Name: Kiprop Amos

Course: DSC14

Date: 01/04/2022

Technical Mentors: Nikita Njoroge, Veronica Isiaho

**Data Report for Autolib Network (Car-Sharing and Recharging).**

1. **Problem statement**
   1. **Overview**

Car rental companies experience seasonal changes in demand for their services by customers. Some periods during the year or week, this service is in such High demand while in other periods the demand diminishes.

The data that will be used is an autolib dataset that contains data on the days in which rental cars were taken or returned. The variables of interest are the number of cars taken or returned at the two different periods of the week

* 1. **Objective**

The objective of this study is to determine whether the demand for car rental services is affected by the part of the week that day is.

* 1. **Questions**
* Is there a difference between the cars taken during a weekend and during those taken during a weekday?
* Is there a difference between the cars returned during the weekends and those returned during a weekday?
  1. **Hypothesis**

**Question 1**

**Ho**: There is no difference between the cars taken during the weekday and those taken during the weekend

**H1:** There is a difference between the cars taken during the weekday and those taken during the weekend

**Question 2**

**H0**: There is no significant difference between the cars returned during the weekday and those taken during the weekend

**H1**: There is a difference between the cars returned during the weekday and those taken during the weekend

* 1. **Justification of Study**

This analysis is very important for the car rental company in order to place them in a position to foresee future changes in demand and put in place systems to maximize the profits brought by a surge in demand of car rental services and minimize operational costs during off periods

1. **Business understanding**

The Autolib Network (Car-Sharing and Recharging) company offers electric car rental service for electric cars in cities all over France. It has three types of cars in its inventory; Blue Cars, Utilib Car, and Utilib-14 Car. They have different stations distributed through the various postal codes in the country.

* 1. **Assessing the Situation**

**Resources**

Autolib Dataset link: [http://bit.ly/DSCoreAutolibDataset](about:blank)

Variable definition link: <http://bit.ly/DSCoreAutolibDatasetGlossary>

Software (Github, Google Collaboratory Notebook, Pandas library, Numpy library, Datetime library, Scipy.stats library, Matplotlib library, Seaborn Library, Microsoft Word)

* 1. **Assumptions**

The data follows a normal distribution

The data collection methods were appropriate and lacked any bias

The data is accurate

* 1. **Constraints**

The were no constraints encountered in the data

* 1. **Success criteria**

A conclusion is arrived at regarding whether the period of the week be it weekday or weekend affects the number of cars returned or taken from their stations

1. **Data understanding**
   1. **Overview**

For this project we are using the data sets availed by the company. The dataset includes:

- The variable definition dataset contains descriptions of the various variables contained in the variable definition dataset

-  Autolib data set; contains data on dates rental cars were taken and returned, the postal codes this car were taken or returned, the day of the week from Monday – Sunday, the part of the week the car was either taken or returned, the number of available slots at different parking stations, the counters noting the movement of the Blue, Utilib, and Utilib-14 cars in and out of the parking station.

* 1. **Data description**
* Postal code- This variable contains information on the postal code of the area (in Paris). Date type: Integer
* Date- Date variable provides information on when a car is returned or taken.
  + The data type is string in the format (dd/mm/yyy)
* Daily Data Points - This variable details the number of daily data points that were available for aggregation, that day.
  + Datatype- Integer
* Day of Week- This variable is the identifier of weekday (0: Monday -> 6: Sunday)
  + Datatype- integer, elements (0, 1, 2, 3, 4, 5, 6)
* Day Type- This variable specifies the part of the week as either weekday or weekend
  + Datatype- string
* Blue Cars Taken Sum- This variable gives the total number of Blue Cars taken that date in that area
  + Datatype - Integer
* Blue Cars Returned Sum - This variable gives the total number of Blue Cars Returned that date in that area
  + Datatype - Integer
* Utilib Taken Sum - This variable gives the total number of Utilib Cars taken that date in that area
  + Datatype - Integer
* Utilib Returned Sum - This variable gives the total number of Utilib Cars returned that date in that area
  + Datatype - Integer
* Utilib-14 Taken Sum - This variable gives the total number of Utilib-14 Cars taken that date in that area
  + Datatype - Integer
* Utilib-14 Taken Sum - This variable gives the total number of Utilib-14 Cars returned that date in that area
  + Datatype - Integer
* Slots Freed Sum- This variable gives the number of recharging slots released that date in that area
  + Datatype-Integer
* Slots taken Sum- This variable gives the number of recharging slots occupied that date in that area
  + Datatype-Integer

