## ***SILVIA WAKASA BARASA***

## ***HYPOTHESIS TESTING REPORT***

### ***Problem Statement***

This dataset was an autolib dataset that contained details about the operation of cars within Paris. It showed a compilation of dates when the blue cars were picked from and returned to the particular addresses. The claim being investigated was whether there was a difference in the average number of blue cars taken on saturday to that of blue cars taken on sunday during that period.

* **Null hypothesis**

*There is no difference in the average number of blue cars taken on saturday and that of the blue cars taken on sunday?*

* **Alternative hypothesis**

*There is a difference in the average number of blue cars taken on saturday to that of the blue cars taken on sunday(CLAIM:)*

Hypothesis testing is important here because it evaluates two mutually exclusive statements i.e the null and alternative hypothesis about a population to determine which statement is best supported by the sample data

## ***Data Description***

The dataset I used for this investigation was an open dataset about cars in Paris. It contains variables like the postal code of the area which was paris, the dates of data collection. The dates ranged between 1st of January 2018 and 19th June 2018. With also had the number of daily data points that were available for aggregation on the particular days of aggregation within the specified time periods. The days of the week were the usual monday to friday with the specifications and special assignments of days. Weekday or weekend the dataset had the specific days within the time period. The blue cars that were taken and returned, the utilib data and the slots set of data were also contained in the dataset. The problem under investigation was on the averages which would make the null and alternative hypotheses.

It was a set of data that was already collected. However, if i were to collect such comprehensive data, i would use my data response team to go out in the field, collect the data and perform the analysis from which conclusions would later on be made.

## ***Hypothesis testing procedure.***

The dataset was large as it contained 16085 rows and 9 columns . Although, from the dataset i was interested with those days that were weekend(sunday,saturday) only which blue car is taken . i located weekend and the total columns to rows were 4541 \* 10. I then picked a sample here using the stratified sampling method. The reason i used the stratified method is because the method ensures each subgroup within the population receives proper representation within the sample. During my analysis and hypothesis testing, something interesting i came across was that the descriptive statistic for blue car taken and blue car returned was almost similar. This was a good sign as it creates a pattern to determine the future trends of business operations relating to blue cars in the Autolib electric car sharing company.

From the stratified sample ,I got a sample of 454 rows and 5 columns .I used the sample to calculate the z score and p value in hypothesis testing. I considered 5%(0.05) as my alpha level of significance.

I also performed a normality test from a sample dataset for a blue car taken to see if it follows a normal distribution or not. I used the Shapiro-Wilk Test and Quantile\_Quantile(Q-Q) Plot. From the normality test I conclude that the distribution of the blue car taken are not normally distributed over the population. This is so because the sample of bluecar taken does not look like gaussian from Shapiro-Wilk Test and also points in Quantile\_Quantile(Q-Q) Plot are not aligned along the red line. The reason why the bluecar is not normally distributed is due to the presence of outliers . The reason why the outliers were not not dropped is because they are an actual representation of the numbers of bluecar taken.

## ***Hypothesis testing results***

From the hypothesis test, we found that there was not sufficient evidence to prove that there is a difference in the average number of blue cars taken on saturday to that of the blue cars taken on sunday. The z score calculated was 0.1293(rounded to 4decimal places) and p values calculated from the z score was 0.5514.

In the interpretation of the p -value, it was greater than the alpha level of significance which was 0.05. Hence i failed to reject the null(accepted the null).The z critical 1.959963984540054 and confidence intervals were between (137.6527 and 171.7658)

## ***Discussion of test sensitivity.***

The true positive rate (also known as the sensitivity of a test) is the percentage of the population that will get a positive result. The total population can be identified by the test to be 100 percent accurate.Although , any population test that is 100 percent sensitive is extremely rare. A test will be conducted.In our case we conducted a hypothesis test for the blue car taken on weekends. In my case the sensitivity was 100-((5)/2) = 97.5

**NOTE**: *Was a two tail test*.

## ***Summary and conclusion.***

In the course of my project , I performed exploratory data analysis using univariate and bivariate analysis and implemented my solution with hypothesis testing.Since my p-value was greater than my alpha level of significance, I failed to reject the null hypothesis because there was not enough proof that there is a difference in the average number of blue cars taken on saturday to that of the blue cars taken on sunday ( CLAIM).