**HYPOTHESIS TESTING REPORT (AUTOLIB PROJECT)**

**Problem Statement**

The goal of this project is to demonstrate the concept of hypothesis testing by investigating a claim about blue cars using a dataset from an electric car sharing service company called Autolib. More specifically, I will be comparing the blue car usage rates in Paris and Hauts-de-Seine, which are two French cities.

**Research Context**

Autolib was an electric car sharing service company in France that was operational between 2011 and 2018. The company had three types of electric cars i.e. blue cars, Utilib cars and Utilib 1.4 cars. Blue cars were most popularly used. These cars were available across various cities and postal codes in France and renters could pick up cars in one station then drop them off at a different station that was closer to their destination. The dataset used in this analysis allows us to understand various electric car usage patterns for the company. The available data contains usage information for various postal codes between January 2018 and June 2018.

For this project, I decided to focus on two cities Paris (department # 75) and Hauts-de-Seine (department # 92). Compared to other cities available in the dataset, these two cities are the top two most densely populated. Even though Hauts-de-Seine is larger than Paris in terms of total area in square kilometers, Paris is much more densely populated and more popular with tourists. As such, I suspect that the average blue car usage rates for those two cities are different. Usage rates are generally higher during the weekends therefore the analysis will focus on this period. Below are the specific hypotheses of interest in this analysis:

***Null Hypothesis -*** For both Paris and Hauts-de-Seine the average number of blue cars picked up during weekends is the same.

***Ho : μ1 = μ2 (where μ1 is the mean for Paris and μ2 is the mean for Hauts-de-Seine)***

***Alternate Hypothesis*** ***-*** The average number of blue cars picked up during the weekend is different for Paris and Hauts-de-Seine.

***Ha : μ1 ≠ μ2***

**Data Description**

The dataset used in the analysis contains records of electric car usage in France from January 2018 to June 2018. This dataset was originally sourced from opendataparis.com. Some of the information in the dataset includes postal code, day of the week and total cars returned or picked up for blue cars, Utilib cars and Utilib 1.4 cars. Given the above null and alternate hypotheses, the main variable of interest is the number of blue cars picked up. This specific attribute is available in the dataset for various postal codes during each day of the week beginning on 1/1/2018 to 6/19/2018.

The original dataset contained 13 different attributes. The full dataset as well as dataset description can be found [here](http://bit.ly/DSCoreAutolibDataset) and [here](http://bit.ly/DSCoreAutolibDatasetGlossary) respectively. Below is a summary of the attributes used in this analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Column Description** | **Data Type** | **No. of Records** |
| Postal code | This attribute identifies the postal code where the observation was recorded | Integer | 16085 |
| Date | Date when the observation was recorded | Date | 16085 |
| Day of Week | The day of the week when the observation was recorded | String | 16085 |
| Day Type | Whether the day was a weekday or a weekend | String | 16085 |
| Blue Cars Taken | Total number of blue cars picked up on a particular day for a particular postal code | Integer | 16085 |
| Blue Cars Returned | Total number of blue cars returned on a particular day for a particular postal code | Integer | 16085 |

**Summary of Data Cleaning and Sampling Technique**

Some of the preliminary steps performed before the hypothesis testing procedure include data cleaning, univariate and bivariate data analysis as well as discussion of the sampling approach. Specific details and results of these steps can be found on this [Python Notebook](https://colab.research.google.com/drive/1aCcah8OjZclG3h9VD0wPIPMcpVQiXhEM#scrollTo=PyBYVWbHJ0y8).

In summary, I created a new attribute called ‘city’ in the dataset. A quick google search shows that postal codes in France are organized by departments (which can also be described as cities). For instance, all postal codes for Paris begin with the number 75. Therefore, I grouped the various postal codes into their representative cities to further compress the dataset.

The hypothesis is tested on a sample of the population. Since I want to compare samples from two different cities i.e usage in Paris versus Hauts-de-Seine, stratified random sampling is the best sampling technique to use. In this case, the ‘city’ becomes the only strata useful in the sampling technique. Pick up and drop off stations are unique to postal codes/city. A pickup station can only exist in one postal code or city. Therefore, a car pickup or drop off recorded in a station belongs to a unique city. Using only one stratum ensures that each record of usage has an equal chance of being selected during sampling. Each stratum has no overlapping sample therefore no bias in the sampling technique. The sample size for each stratum is proportionate to the target population size of that particular stratum.

**Hypothesis Testing Procedure**

To conduct hypothesis testing, I used a two-sample z-test and calculated the p-value in order to either reject or accept the null hypothesis. Below are the reasons why I chose to use the z-test as the appropriate test statistic:

* The sample size is greater than 30.
* Data points are independent from each other.
* The sample data has been randomly selected from a population, so each item has an equal chance of being selected.

The level of significance used in the hypothesis testing is 0.05 or 5%. Therefore, if the p-value calculated from the test statistic is less than 0.05 then we reject the null hypothesis. If the value is greater than or equal to 0.05, we accept the null hypothesis. Since α = 0.05, and the test statistic is a one tailed test, the critical value of z is 1.645 according to the z critical value table.

**Hypothesis Testing Results**

I used *SciPy* and *statsmodels.stats* libraries in python to conduct the z test and calculate both the z statistic and the p-value. The calculated p-value of the two-sample z test was 7.161676119430548e-70, which is much less than 0.05. Therefore, we reject the null hypothesis. This means that the average number of blue cars picked up on weekends is not the same for Paris and that for Hauts-de-Seine. In this case, we accept the alternate hypothesis.

The calculated z statistic is 17.7 against the expected critical value of 1.645. This means that the data points are 17.7 standard deviations away from the mean. This is further evidence that we need to reject the null hypothesis.

**Discussion of Test Sensitivity**

To check the power of the hypothesis test above, I used the *statsmodels* library in python. I calculated the power of the two-sample z test based on the parameters used during hypothesis testing. The statistical power found was 64%. This means that the probability of rejecting the null hypothesis given that it is false is 64%. Ideally, this value should be closer to 90% for a more significant statistical test. Some of the factors that could have affected the statistical power of the test include:

* Sample sizes for the two strata were not the same.
* The population did not follow a normal distribution.

**Summary and Conclusions**

We have successfully defined the null and alternate hypothesis, executed the sampling technique and carried out hypothesis testing which led to the rejection of the null hypothesis. We concluded that the average number of blue cars picked up in Hauts-de-Seine during weekends is not the same as that in Paris.

One factor that needs further investigation is the low power of the test. The high variance in the target population could have impacted the test statistic and contributed to a high likelihood of a Type I error.