Hypothesis Testing Report

Our Data Science team has been approached by Autolib' (an electric car-sharing service company) to investigate a claim about the blue cars from the provided Autolib dataset.

1. Business Understanding

Business Overview

Autolib is an electric car sharing service which started in Paris in 2011. Although it started in Paris, the company has expanded its operations to other cities like Bordeaux and Lyon.
A Car Sharing Service (CSS) is a scheme that provides collectively available cars that can be rented on a pay as you drive basis. These companies are formed on the principle that one needs not to own a car in order to have access to it.
The company maintains a fleet of electric blue cars which are available to the public on a paid subscription basis. The company also maintains a citywide network of charging and parking stations. As of 2016, the company had: 3,980 bluecars; 126,900 registered subscribers; 1,084 stations; and 5,935 charging points.
Autolib blue cars are available to all persons who pay for the subscription and are aged 18 years and above. Additionally, they must hold a valid French driving license or a valid foreign licence. A blue car may be picked up at any rental station and may be returned to any other station. All cars have GPS and so can be tracked by the operations center..

Business Objective

After lengthy discussions with the Autolib Company team we agreed that the main objective was to investigate the electric (bluecars) car usage in Paris by performing Hypothesis Test to see if there is a difference in the means of blue cars taken from two different postal codes selected randomly on weekdays.

Business success criteria

This analysis will be considered a success if we are able to perform exploratory data analysis on the dataset provided and formulate viable conclusions from it and if we are able to successfully test our hypothesis and interpret the results correctly.

Assessing the situation

Requirements

1. Cooperation from the Autolib company.
2. Good Internet connection for our data analysis.

Assumptions

1. The data that the company provided is accurate to the best of our knowledge

Resource inventory:datasets and softwares used

1. Datasets
   We are provided with 2 datasets:
   - Autolib Daily Events Dataset [http://bit.ly/DSCoreAutolibDataset]
   - Column Explanation [http://bit.ly/DSCoreAutolibDatasetGlossary]
2. Autolib Technical support in case of any problems with the data.
3. Google Colab for our Analysis of the data.
4. Github to host the project.
5. We will use Google and other Websites to acquire any information we will need for our analysis.

Risks and Contingencies

Risk	Contingency
Loss of Power	We will have a backup power source to mitigate this risk.
Weak or no Internet Connection	We will have several prepaid service providers sim cards. Therefore if one has weak internet connection, we will immediately switch to another data service provider.
Possible Crashing of Colab Notebook	If this happens, we will use Jupyter Notebook on our local machines.

Implementation Plan

To perform this analysis successfully, we followed the following implementation plan:

Phase	Time-Frame
Formulation of Research Question	10 minutes
Business Understanding	30 minutes
Data Understanding	1 hour
Data Preparation	1 hour
Data Analysis	4 hours
Summary and Conclusions	30 minutes

2. Data Understanding

- **Data Mining goals**
  The data mining goal for this project was to investigate the electric (bluecars) car usage in Paris and perform hypothesis tests on a parameter of the dataset.

- **Data Mining Success Criteria**
  Our data mining process will be considered to be a success if we are able to use the dataset provided to successfully investigate the electric (bluecars) car usage in Paris and successfully perform hypothesis tests on a parameter of the dataset.

- **Data Description**

We are provided with two datasets. One of the dataset is a description of the columns in the other dataset.

1. Autolib Daily Events Dataset- this dataset shows the postal code of a station, the date of data entry. Additionally, it shows the day of the week the data was captured and whether it was a weekday or a weekend. The dataset also shows us the number of Bluecars, Utilib and Utilib14 taken or returned on a given day. The dataset also displays the availability of parking slots in a given station.
2. Column Explanation - This dataset gives us a detailed explanation of information in the Autolib Daily Events Dataset. The columns descriptions are:

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 blue cars taken that date in that area
BlueCars_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

3. Data Preparation

Reading the data

We first imported all the libraries we will need for our analysis and hypothesis testing.
Then we loaded both our datasets into our programming environment and created a data frame (df) to hold the data so that we can analyze it as a dataframe.
After loading the data, we previewed the first 5 and last 5 rows of the data to get a glimpse of the type of information we will be analyzing.

