Independent Project 4: Data Report

Peter Kiragu

Moringa School

**Data Report**

**Problem Statement**

The data available for analysis includes a record of six months beginning from January 2018 to June 2019 from Autolib. The data consists of information about BlueCars, Utilibs, and Utilib 1.4. The random variables that will be the focus of our investigation is BlueCars Taken.

When investigating BlueCars Taken, the goal for hypothesis testing was to assess whether the average number of cars taken during the weekdays is similar to the average number of cars taken during the weekends.

**The null hypothesis is: H0: μ 1 = μ 2**

**The alternative Hypothesis is: H a: μ 1 ≠ μ 2**

This hypothesis is critical because the client wants to understand how the demand for BlueCars varies across the week. It is common knowledge that during weekdays people are busy at work while on weekends, most people want to see their friends and travel. This analysis would allow the client to understand whether people are willing to share a ride when going to work as much as when going for social events.

**Data Description**

The random variable chosen for analysis, BlueCars taken, records the number of BlueCars taken at a particular day of the month and a specific postal code. This variable is critical as it is indicative of the daily, weekly and monthly demand for the BlueCars. The analysis conducted included plotting of boxplot, pie chart, histogram, and distribution plot to understand the characteristics of this variable.

From the analysis, we know that the average number of BlueCars Taken is 49. The BlueCars data is skewed to the right with skewness of 1.854. The mode for BlueCars Taken is 0 and 12. This means that there were a number of days when no BlueCar was taken, and only 12 cars were taken. The pie chart shows that January, March, and April have the highest percentage of BlueCars taken with 19.8%, 19.5%, and 19.1%, respectively. June has the lowest number of cars taken at 12.5%.

**Hypothesis Testing Procedure**

Hypothesis testing was implemented in five stages. The first stage was the determination of the null and alternative hypothesis we are interested in. Null and alternative hypothesis stem from the claim that the average number of BlueCars taken is different during the weekdays and weekends. The logic behind the null and alternative hypothesis is to compare the means from two samples. One sample being random values from the weekend and random values from weekdays.

The second stage was the filtering of the main dataset and generating two separate datasets with weekdays and weekends. This will allow us to randomly select two samples and use a two-sample hypothesis test to evaluate our claim.

The third stage involved determining the sample size, and the Cochran formula was used to determine the ideal sample size. This study used a sample size of 376. The size of the population being estimated is 16,085.

In the fourth stage in hypothesis, random samples of sample size 376 were generated from the dataset for BlueCars taken. These two datasets were categorized by weekdays and weekend

In the fifth stage, the p value and the z-test were determined using the statistic module from Scipy.

A z-test was used in this case because the sample size is above 30. The level of significance or alpha level adopted is 0.05.

**Hypothesis Testing Results**

From the hypothesis test, the following results were obtained. P-value is 0.0353, while the test statistic was -2.104.

At the 5% significance level, the z-score is ± 1.96. This is because this is a two-tail test. The test statistic falls within the rejection region, so we reject the null hypothesis that states that the average number of BlueCars taken on weekdays is the same as the average number of BlueCars taken during the weekends.

Using the p-value, we see that the p-value calculated is less than 0.05, which means that we should reject the null hypothesis at 5% significance level.

**Discussion of Test Sensitivity**

Sensitivity is the ability to identify the true positive correctly. This is where the null hypothesis is correctly predicted. Our sample size has an impact on the sensitivity of our test. The larger the sample size, the higher the ability to detect sensitivity. As such, a larger sample size should be preferred to ensure that type I and type II errors are avoided. That said, it is essential to note when the population exceeds 20,000, the sample size will not change much. To ensure that an ideal sample size was used, Cochran's formula for calculating sample size was calculated. Also, simple random sampling was used to select elements that would be part of the two samples.

**Summary and Conclusions**

This analysis reveals that the demand for BlueCars is not the same during the weekend and weekdays. This could be the same for the other type of cars. The average number of BlueCars taken during weekdays is 118, while the average number of BlueCars taken during the weekend is 148. Further investigation needs to be conducted to understand why the demand for ride sharing is higher on the weekends than on weekdays.