**Report On Hypothesis Testing**

**COOLANT AGENCY**

**Data Link:**<http://bit.ly/DSCoreAutolibDataset>

The Autolib electric car-sharing service company would like to investigate a claim about the blue cars in Europe. They would like the Coolant Agency to try and understand electric car usage in an effort to give appropriate measures on insights in regards to Blue-Cars taken.

The provided data was by the Autolib car sharing services and a direct link to download the data is provided above.

Based on our research question we locked through (. loc) our data to be able to identify two independent samples out of the entire population to answer our research question.

**Problem Statement**

Autolib is a car sharing company that conducts its services in the city of Paris and does lending of cars from various counters, ie Blue car counter, utilib counter and utilib 14 counter. All these counters both record the number of cars taken and returned and documented in a dataset with respect to dates.

Our main focus will be on the blue-cars-taken and each insight recorded below will be in respect to this counter. Our research question is: Is the number of Blue cars taken in the weekend between day 5 and day 6 equal?

This will require one to perform various hypothesis to at least try and evaluate the relationship between different days of the weekend in the data set. This will aid in the identification of which day is the preferred amongst the ones present and also know by what degree is the null hypothesis being rejected or accepted.

Performance of various sampling techniques will be done so as to choose the appropriate method by which a sample will be randomly selected.

The null hypothesis is recorded as H0, while the alternate is Ha.

H0: Is day 5 equal to day 6

Ha: Is day 5 not equal to day 6

Mathematical representation:

H0: Day 5 = Day 6

Ha: Day 5 ≠Day 6

The hypothesis will try to evaluate if there’s a symmetrical difference in the number of cars recorded taken between both days of the weekend and conclude which amongst both of them is the most preferred and to what degree.

**Data Description**

The data had at least 13 columns and 16085 rows. We dropped at least several columns so as to be able to work with a less overcrowded dataset, these are; Utilib returned sum, Utilib taken sum, Utilib 14 returned sum, Utilib 14 taken sum, Slots taken sum and Slots freed sum.

A description of each column is done 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 |
| Day of the Week | Identifier of weekday (0: Monday -> 6: Sunday) |
| Day type | Weekday or weekend |
| Blue Cars taken sum | Number of blue cars taken that date in that area |
| Blue Cars 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 |

***Data Source***

The data was obtained via a secondary means; i.e. it was accessible for download via the Autolib data set. A link to the direct link download of the dataset is provided above as a data link.

***Sampling Approach***

Sampling is the process of selecting certain members or a subset of the population to make statistical inferences from them and to estimate characteristics of the whole population. Different methods of probability sampling were used to try and identify the degree of effectiveness of each sampling method.

Among the ones used are:

• Simple random sampling

• Stratified sampling

• Cluster sampling

• Systematic sampling

**Simple Random Sampling**

This is a technique used to pick the desired sample size and for selecting observations from a population in such a way that the observation has an equal chance of selection until the desired sample size is archived.

Since the whole population had been locked to the weekend in respect to day5\_sample and day6\_sample, a sample of 100 was obtained randomly from the population to describe the dataset. The sample had a census mean of 64.23 blue cars taken and a standard deviation of 58.678521 in respect to day5 and a mean of 53.840 blue cars taken and a standard deviation of 50.519467 to day6.

**Stratified Sampling**

It’s a sampling technique that involves breaking a population into key subgroups and obtaining a simple random sample from each group. The steps involved are;

• Define the population

• Choosing the relevant stratification

• Listing the population according to the stratification

• Choosing your sample size

• Calculating a proportionate stratification

Each of these steps are conducted in the colaboratory notebook. A stratified sample of 106 was obtained according to weekend day, the mean of the blue cars taken was 43.462264 and a standard deviation of 31.056070 for day 5 and the mean of blue cars taken was 43.462264 and a standard deviation of 31.056070 for day 6.

**Cluster Sampling**

This is a sampling method in which the entire population of study is divided into externally homogeneous but internally heterogeneous groups called clusters. We defined our cluster according to location, but this wasn’t possible since the data didn’t have longitudinal and latitudinal columns

**Systematic Sampling**

The technique is used where the sample members from a population are selected according to a random starting point but with a fixed periodic interval. The idea in systematic sampling is that, given the population units numbered from 1 to N, we compute for the sampling interval, given by k=N/n, where n is the number of units needed for the sample.

