**HYPOTHESIS TESTING REPORT ON AUTOLIB CAR SHARING SERVICE**.

**Business Understanding**

**1.1 Business Overview**

**1.2 Research Question**

The main question for this analysis was to investigate whether the use of blue cars was changing over time or if it was constant across different time periods.

**1.3 Business Objective**

The main goal for this analysis is to investigate the electric car usage(blue cars) in Paris to see if there is a difference in the means of blue cars taken from 2 different postal codes by performing a hypothesis testing.

**1.4 Business Success Criteria**

For the analysis to be considered successful,we should perform analysis on the dataset and draw conclusions from it by successfully performing the hypothesis testing and interpret the results correctly.

**1.5 Assessing the situation**

**Resources**

i.Autolib technical support

ii,Data sets

Autolib dataset

Column explanation dataset

iii.Software used

-Python

-Github

-Google Collaboratory

-Google and other websites to get more information.

**Assumptions**

The dataset that the company provided is accurate.

**Constraints**

There could be biases in the data.

**2.Data Understanding**

**2.1 Data Mining Goals**

The data mining goals for this analysis were;

i.Perform exploratory data analysis(Univariate and Bivariate on Autolibs blue car

performance.

ii.Perform hypothesis testing if there is a difference in blue car usage during weekdays

and weekends.

**2.2 Data Description**

We are provide with wit 2 datasets for this project

1.Autolib daily events dataset- This dataset has the postal code and the date of data

entry

It also shows the number of blue cars,utilib and utilib 14 taken and returned at a given

day.It also displays the availability of parking slots and charging slots in different

stations

2.Column explanation dataset.

This dataset provides a detailed explanation of information in the autolib dataset

columns.The column descriptions are;

|  |  |
| --- | --- |
| **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 for aggregation on that day. |
| Day of week | Identifier of weekday (0:Monday->6:Sunday) |
| Day type | Weekday or weekend |
| Blue cars taken sum | Number of blue cars taken that date in that area |
| Blue cars returned sum | Number of blue cars returned that day in that area |
| Utilib taken sum | Number of utilib taken that date in that area |
| Utilib returned sum | Number of utilib returned that day in that area |
| Utilib 14 taken sum | Number of utilib 14 taken that date in that area |
| Slots freed sum | Number of recharging slots released that date in that area |
| Slos taken sum | Number of recharging slots taken that date in that area |

**3.Data Preparation**

**Reading the data**

i.)We first imported all the libraries needed for our analysis and hypothesis testing.

ii.)Then we loaded the datasets into our working environment and created a dataframe.

iii.)After loading the data we previewed the top and last 5 rows of the dataset to understand how it looks.

**Checking the data**

We checked the dataset rows which had 16085 rows and 13 columns.

We also checked the data types of the 13 columns and checked the dataset information and also the summary statistics.

**Data Cleaning**

**i.)**Checking for duplicate values in the dataset.-we checked for duplicate values in the dataset in order to drop them but for our case we did not have duplicate values

ii.)Checking for null values in the dataset in order to work on them but we did not have any null values in our dataset.

iii.)Renaming columns-we removed the white spaces in the column names,replaced them with underscores then we renamed the columns and changed them to lowercase.

iv.) Dropping irrelevant columns-we dropped the columns that were not needed during the analysis.

v.)Checking for outliers-there were many outliers in the blue cars taken and returned but did not drop them since they were many and could affect the results of our analysis.

**3.Analysis**

After the data cleaning,we did data analysis to answer our research questions.

The analysis done were;

i.)Univariate analysis

ii.)Bivariate analysis

iii.)Hypothesis testing

iv.)Parameter point estimate

v.)Confidence interval Construction

**Univariate Analysis**

We plotted a bar graph to understand the car usage between weekdays and weekends.From the bar chart we could see that the blue car usage was higher during weekdays than during weekends.

We then plotted histograms to understand the distribution between blue cars taken and blue cars taken.From both histograms we could see that the distributions were skewed to the right meaning that the mean is greater than the mode.We calculated the mean and mode of the two columns to prove it and found a proof that the mean was greater than the mode in the two columns.

We also went ahead to determine which days were the busiest during weekdays and found out that all days were busy with a slight difference.We did this by plotting a pie chart of the distributions of the blue car during weekdays.

**Bivariate analysis**

We plotted a scatter plot to show the relationship between blue cars returned and the blue cars taken.From the scatter plot we could see that there was a strong correlation between blue cars taken and those that were returned.

**Hypothesis Testing**

We also plotted a scatter plot to show the relationship between slots taken and those released.

From the plot we could see that there was a strong relationship between slots taken and slots released

**Hypothesis Testing**

The analysis sought to investigate the usage of cars if it increased decreased or remained constant,

**Hypothesis Testing Procedure**

We obtained a sample from the dataset since the data was quite large.We grouped the data into postal codes then used random sampling to select a sample from the data.

The hypothesis followed the following procedure;

**Step1:Formulating The null and alternative hypothesis**

HO There is no significant difference between the blue cars taken and the blue cars returned in postal codes 75014 and 94130

HI There is a significant difference between the mean of blue cars taken in postal code 75014 and 94130

Since the alternate hypothesis had a not equal sign ,we concluded that our test would be two tailed.

The hypothesis was done to determine if the average number of ble cars from one postal code was the same in another postal code,

**Step 2:Identifying a test statistic and significance level**

We performed at ttest since our population was relatively small,we used a sample size of 20.We compared the means of two samples,sampled randomly from the dataframe

**Step 3:Computing the test statistic P Value**

After conducting the test in our programming environment the results were;

**-t statistic is: 27.89841085633017**

**p value is: 6.985978734098666e-17**

**Step 4:Analyzing the results to either reject or accept the null hypothesis**

We used the p-value to analyse the results .Since our p-value was less than the stated significance level,we rejected the null hypothesis.

**Step 5:Interpreting the results**

There is a significant difference between the mean of blue cars taken in postal code 75014 and the mean of blue cars taken in postal code 94130.

**Summary and conclusions**

We performed the hypothesis testing to see if there was a difference in the mean of blue cars taken in two different postal codes which were selected randomly.

After performing 2 sample t-tests there is sufficient evidence to conclude that the number of blue cars taken in the 2 postal codes were significantly different.

In order to reduce the chances of errors in our results,we would recommend a replication of these tests with an increased sample size.

The analysis explained in this report was done in a python notebook and can be found here