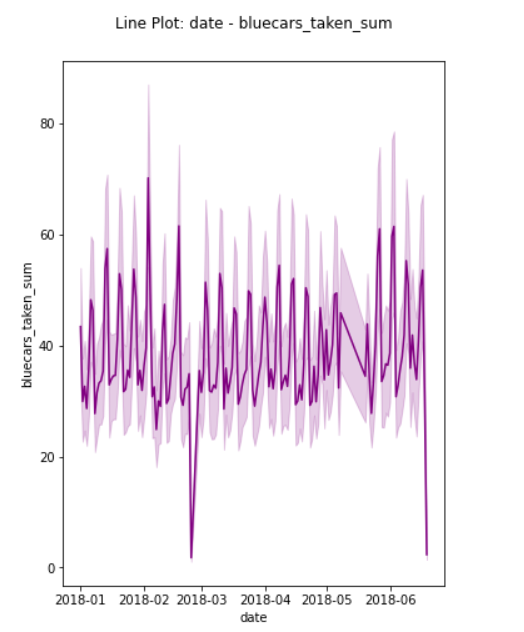
1. **Quantitative Variables**

* From the analysis, Bluecars had the highest mean (38.5) and median (34.0) for cars taken and returned. This was followed by Utilib-14 cars (mean = 2.8, median = 2) and finally Utilib (mean = 1.2, median = 1.0). As observed, Bluecars outperformed the other two. Moreover, the columns were positively skewed as the means were higher than the medians.
* All columns had a mode of zero.Only the data points (1440) and the Bluecar columns (taken = 26, returned = 34) were bimodal. This could be due to either a failure to record values (later filled with zeros), or a lack of customers.
* For the standard deviation, the points column had 33.17, Bluecars taken = 26.57 and Bluecars returned = 26.09. These were the highest standard deviations observed.
* As dor skew, only the points column was negatively skewed. All other variables had a positive skew.
* Finally, after calculation of Kurtosis, all columns were found to be leptokurtic.

### 2. Bivariate Analysis

Significant observations were made in the Numerical-Numerical and Numerical-Categorical columns

1. **Numerical-Numerical Analysis**

* For all cars, the trend observed from January to June of cars taken and returned is constant. However, the use of cars constantly varies. The performance of the service is not constant on a daily basis. This is depicted below, with a plot for Bluecars.
* 
* The number of slots taken and freed are proportional to each other, and for the cars taken and returned.

# Hypothesis Testing Procedure

## Procedure

1. Selection of parametric and non-parametric tests to be applied.
2. Testing the conditions of the selected tests to see if they are suitable for the sample distribution.
3. Selecting a test and carrying out an independent test of the selected samples.
4. Comparing the p-value to the previously selected significance level.
5. Rejection or failure to reject the null hypothesis.
6. Making inferences about the relationship between the samples that were observed.

## Hypotheses

The previously stated null and alternate hypotheses were selected in order to determine the difference, or lack thereof, in the performance of Bluecars in departments of the same region. The original claim is that the samples have the same mean. The hypothesis test is being carried out to disprove this.

## Test Statistic and Acceptance Criteria

1. **Parametric Test**

As stated before, the sample size was 3585, which is greater than 30. Therefore an independent (two sample) Z-test was to be carried out if all conditions for the test held.

**Conditions**

1. The sample size was greater than 30.
2. Data points were to be independent from each other.
3. Normal distribution of data.
4. Random selection of data.
5. Equal sample sizes.

**b) Non-Parametric Test**

If the normality test is failed, a non parametric test will be used. The test of choice is the *Mann Whitney U Test* with the following conditions.

1. The dependent variable should be measured on an ordinal scale or a continuous scale.
2. The independent variable should be two independent, categorical groups.
3. Observations should be independent.
4. Observations are not normally distributed. However, they should follow the same shape (i.e. both are bell-shaped and skewed left).

**Acceptance Criteria**

The significance level = 0.05. This value was to be selected as the test was not strict, but a high level of accuracy was still required.

# Hypothesis Testing Results

The analysis is contained here [[GitHub Repository]](https://github.com/deborahmasibo/Moringa-School-Core-Week-4-IP).

From the Mann Whitney U test, the test statistic was 2610.5. This indicates a significant difference between the samples. The  is less than 0.05. Therefore, the test is significant and there is sufficient evidence to reject the null hypothesis of equal mean.

Population mean for the sum of Blucars taken was 42.53014, while the sample mean of the same was 38.4772. There was a significant difference between the estimated mean and the true mean of 4.0529 (the estimated median was 3.0). This could have been caused by the positive skew present in the data, which persisted even with the removal of outliers.

The sample variable had a positively skewed distribution. Therefore, regular functions could not be used to calculate the confidence interval as they were based on the assumption that the distribution was normal.

# Test Sensitivity

As the variable is positively skewed, normal methods of determining the power of the test would not lead to accurate results. However, the significance level of 0.05 provided a tolerable probability of obtaining a false negative deduction. This was because, as the significance level increased, the type 2 error decreased. However, this led to an increase in the probability of occurrence of a type 1 error. Therefore, an alpha of 0.05 provided a good trade off between the two [[Link]](https://bolt.mph.ufl.edu/6050-6052/unit-4/module-12/errors-and-power/).

The sample size chosen was 10% of the original population, this greatly reduced the chances of getting a type 2 error. This is as the use of Cochran’s and Yamane’s formula would have led to sample sizes of about 400.

The goodness of fit was not determined as the skew of the sample indicates that the original population is also highly skewed. Therefore, the common methods cannot be applied (those that depend on a uniform, normal or chi-square distribution).

# Summary

The entire process began with the initial assumptions of equal mean of the daily sum of Bluecars taken for two departments in the same region. Data was then collected and cluster sampling was done to select random clusters to be used in further analysis. The departments randomly selected were Yvelines, Essonnes and Val-de-Marne. After data preparation and descriptive analysis, it was determined that Val-de-Marne stations had the highest performance. Therefore, the hypothesis test was carried out on Yvelines and Essonnes, which had close population sizes as well as Bluecar service performance. The distribution of the independent variable was skewed, therefore the non-parametric Mann Whitney U test was used.

From the hypothesis test, it was observed that the p-value was much lower than the set significance level of 0.05. Therefore, there was sufficient evidence to reject the null hypothesis of equal daily sum of Bluecars taken. Setting the significance level to be 0.05 and a sample size of 10% ensured that there was a good tradeoff between the type 1 and 2 errors.

# Conclusions

1. The  is less than 0.05. Therefore, the test is significant and there is sufficient evidence to reject the null hypothesis of equal mean for the daily sum of Bluecars taken. Therefore, the daily sum of Bluecars taken in Yvelines differs significantly from that of Essonnes.
2. A significance level of 0.05 and a sample size of 10% ensured that there was a good tradeoff between the type 1 and 2 errors.