Clean Air for African cities

by AirQo - Breath Clean

airqo

Breathe Clean

A initiative by Makerere University, Kampala, Uganda

WE BUILD FOR THE FUTURE

https://gisgeography.com/uganda-map/

South Sudan
Kenya
Northern
Democratic Republic of the Congo
Lake Albert
Lake Kwania
Lake Kyoga
Lake Bisina
Eastern
Central
Western
Kampala
Kenya
Lake Wamala
Lake George
Lake Edward
Lake Victoria
Tanzania
Rwanda
Tanzania
2021 Copyright © Ontheworldmap.com

https://ontheworldmap.com/uganda/city/kampala/

In cooperation with University of Birmingham

UNIVERSITY OF BIRMINGHAM

“9 out of 10 people breathe polluted air”.

The Mission:

Empower communities with accurate, hyperlocal and timely air quality data to drive air pollution mitigation actions.

The team

Dennis Eickhorn

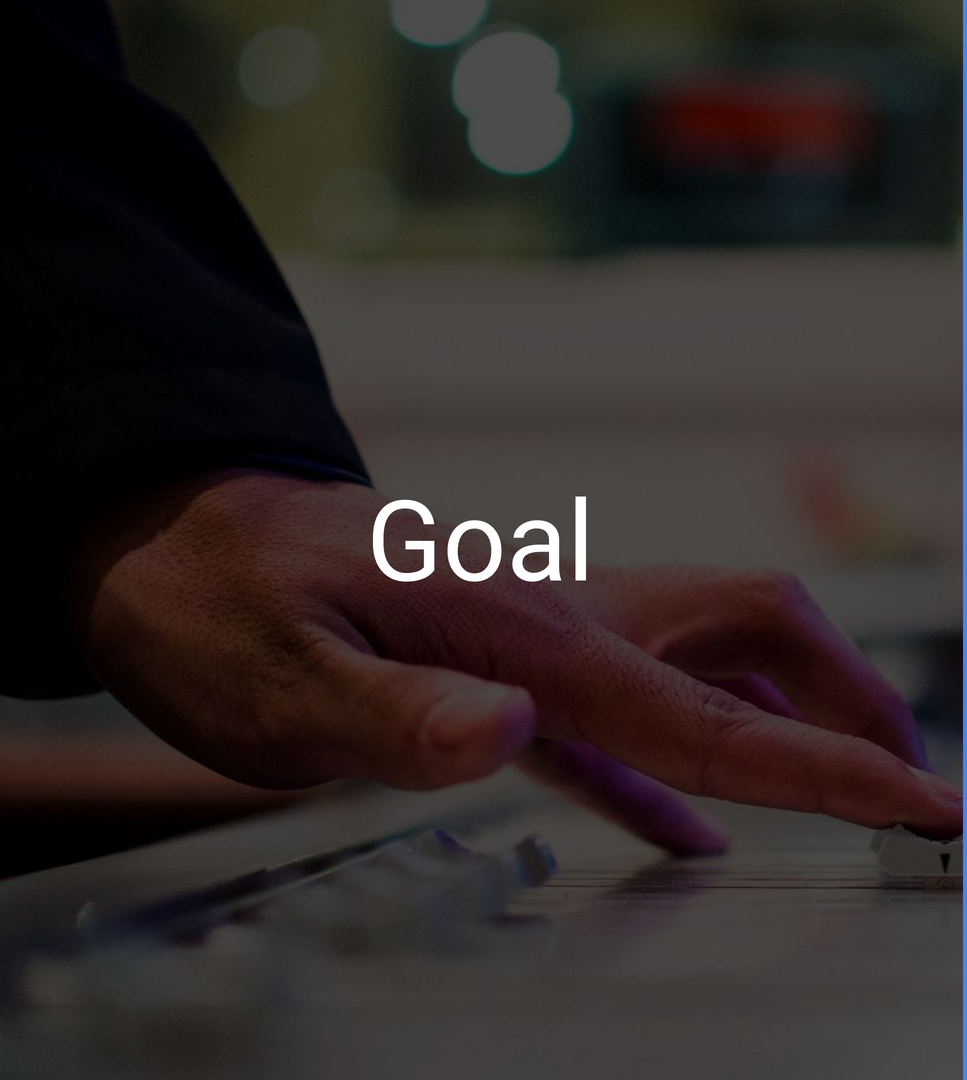
Mathematician

Jeremy Winandy

Historian

Maximilian Kinzler

Philologist



Goal

To accurately forecast air quality for each hour across five locations in Kampala, Uganda.

