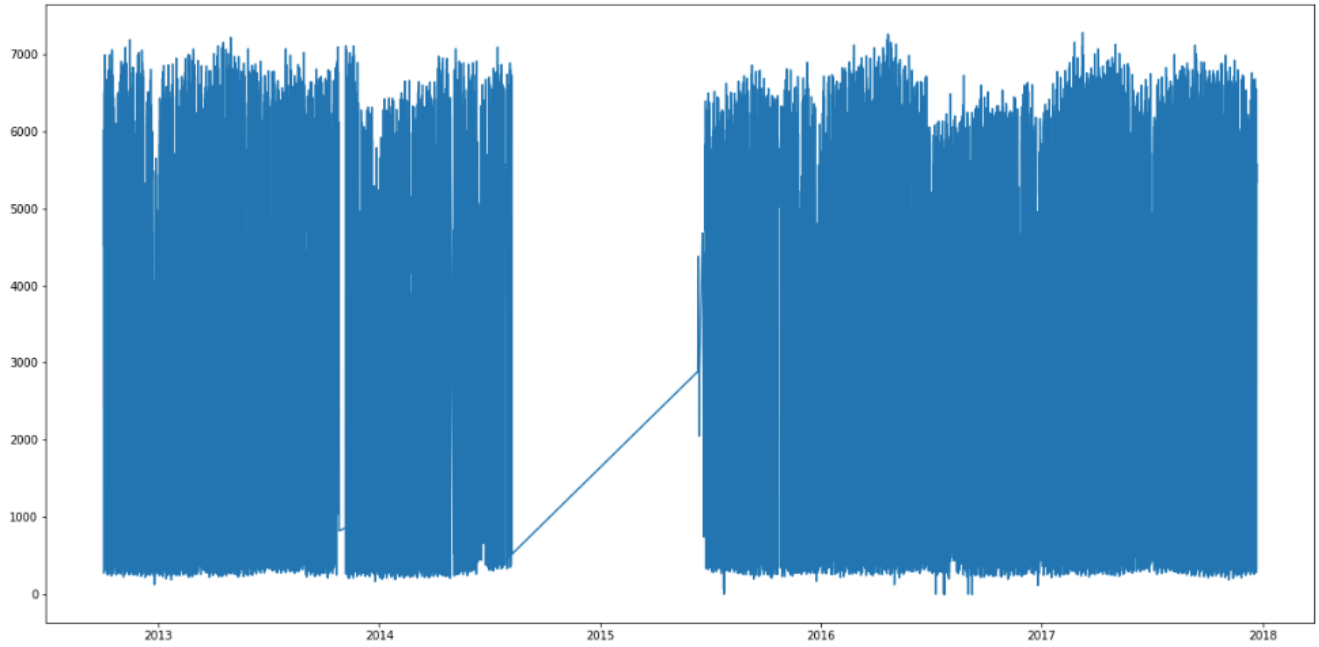
Project 2: Forecasting Highway Car Volumes

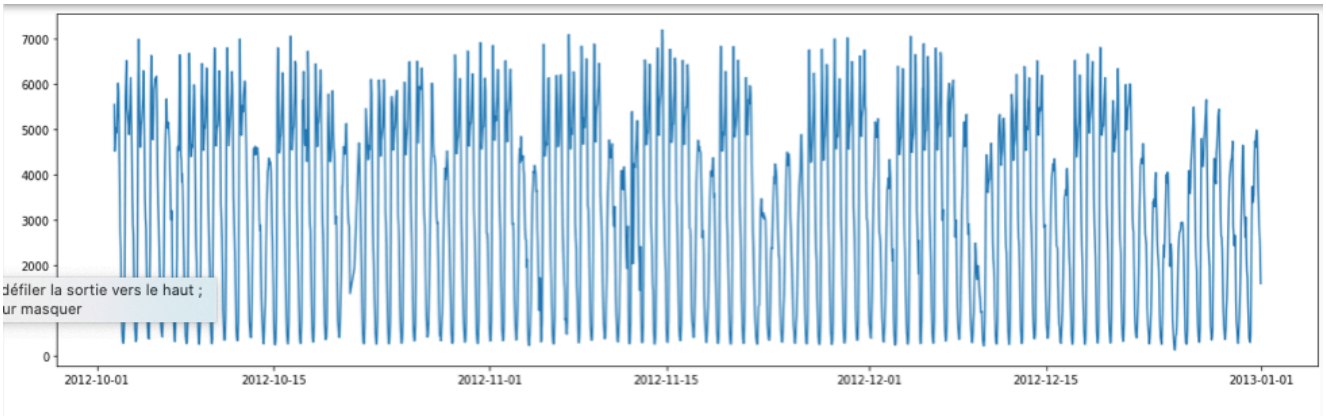
Step 1 Regression on Time

In this task, I need to build a regression model only using the variables Time and TrafficVolume for prediction. To work on this step, I decided to create a sub-dataset where I would initially only have the variables Time and TrafficVolume. After putting Time as the index (which helps for the plots), I extracted from the time the variables 'Year', 'Month', 'Day', 'Hour' and 'Weekday' which will help me to create the model.