A conclusion of pursuing a Simple Random Sampling method was done. This is because the data set had an equal chance of selection and hence no bias and also its simple to interpret the results from the sampling method. Hence using the day5\_sample and day6\_sample data set for parameters selected randomly.

***Descriptive Statistics***

The data set had 16085 rows before being broken down to the scope of study i.e. weekend only day 5 and day 6. Once broken down to weekend the data had 2267 columns for day 5 and 2274 columns for day 6 and no missing values were observed. After checking for outliers and dropping the outliers the data had 1314 rows for day 5 and 1397 rows for day 6. But since we’re using a simple random sampling technique our data has 100 rows that are generated with an equal chance.

No duplicates were found in the data set though standardization was done in order for the row titles to have all an uppercase, with spaces and also have white spaces all around. A brief description of each element in the data set is presented below.

The table below describes data for day 5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Postal code | Daily data points | Day of week | Blue cars taken sum | Blue cars returned sum |
| Count | 100 | 100 | 100.0 | 100.00 | 100.00 |
| Mean | 93028.80 | 1440.0 | 5.0 | 64.230 | 63.340 |
| STD | 871.795 | 0.0 | 0.0 | 58.678521 | 57.628 |
| Min | 91330.00 | 1440.0 | 5.0 | 1.00 | 1.00 |
| 25% | 92285.00 | 1440.0 | 5.0 | 23.500 | 22.00 |
| 50% | 93130.00 | 1440.0 | 5.0 | 44.00 | 43.500 |
| 75% | 93480.00 | 1440.0 | 5.0 | 89.2500 | 87.750 |
| Max | 95870.00 | 1440.0 | 5.0 | 280.000 | 285.000 |

The table below describes data for day 6.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Postal code | Day of week | Blue cars taken sum | Daily data points | Blue cars returned sum |
| Count | 100.00 | 100.0 | 100.00 | 100.00 | 100.00 |
| Mean | 93136.70 | 6.0 | 53.840 | 1440.0 | 54.270 |
| STD | 1034.542845 | 0.0 | 50.519467 | 0.0 | 53.1053 |
| Min | 91330.00 | 6.0 | 0.00 | 1440.0 | 1.00 |
| 25% | 92290.00 | 6.0 | 18.7500 | 1440.0 | 18.750 |
| 50% | 93130.00 | 6.0 | 32.5000 | 1440.0 | 31.500 |
| 75% | 93800.00 | 6.0 | 76.2500 | 1440.0 | 80.000 |
| Max | 95880.00 | 6.0 | 237.00 | 1440.0 | 244.00 |

The interquartile (iqr) range is the difference between the third quartile (q3) and the first quartile (q1). The iqr is shown below.

IQR for day 5:

· POSTAL\_CODE 1195.00

· N\_DAILY\_DATA\_POINTS 0.00

· DAYOFWEEK 0.00

· BLUECARS\_TAKEN\_SUM 65.75

· BLUECARS\_RETURNED\_SUM 65.75

IQR for day 6:

· POSTAL\_CODE 1510.00

· N\_DAILY\_DATA\_POINTS 0.00

· DAYOFWEEK 0.00

· BLUECARS\_TAKEN\_SUM 57.50

· BLUECARS\_RETURNED\_SUM 61.25

Day 5 has a larger sum of blue cars taken compared to day 6, also day 5 has a large sum of blue cars returned compared to day 6.

We have a range of 279 for day 5 and a range of 237 for day 6. The range is calculated as the difference between the maximum and minimum value in the blue cars sum. This shows the spread of our data in the data set between the minimum value and the maximum value.

The variance for day 5 is:

· POSTAL\_CODE 760026.828283

· N\_DAILY\_DATA\_POINTS 0.000000

· DAYOFWEEK 0.000000

· BLUECARS\_TAKEN\_SUM 3443.168788

· BLUECARS\_RETURNED\_SUM 3321.034747

The variance for day 6 is:

· POSTAL\_CODE 1.070279e+06

· N\_DAILY\_DATA\_POINTS 0.000000e+00

· DAYOFWEEK 0.000000e+00

· BLUECARS\_TAKEN\_SUM 2.552217e+03

· BLUECARS\_RETURNED\_SUM 2.820179e+03

Day 5 has a larger variance compared to day 6.