Forecast is based on observations of the past 5 days of hourly air quality measurements at each site.

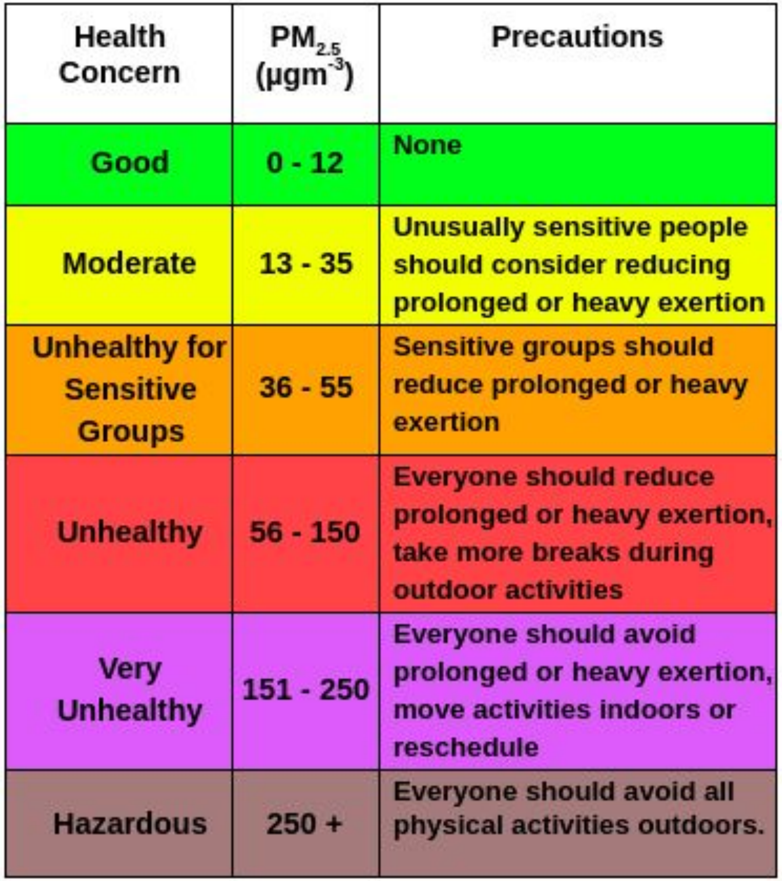
Prediction is done for exactly 24 hours later.

The Target

The target is the mean mass of PM2.5 per cubic metre of air (µg/m3).

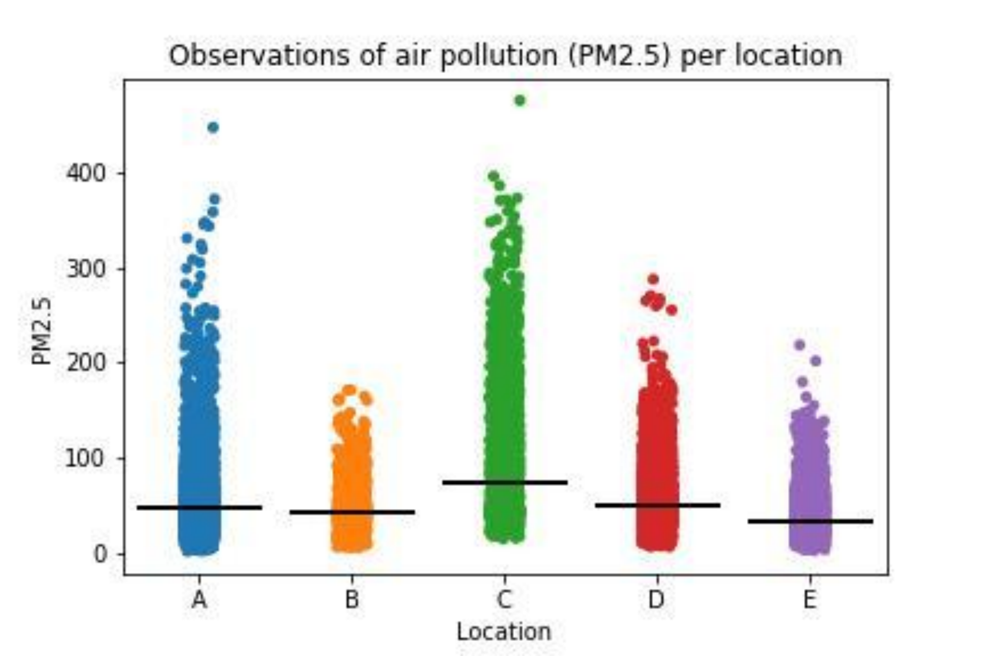
PM2.5 is the particulate matter smaller than 2.5 µm.