I plotting the full data, as shown below:



We can see that apart from 2015 large missing data, the year does not seem to have the biggest influence over the trafic volume. It will as such not be part of the model by itself. We notice however that the trafic volume seems to changes depending on the month, and more specifically, depending to which day of a month we are. As such, the model will depend on day*month. Moreover, we notice on the plots patterns that seems to repeat themselves every 7 days, so the day of the week seems to have an impact on the model. But more specifically, it seems like it is not the day of the week alone that has an impact on the trafic volume, but the time of the day of the week, as shown in the graph below :



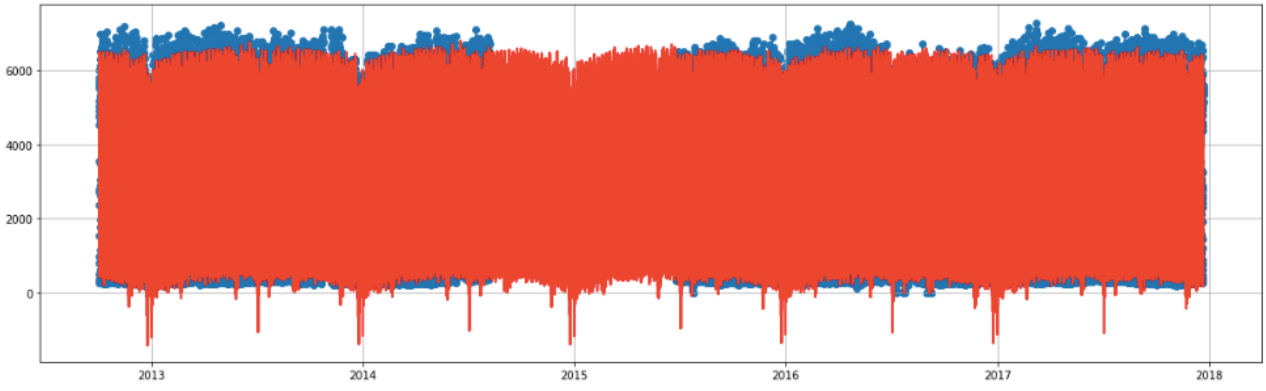
As such weekday*time will be part of the model to predict the trafic volume. I decided to use dummy variables, because as we are working on dates, we work with a variable Time that has many subgroup : Year, Month, Day, Weekday, Hour. Using dummy variables enables to control for time-specific effects, where the impact is restricted to given time period.
The model we are using is as such : 'TrafficVolume~C(Weekday)*C(Hour)+C(Day)*C(Month)'.

We are computing it for 3 dataset : the first goes from the start to the gap of 2015, the second from the end of the gap of 2015 to the end, and the third is the full data set.
This will allow me to make sure that the 2015 gap doesn't have too big of an impact on the choosing of the model, as when there are missing values, the measured values used in computing R_a^2 are 0. Here are the results of the number of degrees of freedom and the R_a^2 for the three models:

	Model 1 2012-10-02 09h - 2014-08-08 0h	Model 2 2015-06-24 0h -	Full model 2012-10-02 09h - 2017-12-22 16h
Degree of freedom	531	532	532
R_a^2	0.954	0.953	0.948

