**DENNIS KIARIE**

**HYPOTHESIS TESTING REPORT**

1. **Problem statement**

This dataset was obtained from Autolib, a car service sharing company , in the city of Paris in France. This dataset contains information about the operation of cars in France. This dataset contains detailed information of dates when Blue cars were taken and returned to the particular addresses.

The claim being investigated was whether or not the average number or blue cars taken was different from the average number of blue cars returned during that period.

**Null hypothesis**: the average number of Blue cars taken is not different that of the Blue cars returned

**Alternative hypothesis:** the average number of Blue cars taken is different that of the Blue cars returned (***CLAIM***)

The interest of this hypothesis is important for understanding the data and its distribution.

1. **Data Description**

The dataset we used for this investigation was obtained from open datasource. This dataset contains information about the Electric cars in the city of france.The provided dataset is a daily aggregation, by date and postal code, of the number of events on the Autolib network (car-sharing and recharging). The dates ranged between January and July of 2018 .The dataset contains other variables such as days i.e Monday to Sunday which are coded with values ranging from 0 to 6. We used days from Monday to Friday and assignment for each day in our investigation.Also, the Blue cars that were taken and returned, the Utilib data, Utilib1.4 data and the Slots set of data.The problem under investigation was on the averages which would make the null and alternative hypotheses.

We used secondary data. However, if we were to collect such comprehensive data, then our data response team would go out in the field, collect the data and perform the analysis from which conclusions would later on be made.

1. **Hypothesis testing procedure**

The dataset we used for investigation consisted of 16, 085 \* 13 record entries. Therefore, from the population provided we picked a sample using a stratified sampling method. Stratified sampling method was mostly preferred because with it the sample would be unbiased. We splitted the data into different sets of strata and then randomly picked a sample from the strata. Since the dataset contained the date when a Blue car was taken and returned, we used that information to determine my hypothesis.

The logic behind the null and alternative hypothesis is that, we used mean to describe our data which in turn assisted us to come up with sampling clusters.It was interesting for us to know if the average number of cars that were taken in a day corresponded to the average number of cars that were returned on the same day, hence making it easier to determine the future trends of business operations relating to Blue cars in the Autolib Electric Car Service Sharing company.

From the stratified sampling, we got a sample of 977 rows × 4 columns entries. This meant that n > 30. As such, we used the z-score to determine the p-value. We performed normality tests on the data since the statistic we used demanded so. The alpha level of significance that we used was 0.05.

1. **Hypothesis Testing Result**

From the hypothesis test, we found that there was no sufficient evidence to prove that the average means of the Blue cars taken and the Bluecars returned are not equal. The z-score value was 1.959963984540054, with the confidence interval being: (27.798305110196864, 70.04611658888194) and as a result, the null hypothesis was not rejected. The p-value was 0.4998761208129123 which was greater than the significance level.

1. **Discussion of Test Sensitivity**

Sensitivity in a statistical test is the measure of performance of a binary classification test. It measures the proportion of the actual positive i.e. the probability of a null hypothesis being true. In this case the sensitivity was 91%.

1. **Summary and Conclusion**

The project was comprehensive and demanding. I performed exploratory data analysis with hypothesis testing as its implementation. Conclusively, I failed to reject the null hypothesis because there was not enough evidence for me to reject the null hypothesis.

Null hypothesis: the average number of blue cars taken is not different that of the blue cars returned

Alternative hypothesis: the average number of blue cars taken is different that of the blue cars returned(*CLAIM*)