Air pollution RATP
AI mini-project 6

Team:
Lara Anna WAGNER (DiSc M2)
Yuliia NIKOLAENKO (DiSc M2)
Mihaela Elena GRIGORE (LeSc M2)

Introduction to our dataset and problem

Quality of air monitoring:

- Underground:
  - From 3 subway stations (RATP)
  - Chatelet, Auber, Frank Roosevelt
  - Features:
    - NO (Nitrogen Monoxide), NO2 (Nitrogen Dioxide) - nitrogen oxides are produced from fuel combustion
    - CO2 - exhaled by commuters
    - PM10 (Particles suspended in air) - friction wheel vs rail, brake, construction material
    - Temperature and humidity

Purpose of gathering this data

- Health risks for commuters - something we must be aware of & estimate
- Ventilation system - project needs for new stations, evaluate satisfaction of needs in old stations, improve
- Are we replacing one source of pollution (surface) with another one (underground)

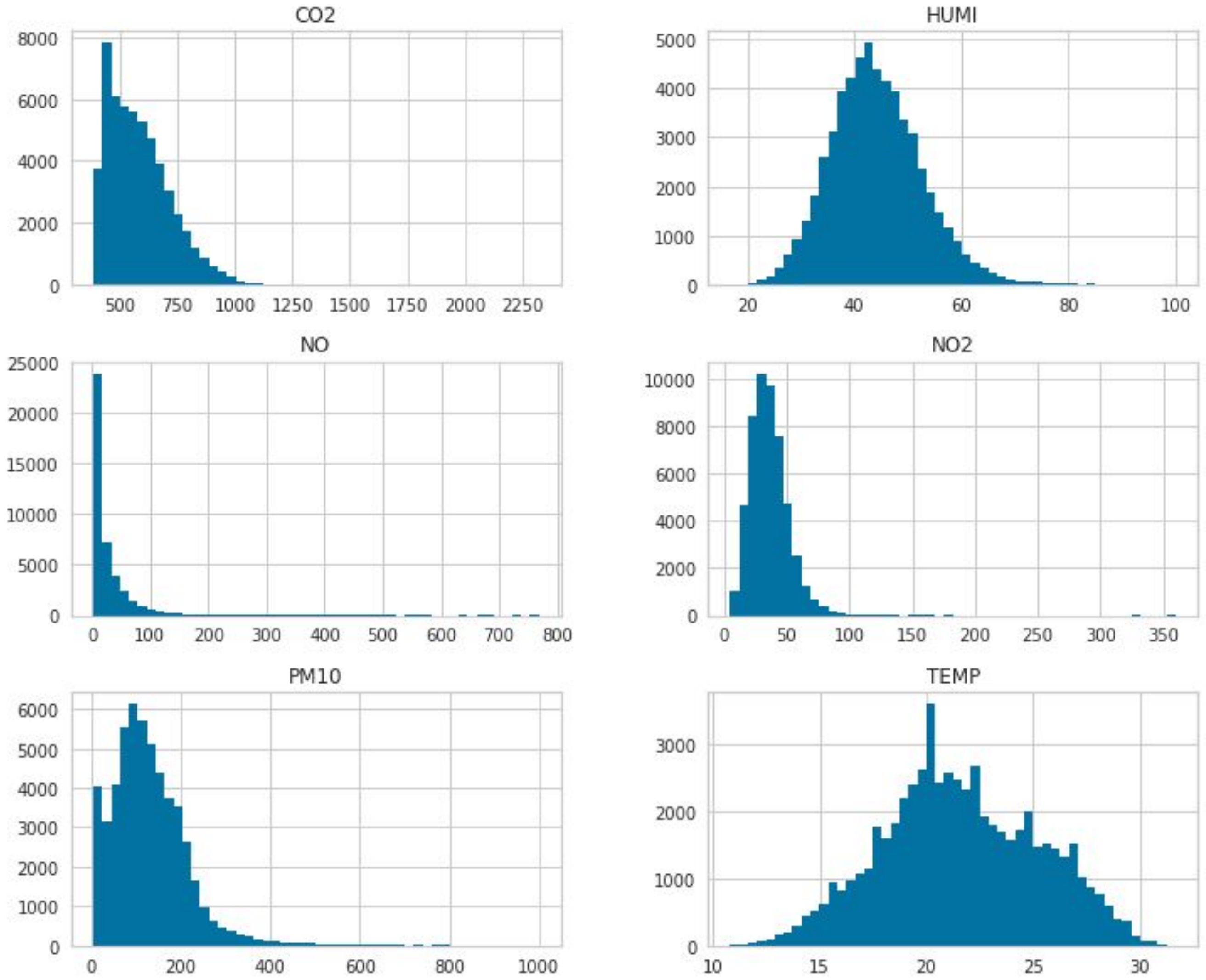
Introduction to our dataset and problem

EDA

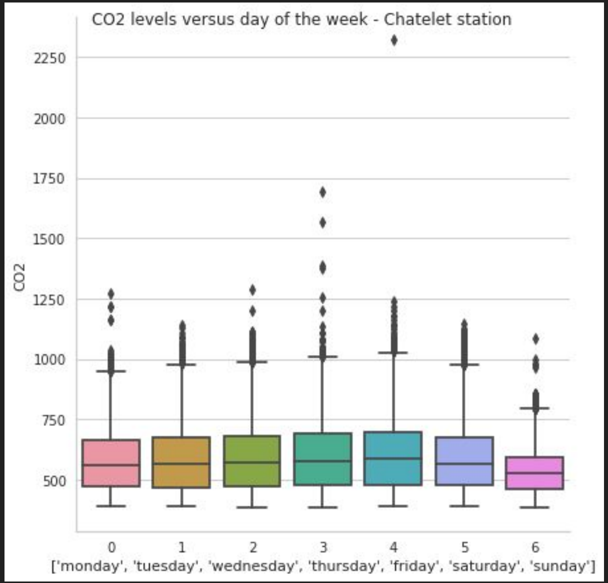
- 7 years of data (from 2013 to Nov 2020)
- randomized
- 1 measurements per hour
- 24/7 measurements

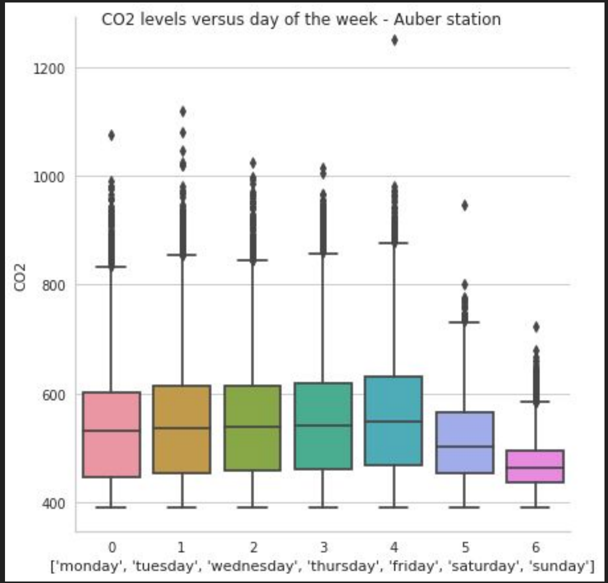
Data engineering

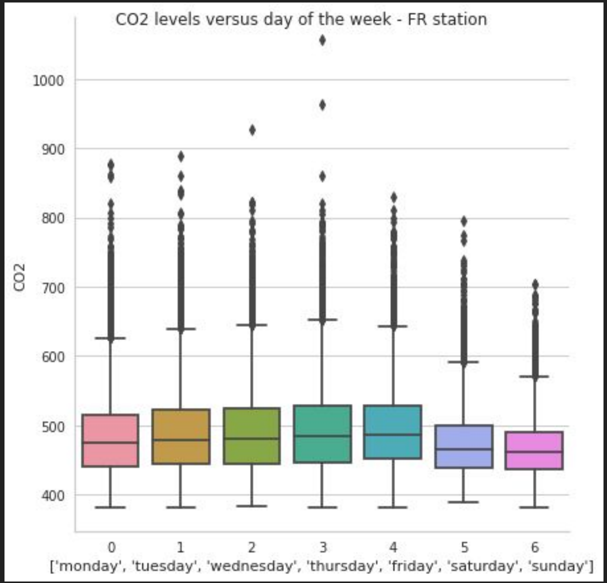
- Detecting and deleting missing values
- Converting the date into the right format: hour, weekday, month
- Creating new attributes weekend 0/1, summer 0/1