Target values are taken exactly 24 hours after the last weather indicators’ reading.



Health Concern	$PM_{2.5}$ (μgm^{-3})	Precautions
Good	0 - 12	None
Moderate	13 - 35	Unusually sensitive people should consider reducing prolonged or heavy exertion
Unhealthy for Sensitive Groups	36 - 55	Sensitive groups should reduce prolonged or heavy exertion
Unhealthy	56 - 150	Everyone should reduce prolonged or heavy exertion, take more breaks during outdoor activities
Very Unhealthy	151 - 250	Everyone should avoid prolonged or heavy exertion, move activities indoors or reschedule
Hazardous	250 +	Everyone should avoid all physical activities outdoors.

Data Distribution – Air quality (PM2.5)

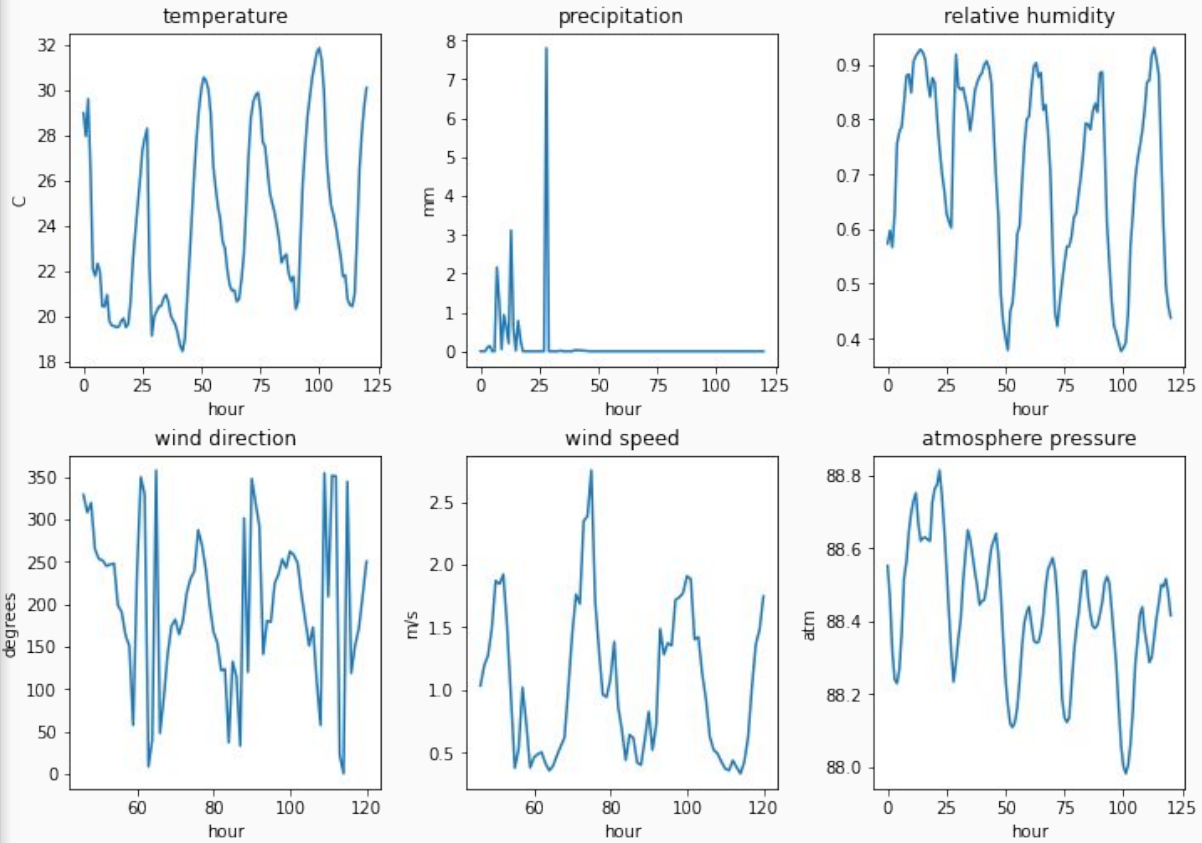


The Data

15000 sets of 5-day series of hourly weather data readings which include

- Temperature
- Precipitation
- Humidity
- Wind direction
- Wind speed
- Atmospheric pressure