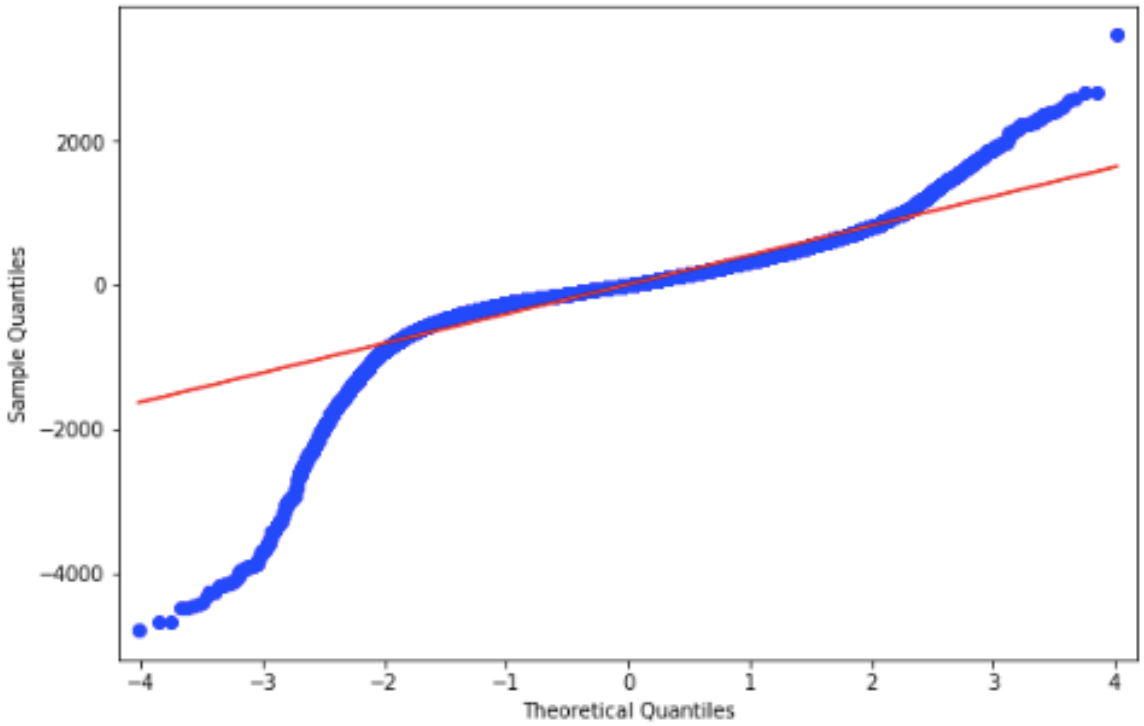
Across the three dataset, we find a R_a^2 of around 0.95, which indicates that the model is quite good at predicting the values.

After plotting the model for the three dataset, we notice that there are some negative traffic volume predictions (impossible !), as shown in the graph of the predictions made on the full dataset (in red) and the original values (in blue):



As we see most of them around the new year period, we can guess that these errors come from the day after low traffic volume from Christmas and New Years Eve. Different approaches could be taken here, but for the sake of keeping this model simple, we will just consider that the negative values are null. It is however an idea to make the model better for step 3.

To male to diagnostic checks, as the results are quite similar for all three datasets, we will just use the model with all the dates (including the 2015 gap).
We notice that some dummy variables have more weight than some others in the model. As such C(Hour)[T.3] has 10 times more weight as C(Hour)[T.4]. This can be explained by the fact that the trafic might not change a lot in between 3am and 4am, and thus, the results for 4am are similar than those for 3am.
To check the normality, from the plots of the residuals (in histogram or scatter form) we can see that most of them are around 0, even though there are some that have higher values. I also plotted the quantiles of the residuals against the theoretical quantiles of a normal distribution and I can see that the residuals stay on a straight line on the normal quantile quantile plot for quantile in between -2 and 2, but they do not follow the straight line for other values of the quantile, as such, the residuals do not follow a normal distribution :



Nevertheless, the model seems to be working well and the data is not dependent spatially, so we can still use this model.
The partial autocorrelation plot also shows that the model is not completely stationary, but it does not seem to be completely out of bounds, so we can accept that difference.

Step 2 Exponential Smoothing

In this task, I will apply appropriate exponential smoothing to make forecasting.

As the trafic volume is very dependent on the time of the day, the day of the week and the day of the month, we can say that it is seasonal data.

As we do not particularly see a linear growth trend in the data, we will not use double exponential smoothing using Holt's method.

If we use a simple exponential smoothing, we can see that the best alpha is 1, as the error follows a decreasing function. However alpha = 1 means that our prediction is the observation, and as such there is no forecasting.

Similarly, the moving average yields the best result for the smallest possible number of observations used to forecast.