1. **Data preparation**
   1. **Loading datasets**

The data was loaded into the Google Collaboratory, visualized and previewed

* 1. **Data Cleaning**
     1. **Checking for outliers**

This step entailed removal of outliers from the different steps in the outliers.

The number of columns in the data frame after removing outliers;

* Columns 14908
* A total of 1177 records as outliers
* The percentage of records lost as outliers was 7.317%
  + 1. **Checking for null Values**

This step entailed checking for any missing data in the data frame. No null values were found

* + 1. **Checking for duplicate values**

This step entailed checking for similar values or columns within the data frame. No duplicate values were found in the data frame

* + 1. **Validity and other anomalies**

The anomalies looked for were unusual data types in the different columns in the data frame, poorly named columns.

* The only column with an anomaly was the date column which was changed from string to datetime kind of data
* Most of the columns had been named poorly hence the final step in cleaning the data entailed renaming those columns appropriately
* The cleaned data frame had 14908 columns and 13 rows. A total of 1177 records were lost which represented 7.317% of the total data which is insignificant.

1. **Analysis**
   1. **Univariate Analysis**
      1. **Numerical analysis**

The means, standard deviations, median values, skewness and kurtosis of the various variables in the data frame was determined

| Variable | Mean | Std | Median | Skewness | Kurtosis |
| --- | --- | --- | --- | --- | --- |
| Daily Data Points | 1435.97 | 12.80 | 1440.00 | - 4.01 | 15.90 |
| Bluecars Taken Sum | 95.75 | 124.89 | 43.00 | 1.98 | 3.27 |
| Bluecars Returned Sum | 95.72 | 124.84 | 43.00 | 1.99 | 3.32 |
| Utilib Taken Sum | 2.78 | 3.93 | 1.00 | 2.04 | 4.10 |
| Utilib Returned Sum | 2.78 | 3.93 | 1.00 | 2.04 | 4.04 |
| Utilib-14 Taken Sum | 6.55 | 8.69 | 3.00 | 2.02 | 3.89 |
| Utilib-14 Returned Sum | 6.55 | 8.69 | 3.00 | 2.02 | 3.89 |
| Slots Freed Sum | 14.03 | 34.81 | 0.00 | 2.67 | 6.14 |
| Slots Taken Sum | 14.04 | 34.87 | 0.00 | 2.68 | 6.14 |

The key takeaways from this analysis were;

* All variables aside from the day of the week variable had a positive skew. This implied that the variables had more tails than the normal distribution
* The day of the week had a negative kurtosis. This implied it had less tails than a normal distribution.
* The blue cars had the highest variation in cars taken or returned
  + 1. **Visualizations**

Two visualizations were plotted in the univariate analysis. One was to compare the cars taken or returned during the weekdays and the weekend. The other visualization was a bar chart comparing the cars taken or returned at different days of the week. The following features were observed;

* Close to three times more cars were taken/returned during the week compared to the weekend
* The most cars were taken/returned during on a weekday on the first day of the week (Day 0, Monday)
* The least cars were taken/returned on the sixth day of the week (Day 5, Saturday) a weekend
  1. **Bivariate Analysis**

A bivariate summary of all numerical variables in the data frame was plotted to compare all these variables. To further get a clearer picture of the correlation between the variables, a correlation plot was plotted. The following was observed;

* The following columns showed a high correlation of 0.6 and above amongst each other; Blue Cars Taken Sum, Blue Cars Returned Sum, Utilib Taken Sum, Utilib Returned, Sum, Utilib-14 Taken Sum, Utilib-14 Returned Sum, Slots Freed Sum, and Slots Taken Sum.