CO2
HUMI
NO
NO2
PM10
TEMP

Weekday and CO2 level for each station







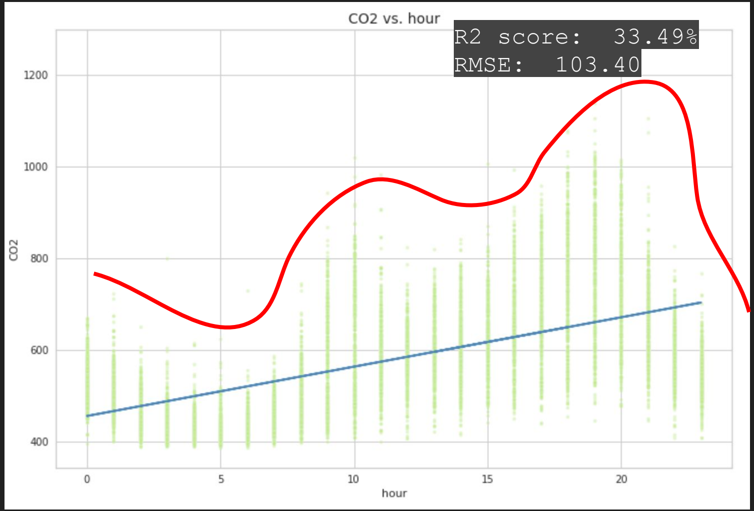
Part 1

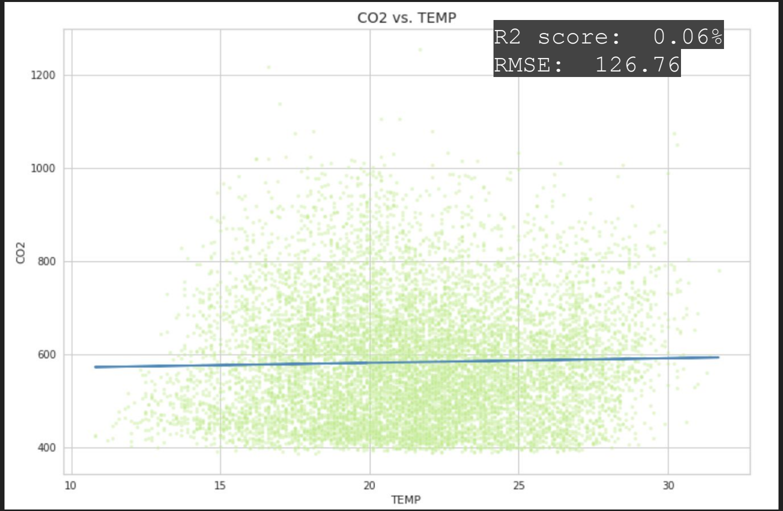
Predict **CO2** level in Chatelet from **time of day**, **temperature and humidity**

