**WEEK 4 CORE IP**

**PROBLEM STATEMENT:**

The dataset we are using and the glossary for this dataset is:

<http://bit.ly/DSCoreAutolibDataset>

<http://bit.ly/DSCoreAutolibDatasetGlossary>

In this dataset we have the postal codes of different areas in Paris. We also have information on the time and the number of blue cars, utilib and utilib 14’s taken and returned on those days and the recharging slots released in that area.

With this data, we shall be investigating information on the number of blue cars returned and received.

Following this our null hypothesis and alternate hypothesis will be as follows:

Null hypothesis- the number of blue cars taken is more than those returned on weekdays

Alternate hypothesis- the number of blue cars returned is more than those taken on weekdays.

The reason for this is because we want to investigate if on weekdays, people prefer to take blue cars than return them on the days people are mostly known to go for work/ school.

**DATA DESCRIPTION**

The dataset contained 16085 rows and 13 columns.

The columns included the postal codes for different areas in Paris, the Dates which ranged from January to July of 2018. It also contained the number of data points available for aggregation on those days, the days of the week ranging from Monday to Sunday and stated if the days were weekends or not. Moreover, we are also given the number of blue cars, utilib and utilib 14’s taken and returned and taken and the slots available for recharging.

Having this in mind, the data collection technique I would use to collect such data is observation. I would send a team of researchers to different addresses in Paris and collect the data. With this method of data collection, I would use instruments to be able to observe how many cars are parked at a particular time and date and record this data.

Using this data, I used measures of frequency, measures of dispersion and measures of central tendency to analyze my data.

Measures of frequency - I used the count function to display the number values that we have been provided in various fields. Moreover, the count function helped us make computations and visualizations on some of the analysis like constructing a bar chart

Measures of central tendency - I used the mean function to help me calculate the z-score for purposes of hypothesis testing

Measures of dispersion - I used the standard deviation to help me calculate the z-score for purposes of hypothesis testing

I made sure that the data was consistent and accurate by cleaning the data and dealing with anomalies. I dropped the null and duplicate values, making sure that the column names are consistently labeled, dropping columns that we do not need to work with and checking and dropping outliers in our dataset.

**HYPOTHESIS TESTING PROCEDURE**

I started by separating my data into weekdays since this is a target variable.

Thereafter, I sampled my population using stratified random sampling because I wanted a representation of all the listed postal codes and divided the population into strata in regard to the postal code and got a sample size of 1103.I calculated the population mean followed by the sample mean and the standard deviation.

Having this data, I calculated the z-score and derived the p-value from the z-score.

The reason for using the z-score is because we had a large sample size which is greater than 30. The alpha level I used is 0.05.

**HYPOTHESIS TESTING RESULTS**

I concluded that we should fail to reject the null hypothesis.

I calculated the z-score to be -0.067

The calculated p-value is 0.473 which is greater than alpha.

The population mean is 129.692 and the sample mean is 118.315 with a standard deviation of 117.82

There is a 95% chance that the confidence interval contains the null hypothesis

**SUMMARY AND CONCLUSIONS**

In this project, I:

Imported the necessary libraries for analysis

Read the data in the dataset

Carried out data cleaning

Carried out univariate analysis

Carried out bivariate analysis

Carried out hypothesis testing

Wrote the project report

Suggesting necessary conclusions and recommendations

The null hypothesis was the number of blue cars taken is more than those returned on weekdays while the alternate hypothesis was the number of blue cars returned is more than those taken on weekdays.

We failed to reject the null hypothesis meaning that there is a 95% chance that the blue cars taken are more than those returned.