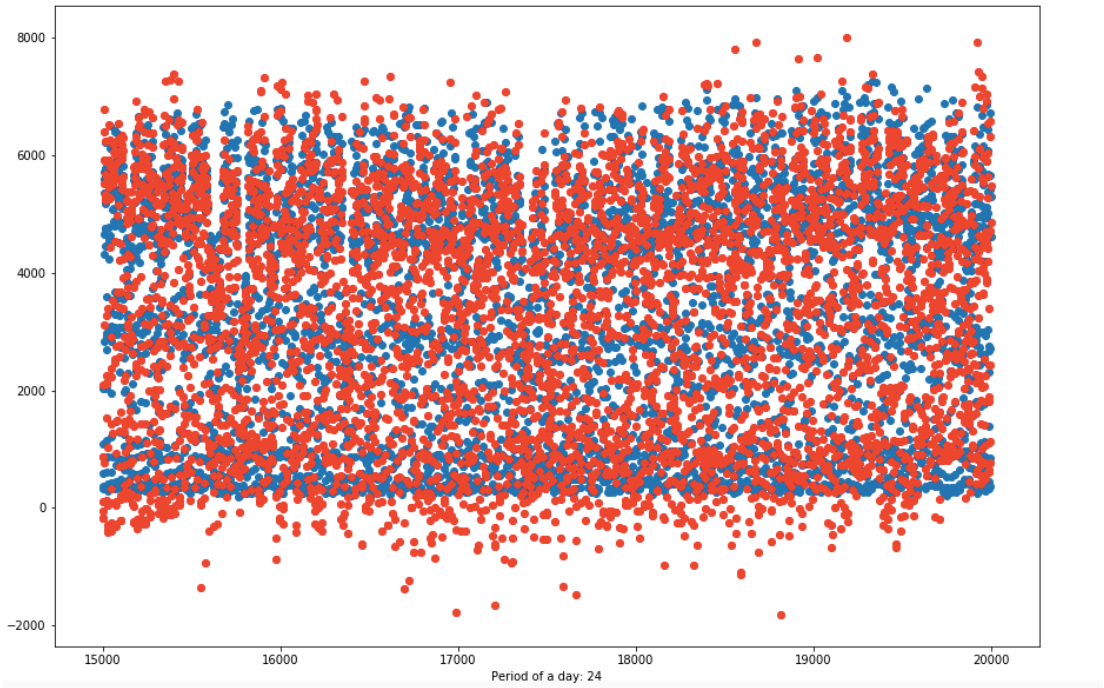
This because of the seasonality, and for that reason, we will apply the Holt's Winter Method which accounts for seasonality.

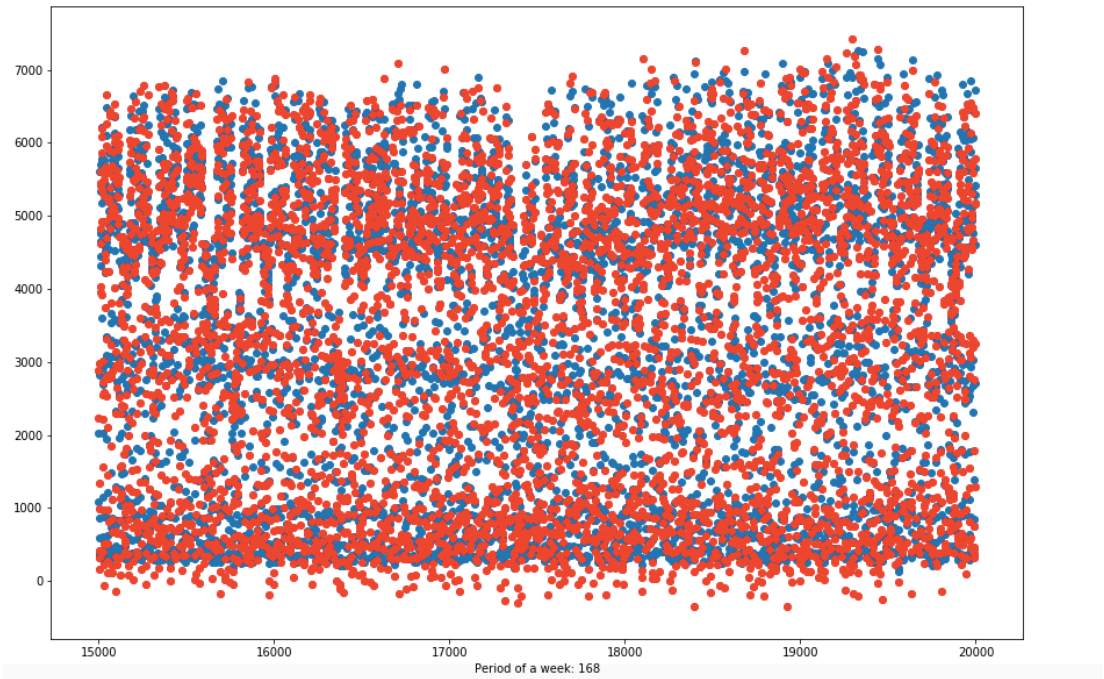
Using statsmodels.tsa.holtwinters allows me to model on the data the Holt's Winter Method. As our dataset does not have a linear growth trend, I set the trend on « None » to only have an additive seasonal model. I did not give the program any alpha and beta value, as it has the possibility to optimise their value, which is ideal here. I tried two different periods for this seasonal modelling :