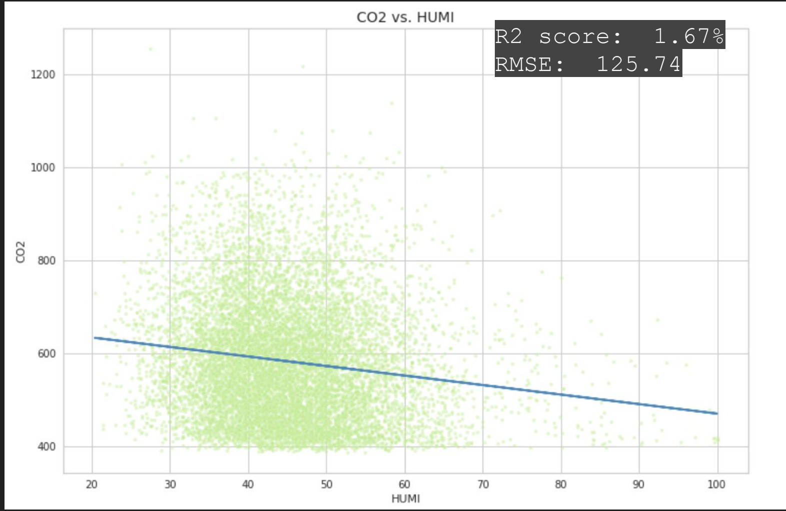
Correlations

Indep:
TEMP, HUMI, Weekend, Hour

Dep: CO2



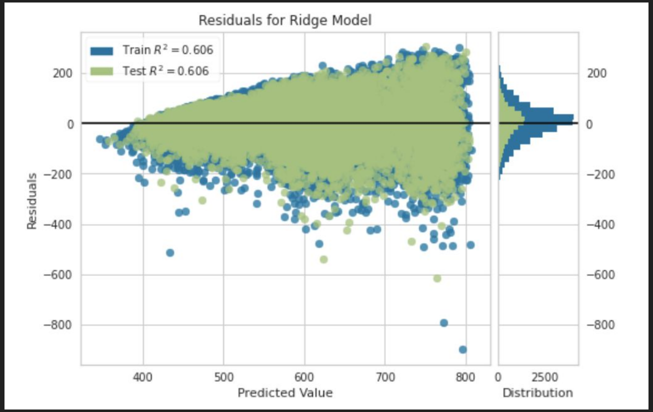


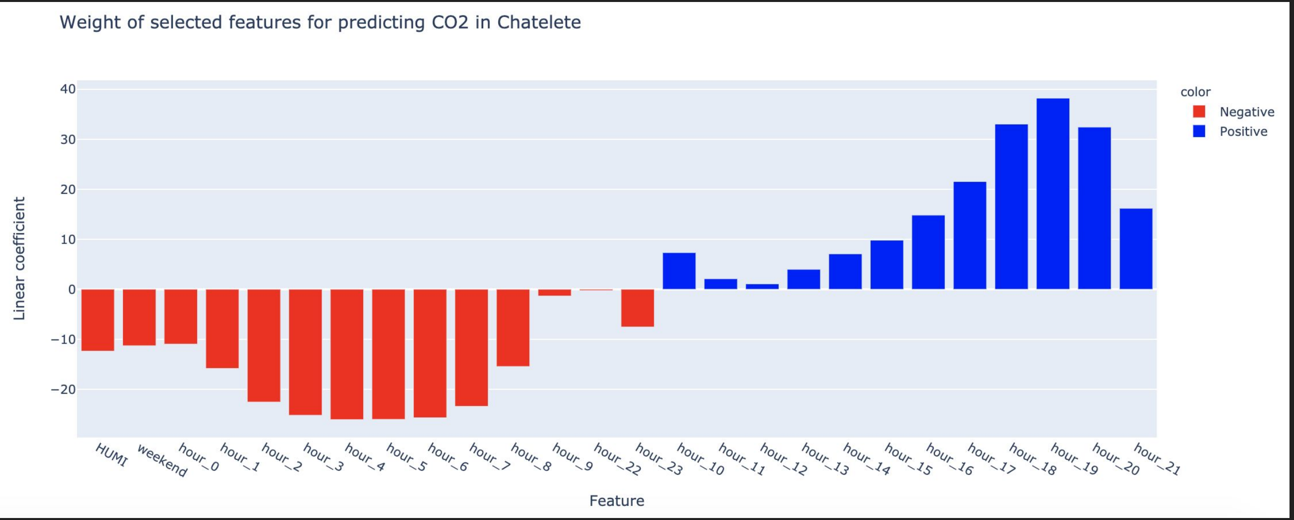


Feature	SD
CO2	128.52

Evaluation CO2 in Chatelet

Intercept	582.79
R2 Mean Score	55%
RMSE	80.45





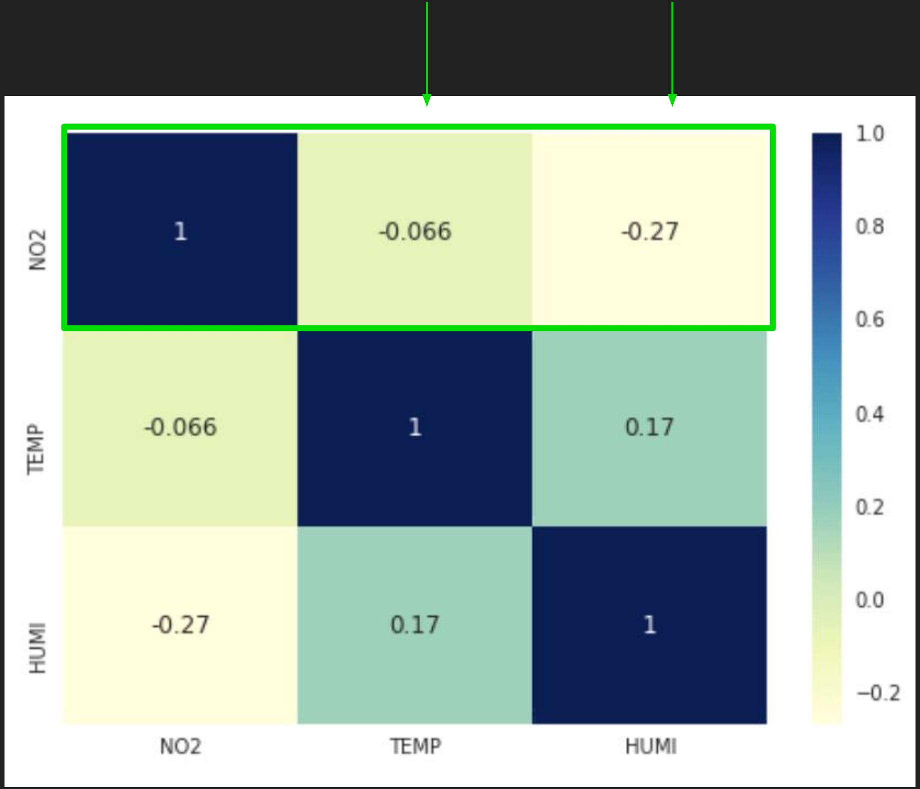
Indep:
HUMI, Weekend, Hour
(ALL standardized)

Dep: CO2

Part 2

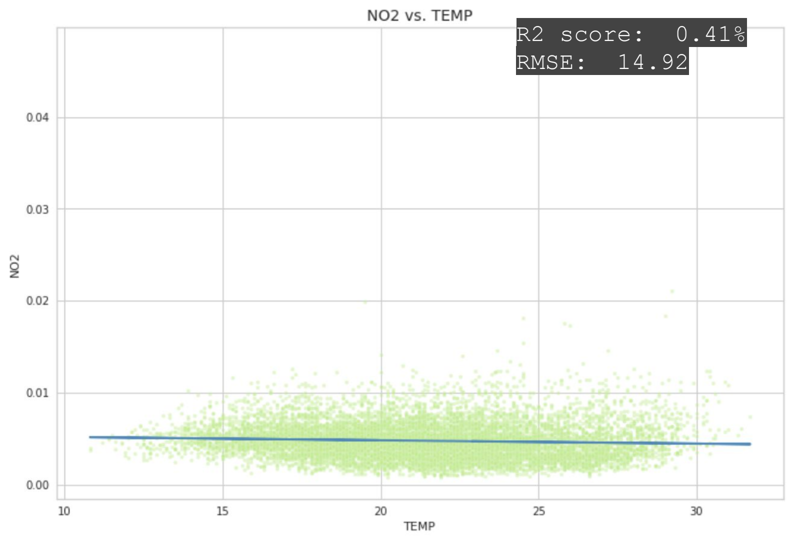
Predict **NO2** level in a Chatelet from **past values**, **temperature** and **humidity + CO2**.