One example observation

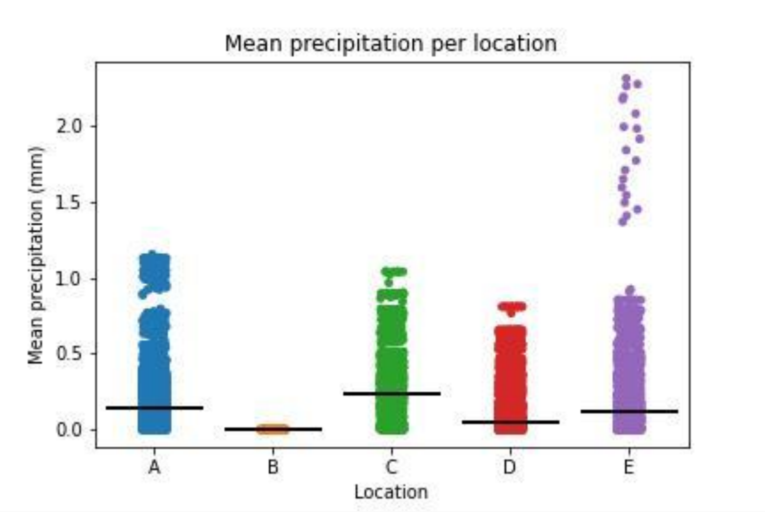


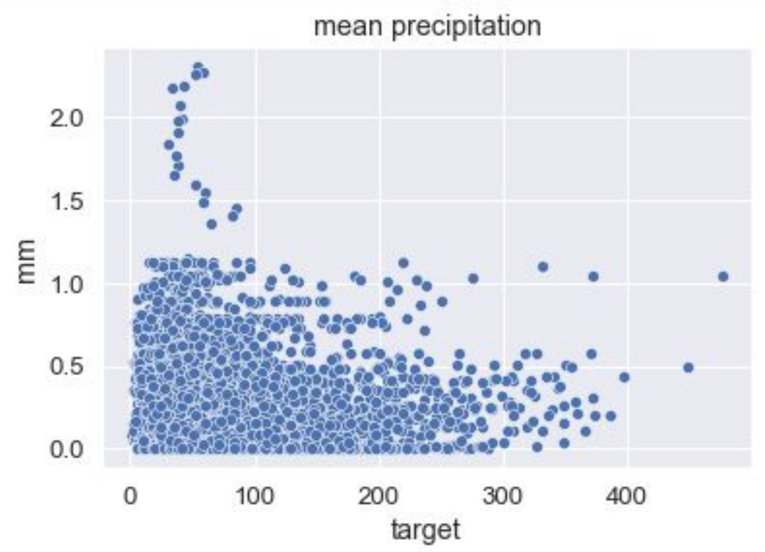
Data Distribution – Average values

Example: mean of precipitation

Extreme values of location E

→ outliers in overall mean



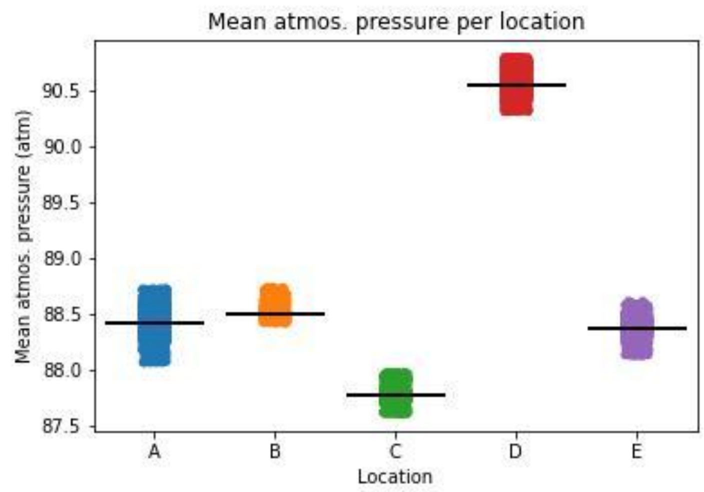


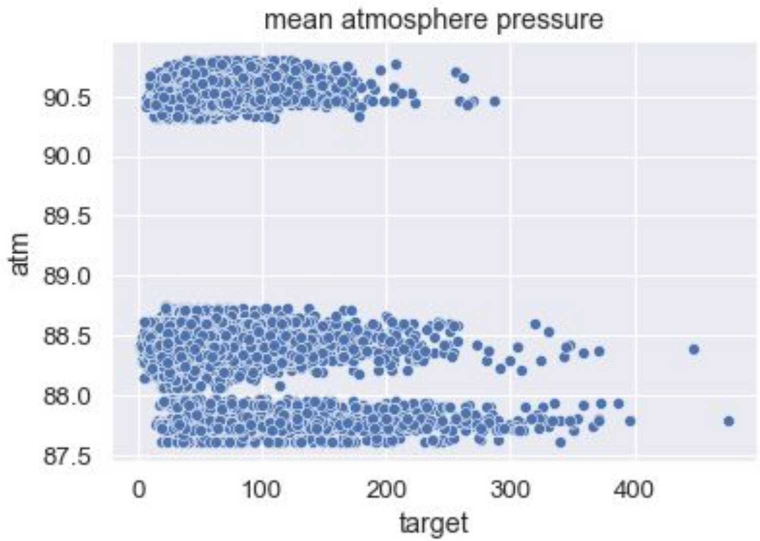