Skewness is the asymmetry in a symmetrical bell curve in a data set. The skewness is shown below.

Skewness for day 5:

· POSTAL\_CODE 0.473827

· N\_DAILY\_DATA\_POINTS 0.000000

· DAYOFWEEK 0.000000

· BLUECARS\_TAKEN\_SUM 1.511081

· BLUECARS\_RETURNED\_SUM 1.580145

Skewness for day 6:

· POSTAL\_CODE 0.836580

· N\_DAILY\_DATA\_POINTS 0.000000

· DAYOFWEEK 0.000000

· BLUECARS\_TAKEN\_SUM 1.505175

· BLUECARS\_RETURNED\_SUM 1.477842

Day 5 records a larger skewness compared to day 6.

The kurtosis is a statistical measure used to describe the distribution of observed data around the mean.

Kurtosis for day 5:

· POSTAL\_CODE -0.206720

· N\_DAILY\_DATA\_POINTS 0.000000

· DAYOFWEEK 0.000000

· BLUECARS\_TAKEN\_SUM 2.241840

· BLUECARS\_RETURNED\_SUM 2.631713

Kurtosis for day 6:

· POSTAL\_CODE 0.538392

· N\_DAILY\_DATA\_POINTS 0.000000

· DAYOFWEEK 0.000000

· BLUECARS\_TAKEN\_SUM 2.019683

· BLUECARS\_RETURNED\_SUM 1.759207

For day 5, postal code is platykurtic since it has a Kurt lass than o (Kurt < 0), while daily data points and day of the week are mesokurtic since its equal to 0 (Kurt = 0) and blue cars taken and blue cars returned are leptokurtic since they’re greater than 0 (Kurt > 0).

For day 6 postal code is platykurtic since it has a Kurt lass than o (Kurt < 0), while daily data points and day of the week are mesokurtic since its equal to 0 (Kurt = 0) and blue cars taken and blue cars returned are leptokurtic since they’re greater than 0 (Kurt > 0).

**Hypothesis Testing Computation**

According to our research question; Is the number of Blue cars taken on the weekend between day 5 and day 6 equal?

The null hypothesis is recorded as H0, while the alternate is Ha.

Mathematical representation:

H0: Day 5 = Day 6

Ha: Day 5 ≠ Day 6

Let’s set our alpha (α) to be 0.05 at a 5% significance level. But since it’s a two sample hypothesis the α is divided by 2 (α / 2).

We can calculate both the population standard deviation and the sample standard deviation to be used to test for a two sample z test and be in favor due to the sample size of 100. Though this wasn’t the case. Independent T calculated for each sample between day 5 and day 6.

Day5…….t= (64.23000-125.926951)/ (58.678521/sqrt (100)) = -10.51440117

Day6…….t= (53.84000-125.926951)/ (50.519467/sqrt (100)) = -14.26914491

The T tabulated from the T tables is t 0.05/2, df = 1.984

The entire data wasn’t normally distributed hence the normal t test calculations were done but the generated p value was biased, hence used the Mann Whitney U test to find the p value of the test static.

A calculation of the two sample independent t tests using python generated the following result; Ttest\_indResult (statistic=1.3418593487011778, p-value=0.18117840224189533). Since the p- value is greater than alpha we fail to reject the null hypothesis. (p > α/2)

But using the Mann Whitney U test a different test result was obtained. The test result is; stat=0.991, p=0.093 and the data had the same distribution. This is because the alpha used was less to the p value computed.

We settled for the Mann Whitney U test since the hypothesis p-value wasn’t biased and at least had the same distribution. Hence we fail to reject the null hypothesis.

**Hypothesis Testing Interpretation**

When your p-value is greater than your significance level, you fail to reject the null hypothesis. Your results are not significant. The level of statistical significance is often expressed as the so-called p-value. Depending on the statistical test you have chosen, you will calculate a probability (i.e., the p-value) of observing your sample results (or more extreme) given that the null hypothesis is true.

There wasn’t enough evidence to support the claim that the number of Blue cars taken in the weekend between day 5 and day 6 equal, hence we failed to reject the null hypothesis. Calculating a separate p value and a statistic generated the following result, Statistics=0.268, p=0.789 hence the conclusion being Same distribution (fail to reject H0). A sample of 100 was still maintained while calculating the p-value with the process of comparing their distributions between day 5 sample and day 6 sample.