1. **Hypothesis testing**
   1. **Hypothesis Testing Procedure**
      1. **Procedure**

* Created two data frames from the existing data frame one containing the information on vehicle returns and pickups
* Summed up the information on car returns for blue, utilib, and utilib-14 cars in one newly created column
* The same was repeated for the column with dataset containing information on car pickups
* Stratified random sampling was then applied to the two datasets to ensure the obtained samples had proportions of weekdays and weekends equal to the proportions in the original data frame
* A small sample size of 50 was chosen
* The two datasets were then subjected to two separate t-tests since the samples were small (50 elements) after grouping the sum totals of returns and pickups in the two datasets by type of day
* The obtained result was a t-statistic and a p-value.
* The p value was used to either reject or accept the null hypothesis
  + 1. **Justification of hypothesis**

The null hypothesis seeks to show that there isn’t enough evidence to show that car returns and pickups is affected by the type of day of the week

The alternate hypothesis in this study seeks to show that there is enough evidence that the car returns and pickups are affected by time of day of the week

* + 1. **T-test justification**

The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale.

The data also appears be normally distributed

Also, a small sample size of 50 elements has chosen from the dataset hence it will have a high accuracy

* + 1. **Confidence level**

The confidence level chosen for this study is 5%. It is stringent enough but also not too stringent for this study

α = 0.05

* 1. **Hypothesis Testing Results**

**Question 1**

On the first question regarding whether there is difference between the cars taken during a weekend and during those taken during a weekday the following results were obtained;

**Result;**

- T-statistic = 0.91503

- P-value = 0.36167

- Fail to reject the null hypothesis.

- The difference in cars taken over the weekend and during weekdays was not statistically significant

- There is no significant difference between the cars taken during a weekday compared to those take during a weekend (p = 0.36167)

**Point Estimator**

- The mean of the dataset with data on car pickups before sampling was *105.071*

- Based on a sample of 149 pickups our estimator underestimates the true mean by *11.427*

- We can conclude that we can get a fairly accurate estimate of a large population from a fairly small subset.

**Confidence interval**

* lower limit 102.88488483425692
* upper limit 107.25691822450347

**Question 2**

On the second question regarding whether there was a difference between the cars returned during the weekends and those returned during a weekday, the following result was obtained;

**Result**

- T-statistic = -0.18818

- P-value = 0.85100

- Fail to reject the null hypothesis.

- The difference in cars returned over the weekend and during weekdays was not statistically significant

- There is no significant difference between the cars returned during a weekday compared to those returned during a weekend (p = 0.85100)

**Point Estimator**

- The mean of the dataset with data on car returns before sampling was *105.046*

- Based on a sample of 149 returns our estimator overestimates the true mean by

*- 4.343*

- We can conclude that we can get a fairly accurate estimate of a large population from a fairly small subset.

**Confidence Interval**

* lower limit 102.8612817657061
* upper limit 107.23168845162688

1. **Summary**

In summary, the project entailed testing to see whether specific periods in a week affect car returns and pickups. The problem statement and hypothesis were first defined. The next step entailed detailing information on the data to be handled and understanding the overall nature of the business

After these stages, the dataset provided was cleaned to deal appropriately with missing values, null values, duplicates, anomalies and outliers that might be found in the data. Analysis was then performed on the cleaned data set

The analysis entailed bivariate analysis which entailed various visualizations and numerical analysis on the data. The other was bivariate which entailed comparing two variables in the dataset.

The final stage involved hypothesis testing. Here the two questions raised in the problem statement were answered by calculating T-statistic and p-values for each question. This obtained data was used to make inferences on the question. Point estimation and confidence intervals were also determined to confirm the accuracy of our sampling and data frames

A report was then written summarizing all the steps taken in the google Collab and the conclusions arrived at after analysis.

1. **Conclusion**

In conclusion, at a confidence level of 0.05, there is not enough evidence to show that car pickup and return to and from the stations is affected by the type of day of the week.