**TO IDENTIFY THE TOP THREE POSTAL CODE/AREAS THAT HAVE THE HIGHEST NUMBER OF BLUE CARS TAKEN DURING THE WEEKDAY AND DETERMINE IF THEY ARE THE SAME FOR THE 75015 & 75017 POSTAL CODE AREAS.**

**By: David Muna Mwangi**

1. **Problem Statement**

The data being described in the Hypothesis test is obtained from Autolib electric car-sharing service company. The data was collected over a period of 6 months in the year 2018 for the sum of cars taken in a given area with a specific postal code. The data contains the following columns with their respective descriptions.

1. **Random Variable**

The random variable we’re interested in is the number of blue cars taken. The areas of interest as seen in the analysis notebook are postal code 75015 and 75017 over the six month period.

|  |  |
| --- | --- |
| **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, that day |
| dayOfWeek | identifier of weekday (0: Monday -> 6: Sunday) |
| day\_type | weekday or weekend |
| BlueCars\_taken\_sum | Number of blue cars taken that date in that area |
| BlueCars\_returned\_sum | Number of blue cars returned that date in that area |
| Utilib\_taken\_sum | Number of Utilib taken that date in that area |
| Utilib\_returned\_sum | Number of Utilib returned that date in that area |
| Utilib\_14\_taken\_sum | Number of Utilib 1.4 taken that date in that area |
| Utilib\_14\_returned\_sum | Number of Utilib 1.4 returned that date in that area |
| Slots\_freed\_sum | Number of recharging slots released that date in that area |
| Slots\_taken\_sum | Number of recharging slots taken that date in that area |

**b. Stating the Null and Alternative Hypotheses**

The points below show the Null Hypothesis (Ho) and the Alternative Hypothesis (Ha).

\* Ho : The number of Blue Cars taken in area/postal code 75015 and 75015 are the same

\* Ha : The number of Blue Cars taken in area/postal code 75015 is different from Blue Cars taken in area/postal code 75017 (CLAIM)

Writing the Ho and Ha in mathematical form noting that Null Hypothesis is always identified by equality.

1. Ho : Blue Cars Taken in 75015 = Blue Cars Taken in 75017

2. Ha : Blue Cars Taken in 75015 != Blue Cars Taken in 75017(CLAIM)

1. **Importance of Hypothesis Testing**

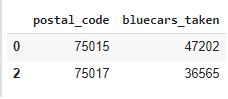
Carrying out this hypothesis test is important because it allows Autolib to backup claims that may be true or untrue. When a claim is made and the evidence from the hypothesis test is statistically significant and sufficient to back up the claim, then business decisions can be made based on a claim that was made and statistically backed.

1. **Data Description**
2. **Information about the data**

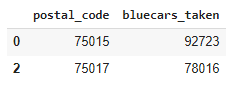
The data that is necessary to understand the report and the analysis notebook, is the random variable, Blue Cars Taken, and the specific area or postal code where the blue cars were taken. These areas are limited to postal codes 75015 and 75017. The days of interest during the week are the weekdays. This is because from the univariate and bivariate findings, weekdays tend to have more blue cars taken than weekends.

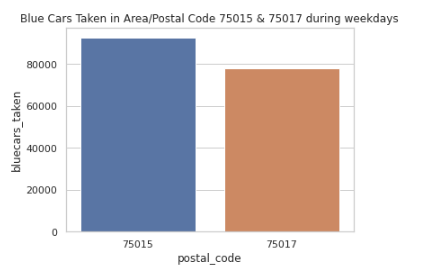
The dataset was obtained from the following link provided by Autolib Car Sharing Service Company. [Here](http://bit.ly/DSCoreAutolibDataset) . Data collection procedures involved Autolib recording the cars taken and returned from different postal codes using a meter that automatically records the activities within the different stations. For the data to be usable, a number of procedures were carried out to achieve this.

* The columns were renamed columns, fixed from syntax errors and put in lower case for standardisation
* The date column was changed from object data type to datetime
* Unnecessary columns were dropped. Additional columns were dropped along the way as the need arose.
* The date column was converted to the dataframes index so that time series analysis can be performed.
* The data was checked for anomalies
* The data was checked for outliers. The random variable had many outliers and the decision to retain them stemmed from the fact that removing them would drastically change the outcome of the analysis and the hypothesis test.