- Period is 168, which is 7*24, so the number of hours in a week. It seemed to be a good possible period as when looking at the initial plot of the trafic volume, you could see variations each 7 days. The error for this period is 1601.64.
- Period is 24, which accounts for a day. As there is always less trafic at night than in the morning of the late afternoon, the seasonality very clearly depends on the time of the day. The error for this period is 1529.17.

Both errors for this method are much better than the errors obtained for the moving average (which is 1901.29) and for the exponential smoothing, where the error went below 1500 for an alpha bigger than 0.5 (so the observation has the biggest weight).

If we choose the period that yields the smallest error, then the best period to model the seasonal data using exponential smoothing is a period of 24, so a period that accounts for each day. However, when looking at the plot of the predicted value, we can notice that this period creates a lot of negative results (which are impossible), compared to the model with period 168. Then in that regard, the model with period 168 would be better :





The big risk with this model, is that as our data has some missing values, the season might shift over time and not be exactly the same for each new period. We are also not able to access the optimised alpha and beta, which could be useful to understand better the model.

Step 3 Free form forecasting

In this task, I will build your own model to predict the traffic volumes.

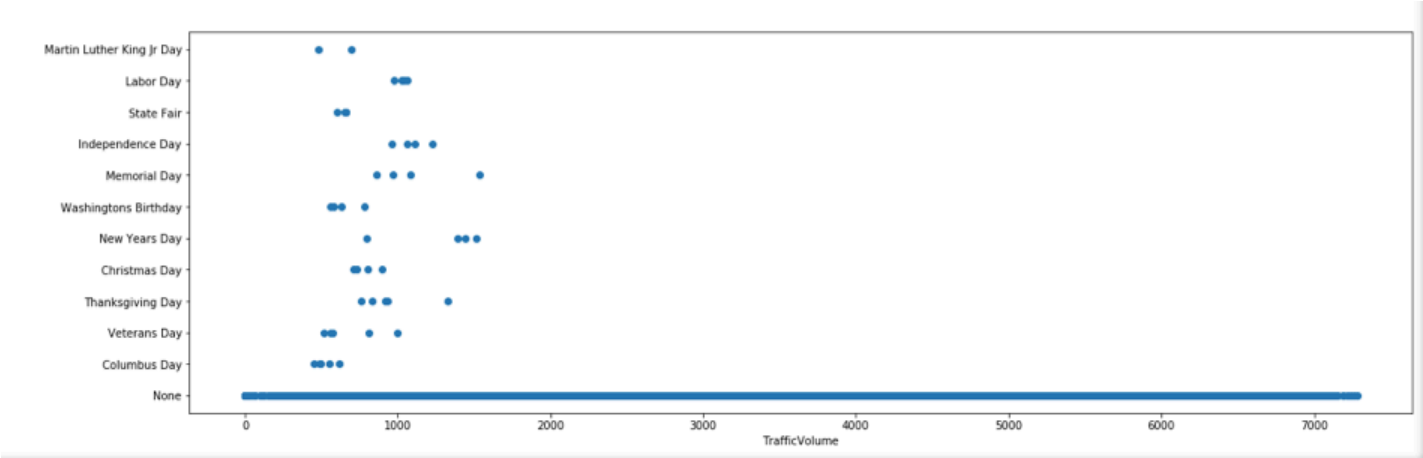
For this step, I can use every variable provided by the measuring station. They are :

