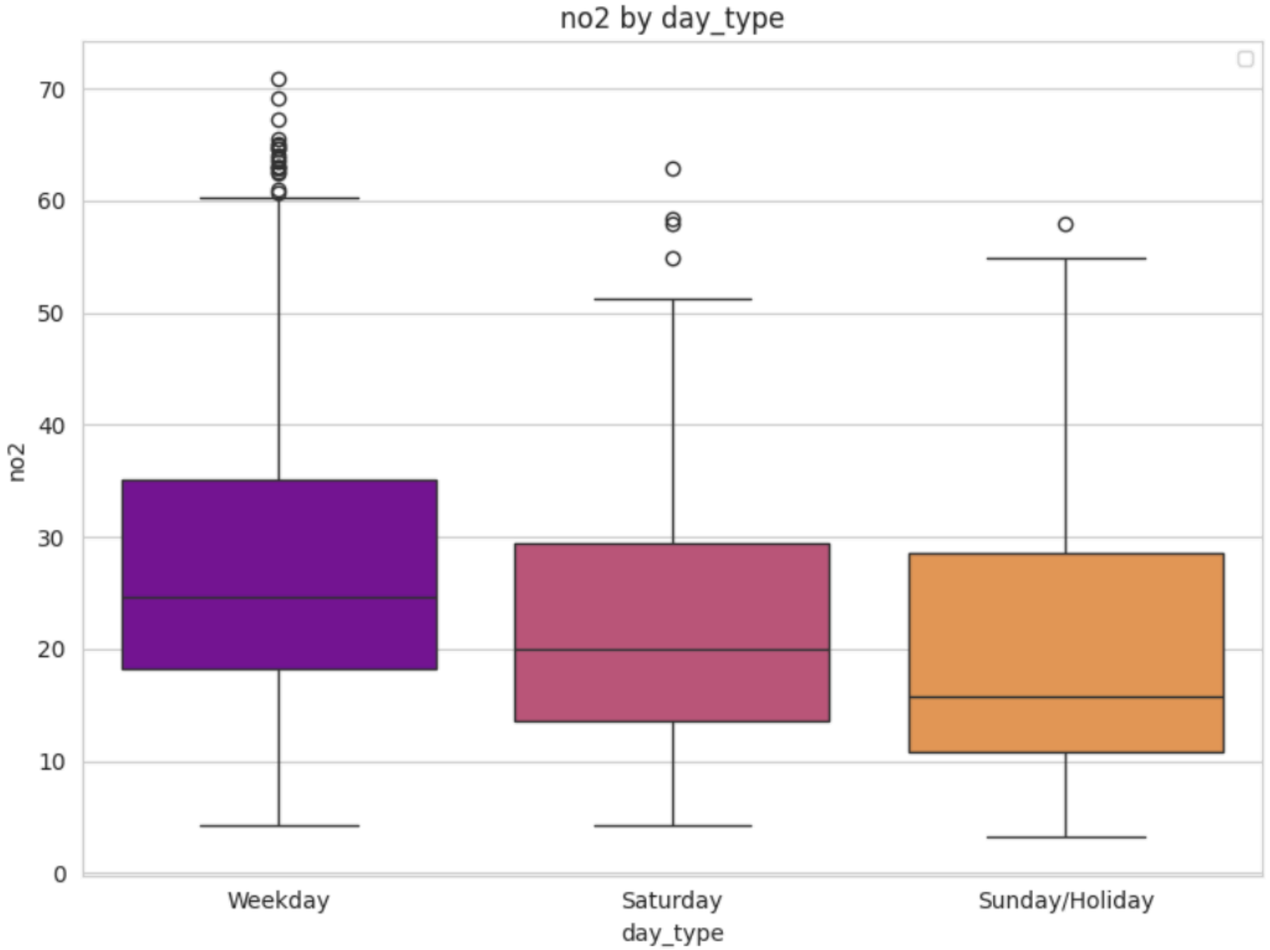
