Autolib Electric Car Sharing Hypothesis Testing

1. **Business Understanding**
2. **Business Overview**

Autolib is an electric car-sharing service that started in Paris in December 2011. With time it has enlarged its enterprise to other cities such as Lyon. The company provides a fleet of electric blue cars that are available to the public on a paid subscription as well as providing stations for charging and parking. The Autolib cars are available to all persons over the age of 18 years and have a valid French driving license. Moreover, it allows foreigners with valid driving licenses to have access to their services. Any available car could be collected for use from any rental station and returned to any other rental station. Each car had onboard GPS capabilities and can be tracked by the system's operations center.

1. **Business Objective**

The Autolib company has conducted our team to investigate a claim about their blue cars. They have provided us with the Autolib dataset. The objective of this study is to investigate electric (blue car) car usage. This will be achieved by performing Hypothesis testing to investigate if there is any similarity between the mean number of blue cars returned from postcodes starting with "92" and the rest of Paris postcodes.

1. **Business Success Criteria**

The study will be termed as a success if we can conduct exploratory analysis from the data and come up with viable conclusions and recommendations and have successfully conducted hypothesis testing on the claim and interpreted the results correctly.

1. **Assessing the situation**
   1. **Assumptions**

The data provided is accurate.

1. **Resource inventory**
2. **Datasets**

We have been provided with two datasets:

* Autolib Daily Events Dataset [<http://bit.ly/DSCoreAutolibDataset>]
* Column Explanation [[http://bit.ly/DSCoreAutolibDatasetGlossary]](http://bit.ly/DSCoreAutolibDatasetGlossary%5D)

1. **Software**

* GitHub
* Google collab

1. **Implementation plan**

|  |  |
| --- | --- |
| **Phase** | **Time-Frame** |
| Formulation of Research Question | 30 minutes |
| Business Understanding | 1 hour |
| Data Understanding | 1 hour |
| Data Preparation and Cleaning | 5 hour |
| Data Analysis | 1 hour |
| Summary and Conclusion | 1. inutes |

1. **Data Understanding**

* **Problem Statement**

The data we shall be using the Autilib dataset that the company has provided. It contains 13 columns and 16085 rows. The primary variable that we shall be investigating is the postal code.

To investigate this, our hypothesis will be:

*1.*  ***Null hypothesis*** *is that the mean of blue cars returned in postcodes starting with "92" is greater than or equal to that of all the Paris postcodes.*

*2.* ***Alternative hypothesis*** *is that the mean of blue cars returned in postcodes starting with "92" is less than that of all the Paris postcodes.*

The company is conducting a study to understand electric car usage in Paris. It will seek to understand if the mean number of cars returned from the stations is different. The study will help the company understand how the usage of their services is affected by the geographical locations of the stations. The insights will be keen to the company to understand how to distribute their resources effectively meeting all customer's needs.

* **Data Description**

The company has provided two datasets:

1. **Autolib\_daily\_events\_postal\_code.**

This is the main dataset. It contains the postal codes of different stations around Paris city, date, day of the week encoded with numbers, and day type whether weekday or weekend. In it are the number of Blue cars, Utilib and Utilib14 taken and returned on a given day. The dataset also displays the availability of parking slots in a given station.

1. **Columns\_explaination**.

This dataset gives detailed information about our dataset columns.

|  |  |
| --- | --- |
| **Column Name** | **Explanation** |
| Postal code | Postal code of the area (in Paris) |
| Date | Date of the row aggregation |
| n\_daily\_data\_points | Number of daily data points that were available |
| day of week | Identifier of the week(0: Monday -> 6: Sunday) |
| Day\_type | Weekday or weekend |
| BlueCars\_taken\_sum | Number of Blue cars taken that date in that area |
| BlueCars\_returned\_sum | Number of Blue cars returned that date in that area |
| Utilib\_taken\_sum | Number of Utilib taken that date in that area |
| Utilib\_returned\_sum | Number of Utilib returned that date in that area |
| Utilib\_14\_taken\_sum | Number of Utilib 1.4 taken that date in that area |
| Utilib\_14\_returned\_sum | Number of Utilib 1.4 returned that date in that area |
| Slots\_freed\_sum | Number of recharging slots releases that date in that area |
| Slots\_taken\_sum | Number of recharging slots taken that data in that area |

1. **Data Preparation**
2. **Loading and reading the datasets**

We first imported the required libraries to help in the analysis. After, we loaded our two datasets and created data frames for each. We then previewed both the head (first 5 rows) and tail (last 5 rows) to get an understanding of the information we shall be analyzing.

1. **Exploring the data**

Afterward, we checked the shape of our dataset. Our data contains 13 columns and 16085 rows. We checked for the information of each column, to understand the datatypes of each, as well as a statistical summary of the data.

