### 

### 

### 

### 

### 

### 

### **AUTOLIB DATASET.**

### **HYPOTHESIS TESTING REPORT**

### **BY**

### **RUTH NDUTA. M**

### **1)** **Problem Statement**

### Autolib is an electric car sharing company. The company has 3 main types of electric cars: blue cars, utilib counter and the utilib 4 counter. These cars operate in Paris, France.

### The claim is that the mean number of blue cars returned in a certain area is equal to the mean of the returned blue cars from another area during weekends.

### The company has tasked us to do an investigation to clearly determine the truth in this so as to aid in planning the marketing strategies and segmentation in different areas and maximize profit.

### **Null hypothesis:**

### In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups.

### *Ho: The mean number of blue cars taken on the weekends from different postal codes is equal. Ho: U1 = U2*

### **Alternative hypothesis:**

### The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

### *H1: The mean number of blue cars taken on the weekends from different postal codes is not equal. Ho: U1 ≠ U2*

### **Level of significance:**

### Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%. This is normally denoted with alpha and generally it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.

### **Data description.**

### Usage of electric cars has seen growth with the advancement of technology. Unlike the fuel cars, electric cars are more environmentally friendly and thus sustainable. An increase in electric car usage will mean a well-kept environment due to reduced air and noise pollution and a lot of other factors.

### The data set is provided by the Autolib car sharing company. It contains a daily aggregation, by date and postal code, of the number of events on the Autolib network.

### We performed exploratory data analysis on the data set where we found that it contains 14 columns and 16065 rows which are as described below.

### 

|  |  |
| --- | --- |
| **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 |

### The data set had neither missing values nor duplicated data. Most of the data types were integers. We dropped some columns that we were not going to use for our analysis. Any syntax errors were corrected and all the column names changed into lowercase. The date column which was an object data type was converted into date-time format.

### We observed quite a number of outliers in each of the variables implying that it comprised a huge part of the data thus removing them would result in removal of a huge chunk of the dataset. Having acknowledged their presence left them for further observation.

### For the descriptive analysis we performed in univariate, bivariate and multivariate analysis on the data set. Multivariate analysis, we used linear discriminant analysis for dimensionality reduction where we got an accuracy of 72.2%.

### 

### **2)** **Hypothesis Testing Procedure**

### For the sampling method, we used both the random sampling and stratified sampling methods. The random sampling method was used to pick any two random postal codes which we were to use to compare against each other.As for the stratified sampling method, we used the postal codes generated from the random sample to group the data frame into 2 using those postal codes and separated them to create two different data frames which will be used for hypothesis testing.

### The test statistic to be used is the z-statistic. We chose this because the dataset met all conditions suitable for a z-test i.e.

### The sample size is greater than 30.

### Data points should be independent from each other. i.e, one data point isn’t related or doesn’t affect another data point.

### The data should be randomly selected from a population, where each item has an equal chance of being selected.

### Sample sizes should be equal if at all possible.

### The significance level chosen is 0.05. Research allows a 5% error This means there is a 5% risk that we will be rejecting null when it’s true.

### Having done the sampling successfully, we then looked for the mean of the data for each of the postal codes, the standard deviation and the sample size for each of the postal codes. Using this information, we applied the Z-statistic formula to get the z-score. After which, using the Python functions we got the p-value which was then compared to the significance level to either reject or accept the null hypothesis.

### **3)** **Hypothesis Testing Results**

### The null hypothesis is rejected when: P-value < Significance level otherwise we accept the null hypothesis.

### After the analysis, we found that the P-Value is less than the significance level i.e. 0.0.0134 < 0.05. In this case we rejected the null hypothesis and concluded that: On the weekends, the number of blue cars taken in two different postal codes is not equal on the weekends.

### **4)** **Discussion of Test Sensitivity.**

### The power of the test is 1-0.05=0.95 which we used as our level of significance.

### When the sample size is increased, the statistical power of the hypothesis test (1.00 minus Beta) also increases. This means with a bigger data set we get more accurate results. This on the contrary would increase the probability of committing a type 2 error (not rejecting null when it's false).

### **5)** **Summary and Conclusions**

### From our hypothesis test there is no significant difference between the blue cars availability in the 2 postal codes. On the weekends, the number of blue cars taken in two different postal codes is not equal on the weekends. The alternative analysis prevails.

### This goes to show that there is a need for further analysis in order for Autolib to be able to fully understand the customer distribution and demand especially between different postal codes on the weekends in order to maximize profits.

### 