Checking the Data

We determined that the dataset has 16085 rows, and 13 columns.
We checked all the datatypes of the 13 columns and accessed some general information and summary statistics(like mean and standard deviation)of the data.
Additionally, we Checked the entire profile of the dataframe.

Data Cleaning

1. Checking for duplicate values- we decided to check if there were duplicates in our dataframe so that we can drop them. In our case, we found no duplicate values.
2. Dropping Irrelevant columns-from our business objective, we can see that we have been tasked with analysing blue car usage in paris therefore, we decided to drop the Utilb and Utilb 14 taken and returned columns as they were unnecessary for our analysis.
3. Removing syntax Errors- we removed any white spaces that may be present in our columns and changed all the column names to lowercase so as to make the data visually appealing.
4. Counting missing values - we decided to check for any missing values in our dataframe so as to decide the appropriate way to deal with them. Fortunately for us we found that our data frame had no missing values.
5. Checking for outliers - An outlier is defined as a data point that significantly differs from the other observations. For our blue cars taken and returned columns, we found that there were many outliers. We decided not to drop them because dropping such a large part of our data would seriously affect the validity of our results.

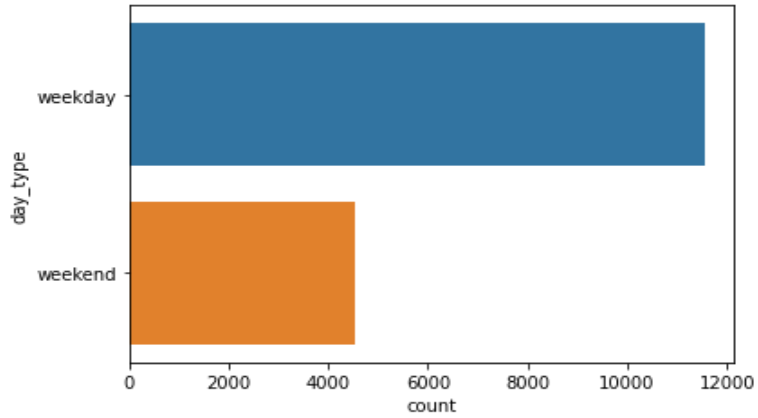
4. Analysis

With our dataframe cleaned, it was time to do our analysis in order to answer our research question.
Detailed Analysis was conducted in a python notebook using various python libraries and can be found here https://github.com/LynnNjoroge/Electric-car-sharing-Hypothesis-Testing
Our analysis is divided into the following sections:

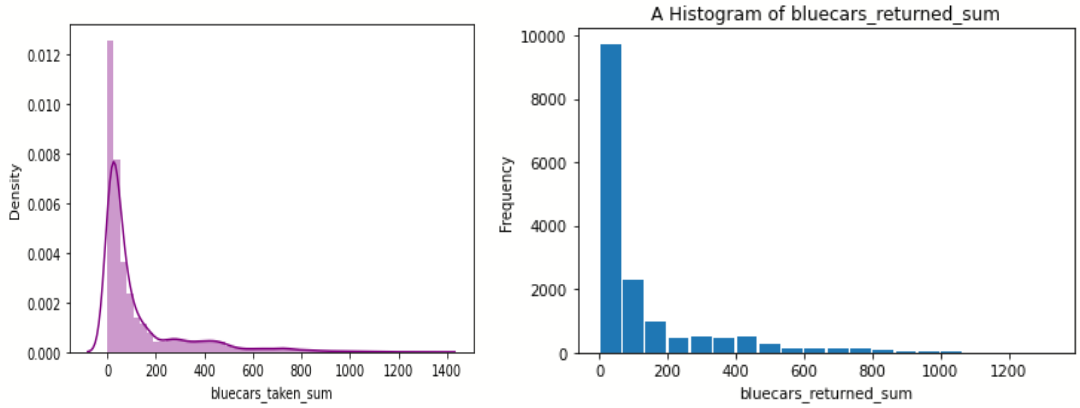
- Univariate Analysis
- Bivariate analysis
- Hypothesis Testing
- Parameter Point estimate
- Confidence Interval Construction