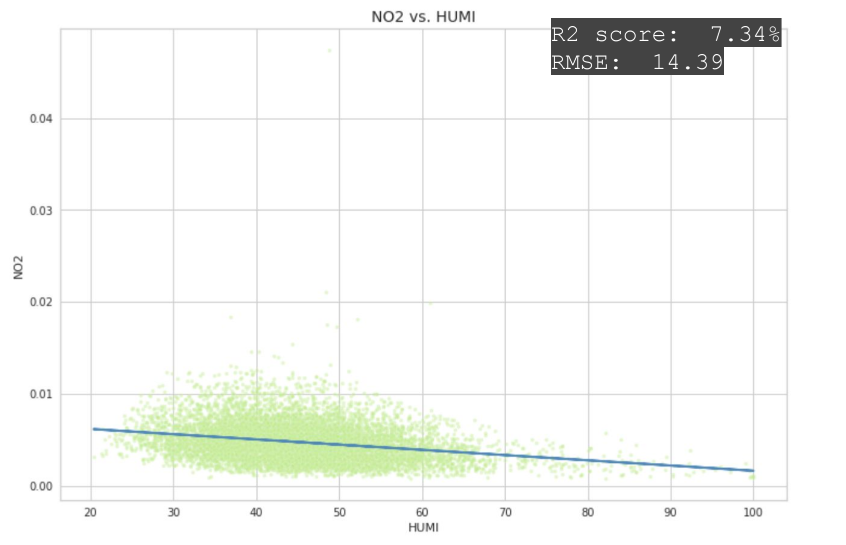
Correlations



Indep:
TEMP, HUMI

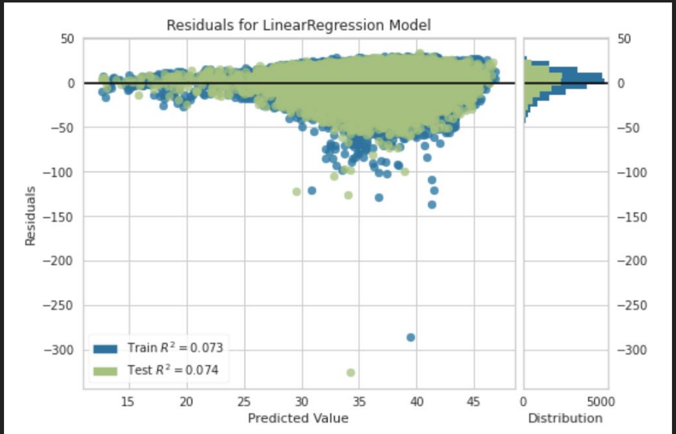
Dep: NO2





Feature	SD
NO2	14.84

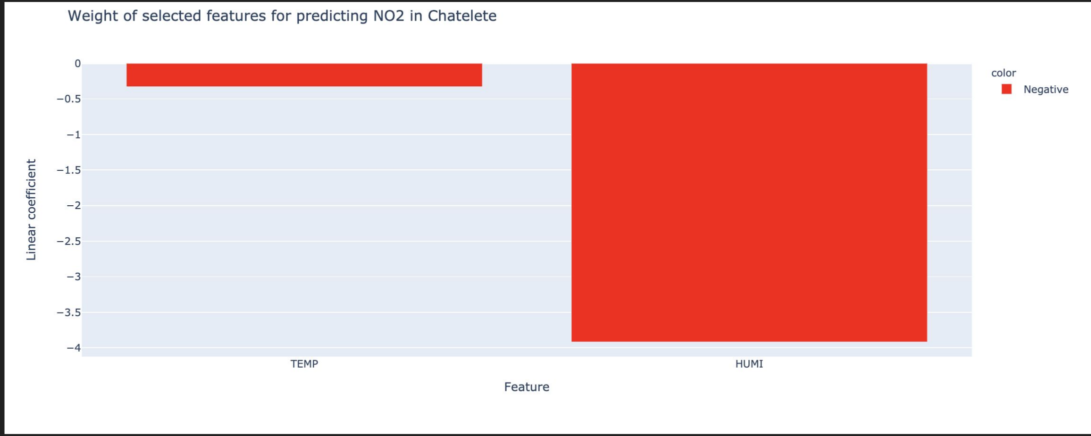
Evaluation NO2 in Chatelet



Indep:
HUMI, TEMP

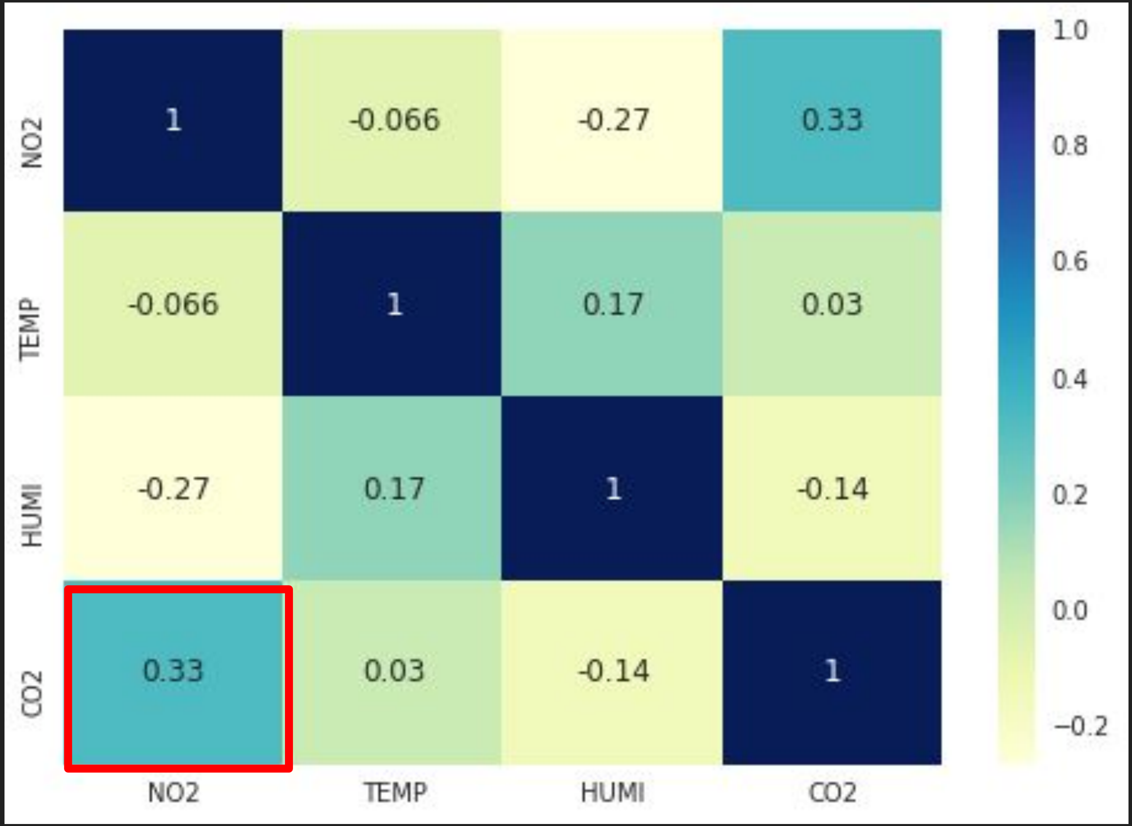
Dep: NO2

Intercept	36.17
R2 Mean Score	4.66%
RMSE	14.39

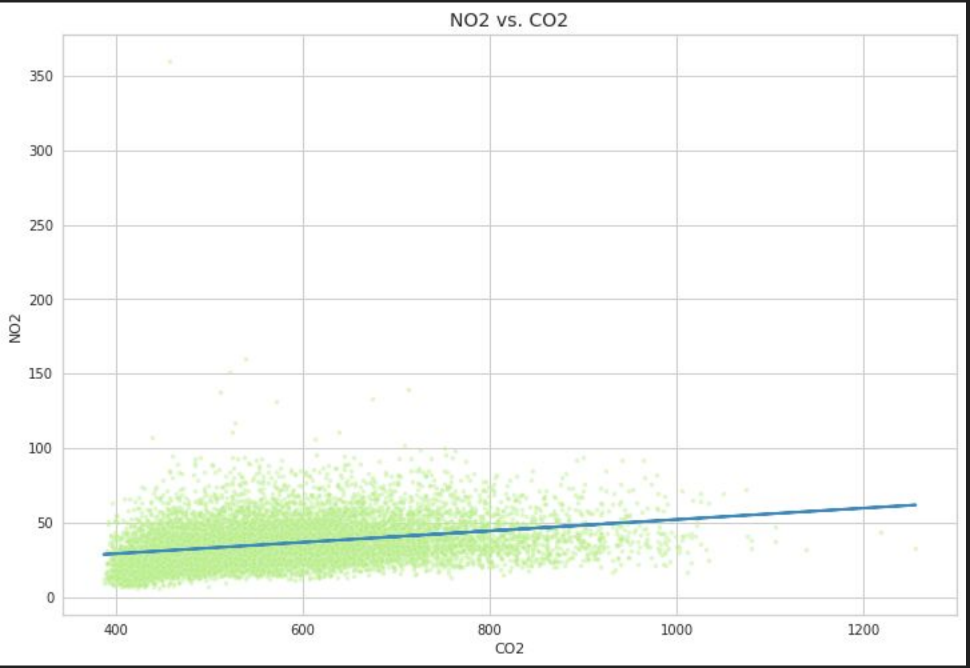


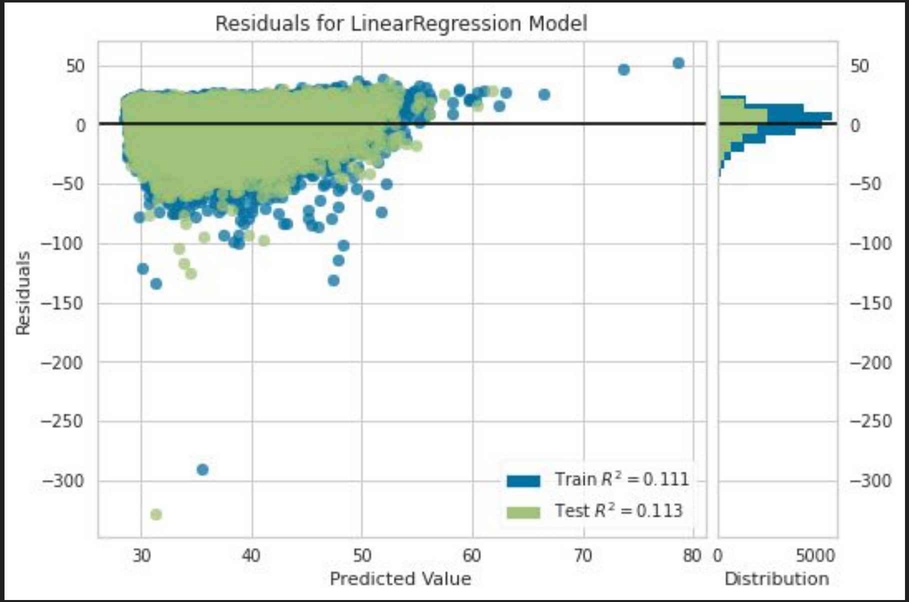
Feature	Coefficient
TEMP	-0.33
HUMI	-3.92

CO2 and NO2 correlation



CO2 and NO2 correlation





Comparison of the models predicting NO2 for Chatelet station

Model	The Explained Variance	The Mean Absolute Error	The Median Absolute Error	RMSE	R2 score	R2 mean cross-valid
Without CO2	0.07	10.81	8.85	14.39	0.07	0.04
With CO2	0.17	10.05	8.43	13.13	0.17	0.14

Better fit with the CO2 level as the the model shows higher accuracy and less errors probability.