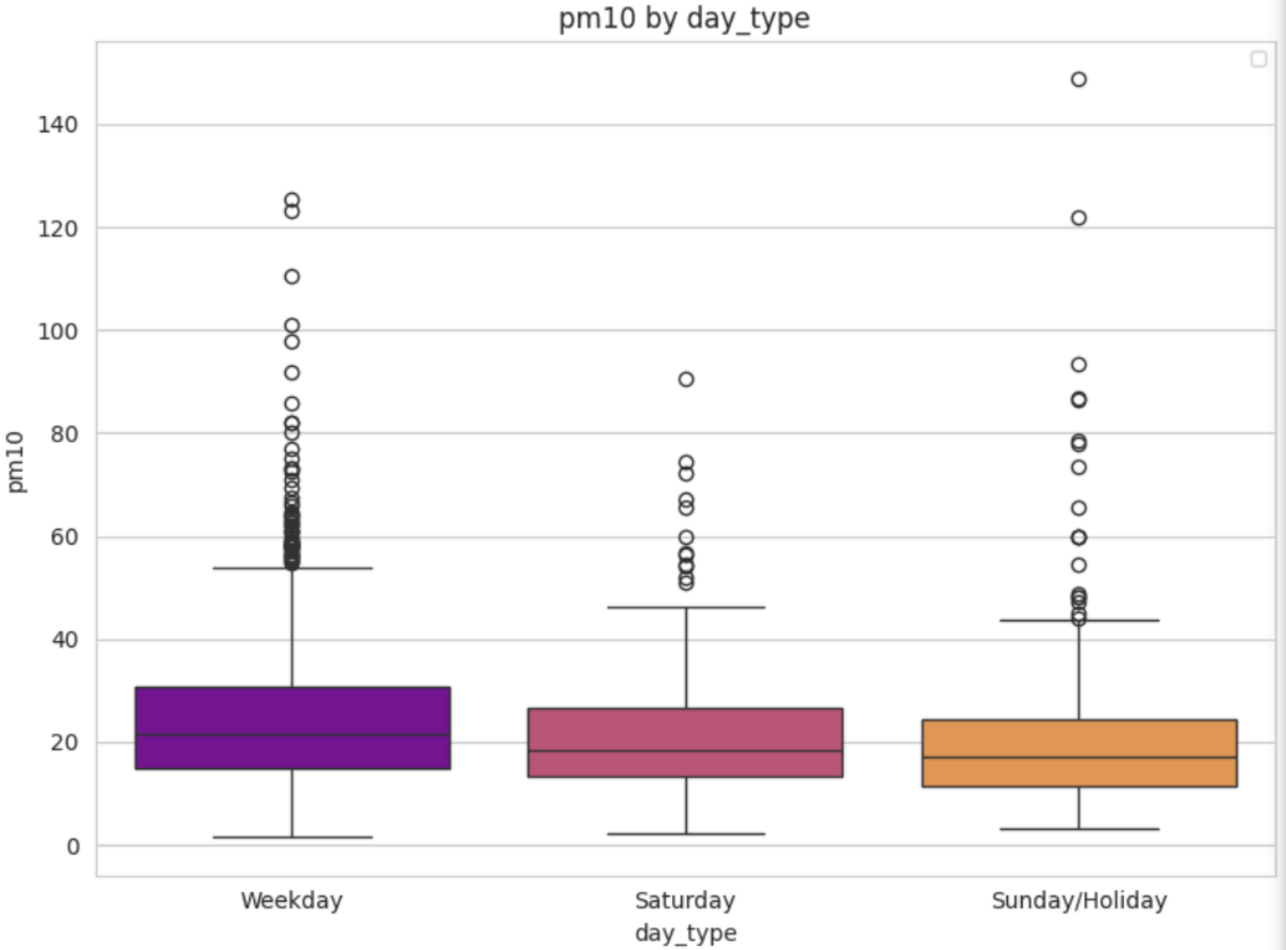
Assumption

