Adding Challenge

Software development

written by:
CORDOVAL Chloë
Master's degree in information and decision-making mathematics

2020/2021

Git Code link

Script Python Prediction

Git link Visualization

Script Python Visualization

1 Introduction

This is a small challenge based on real-time data. We are going to study the database which represents the community readings of the Albert 1er totem pole. Totems Montpellier is a collective which collects daily data from the "bicycle counter" installed at Albert 1er in Montpellier.

Since March 2020, the cycle path behind the “place Albert-1er - Saint-Charles” tram stop, along the quai des tanneurs, has been welcoming the 1st cycling counting totem in the metropolis of Montpellier.

We need to predict **how many bikes will pass** between 12:01 am and 9:00 am on Friday April 2, 2021

2 Prediction Part

For this part, you will use the data from

[https://docs.google.com/spreadsheets/d/1ssxsl9AIobDofXFohvwxqCPF0tn6dgXpixhiDzus0iE/editgid=59478853]

to predict the number of bicycle passing between 00:01 am and 09:00 am on Friday, April 2nd.

2.1 Dataset

At first, we will reformat the data of our Dataframe finally to remove the "NAN" and replace them by the value 0 in our Dataset. We have therefore removed the last two columns as well as the first two rows.

We therefore have at the end a table in which takes as columns, the Date, the Hour and Todaystotal.

In a second, we will separate the date that was expressed until now Year-Month-Day and we will therefore have a new dataframe which will take as column the Date, Time, Todaystotal, DayOfWeek and Year.

Now, we are going to create a sub dataframe of our basic dataframe by making particular maneuvers on the Time format because it is considered as a character string which is unusable.

In this new dataframe, we have the columns: DayOfWeek, Hour, Todaystotal, Date and Year except that we have imposed on our dataframe, to express that the times which are between 00:01 and 08:00 (because it is equivalent to time 8 am-8:59 am).

Which is favorable to us for the interpretation.

To finish our work on the DataSet, we will divide our new dataframe into Blocks per hour.

This will give us all the values between each hour slice; ie the days and the Todaystotal.

2.2 Prediction

Our prediction work consists of taking the average and then the median for each hourly slice.

This is **the robust median method** which is a two-dimensional linear regression method.

It is simply a question of applying on $X_i = \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_9 \end{pmatrix}$ which represents the hours and $\lambda_i = \begin{pmatrix} \lambda_1 \\ . \\ . \\ . \\ \lambda_9 \end{pmatrix}$ which represents the number of bikes per hour.

And then we know the formulas for the mean and the median:

$$mean = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$median = \begin{cases} \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \text{si n est pair} \\ X_{\frac{n+1}{2}} \text{ si n est impair} \end{cases}$$

However, when we sum the means and the medians, we find $mean = 408,8666$ and $median = 401$.

There are outliers due to the health situation and government decisions.

Our Dataframe starts on March 12 which is 4 days before national containment so there hasn't been a lot of traffic.

Then, on May 11, there was the start of deconfinement so the circulation of bicycles in the city center was able to resume.

The summer vacation period brings a set of values too large for the April 2 prediction.

In October, there was a partial confielent which lowered the numbers of passages.

In December, on the 15th, a curfew was put in place from 8 p.m. until 6 a.m. and since January 2, 2021, the curfew has been increased from 6 p.m. to 6 a.m.

It is therefore more interesting to look at the data since the start of the curfew, i.e. January 2, 2021.

We are therefore going to put a filter on the "Year" data, in order to redo all our calculations of the mean and the median without the outliers of confinement, summer holidays and the curfew at 8 p.m.

So we have $mean = 211,833$ and $median = 220$.

So we can make as a prediction, that there will be approximately **220 bicycles** which will pass between 00:01 and 09:00 in the morning.