Part 3

Model Comparison

Comparison of the models predicting CO2 and NO2 for all stations

	Chatelet		Auber		Roosevelt	
Prediction of	CO2	NO2	CO2	NO2	CO2	NO2
The Explained Variance	0.24	0.32	0.31	0.38	0.43	0.38
The Mean Absolute Error	85.72	9.16	114.31	15.75	33.2	13.2
The Median Absolute Error	70.22	7.65	108.79	12.75	70.22	7.65
RMSE	112.16	11.92	137.08	21.23	44.59	17.23
R2 score	0.24	0.32	-0.62 ?	-0.1	0.43	0.38
R2 mean cross-valid	0.2	0.3	0.3	0.37	0.4	0.35

Part 4

Based on data from two stations, can we predict air quality in the third one ?

Revisit correlations - what features to consider

Chatelet heatmap

	NO	NO2	PM10	CO2	TEMP	HUMI	hour	weekday	weekend	month	summer
NO	1.00	0.47	0.26	0.33	-0.15	-0.04	0.04	-0.14	-0.19	0.05	-0.18
NO2	0.47	1.00	0.21	0.36	-0.02	-0.25	0.19	-0.11	-0.17	-0.11	-0.17
PM10	0.26	0.21	1.00	0.40	0.18	-0.11	0.20	-0.19	-0.23	-0.06	-0.03
CO2	0.33	0.36	0.40	1.00	0.08	-0.12	0.58	-0.06	-0.09	-0.04	-0.12
TEMP	-0.15	-0.02	0.18	0.08	1.00	0.17	0.12	-0.00	-0.01	0.35	0.61
HUMI	-0.04	-0.25	-0.11	-0.12	0.17	1.00	-0.06	0.01	0.01	0.21	0.21
hour	0.04	0.19	0.20	0.58	0.12	-0.06	1.00	-0.00	-0.00	-0.00	0.00
weekday	-0.14	-0.11	-0.19	-0.06	-0.00	0.01	-0.00	1.00	0.79	0.00	0.00
weekend	-0.19	-0.17	-0.23	-0.09	-0.01	0.01	-0.00	0.79	1.00	0.00	0.00
month	0.05	-0.11	-0.06	-0.04	0.35	0.21	-0.00	0.00	0.00	1.00	0.10
summer	-0.18	-0.17	-0.03	-0.12	0.61	0.21	0.00	0.00	0.00	0.10	1.00

Auber heatmap

	NO	NO2	PM10	PM2.5	CO2	TEMP	HUMI	hour	weekday	weekend	month	summer
NO	1.00	0.50	0.15	0.17	0.27	-0.11	0.02	0.03	-0.10	-0.15	0.06	-0.10
NO2	0.50	1.00	0.03	0.06	0.18	0.10	-0.17	0.15	-0.01	-0.04	-0.05	0.06
PM10	0.15	0.03	1.00	0.91	0.49	0.12	0.13	0.24	-0.11	-0.15	-0.07	-0.00
PM2.5	0.17	0.06	0.91	1.00	0.47	0.13	0.12	0.24	-0.10	-0.14	-0.06	-0.00
CO2	0.27	0.18	0.49	0.47	1.00	0.02	0.03	0.51	-0.19	-0.26	-0.04	-0.04
TEMP	-0.11	0.10	0.12	0.13	0.02	1.00	0.22	0.09	-0.01	-0.03	0.33	0.65
HUMI	0.02	-0.17	0.13	0.12	0.03	0.22	1.00	-0.04	-0.03	-0.03	0.11	0.16
hour	0.03	0.15	0.24	0.24	0.51	0.09	-0.04	1.00	-0.00	-0.00	-0.00	0.00
weekday	-0.10	-0.01	-0.11	-0.10	-0.19	-0.01	-0.03	-0.00	1.00	0.79	0.00	0.00
weekend	-0.15	-0.04	-0.15	-0.14	-0.26	-0.03	-0.03	-0.00	0.79	1.00	0.00	0.00
month	0.06	-0.05	-0.07	-0.06	-0.04	0.33	0.11	-0.00	0.00	0.00	1.00	0.10
summer	-0.10	0.06	-0.00	-0.00	-0.04	0.65	0.16	0.00	0.00	0.00	0.10	1.00

F.R. heatmap

	NO	NO2	PM10	CO2	TEMP	HUMI	hour	weekday	weekend	month	summer
NO	1.00	0.57	0.25	0.42	-0.28	0.18	0.07	-0.09	-0.14	0.09	-0.22
NO2	0.57	1.00	0.15	0.37	-0.08	-0.09	0.24	-0.06	-0.11	-0.08	-0.10
PM10	0.25	0.15	1.00	0.55	0.12	-0.04	0.17	-0.16	-0.21	0.07	0.04
CO2	0.42	0.37	0.55	1.00	-0.06	-0.04	0.42	-0.10	-0.16	0.01	-0.11
TEMP	-0.28	-0.08	0.12	-0.06	1.00	-0.21	0.05	0.01	-0.01	0.29	0.65
HUMI	0.18	-0.09	-0.04	-0.04	-0.21	1.00	-0.07	0.00	-0.00	0.10	-0.09
hour	0.07	0.24	0.17	0.42	0.05	-0.07	1.00	-0.00	-0.00	-0.00	0.00
weekday	-0.09	-0.06	-0.16	-0.10	0.01	0.00	-0.00	1.00	0.79	0.00	0.00
weekend	-0.14	-0.11	-0.21	-0.16	-0.01	-0.00	-0.00	0.79	1.00	0.00	0.00
month	0.09	-0.08	0.07	0.01	0.29	0.10	-0.00	0.00	0.00	1.00	0.10
summer	-0.22	-0.10	0.04	-0.11	0.65	-0.09	0.00	0.00	0.00	0.10	1.00

The new dataframe

[feat_1_station_1, …, feat_n_station_1, feat_1_station_2, …., feat_n_station_2, dependent_variable_1_station_3]

