Bike challenge

Foux Quentin

April 2021

1 Introduction

For this prediction part, we had to predict the number of bicycles passing between 00:01 AM and 09:00 AM on Friday, April 2nd, by Albert 1er street in Montpellier. A url containing data, in a csv format, was provided. In this file, I am going to explain my approach, which leads to an estimation of that desired bike number.

First, I will present how I sorted my data, and then I will explain my two prediction methods: the first one is simple and easily works, giving a "good" estimation according to very last values, and the second one which is harder but does give an atypical value.

2 Data sorting

After downloading the file given in the url, I loaded it into a pandas dataframe so as to manipulate columns with convenience.

At first, I deleted rows filled only with NA values and I removed columns, from my pandas dataframe, which would have not been useful to my analysis.

Then, I decided to keep only the rows whose hour was between 8:30 AM and 9:30 AM. My idea was to use only data close to 9:00 AM, which is the time of when we have to give our prediction number.

Besides, certain data had huge circumstance differences, some of them were taken during the first lockdown, some were taken during non-lockdown periods, and others during curfews. In consequence, after observing values per day, I made the choice to keep only dates after first lockdown. Indeed, during first lockdown, values were really low compared to others. The others, apart from few atypical values, were pretty similar.

Finally, I noticed week-end values were far lower than weekday ones while weekdays are quite similar between them. So, I kept only weekdays since friday is a weekday.

3 Mediane method

I first tried a simple idea: create a pandas dataframe composed of weekdays only, then calculate with median() function the mediane of the column Day's total. Indeed, here we still have some atypical values, so the mediane might be better than a soft mean. With only weekdays, I got a prediction of 248 bikes.

However, I thought that take only fridays could be a good idea as well, since weekdays are unique: people who work fridays but not mondays for instance or the contrary, students who have not class early friday mornings but another weekday morning (when not quarantined)... we could get circumstances more precise taking just fridays.

In the same way than weekdays, I created a friday pandas dataframe. And that time, using median() function, I got a prediction of 279 bikes. Furthermore, the last two weeks, values were rather close to that prediction.

That prediction might be a good one to start with.

4 Linear regression

In that part, results were not satisfying. I wanted to make a linear regression of weekdays total compared to dates, so as to get the linear regression line and its equation. With the coefficients of this line, I wanted to predict weekdays total the April 2nd, and then I just had to do the difference between that prediction and the real value of the weekdays total the day before that prediction.

In that purpose, I filled missing dates, then I created the linear regression line appearance (with seaborn) and displayed the line coefficients (with linear_model and statsmodel for more information). Nethertheless, I got an estimation of 660 bikes, which is a lot...

That method is probably not a good idea for our prediction, or is not used the right way. Consequently, I didn't try with only fridays.

5 Conclusion

Linear regression method seems to give atypical values that cannot be trusted. Even if that method was a failure, we still can keep mediane method that looks reliable.

Hoping that Meteo France is right and it's not raining 2nd April early in the morning, 279 bikes prediction might be close to the real value. That's my prediction.