Univariate Analysis

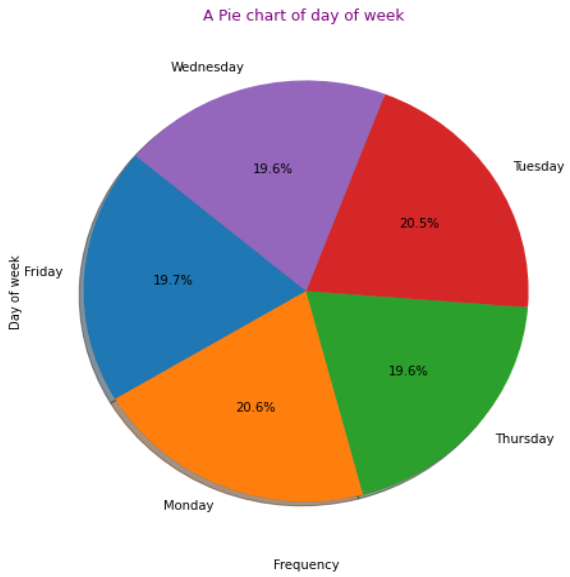
We plotted a bar graph to visualize comparison of blue car usage between weekday and weekend.The bar chart showed that blue cars were used more on weekdays than on weekends.



We then plotted histograms to show the distribution of bluecars taken and bluecars taken. Both histograms showed that the distributions were skewed to the right meaning that the mean is greater than the mode. To prove that this was true, we calculated the respective means and modes of the two columns and found definitive proof that indeed the means were greater than the modes.

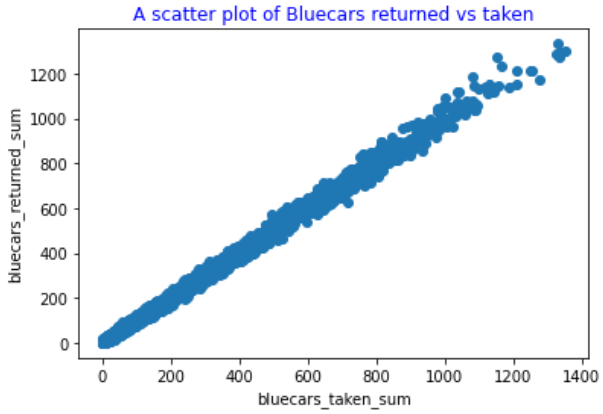


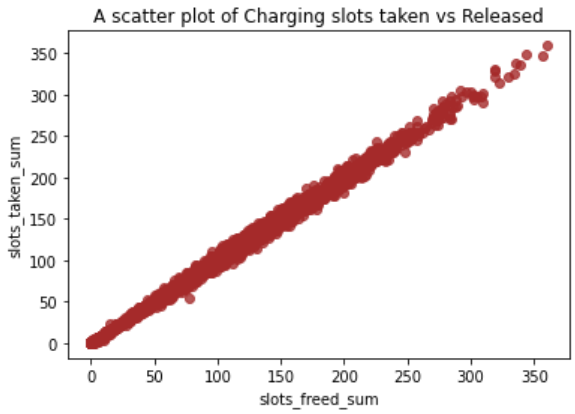
Since we now know that weekdays are the time when bluecars are used the most. We went ahead to find out which specific weekday was the busiest for the company. We plotted a pie chart to visualize this. From this pie chart we can see that the days of the week are all relatively busy with only some slight variations.