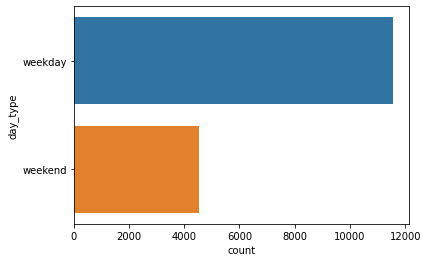
1. **Data cleaning**
2. **Uniformity**: We checked for uniformity in naming the names of the columns and remove any white spaces in our columns as well as renaming our column names to lowercase.
3. **Missing Values**: checked for any null value in our data. Fortunately for us, we found that our data had no missing values.
4. **Duplicates**: We checked if there were any duplicates in our dataset. Fortunate enough, our data contains no duplicates
5. **Irrelevant Columns**: From our business objective, we are seeking to understand blue car usage. This implies that Utilib and Utilib 1.4 columns were irrelevant to our analysis, hence we dropped p the Utilb and Utilb 14 taken and returned columns.
6. **Outliers**: We decided to check for outliers in our dataset. Our dataset had a lot of outliers. We decided not to drop them as dropping them would affect the accuracy of the data analysis leading to results being inconclusive and incorrect.
7. **Analysis**

After cleaning our dataset, it was time to analyze to answer our research objective. Detailed Analyses were conducted in google collab notebook and can be found in my GitHub [[https://github.com/Rino04/Autolib-Electric-Car-Sharing\_Hypothesis-Testing]](https://github.com/Rino04/Autolib-Electric-Car-Sharing_Hypothesis-Testing%5D)

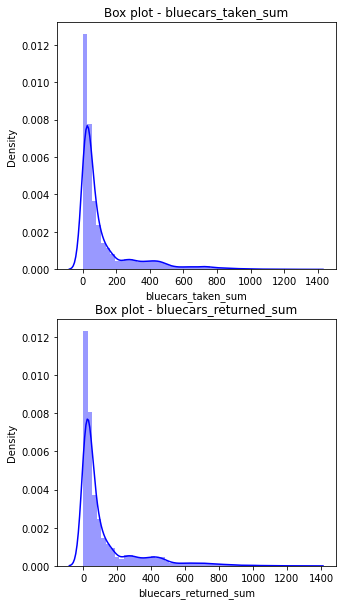
Our analysis was divided into:

1. **Univariate Analysis**

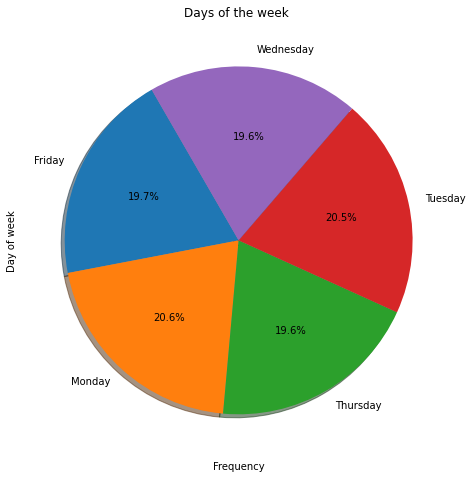
We plotted a bar plot to visualize the blue car usage on both weekdays and weekends. The chart showed that more blue cars were used on weekdays more than on weekends.



We then plotted histograms to visualize the distribution of blue cars taken and returned. It was evident that the distribution was skewed to the right which we confirmed by calculating both the measures of central tendency and dispersion.