Blue Cars taken over the weekend

Blue Cars taken over the weekdays





Comparison of Blue Cars taken in 75015 and 75017 over the weekdays

One assertion about this model is that the Null Hypothesis stated above creates room for Type II Errors because from the images above, there’s an indication that the Null Hypothesis might be false because there seems to be a difference between blue cars taken in the different postal codes.

Additionally, the bluecars\_taken variable shows skewness to the right of the distribution which is above the normal skew range of +/-1. This indicates that the distribution is slightly not normal. Therefore, the hypothesis test will be conducted with precaution.

Below are some of the summary statistics for the random variable:

* Population Variance 34383.016113
* Population Median 46.0
* Population Mode 46.0
* Population Skew 2.406355
* Population Kurtosis 6.172692
* Population Mean 125.92
* Population Standard Dev. 185.42

1. **Hypothesis Testing Procedure**

To start off, a sample will be obtained from the dataset to represent the population of interest. To do this, the Stratified Random Sampling Technique was used. The records for postal codes 75015 and 75017 were isolated from the original dataset separately, then merged into one dataset that contains the two postal codes. From that dataset, stratified random sampling was conducted to arrive at the sample size of 64 with a 30% ratio to the population. The proportions of the two area codes were 50-50 in the population and the sample as well.

The procedure followed in executing the Hypothesis test was as follows:

* Stating the Null Hypothesis
* Stating the Alternative Hypothesis
* Calculating a Test Statistic and determine the Level of Significance
* Calculating the P Value
* Drawing a conclusion about the Null Hypothesis based on the P Value

The logic behind the null hypothesis is that the null hypothesis is always identified using measures of equality and the alternative hypothesis is always identified using measures of inequality. This is how the hypotheses were constructed. The hypotheses constructed are interesting because the data suggests that the alternative hypothesis which is the claim, is more likely to be true.

The test statistic chosen for this test is the z-score since we already have the mean and standard deviation of the population obtained from the summary statistics in univariate summaries, and the sample size , n=64. The distribution of the population also appears to be normal with a slight skew to the right side.

The following assumptions were met before calculating the z-score;

* The population from which the sample was obtained was somewhat normally distributed
* The population standard deviation is known
* The population and sample mean is known
* The sample size is larger than 30
* The sample is a simple random sample of the population (Though the population was first stratified then randomly sampled) . This allows each record to have an equal chance of being selected in the sample.
* The data is continuous since we indexed it by date

The alpha level chosen was 5% level of significance.

1. **Hypothesis Testing Results**

The z-score and p-value of the hypothesis test was found out to be 3.48 and 1.99 respectively. From the above results, the following interpretation and conclusion was made;

* + - * We fail to reject null hypothesis
      * There's no sufficient statistical evidence to support the claim that Bluecarstaken in 75015 != Bluecarstaken in 75017 at the 5% level of significance
      * This is a Type II error because the Null Hypothesis is False

The confidence interval constructed around blue cars taken is;

* The start is : 747.0
* The end is : 799.0

1. **Discussion of Test Sensitivity**

When manually changing the size of the sample from a fraction of 0.3 to 0.1 or 0.5 to alter the sample size, there is no significant change in the P-Value. The only changes noticed are in the value of the Z-Score which changes negligibly with a change in the size of the sample.

1. **Summary and Conclusions**

The process of hypothesis requires one to take precautions and test whether the population observed a normal distribution. It is also important for the assumptions required in the hypothesis test and test statistic to be met so that the results of the test can be statistically significant.

To conclude, the Null Hypothesis with regards to this hypothesis test is False judging from the results. This indicates that we have a Type II Error in our hypothesis test where we fail to reject the null hypothesis when in real sense it is false.

Results indicate a P-Value of 1.99 which is greater than our significance level of 0.5, hence the reason we failed to reject the null hypothesis. The test is not so sensitive since changing the sample size to a fraction of 0.1 or 0.5 of the population does not alter the P-Value.