Case: Graz-West
(more industrial location)

no2 pollutant -> more influenced by human activity than pm10

Prove by R-squared score





pp created by Egor Chumakov, presented by Ghimire Ayusch, Lamsak Ayusch, Sapkota Arogya

NO2 - RANDOM FOREST

{'max_depth': 9, 'min_samples_leaf': 2}
optimal params by GridSearchCV algo

non-human feature importance

temperature: 0.39
windspeed: 0.09
precipitation: 0.01
humidity: 0.02
inversion temperature: 0.50

temperature, windspeed, precipitation, inversion temperature

R-squared for train set: 0.84
R-squared for test set: 0.72

feature selection

peak velocity removed due to p-value of 0.504 - identified with Linear Regression Diagnostics

temperature, windspeed, precipitation, inversion temperature + day type

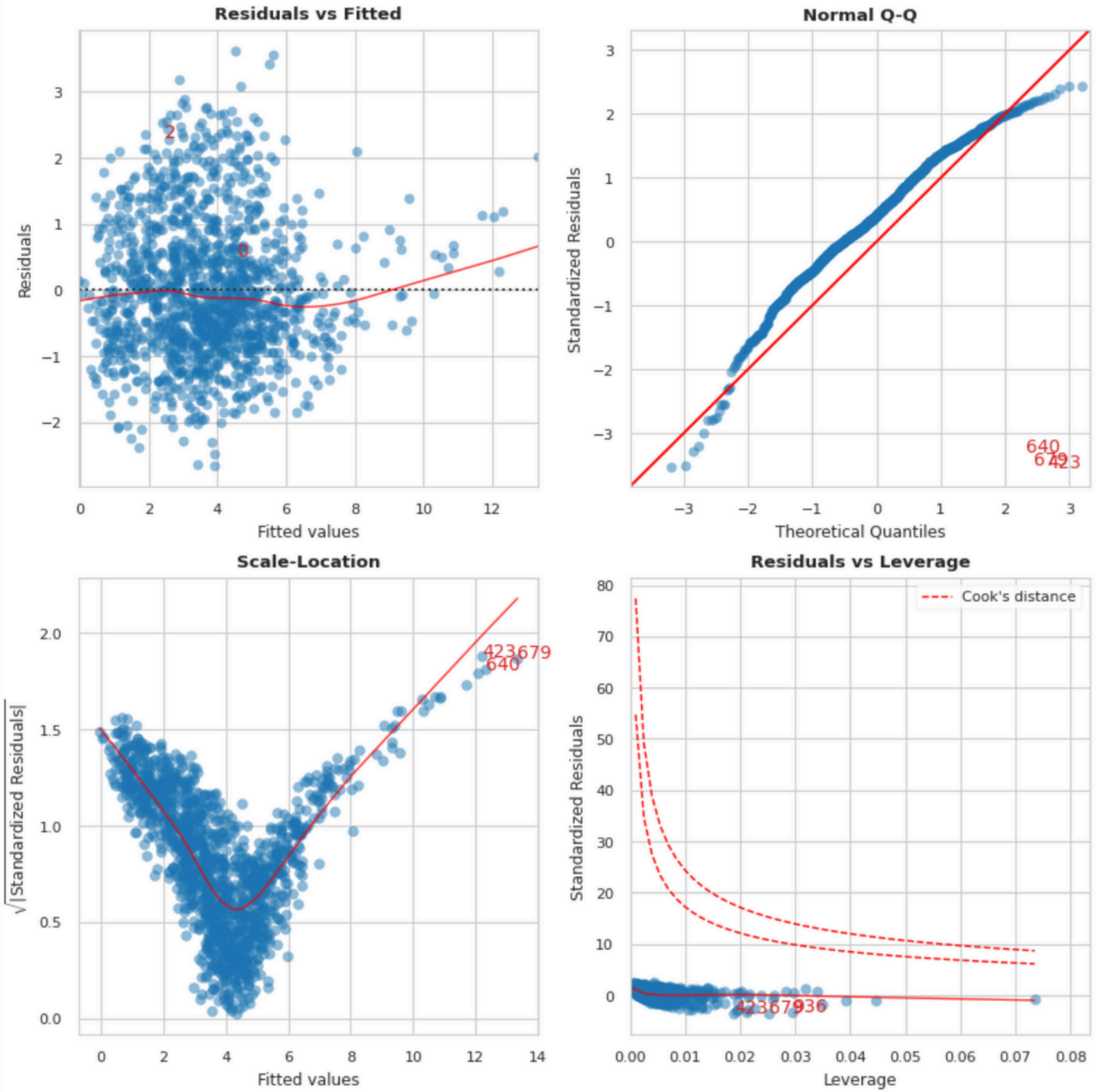
0.94 :R-squared for train set
0.82 :R-squared for test set