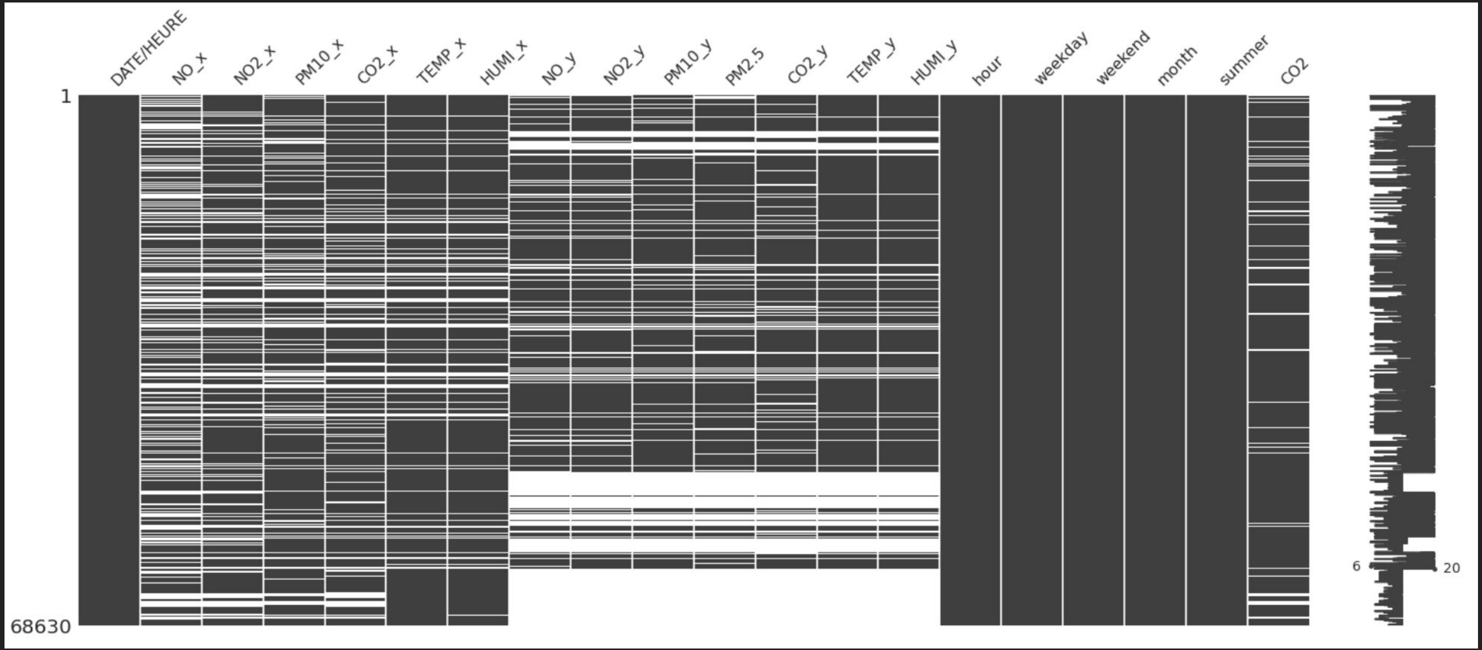
We inner join df_chatelet, df_auber and df_roosevelt.CO2 on the DATE/HEURE colum =>

- NO
- NO2
- PM10
- TEMP
- HUMI
- Hour 0:23
- Weekday 0:6
- Month 1:12
- Weekend 0/1
- Summer 0/1

one set for each station (Chatelet and Auber)

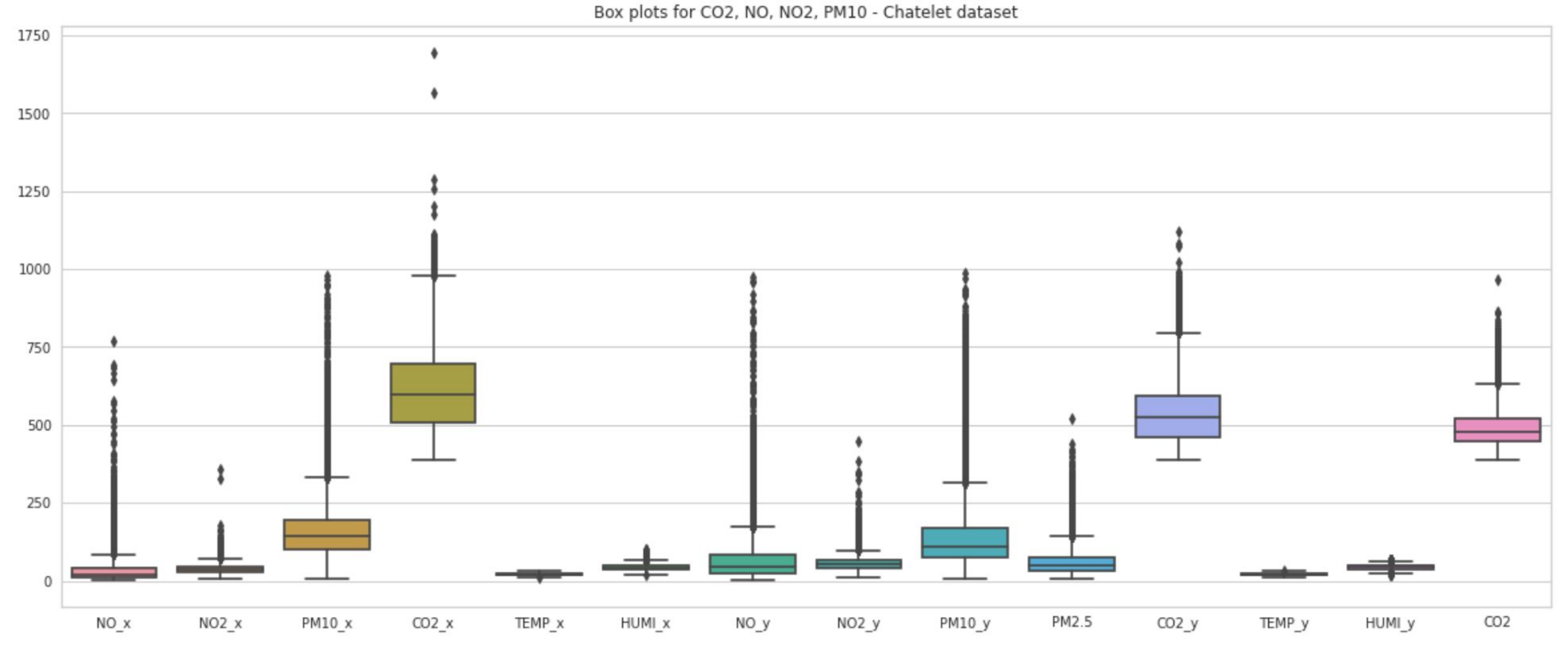
Wrangling and feature engineering

- We remove all rows with at least one missing value.
- From ~68.000 entries, we are left with ~20.000



Outliers

- Remember we saw outliers in previous slides
- Let’s see what remained after we removed ⅔ of our data (na removal)



MLR model coeff

Independent
NO, NO2, PM10, TEMP, HUMI

Hour, Weekday, Month, Weekend, Summer

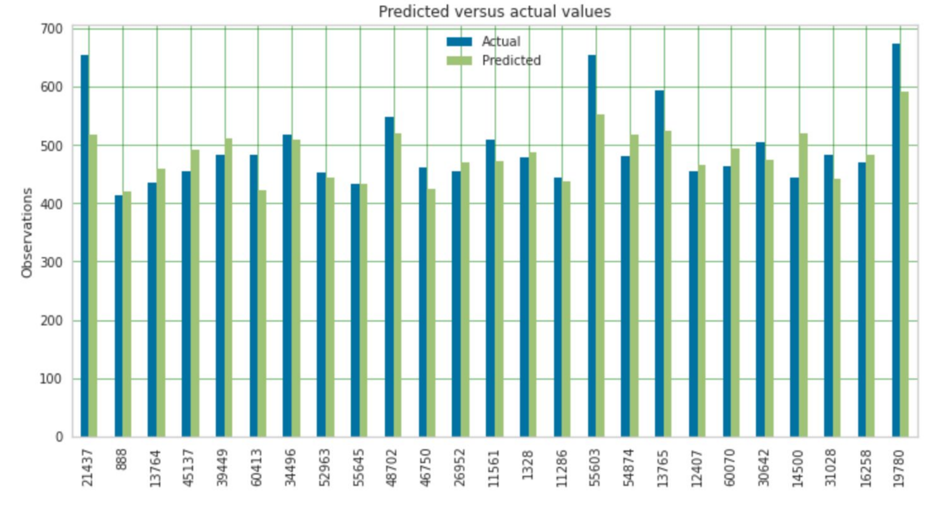
ALL standardized

Chatelet + Auber

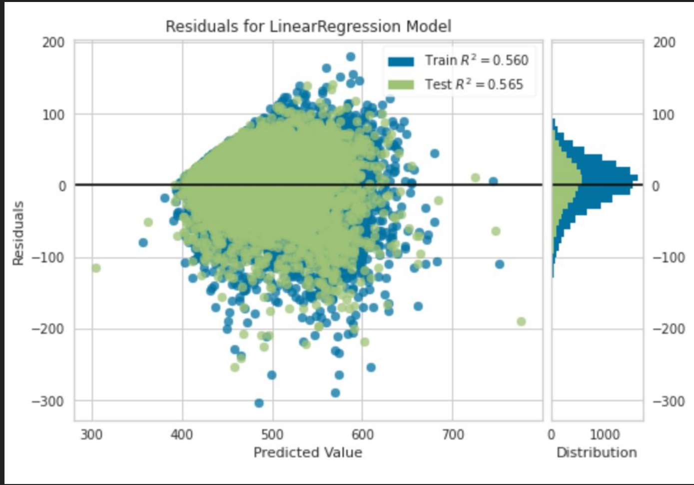
Dep: CO2 F.R. station

https://stats.stackexchange.com/questions/463690/multiple-regression-with-mixed-continuous-categorical-variables-dummy-coding-s

