# HYPOTHESIS REPORT DOCUMENTATION

## *Problem Statement*

https://bit.ly/DSCoreAutolibDataset is the data to be investigated.

https://bit.ly/DSCoreAutolibDatasetGlossary explains the columns of the dataset.

We will be investigating whether the day type affects the population mean of the *sum of blue cars taken* whether it is different from the sample mean.

***Null hypothesis***

The population mean is equal to the mean of the sum of blue cars taken.

***Alternative hypothesis***

The population mean is not equal to the mean of the sum of blue cars taken.

The investigation will help show whether what we think about the dataset is correct. The goal is to determine whether our guess is right or wrong.

## *Data Description*

The data comes from the autolib car-sharing company located in France for the city of Paris year 2018. It gives information about the postal codes of the areas in Paris, cars taken and utility data for the corresponding areas. The dataset is a daily aggregation, by date and postal code, of the number of events on the Autolib network (car-sharing and recharging). The column we are interested in is the blue cars taken which shows the number of cars taken that date in that area. The table below shows a brief descriptive statistic for this column:

| Statistics | BlueCars taken column |
|---|---|
| Count | 16085 |
| Std | 185 |
| Min | 0 |
| Max | 1352 |
| Mean | 125 |
| 25% | 20 |

| | |
|---|---|
| 50% | 46 |
| 75% | 135 |

## *Hypothesis Testing Procedure*

1. Begin by asking a question: for our case, we could ask what is the mean for the blue cars taken.
2. Develop an experiment to answer our question: for our case, we mathematically find the formula for the mean.
3. Gather data to answer the question: we first load our dataset and do some descriptive statistics to get our results.
4. Create our hypothesis: working with this data we believe that the mean is pop_mean. (this becomes our null hypothesis)
5. Analyze the data: based on the dataset we find more information about the data we are working with, what kind of distribution does it have, the correlation to the other columns.
6. Challenge our hypothesis: this is we do not believe that the mean is pop_mean. (this becomes our alternate hypothesis)
7. Perform a z  test.
8. Perform a t-test. (as a curious sentiment)
9. Find the p-values from both the z test and t-test: this is to help us find out if there are similarities to the results we get.
10. Draw conclusions from the results we get and either fail to reject our null hypothesis or accept our alternate hypothesis.

The logic behind our null hypothesis is that we have run some descriptive statistics on the data and believe that the mean is in fact pop_mean. But we are interested in finding out whether this is accurate.
For our alternate hypothesis, we are challenging the null hypothesis by claiming that the mean is not pop_mean. This is what we are attempting to demonstrate in an indirect way by the use of our hypothesis test.
We will perform the z-test since our sample is greater than 1000 and we have all the values we need to place in our formula.

We have made assumptions that are as follows:

1. Distribution is normal.
2. The sample is randomly selected from the population.
3. Sample data is representative of the whole population and can, therefore, estimate population parameters.

4. Approximately equal variances of the population and sample.

The distribution does not appear to be normal by drawing a distribution plot for our data.
We will use an alpha level of 0.05 if the p-value is smaller than this we reject the null hypothesis otherwise we fail to reject the null hypothesis.

## *Hypothesis Testing Results*

| | |
|---|---|
| z-score | 0.0221721068649974 |
| p-value | 0.4912 |
| Reject or Accept | Fail to reject the null hypothesis |

After running the z test we got a z-score of 0.0221721068649974 and a p-value of 0.4912. With our significance level of 0.05, the p-value was greater than the significance level thus we fail to reject our null hypothesis.

After running the point estimate we got the mean as 56.5829 with a difference of 0.7515378480662261 with the overall population mean.

We then construct a confidence interval of 95% and 98% which gives us the following results:

| | **95% CI** | **98% CI** |
|---|---|---|
| Lower | 54.46749551794249 | 53.82520878086015 |
| Upper | 61.32532666802815 | 61.967613405110484 |

We did perform the t-test but since our sample size is large then we do not feel like the results are worth adding in the report.

## *Summary and Conclusions*

In summary, we defined our null and alternate hypothesis, loaded our data, performed statistical analysis on the data, divided the data into population and sample data which gave us the values we needed for the calculation of the z-score.
The z-score (0.0221721068649974) gives us a p-value (0.4912) that tells us not to reject the null hypothesis that the population is pop_mean at the significance level of (0.05).