Evaluation of Linear Model created with train set:

- y (response) no2
- X - predictors

All VIF values are well below 5 → predictors are not highly correlated → all features can be included in regression analysis

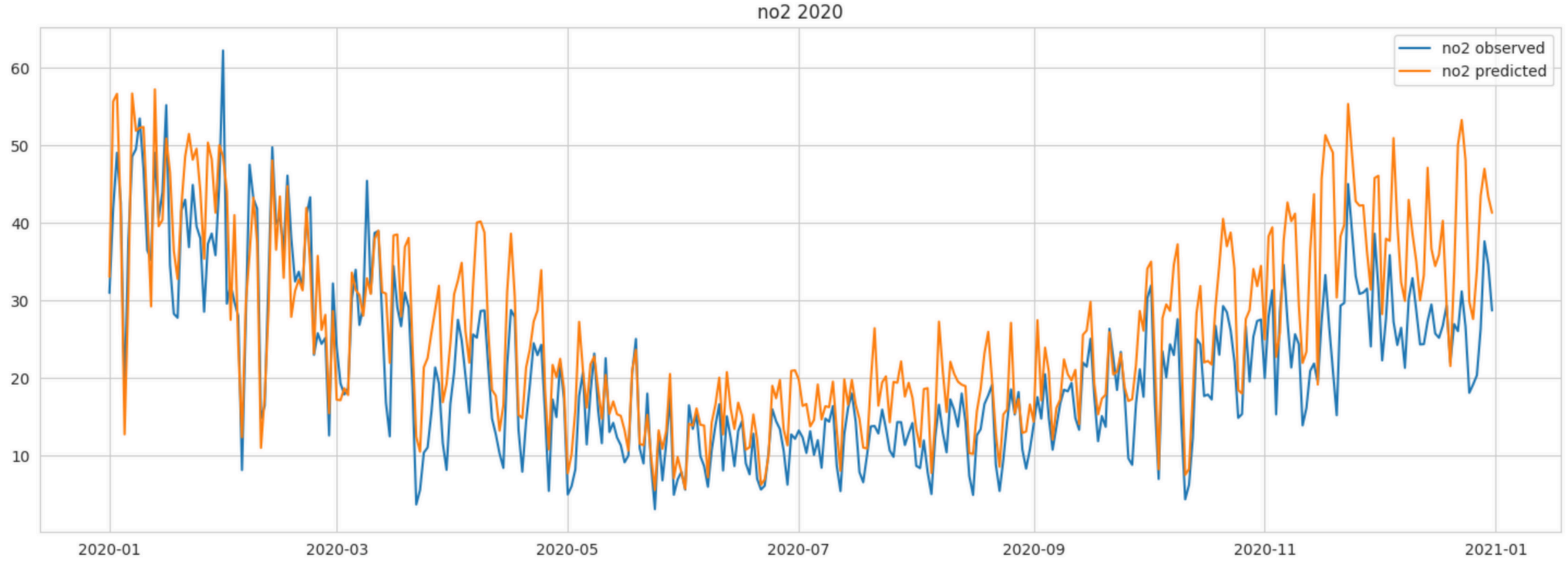
	Features	VIF Factor
4	day_type_Saturday	1.15
2	prec	1.18
5	day_type_Sunday/Holiday	1.19
3	inversion_temp	2.12
0	temp	2.56
1	windspeed	3.20



