**HYPOTHESIS TESTING REPORT**

**PROBLEM STATEMENT**

The goal of this hypothesis is to investigate a claim about blue cars from the Autolib dataset . I chose to compare the number of blue cars taken by two different postal areas, that is , 75001 and 93260. I suspect that the average number of blue cars taken, specifically during the weekend , between the two areas is different. This will help the Autolib electric car-sharing service company have insight on how blue cars fair in the two areas. From the analysis, since most of the cars were used during the weekend, I used that as my period of interest.

**Null Hypothesis:**

Is the average number of blue cars taken during the weekend for area 75001 equal to that of area code 93260.?

**H0 : µ1 = µ2**

**Alternate Hypothesis:**

Is the average number of blue cars taken during the weekend for area 75001 not equal to that of area code 93260.?

**Ha : µ1 ≠ µ2**

**Data Description**

The data was collected from between January 2018 and June 2018 and had a total of 13 columns and 16085 records. Since my area of interest was the blue cars , I dropped the columns that were not needed and remained with 6 columns:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Description** | **Data Type** | **Records** |
| Postal code | postal code of the area | Int64 | 16085 |
| date | date of the row aggregation | datetime64[ns] | 16085 |
| dayofweek | identifier of weekday (0: Monday -> 6: Sunday) | int64 | 16085 |
| day\_type | weekday or weekend | object | 16086 |
| bluecars\_taken\_sum | Number of bluecars taken that date in that area | int64 | 16085 |
| bluecars\_returned\_sum | Number of bluecars returned that date in that area | int64 | 16085 |

**Data Cleaning**

I checked for any missing values in the data but found there were no missing values. While looking for outliers, I found that the columns bluecars\_taken sum and bluecars\_returned\_sum showed presence of outliers. I decided to keep them as they would play a role in our analysis. The data did not have duplicate values. I deemed the data clean and safe enough to be begin my analysis

**EDA Process**

From my univariate analysis, I came up with some conclusions about the data:

1. More data points were collected during weekdays compared to the weekend
2. Cars borrowed during the week were evenly distributed.
3. Distribution of blue cars taken is skewed to the right.
4. The mean of Blue cars taken is 125, mode is 12 and median is 46.
5. Distribution of blue cars returned is skewed to right.
6. The mean of Blue cars returned is 125 , mode is 13 and median is 46.

From my bivariate analysis,I came up with these conclusions:

1. Blue cars taken and Blue Cars Returned are positively related.
2. Cars taken and returned during the week were mostly fairly distrubuted but significantly higher on Saturday.
3. Blue Cars Returned and Taken were significantly higher during the weekend

**Hypothesis Testing Procedure**

1. **Stating our Null and Alternate hypotheses.**

I had already stated the null and alternate hypothesis at the beginning of this document.

1. **Set significance level**

I set the significant level(α) at 0.05.

1. **Choosing sample size**

Since our area of focus was 75001 and 93260 , I created two dataframes. One containing information about postal code 75001 during the weekend and the other for postal code 93260 during the weekend. Our data had a total of 44 records from each dataframe. I did not want use the entire data and wanted a sample from each. I chose the simple random sampling procedure as it was easier to implement and picked half of my data as the sample.

1. **Hypothesis Test Results**

**Test statistic and P-value**

I chose to use the two sample T-test unequal variance to perform this hypothesis. This is because the samples had less than 30 records and don’t have equal variance. The distribution also looks like a normal distribution.

Test statistic:

I found our test statistic as 18.466 and our p-value to be 0.00

Since our p-value was less than significant value, we rejected the null hypothesis.

**Point estimation and Interval estimation.**

For point estimation, I found the difference between the population mean and sample mean and concluded:

* The sample mean of blue cars taken from area 75001 underestimates the true mean by -4.2272.
* The sample mean of blue cars taken from area 93260 underestimates the true mean by -0.4318.

For Interval estimation, I set the confidence level to 95% and concluded that:

* The population mean of blue cars taken for area 75001 lies between 170.13 and 195.77
* The population mean of blue cars taken for area 93260 lies between 53.39 and 60.87

1. **Test Sensitivity**

I calculated the power of two sample t test based on the parameters used during the hypothesis test. Statistical power of test was found to be 1.0. This means that the probability of making a Type II error is very low. This means the statistical test is significant.

1. **Conclusion**

Since the p-value < α= 0.05, I rejected the Null Hypothesis and conclude that there is sufficient evidence that the average of blue cars taken during the weekend for the two areas is not equal at 0.05 level of significance.

I chose the two areas at random but if I had taken more time to investigate which particular area each code represents in Paris, my findings would have made more impact.