MLR model - evaluation




```
The Explained Variance: 0.56
The Mean Absolute Error: 30.33
The Median Absolute Error: 23.46
Mean squared error: 1652.41
Root mean squared error: 40.65
R2 Score: 0.56
Intercept: 490.64
```



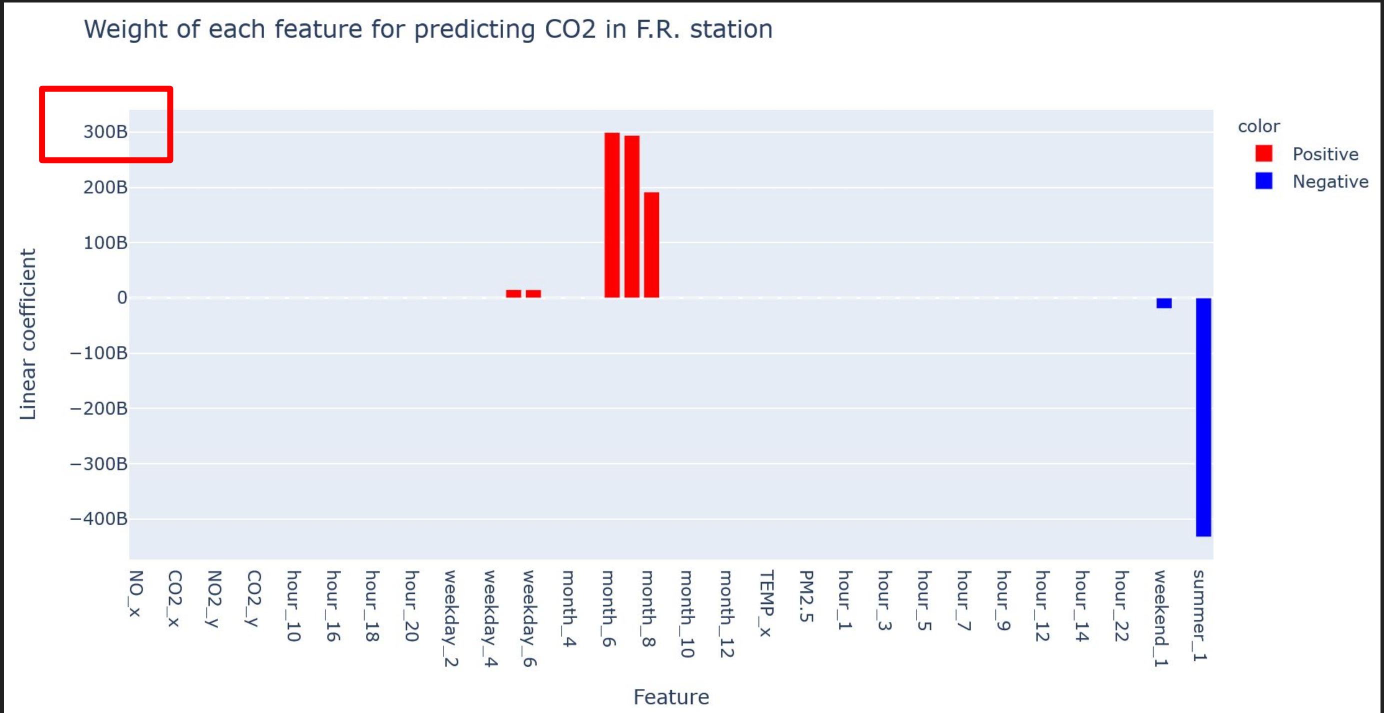
The Effect of regularization

New dataframe: one hot encoding

We inner join df_chatelet, df_auber and df_roosevelt.CO2 on the DATE/HEURE colum =>

○	NO	
○	NO2	
○	PM10	
○	TEMP	one set for each station (Chatelet and Auber)
○	HUMI	
○	Hour	0:23 ---|
○	Weekday	0:6 ---|---> categorical variables or not ? → one hot encoding
○	Month	1:12 ---|
○	Weekend	0/1
○	Summer	0/1

New dataframe: one hot -> exploding coefficients



NO
NO2
PM10
TEMP
HUMI
Hour 0:23 ---|
Weekday 0:6 ---|-one-hot
Month 1:12 ---|
Weekend 0/1
Summer 0/1

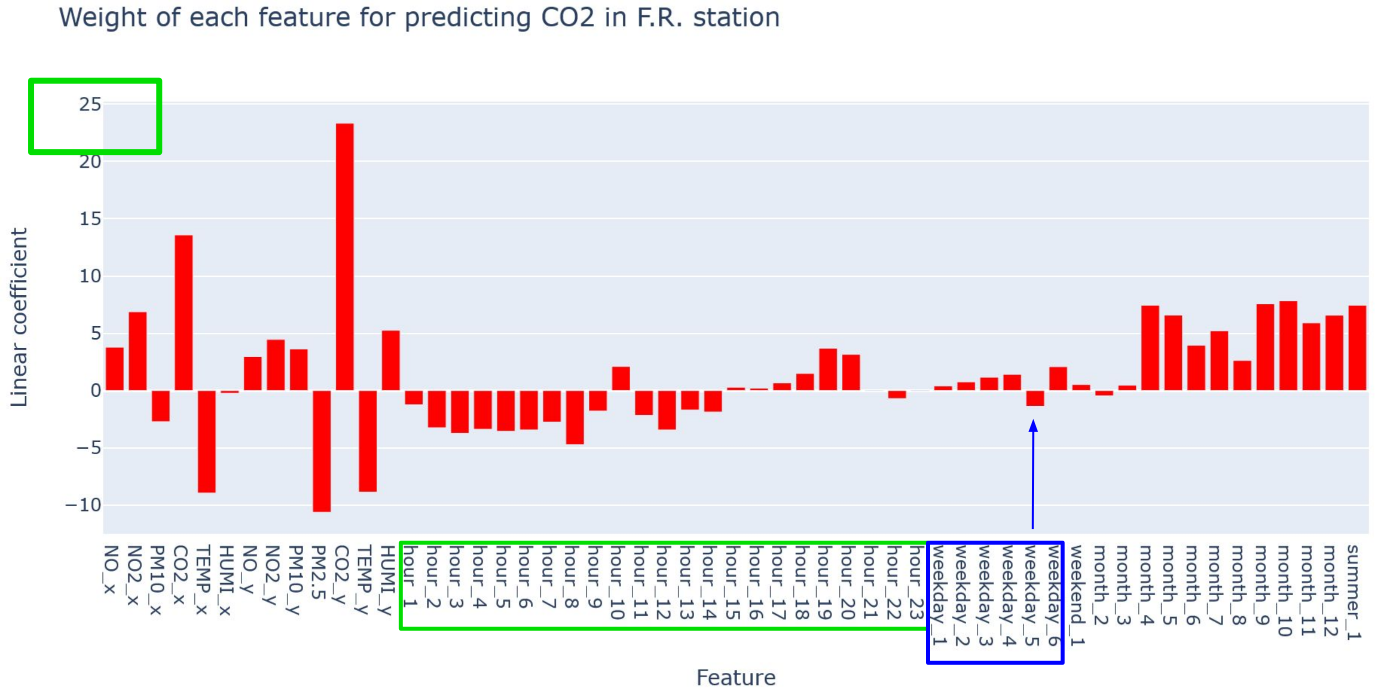
ALL standardized

Even the one hot encoded variables

Conclusion: must prevent large coefficients

https://stats.stackexchange.com/questions/463690/multiple-regression-with-mixed-continuous-categorical-variables-dummy-coding-s <- when to standardize dummy variables