no2 prediction

Performed with RandomForest tree

Predicted variance is generally higher after first COVID-19 lockdown



PM10 - RANDOM FOREST

{'max_depth': 9, 'min_samples_leaf': 5}
optimal params by GridSearchCV algo

non-human feature importance

temperature: 0.20
windspeed: 0.08
precipitation: 0.06
peak velocity : 0.42
inversion temp : 0.25

temperature, windspeed, precipitation, inversion temperature

R-squared for train set: 0.76
R-squared for test set: 0.59

feature selection

humidity was removed because p-value of 0.32 - identified with Linear Regression Diagnostics

temperature, windspeed, precipitation, inversion temperature + day type

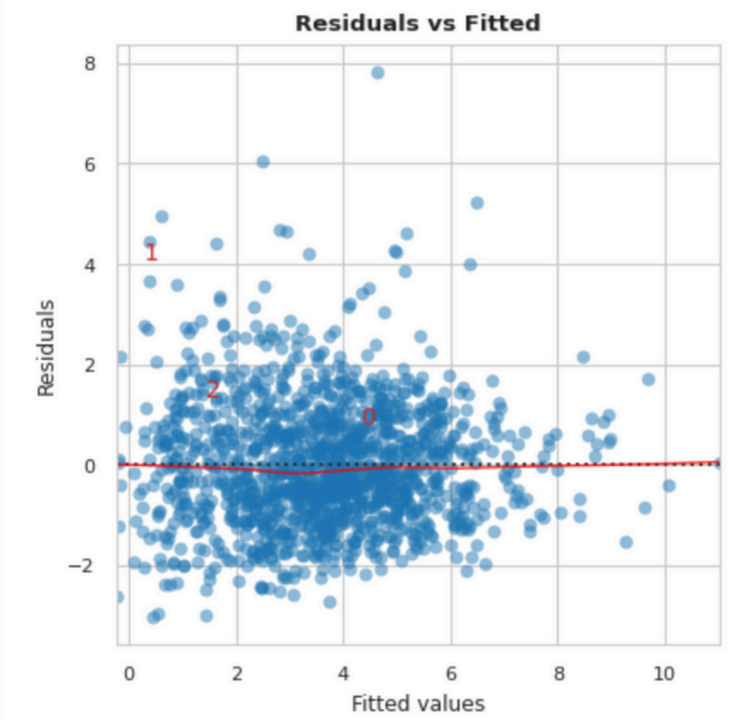
0.77 :R-squared for train set
0.60 :R-squared for test set

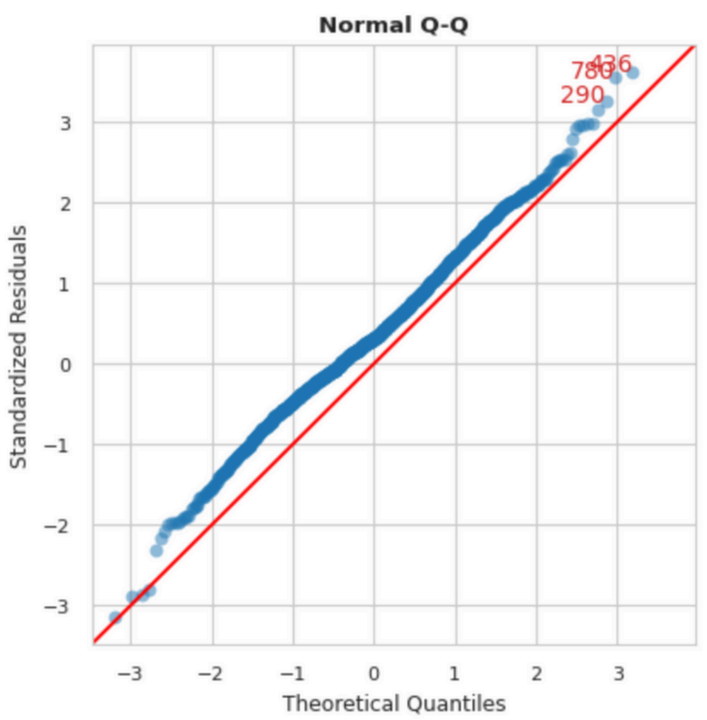
Evaluation of Linear Model created with train set:

