**HYPOTHESIS TESTING REPORT**

**1.BUSINESS UNDERSTANDING**

**Business Overview**

Autolib is an electric car sharing service in Paris, France and was founded in December 2011. The system electric cars are supplied by the Bolloré industrial group.Construction of the Autolib' stations began in mid-2011, and 66 of the Bolloré Blue cars were deployed for a two-month preliminary trial period between October and December 2011.

By July 2012, 650 parking and charging stations had been deployed around Paris and by February 2013 there were 4,000 charging points.

Our Data Science team will work for the Autolib electric car-sharing service company to investigate a claim about the blue cars from the provided Autolib dataset.

**Business Objective**

Our main objective is to investigate the blue car usage in Paris by performing hypothesis testing to see if the sum of blue cars taken from postal code 75001 is greater during weekdays or weekends.

**Business Success Criteria**

We should be able to perform an analysis with the provided dataset by testing our hypothesis successfully and coming up with appropriate conclusions and recommendations.

**Assessing the Situation**

**Assumptions**

The data provided is accurate and there are no missing values

**Resource Inventory**

We have been provided with one dataset and a glossary to explain the variable meanings.

* [<http://bit.ly/DSCoreAutolibDataset>]
* [[Link]](http://bit.ly/DSCoreAutolibDatasetGlossary)

**Technology Used**

* Google Colaboratory(Jupyter Notebook)
* Github

**Implementation Plan**

|  |  |
| --- | --- |
| Phase | Time Frame |
| Formulation of Research Question | 15minutes |
| Business Understanding | 45minutes |
| Data Understanding | 1hour |
| Data Preparation | 4hours |
| Data Analysis | 3hours |
| Summary and Conclusion | 1hour |

**2.DATA UNDERSTANDING**

**Data Mining Goals**

Our data mining goal for the project is to perform hypothesis testing to see if the sum of blue cars taken from postal code 75001 is greater during weekdays or weekends.

**Data Description**

The description of the variables(columns) in the dataset is available in the below link:

* [[Link]](http://bit.ly/DSCoreAutolibDatasetGlossary)

**Data Preparation**

Reading the data

We imported the libraries we needed the for the analysis to the jupyter notebook

We then loaded the dataset and the glossary for easier accessibility and viewing of the variable description.

**Checking the data**

The data set has a total of 16085 rows and 13 columns

**Data Cleaning**

We checked for null, missing and duplicate records and there were none. There were many outliers and we decided to work with them in order to get accurate results in the analysis.

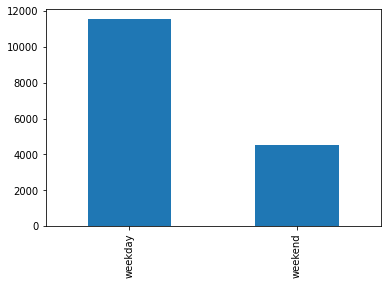
**3.DATA ANALYSIS**

A thorough analysis of the project is available in the Github Repository below:

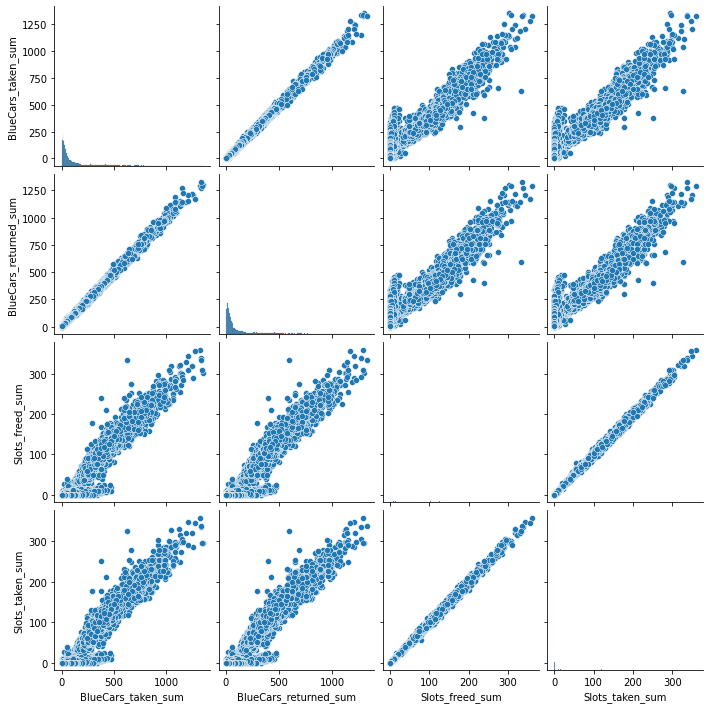
We did Exploratory Data Analysis which was divided in the following areas:

* Univariate
* Bivariate
* Hypothesis Testing

For the univariate analysis we plotted a bar graph to which day types were busiest and found that there was more activity during the weekdays.



For the bivariate analysis we plotted a pairplot with seaborn to check for correlation between the numerical variables. There's a positive correlation between the numerical variables.



**4.HYPOTHESIS TESTING PROCEDURE**

Our hypothesis were as follows:

1.) The Null hypothesis: The sum of blue cars taken from postal code 75001 during the weekdays is greater than or equal to the sum of those taken during the weekends.

2.) The Alternative hypothesis: The sum of blue cars taken from postal code 75001 is less during weekdays than weekends

The idea behind the null hypothesis is to see if there are any differences between blue cars taken during the weekend and weekday of a sample postal code which should be a representation of the overall population.The alternative hypothesis is the opposite of the null hypothesis and will be applied if the null hypothesis is rejected.

For the sampling method, we used stratified sampling to come up with the samples since stratified sampling is unbiased and is appropriate for large datasets.

For the test statistic we will use z score because:

it is a more sensitive descriptor and will provide more accurate results in regards to our hypothesis.

Our sample is reasonably random.

The data is not normally distributed but is from a large sample and z-score can be used for either.

The alpha used is 0.05 to allow for a 5% margin of error.

**5.HYPOTHESIS TESTING RESULTS**

We do not reject the null hypothesis.

From our analysis and results, the z-score = 0.144 which is less than the alpha which is the confidence level hence the null hypothesis is true whereby the sum of blue cars taken from postal code 75001 during the weekdays is greater than or equal to the sum of those taken during the weekends.

The point estimate is our sample mean = 125.92

And from the above point of estimate, we’re 95% confident that our population mean lies between 152.65 and 149.55 which are our sample means measured with different sample sizes.

**Discussion of test sensitivity**

For the test sensitivity we measured the effect of changing the sample sizes and the sample mean did change with a negligible difference.

**Summary and Conclusions**

From the analysis and hypothesis testing report, we can conclude that the sum of blue cars taken from postal code 75001 during the weekdays is greater than or equal to the sum of those taken during the weekends. This forms an accurate assumption that in general in most if not all the stations, the uptake of blue cars is greater during the weekdays than weekends bearing in mind there are more days during the weekdays. Investment is therefore encouraged for more bluecars to present during the week to meet the demand.