**Autolib Dataset Hypothesis Testing Project Report**

1. **Problem Statement**

The Autolib car-sharing program was introduced in Paris, France on December 5, 2011 and it has enjoyed substantial popularity to date. The company serves people in multiple areas across France and is keen to expand its services in areas with the highest demand for rental cars. According to company officials, the BlueCars have been increasingly popular across different parts of France so they will be the focus of this expansion project. As the data scientist for Autolib car-sharing program, I have been tasked with identifying the top 2 postal codes with the highest number of BlueCars rented on average, and also, to inform whether the difference, if any, is statistically significant.

**Hypothesis**

The null hypothesis (H0) is that there is no difference between the average number of BlueCars rented in the top two postal areas.

The alternate hypothesis (H1) is that there is a difference between the average number of BlueCars rented in the top two postal areas.

H0: Mean postal code 1 == Mean postal code 2

H1: Mean postal code 1 =/ Mean postal code 2

**Importance of Hypothesis Test**

The Autolib car-sharing program has a fixed amount of money it can use for expansion projects. Ideally, the money is equally shared between the two postal areas if their means are similar. If not, a larger share of investment is directed toward the postal area with the higher average hence the need for a hypothesis test to establish whether the means are significantly different to warrant a biased split.

1. **Data Description**

The data set contains time-series data which was collected between 1st January 2018, and 19th June 2018, a total of 156 days. The data set has 104 unique postal codes which will be used to group the data, and identify the postal areas with the highest average of BlueCars taken which is the main variable in this project. An overview of the other variables can be seen 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 bluecars taken that date in that area |
| BlueCars\_returned\_sum | Number of bluecars 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 Autolib program has a designated database in which they keep daily records of the number of cars that are rented out for a particular type along with the subsequent number of slots that are taken or freed as a result. Due to the large size of the data, Autolib was kind enough to share six months worth of data from January to June 2018 which was to be used for analysis. This data was collected on all days in the aforementioned period, and simultaneously in all the postal areas amounting to a total of 16085 records. The type of day is also an important variable since Autolib wanted to compare the postal areas on weekdays and weekends.

From the data set that was provided, I identified postal codes 75015, and 75016 as the two areas with the highest portion of BlueCars that were taken during the survey period. After splitting the data set into weekdays and weekends, there were some outliers in the sum of BlueCars taken on both day types. To maintain the consistency of the data sets, I decided to drop all rows whose dates had outliers in the BlueCars taken column on both the weekday, and weekend data. Since that data that was provided was based on a particular time, my biggest concern is that it could be defined by certain market trends that are not easily observable outside the full year cycle.

1. **Hypothesis Testing Procedure**

Step 1

I grouped the data set according to the unique postal codes and sorted the result to identify the postal codes with the highest averages for BlueCars taken which were 75015 and 75016. For each of the two postal codes, I created a new data set which I will use to compare the means. A key component of this project was the need to distinguish between weekends and weekdays which I did for both postal codes so I had four unique data sets that compared weekday, and weekend data in postal code 75015 and 75016.

Step 2

To perform the hypothesis test, I decided to use a sample of the six-month data rather than the entire data set which I was concerned could be affected by seasonality trends I used stratified random sampling to collect data points to be used which, theoretically, would help overcome the influence of seasonality.

Step 3

I computed the test statistic for the mean difference over the weekdays along with its p-value. I also computed the test statistic for the mean difference over the weekends along with its p-value.

The hypothesis test is meant to establish whether the average rental rates of BlueCars is equal in postal areas 75015 and 75016. This test functions to inform on the investment funds that are to be directed at expanding services in the identified areas on weekdays and weekends respectively depending on the test results.

The main reason for selecting a T-test of independent means was because the population standard deviation for the sum of BlueCars taken was unknown considering that only six months data was provided. Additionally, I also assumed that the populations are not related to each other considering they are from different areas. The suitability of the test was also highlighted by other factors including: no outliers in the data and normal distribution of data. With regards to mean variances, I established that the weekday data had different variance while the weekend had equal variances. The status of variances was also noted while carrying out the test as I ensured appropriate formulas were used to reflect this.

The test was to be determined at a significance level of 0.05.

1. **Hypothesis Testing Results**

**Weekdays**

With an alpha of 0.005, we fail to reject the null hypothesis that postal areas 75015 and 75016 have equal, average number of BlueCars taken over the weekdays. The p-value for the weekday data is 0.1041 which is higher than the level of significance hence the failure to reject the null hypothesis.

From a sample size of 0.4 of the entire dataset,we can be 95% certain that the difference between the population means of postal code 75015 and 75016 is between 28.62 and 111.79.

From a sample size of 0.2 of the entire dataset,we can be 95% certain that the difference between the population means of postal code 75015 and 75016 is between 35.27 and 145.68.

**Weekends**

With an alpha of 0.005, we reject the null hypothesis that postal areas 75015 and 75016 have equal, average number of BlueCars taken over the weekends and accept the alternate hypothesis that the averages are significantly different. This decision is supported by the p-value for the weekend sample data which is 0.013 which is lower than the level of significance.

From a sample size of 0.4 of the entire dataset,we can be 95% certain that the difference between the population means of postal code 75015 and 75016 is between 101.52 and 259.22.

From a sample size of 0.2 of the entire dataset,we can be 95% certain that the difference between the population means of postal code 75015 and 75016 is between 39.32 and 305.93.

1. **Discussion of Test Sensitivity**

While computing the confidence interval, I varied the size of the sample between 10, 20, and 30% of the entire datasets. Varying the sample size had an effect on the power of the test since there was an increased likelihood of the rejecting the null hypothesis that the two means, whether from weekend or weekday data, are equal. I did not carry out a goodness of fit test since my sample had an equal number of days sampled.

1. **Summary and Conclusions**

Based on the findings, we can conclude that postal areas 75015 and 75016 have the top two highest average rental rates for BlueCars. Comparing the two postal areas, 75015 had a higher average of BlueCars taken than 75016 on both weekdays and weekends. The hypothesis test indicated that, although area 75015 had a higher average of BlueCars taken than 75016 on weekdays, the difference between their means was not statistically significant. On the other hand, the difference between the means of the two postal areas over the weekend was significantly different.

With regards to Autolib’s investment strategy, the investment funds reserved for expanding services on weekdays will be shared between postal areas 75015 equally since their means lacked a significant difference. On the other hand, postal code 75015 will receive a greater share of the investment funds reserved for expanding weekend services since its mean is significantly higher than that of 75016.