Ridge regularization prevents large coefficients

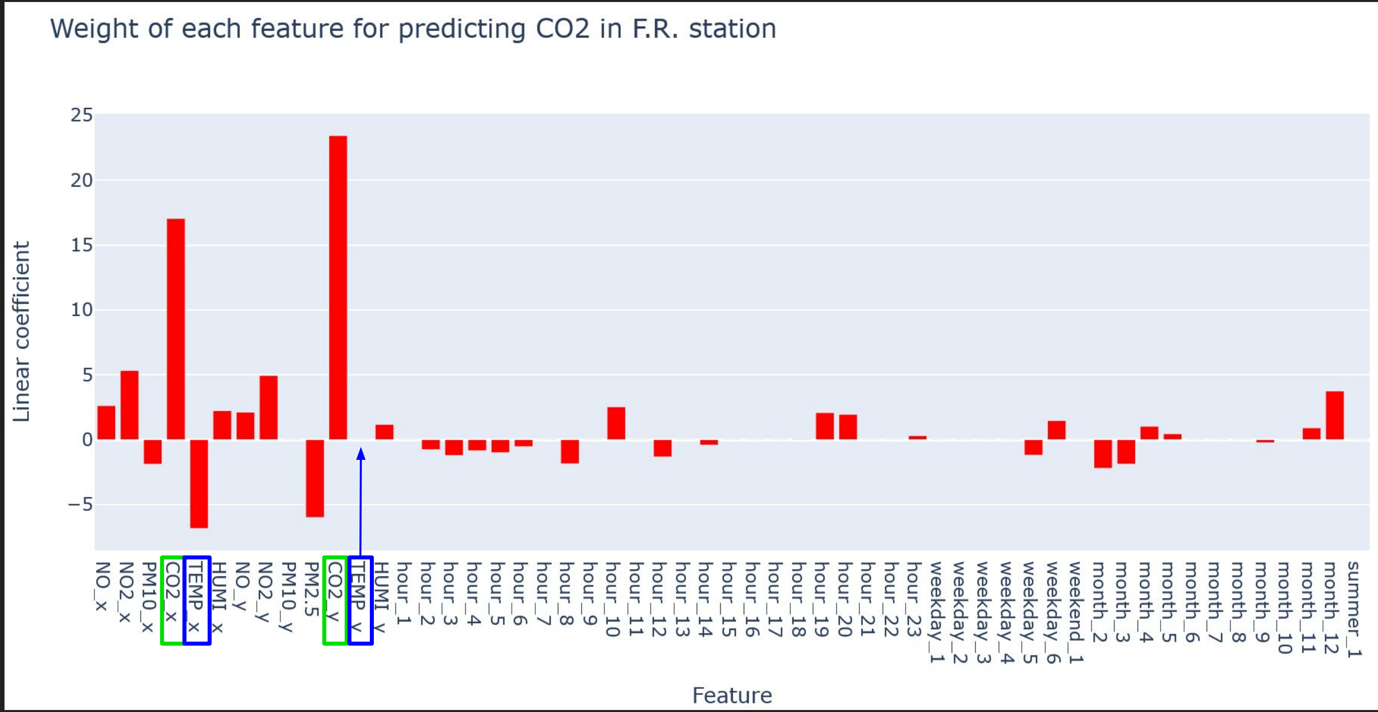


Observations:

- We don't find the same relationship between increasing hour and coeff size / sign
- Difficult to interpret
- We try Lasso next !

```
The Explained Variance: 0.57
The Mean Absolute Error: 29.81
The Median Absolute Error: 23.62
Mean squared error: 1579.47
Root mean squared error: 39.74
```

Lasso regularization -> to reduce # of features



alpha = 0.8 (where alpha = 0 is equivalent to ordinary least square)

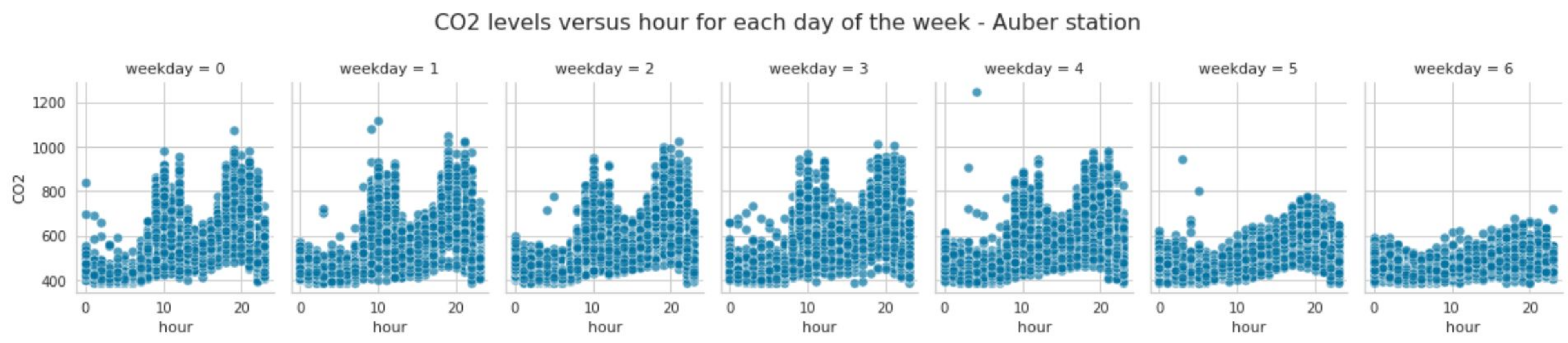
Observations:

- CO2 in Chatelet and Auber most important features when using Lasso regularization
- We sacrificed performance ?

```
The Explained Variance: 0.56
The Mean Absolute Error: 30.24
The Median Absolute Error: 23.96
Mean squared error: 1635.66
Root mean squared error: 40.44
R2 Score: 0.56
Intercept: 490.622075
```

Non - linear relationship - CO2 versus hour per day

- We see here why hour of day (0:23) would not be a good candidate to predict CO2 on through a linear model



Models comparison

Model comparison: same features,different approach

Independent: NO, NO2, PM10, TEMP, HUMI, hour, weekday, month, weekend, summer in Chatelet & Auber

Dependent: CO2 in F. Roosevelt

Model 1: all features numeric, scaling, no regularization

Model 2 : all features, one hot encoding + scaling + Ridge regularization

Model 3: all features, one hot encoding + scaling + Lasso regularization

Question: differences in outcome ?

	Model 1	Model 2	Model 3
R2 score	.56	.57	.56
RMSE	40.65	39.74	40.44

Model comparison: 5-cv

Independent: NO, NO2, PM10, TEMP, HUMI, hour, weekday, month, weekend, summer in Chatelet & Auber

Dependent: CO_2 in F. Roosevelt

Model 1: all features, one hot encoding + scaling no regularization

Model 2 : all features, one hot encoding + scaling + Ridge regularization

Model 3: all features, one hot encoding + scaling + Lasso regularization

Question: differences in outcome ?

```
R2 mean score simple lr: 0.580
R2 mean score ridge: 0.580
R2 mean score lasso: 0.565
```

Next

RFE to choose features

Data engineering:

- Hour and month as sine (keep cyclical nature)
- Or encode hour as rush hour / not rush
- Imputation for missing values