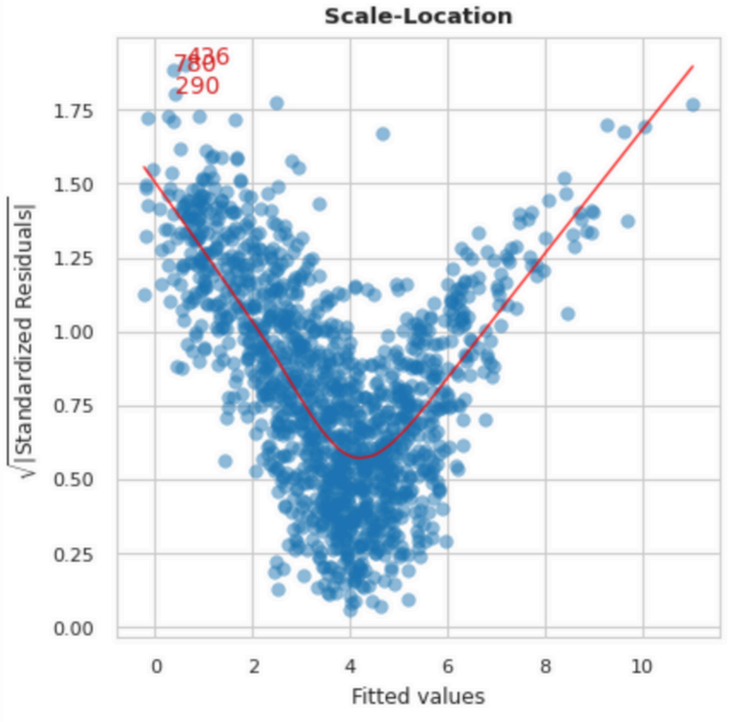
- y (response) pm10
- X - predictors

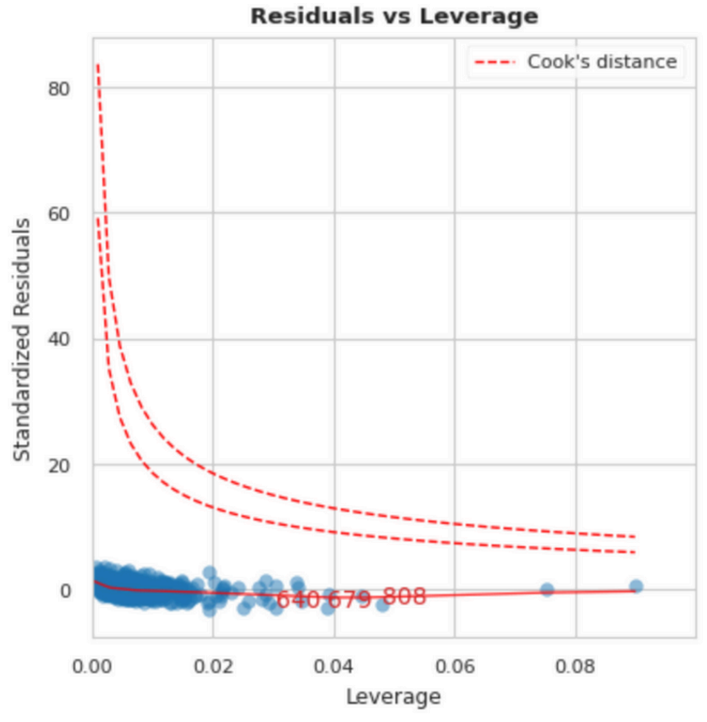
Some VIF values are higher than 5 → windspeed and peak_velocity predictors are highly correlated → better remove either of them

