**AUTOLIB ELECTRIC LTD HYPOTHESIS TESTING REPORT**

* **PROBLEM STATEMENT:**

Autolib electric car sharing company has provided us with data regarding the car sharing service that contains details as to the number of cars returned and taken based on 3 types of cars available. The null hypothesis I will be testing is whether the mean number of cars returned/taken is the same on both Saturdays and Sundays. I have chosen this hypothesis in order to find out whether both days of the weekend have the same high activity level or they differ.

* **DATA DESCRIPTION:**

The dataset has been provided by Autolib electric car sharing company and it contains 16,085 rows and 13 columns. Of these 13 columns, we shall be focusing our attention primarily on the **“BlueCars\_returned\_sum”** and the **“BlueCars\_taken\_sum”** columns since Autolib expects us to investigate claims about the blue cars. The data contains no null values. The mean and standard deviation of our primary columns is:

|  |  |  |
| --- | --- | --- |
| **COLUMN** | **MEAN** | **STANDARD DEVIATION** |
| **BlueCars\_taken\_sum** | 125.926951 | 185.426579 |
| **BlueCars\_returned\_sum** | 125.912714 | 185.501535 |

* **HYPOTHESIS TESTING PROCEDURE:**

I began by formulating a secondary hypothesis that led to the main hypothesis by looking into whether there was any difference between the mean sum of Blue Cars returned/taken on weekdays and weekends. After this I focused my attention on the main hypothesis by trying to ascertain whether there was any difference between the mean sum of Blue Cars returned/taken on Saturdays and Sundays. This would help inform us whether both days of the weekend experienced the same activity level or they had varying activity level.

|  |  |
| --- | --- |
| **SECONDARY HYPOTHESIS** |  |
| **Null Hypothesis(H0)** | Mean number of blue cars taken/returned on the weekday and weekend is the same |
| **Alternative Hypothesis(H1)** | Mean number of blue cars taken/returned on the weekday is different from that of the weekend |

|  |  |
| --- | --- |
| **PRIMARY HYPOTHESIS** |  |
| **Null Hypothesis** | Mean number of blue cars taken/returned on Sunday is the same as that of Saturday |
| **Alternative Hypothesis** | Mean number of blue cars taken/returned on Sunday is different from that of Saturday |

I used a z test in this test since the data as plotted on the QQ plot showed that it was not adhering to a normal distribution. Also we could calculate the population standard deviation and the size of the sample was greater than 30. Hence all these factors made it seem reasonable to use a z test. I set the significance level at 5% or 95% confidence level.

* **HYPOTHESIS TESTING RESULTS:**

|  |  |  |
| --- | --- | --- |
| **DETAILS** | **ACCEPT/REJECT NULL** | **CONCLUSION** |
| HYPOTHESIS 1 | REJECT NULL & ACCEPT ALTERNATIVE | P value of 0 is smaller than our significance level of 0.05 hence there is sufficient evidence that there’s a difference between the mean of blue cars returned on weekdays and that of weekends |
| HYPOTHESIS 2 | REJECT NULL & ACCEPT ALTERNATIVE | P value of 0 is smaller than our significance level of 0.05 hence there is sufficient evidence that there’s a difference between the mean of blue cars taken on weekdays and that of weekends |
| HYPOTHESIS 3 | FAIL TO REJECT NULL | P value of 0.08035 is greater than the significance level of 0.05 hence there is not sufficient evidence that there’s a difference between the mean of blue cars returned on Sundays and that of Saturdays |
| HYPOTHESIS 4 | REJECT NULL AND ACCEPT ALTERNATIVE | P value of 0.00349 is smaller than our significance level of 0.05 hence there is sufficient evidence that there’s a difference between the mean of blue cars taken on Saturdays and that of Sundays |

* **SUMMARY AND CONCLUSIONS:**

The steps performed throughout the hypothesis test include:

1. Loading the relevant libraries
2. Loading and previewing the dataset
3. Determining the shape of our data i.e. the number of rows and columns
4. Checking the relevant data types of each column
5. Changing the data column data type to the correct type
6. Checking for outliers in all the relevant columns using boxplot
7. Identifying any missing values in our data set
8. Getting the Pearson correlation coefficient for all columns
9. Plotting scatter plots to show relationship between different columns
10. Counting the data points for respective days of the week column
11. Performing cluster sampling based with the days of the week acting as 7 clusters
12. Plotting the data to show the respective clusters for both blue cars returned and taken
13. Plotting a Q-Q plot to see whether the blue cars returned/taken columns follow a normal distribution
14. Performing 4 different hypothesis tests based on the blue cars returned/taken column

**CONCLUSION:**

From my analysis we can be 95% confident that the population mean interval estimate for blue cars returned is between 123.046 and 128.779 while we can also be 95% confident that the population mean interval estimate for blue cars taken is between 123.061 and 128.793. Also we find as illustrated in the prior tables hypothesis 1, 2 and 4 feature the rejection of their null hypotheses owing to their p values being less than the 0.05 significance level. In hypothesis 3 is where we find the p value being greater than the significance level hence the decision to fail to reject its null hypothesis.