Autolib Electric Car Sharing System

Hypothesis report

## Overview

We will work as a Data Scientist for the Autolib electric car-sharing service company to investigate a claim about the blue cars from the provided Autolib dataset.

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.

To work on this project, we will perform the following analysis with Python:

1. Find and deal with outliers, anomalies, and missing data within the dataset.

2. Plot appropriate univariate and bivariate summaries recording our observations.

3. Implement the solution by performing hypothesis testing.

The data set I am working with is the Autolib dataset that is produced on the Moringa School LMS. It contains information concerning 3 brands of electric cars belonging to the Autolib company: the Bluecar, the Utilib, and the Utilib 14.

### 1.1 Research Question

To investigate whether there is a viable difference between Bluecar usage on different day types i.e. weekdays or weekend

### 1.2 Objective

The null hypothesis we will test is that the mean of the number of Bluecars cars taken on the weekdays is equal to that of the Bluecars taken on the week, i.e., there is no difference between the two means.

* μ₁: mean number of bluecars taken during normal weekdays
* μ₂: mean number of bluecars taken during the weekends

Null and Alternative hypothesis

* Ho: μ₁ = μ₂
* H1: μ₁ ≠ μ₂

We will choose a test to analyze the hypothesis based on the distribution of the data

### 1.3 Success Criteria

The research will be successful when below are accomplished.

* Data is successfully explored
* Null and Hypothesis tests are well documented
* Sampling done is representative of the population
* Hypothesis testing is done
* Recommendation and conclusion are drawn

### 1.4 Assessing the situation

1. Resources

* Project dataset. Data was originally extracted from opendataparis.com by Dalberg Data Insights.

1. Dataset extracted from <http://bit.ly/DSCoreAutolibDataset>
2. Data description provided here <http://bit.ly/DSCoreAutolibDatasetGlossary>

* Python Notebook
* Github
* Moringa school canvas access for data and problem statement access

1. Assumptions

* Data sampled will be an accurate representation of the entire dataset

1. Constraints

* The dataset used is large and may make task stall

1. Risks and Contingencies

* The dataset provided may lack vital information

## 2. Data Understanding

### 2.1 Data Description

The dataset has 13 columns and 16805 rows

|  |  |  |
| --- | --- | --- |
| **Column name** | **Explanation** | **Datatype** |
| Postal code | postal code of the area (in Paris) | Object |
| date | date of the row aggregation | integer |
| n\_daily\_data\_points | number of daily data poinst that were available for aggregation, that day | Integer |
| dayOfWeek | identifier of weekday (0: Monday -> 6: Sunday) | Integer |
| day\_type | weekday or weekend | Object |
| BlueCars\_taken\_sum | Number of bluecars taken that date in that area | Integer |
| BlueCars\_returned\_sum | Number of bluecars returned that date in that area | Integer |
| Utilib\_taken\_sum | Number of Utilib taken that date in that area | Integer |
| Utilib\_returned\_sum | Number of Utilib returned that date in that area | Integer |
| Utilib\_14\_taken\_sum | Number of Utilib 1.4 taken that date in that area | Integer |
| Utilib\_14\_returned\_sum | Number of Utilib 1.4 returned that date in that area | Integer |
| Slots\_freed\_sum | Number of recharging slots released that date in that area | Integer |
| Slots\_taken\_sum | Number of recharging slots taken that date in that area | Integer |

#### Observations

* The date column has an integer Datatype. This needs to be changed to datetime dtype
* The column names need to be renamed
* Some columns are unneeded and are to be dropped

### 2.2 Data Loading

1. We load the data to a python notebook using pandas.
2. We check for the data description and check for errors in the dataset

### 2.3 Data Cleaning

i) Check for missing values and duplicates

* No null values or duplicates exist hence no cleaning action.

ii) Rename column names

* Drop ‘\_sum’ suffix
* Rename ‘daysofweek’ to ‘day\_week’

iii) Data Columns Selection/Reduction

By visual data exploration, the columns that are not of interest to answering of the hypotheses are identified and dropped. The selected columns are:

['postal code','n\_daily\_data\_point','utilib\_taken', 'utilib\_returned', 'utilib\_14\_taken', 'utilib\_14\_returned', 'slots\_freed', 'slots\_taken']

iv) Constructing data set/Feature Engineering

* No new column created.

v) Outliers dropping

-‘BlueCars\_taken\_sum’ and 'BlueCars\_returned\_sum' have outliers. The outliers are observed to be valid data. Therebeing, they were not in the first iteration.

Therefore, the outliers were dropped using the IQR approach.

## 3. Hypothesis Testing

### 3.1 Problem statement

The null hypothesis we will test is that the mean of the number of Bluecars cars taken on the weekdays is equal to that of the Bluecars taken on the week, i.e., there is no difference between the two means.

* μ₁: mean number of bluecars taken during normal weekdays
* μ₂: mean number of bluecars taken during the weekends

Null and Alternative hypothesis

* Ho: μ₁ = μ₂
* H1: μ₁ ≠ μ₂

### 3.2 Hypothesis Testing procedure

i) Check for normality. This was done using The Shapiro-Wilk Test and D’Agostino’s K^2 Test

ii) Choose appropriateTest to test the hypotheses

iii) Choose an appropriate sample size

iv) Calculate the statistical significance and p - value

### 3.3 Hypothesis Testing Results

A wilcoxon T test was used to analyse the mean of the two strata. A sample of 1382, which represented a 10% representation of the population. We used a 95% significance level to test the hypothesis.

The p value of 0 is less than alpha, 0.05. Therefore we reject the null hypothesis. This shows that the mean of the number of bluecars rented on the weekends (65.600794) differed with the mean of the number of bluecars rented on the weekdays (59.376767)

We challenge the solution using a smaller sample size. We also reject H0

## 4. Summary and conclusions

#### Univariate analysis

* The number of Bluecars taken and that returned seems to have a relationship. The mean, mode and Median numbers of cars taken and those returned is practically the same.
* Both distributions (Bluecars taken and Bluecars returned) are highly skewed to the right

#### Multivariate analysis

We calculate the correlation coefficient between Bluecars rented and Bluecars returned

-The correlation coefficient of 0.995543 indicates a very strong relationship between the two variables. A scatter plot confirms the relationship

- A line graph shows the trend of cars picked up per year. There are two dips in which they show times cars were picked in a low amount. It can be imagined it was an holiday

#### Hypothesis testing

In conclusion, the study shows significant difference between the number of cars taken during weekdays and during weekends

The autolib management can therefore make strategic decisions:

1. to increase the number of cars available on weekends from this findings.
2. increase the staff roster over the weekend compared to the weekday