	Features	VIF Factor
5	day_type_Saturday	1.15
6	day_type_Sunday/Holiday	1.19
2	prec	1.26
4	inversion_temp	2.14
0	temp	2.84
1	windspeed	9.71
3	peak_velocity	10.27





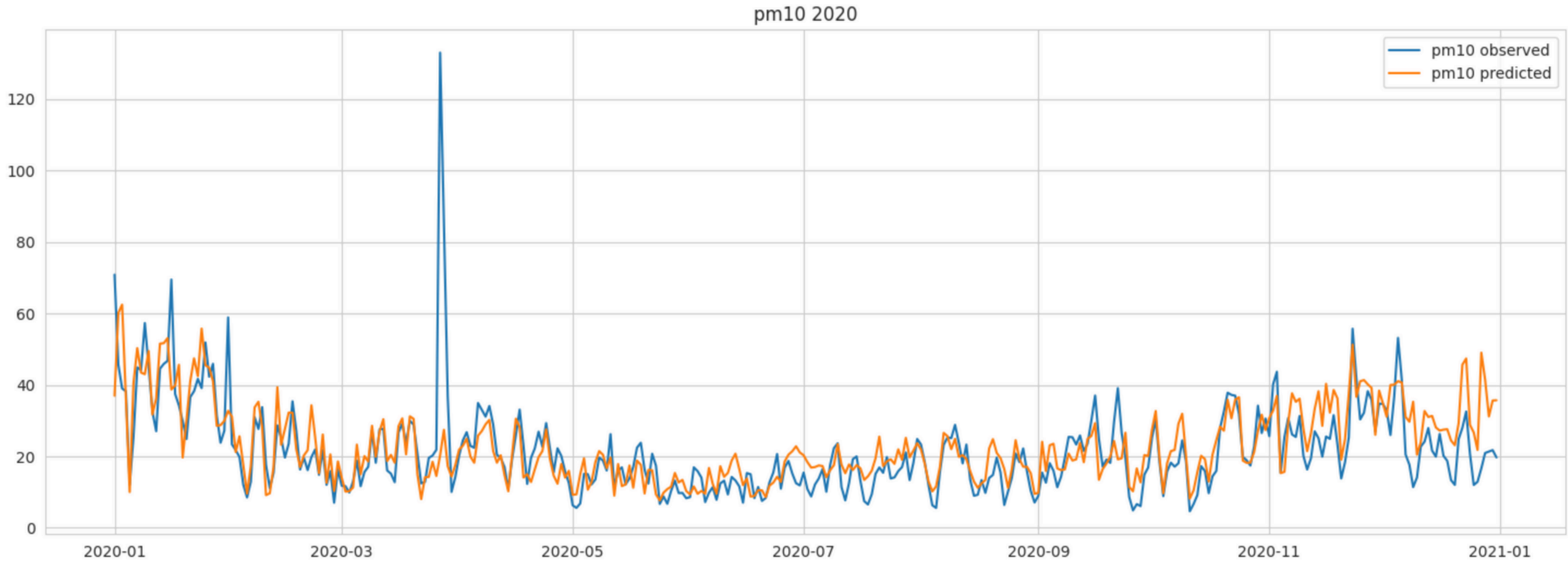




pm10 prediction

Performed with RandomForest tree

Predicted variance is generally higher after first COVID-19 lockdown



Assumption proved

When excluding day_type predictor from the model:

- no2 response R-squared score decreased in train set from 0.94 to 0.84 (similarly in test set from 0.82 to 0.72)
- pm10 response R-squared score have not shown significant changes to be considered in train set from 0.77 to 0.76 (similarly in test set from 0.60 to 0.59)

Conclusion:

- human factor (predictor day_type) included in the model improves the ability of model to explain the variance of data concerning no2 response
- no2 pollutant, opposed to pm10, is directly produced by human activity in manufactory, thus result is logically valid → less human activity, less no2
- supported by plots in assumption (difference in mean, median variance per day type