Variable	Type	Description	Could have an impact on trafic ?
IsHoliday	Categorical	whether it is a public holiday or not	Yes
Temp	Numeric	average temperature during the hour in Kelvin	No necessarily, the decrease in temperature could be linked with the time (colder at night) and as such the temperature would not be needed for predict the trafic (time is enough)
Rain1h	Numeric	amount in mm of rain that occurred in the hour	Yes, probably less cars when it rains
Snow1h	Numeric	amount in mm of snow that occurred in the hour	Yes, probably less cars when it rains
CouldsAll	Numeric	percentage of cloud covering the sky	Not necessarily, I don't think the number of clouds impact the trafic
WeatherMain	Categorical	short textual description of the current weather	Could be linked with Rain1h, Snow1
WeatherDescription	Categorical	longer textual description of the current weather	Not necessarily as WeatherMain will probably explain what is needed
Time	DateTime	Hour of the data collected in local time	Yes
TrafficVolume	Numeric	hourly reported westbound traffic volume	-

First we will have a look at the missing data (none), and for the data whose value seems absurd. I found that Temp had 0 values, which for a temperature in Kelvin is impossible on earth. I replaced these temperatures with the temperature from the previous hour, to minimise the error in forecasting. There also was an absurd value of rain where it would have rained 9m in hour, I replaced this value by 98mm to keep an idea of the original value and as the description indicated « Heavy Rain » .

I then plotted the different variables above (except time as it was done previously in the code already) according to « TrafficVolume » to see if there was any relationship between them.

At a first glance, IsHoliday seems to mean that there is less trafic on that day :



For the other variables, none seem to have a direct relationship with « TrafficVolume », so I will have to proceed differently to improve my model.

I started from the model in Step 1 and tried to add the new parameters. As 'IsHoliday' is repeated each year, I am liking it the to the year, and the other variable are more dependent on the hour, I tried the model: 'TrafficVolume~C(Weekday)*C(Hour)+C(Day)*C(Month)+C(IsHoliday)*C(Year) +C(WeatherDescription)*C(Hour)+C(WeatherMain)*C(Hour)+CloudsAll*C(Hour)+Snow1h*C(Hour) +Rain1h*C(Hour)+Temp*C(Hour).
In this model, the variable cloud has very little impact on the prediction, as its coefficients are always under 1. Similarly, the temperature doesn't seem to have a large impact on the model as its coefficients are in the order of 1. The variables « Rain1h »,« WeatherMain » and « WeatherDescription » seem to have a little more impact as its coefficients are in the order of 10^1 but it still a rather small impact compared to other variables such as « Snow1h » which has coefficients are in the order of 10^2 and « IsHoliday » which has coefficients in the order of 10^3.
This model had an adjusted R^2 of 0.951, which is already really good, but I would like to improve it by at least simplifying it and having the best RMSE possible, as it is now of 431.94, which is quite big compared to the average of the trafic volume which is of 3289 (RMSE is 13% of the average trafic volume now).

I will now try a model which doesn't have the insignificant variables from above : 'TrafficVolume~C(Weekday)*C(Hour)+C(Day)*C(Month)+C(IsHoliday)*C(Year)+Snow1h*C(Hour)'. This model yield a RMSE of 442.45, which is actually worse than the previous model !

I also tried this model : 'TrafficVolume~C(Weekday)*C(Hour)+C(Day)*C(Month)+C(IsHoliday) +Snow1h*C(Weekday)+Rain1h*C(Weekday)' to see if there was a link between the trafic and the rain and snow that falls at different moments of the week, but it did not seem to be any better as the RMSE of this model is 445.23.

As I did not get very conclusive results from this technique, I decided to try a Box-Jenkins methodology for seasonal data : SARIMA, with a period of a day (=24). This model yields a RMSE of 779.60, which is even larger than the previous models.

As such to make the predictions on the test dataset, I will use the first model which has all the variables included. We are deleting the link between the year and the holiday and just leaving the holiday as when we are predicting the test data, there is the year 2018 which is not in the train data and it raises an issue. I also had to delete the variable 'WeatherDescription' as it had a new description in the test dataset. This new model is : 'TrafficVolume~C(Weekday)*C(Hour) +C(Day)*C(Month)+C(IsHoliday)+C(WeatherDescription)*C(Hour)+C(WeatherMain)*C(Hour) +CloudsAll*C(Hour)+Snow1h*C(Hour)+Rain1h*C(Hour)+Temp*C(Hour)'. It has an RMSE of 438.97 and an R_a^2 of 0.950.
My results are in two .csv files : one that contains the predictions only for the index values, and one that has the prediction values for all the values for the test dataset.

To improve this model, I could have found a way to keep the variable 'Year' and 'WeatherDescription' involved in the model as they helped it get a better score. I am also not completely satisfied with the use of the new variables (compared to step 1) as I feel like they could have added more to the model.