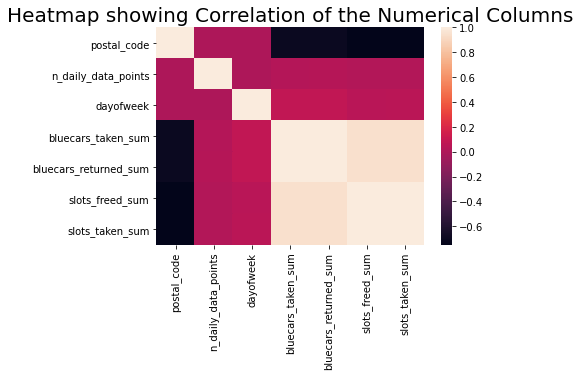


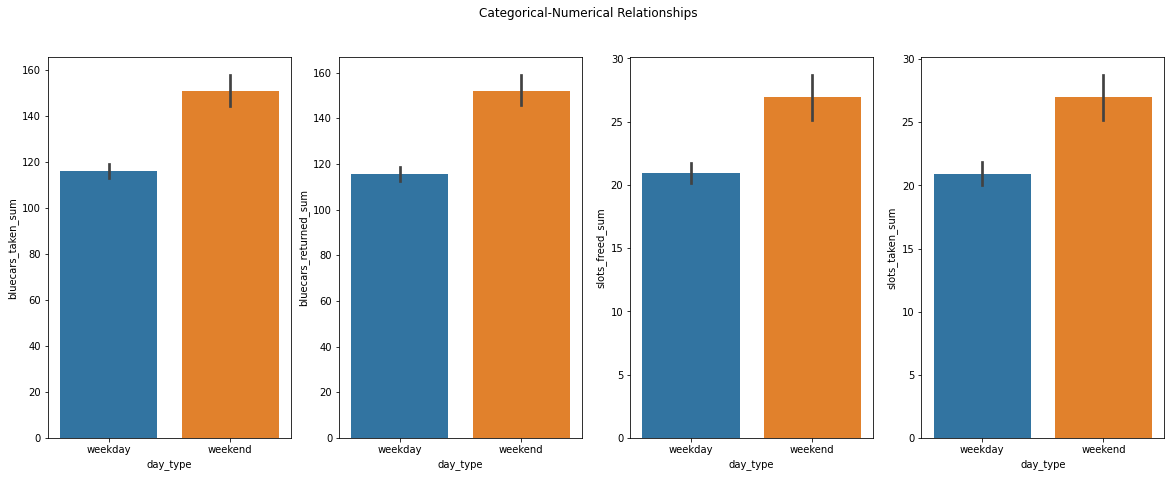
After understanding that weekdays are when most blue cars are used, we wanted to understand which specific day was the busiest. Plotting a pie chart, we were able to conclude that all the days were relatively busy with very slight variation.



1. **Bivariate Analysis**

After understanding the characteristics of each column, we wanted to understand the relationship between the columns. To understand the correlation between each other, we used the heat maps which showed that there was a positive correlation between blue cars taken and returned as well as slots freed and slots taken.



We then went ahead to understand the relationship between categorical variables and numeric variables. It was noted that during the weekends there was more activity than on the weekdays. This might be because many people have a lot of free time during these days. 

1. **Hypothesis Testing**

It's testing an assumption of the population from a sample. In Our case, we were testing the assumption that there is a difference between the blue cars returned for postal codes that begin with "92" and that of the rest of the postal codes in Paris city. To be able to achieve this, we used a stratified sampling technique. We filtered our population to only data containing postal codes that begin with "92" which comprised our strata. We went forward and used simple random sampling in selecting the sample of 30 from each stratum.

To perform hypothesis testing we following steps:

a. **Step 1: Formulate the null hypothesis and the alternative hypothesis.**

***The null hypothesis*** *is that the mean of blue cars returned in postcodes starting with "92" is greater than or equal to that of all the Paris postcodes.*

***The alternative hypothesis*** *is that the mean of blue cars returned in postcodes starting with "92" is less than that of all the Paris postcodes.*

The company is conducting a study to understand electric car usage in Paris. It will seek to understand if the mean number of cars returned from the stations is different. The study will help the company understand how the usage of their services is affected by the geographical locations of the stations. The insights will be keen to the company to understand how to distribute their resources effectively meeting all customer's needs.

b. **Step 2: Identify a test statistic and significance level.**

Each stratum contained rows varying from 147-156. We decided to take our sample size for each stratum to be 30, implying our test statistic would be z-test.

To perform this test, the following assumptions, have to be met:

1. The data follow the normal probability distribution.
2. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. We have used simple random sampling to select the rows in each postal code.
3. The population standard deviation is known. We were able to calculate it from the population.

Having met all the conditions, we choose the significance level of 5%. This meant that there was a 5% chance of committing a type 1 error.

c. **Step 3: Computing the test statistic and p-value**

After calculating the z-test, the z-score was -0.28176232285079406 and the p-value was 0.6109371185721479

d. **Step 4: Analyze the results**

Comparing our p-value to the alpha value we determined that it was greater than alpha. We failed to reject the null hypothesis and accepted the alternative hypothesis. This was also confirmed by comparing the z-score gotten to the z-critical (2.65) which was less hence failing to reject the null hypothesis.

e. **Step 5: Interpret the results**

The p-value is greater than the alpha therefore, it is not statistically significant. This indicates strong evidence for the null hypothesis hence we fail to reject the null hypothesis and conclude that the mean blue cars returned of postal codes starting with "92" is greater than or equal to the mean blue cars returned of all Paris postal codes.

**Point Estimate:**

Our point estimate is the population means of blue cars returned from postal codes starting with "92".

**Confidence interval**

We constructed a confidence interval around the sample mean using a 95% confidence interval. Using the standard error and sample mean, we found our confidence interval lies between 69.30468078923072 and 77.98605995151003. Our sample mean was 73.64537037037037 which was between the range.

**Test Sensitivity**

To perform test sensitivity, we changed our sample size to 60 and increased our random state to 30. We found that the p-value was still greater than alpha hence failing to reject the null hypothesis. This determines the rigidity of our method.

**5. Summary and Conclusion**

From our analysis, the mean for the postal code starting with '92' was higher than the mean of all the postal codes. The z-score (-0.281) is less than the z-critical (1.625), as well as the p-value (0.61) is higher than alpha value (0.05) hence we failed to reject the null hypothesis. We have also conducted a test sensitivity to determine the rigidity of the method. We did this by changing the sample size to 80 and the random state to 30 and the results still showed that there was no means greater. This means that the postal codes starting with '92' have a higher mean of the blue cars being returned daily.