Bivariate Analysis

We plotted a scatter graph to show the relationship between blue cars taken and returned.
The plot showed there was a strong positive correlation between blue cars taken and returned.





We also plotted a scatter graph to show the relationship between slots taken and released.
The plot showed there was a strong positive correlation between slots taken and released.

Hypothesis Testing

Hypothesis testing is a statistical method that is used in making statistical decisions using data.It is an assumption that we make about the population parameter from a sample.
In our case we want to test whether there is a difference between the means of blue cars taken from 2 different postal codes.
Before we started our hypothesis testing, we filtered our dataset to only have data on weekdays. Then we used simple random sampling to select a sample of two postal codes from our filtered weekdays data frame.

For our Hypothesis testing, we followed the following steps:

***Step 1**: Formulate the null hypothesis and the alternative hypothesis.*

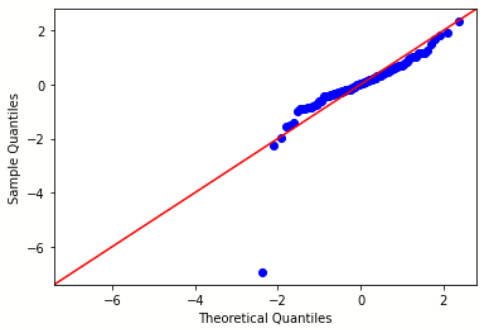
The hypotheses we have formulated are:
$H0$: There is no significant statistical difference between the mean of blue cars taken in postal code 75014 and 94130
$H1$: There is a significant statistical difference between the mean of blue cars taken in postal code 75014 and 94130
In statistical terms:
$H0 : \mu 1 = \mu 2$
$Ha : \mu 1 \neq \mu 2$
Because the alternate hypothesis had a not equal to sign, we concluded that our test would be two tailed.
This hypothesis was chosen because we wanted to find out if the average number of blue cars taken from one postal code was the same as another postal code.
This would be interesting and beneficial to the company because if we end up discovering that the postal codes have different numbers of blue cars taken, the company can use this knowledge to ensure that the busiest areas always have sufficient resources to increase customer satisfaction.

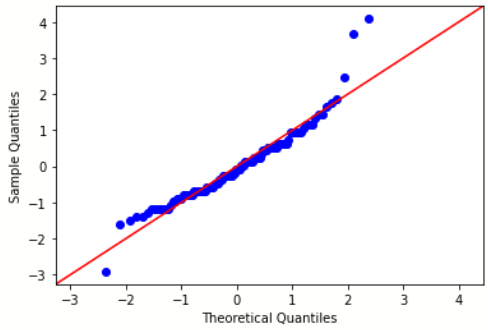
***Step 2**: Identify a test statistic and significance level.*

Because our population of the two postal codes is relatively small i.e. 112 rows we chose to use a sample size of 25. Being that our sample size is less than 30, it means that we will perform a student t-test. We are comparing the means of two samples from our dataframe, therefore we will be performing a two sample t-Test.

Before conducting a valid two-sample t-test, the following requirements must be met:

1. Sample Size should be less than 30.
   Our sample size is 25, which is less than 30. (this condition has been satisfied)
2. Data values must be independent. Measurements for one observation do not affect measurements for any other observation.
   The data was obtained from two independent postal codes whose bluecars taken from one postal code were not dependent on the other. (this condition has been satisfied)
3. Data in each group must be obtained via a random sample from the population.
   We used simple random sampling to select a sample of two postal codes from our filtered weekdays data frame. (this condition has been satisfied.).
4. Data in each group are normally distributed. To test for normality of our independent groups, we plotted a q-q plot. The plots verified that samples are normally distributed.





Now that all our conditions are satisfied, we can choose a significance level. The significance level (α), is a measure of the strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically significant. We chose a significance level of 0.05 since it is the most commonly used in statistical tests. This means that there is a 5% risk of concluding that a difference exists when there is no actual difference.

Step 3*: Computing the test-statistic and P-value.*

After conducting the test in our programming environment, the results were:

t-statistic is: 46.60397412679942

p value is: 4.705717155912204e-25

Step 4*: Analyze the results and either accept or reject the null hypothesis.*

We used the p-value to analyze the results. Since our p-value was less than the stated significance level, we rejected the null hypothesis.

Step 5*: Interpreting the Results.*

Rejecting the null hypothesis means that we have enough statistical evidence to state that, There is a statistical significant difference between the mean number of blue cars taken in postal code 75014 and the mean number of blue cars taken in postal code94130.

Point Estimate

Point estimation is the process of finding an approximate value of some parameter of a population from a random sample of the population. For us, the population parameter is the mean (average) of the blue cars taken from two different postal codes.

A point estimate for the difference in two population means is simply the difference in the corresponding sample means.

Our calculated point estimate between the two population means was 412.88.

This means that, we estimate that the average number of bluecars taken from postal code 75014 is 412.88 points higher than it is for the average number of bluecars taken from postal code 94130.

Confidence Interval

Confidence intervals are an important part of inferential statistics, upon which most market research is based. We constructed a confidence interval around the parameter which is population mean. We calculate a 95% confidence interval for our mean data for both postal code data. From the results we can be 95% certain that the population mean data for blue cars taken from postal code 75014 is between 433.62 and 455.98. Our sample mean of bluecars taken from this area is 443.04 which lies within this interval.

Additionally, we can be 95% certain that the population mean data for blue cars taken from postal code 94310 is between 28.88 and 32.56. Our sample mean of bluecars taken for postal code 94130 is 30.16 which lies within this interval

Discussion of Test Sensitivity

After completion of an analysis, the researcher has to evaluate the chance of making errors in his hypothesis test results. The analyst establishes the maximum chance of making type I and type II errors.

The probability of committing a type I error (rejecting the null hypothesis when it is actually true) is called α (alpha) which is the level of statistical significance. We had already stated our α as 0.05 before starting our tests. This means that 5% is the maximum chance of incorrectly rejecting the null hypothesis (and erroneously inferring that there is a statistical difference between the mean number of blue cars taken from two postal codes). This further means that we are 95% confident in the results of our tests.

The probability of making a type II error (failing to reject the null hypothesis when it is actually false) is called β (beta). The quantity (1 - β) is called power. Here we set our β as 0.10 meaning that we are willing to accept a 10% chance of missing a difference in the mean number of blue cars taken from two postal codes.

Ideally *alpha* and *beta* errors would be set at zero, eliminating any possibility of Type I or Type II error. In reality though, this is not possible, and the goal is to make them as small possible. However, in order to reduce *alpha* and *beta,* one would need to increase the sample size. Therefore the effect of increasing our sample size (to say a higher number-greater than 30) would reduce the possibility of committing either a Type I or Type II error.

5. Summary and Conclusions

We set out to investigate the electric (bluecars) car usage in Paris by performing Hypothesis Test to see if there is a difference in the means of blue cars taken from two different postal codes selected randomly on weekdays. To perform our analysis, we loaded our dataset into our programming environment and cleaned it in preparation for analysis. During analysis we plotted bivariate and univariate charts to better understand the information in our dataset. Finally we performed a hypothesis test to check whether there was sufficient evidence to say that the mean number of bluecars taken from one postal code was different from the mean number of bluecars taken from another postal code.

After performing a two sample t-test we found that yes, there is sufficient evidence to conclude that the mean number of blue cars taken from postal code 75014 is statistically significantly different from the mean number of blue cars taken in postal code 94130.

Although we correctly performed our hypothesis testing, there is always the chance of committing either a Type I or Type II error. Therefore, in order to minimize the chance of errors in our results, we would recommend a replication of these tests with an increased sample size.