Data Distribution – Average values

Example: mean of atmospheric pressure

Values of location D very (and location C a little bit) different from rest

→ two (to three) clusters in overall mean

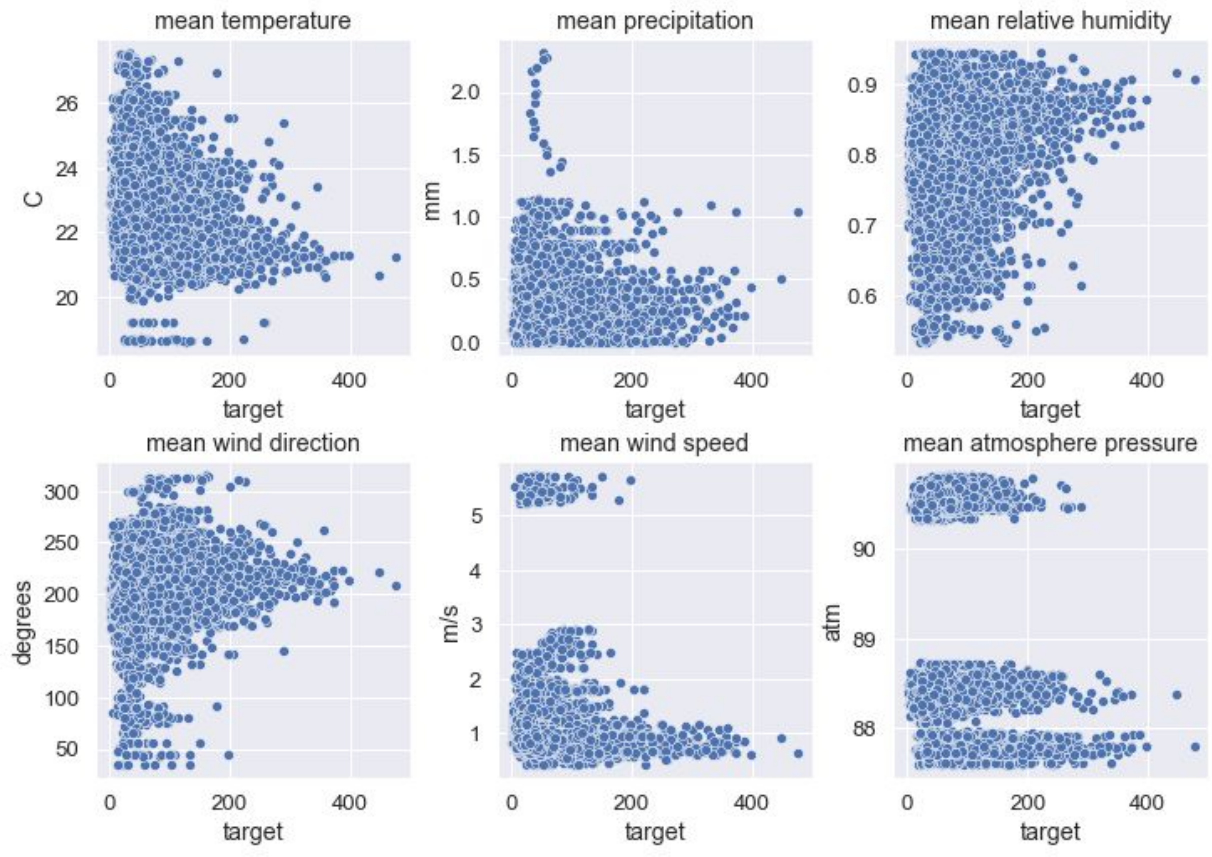




Overview of feature (means)

For each observation we have the following features:

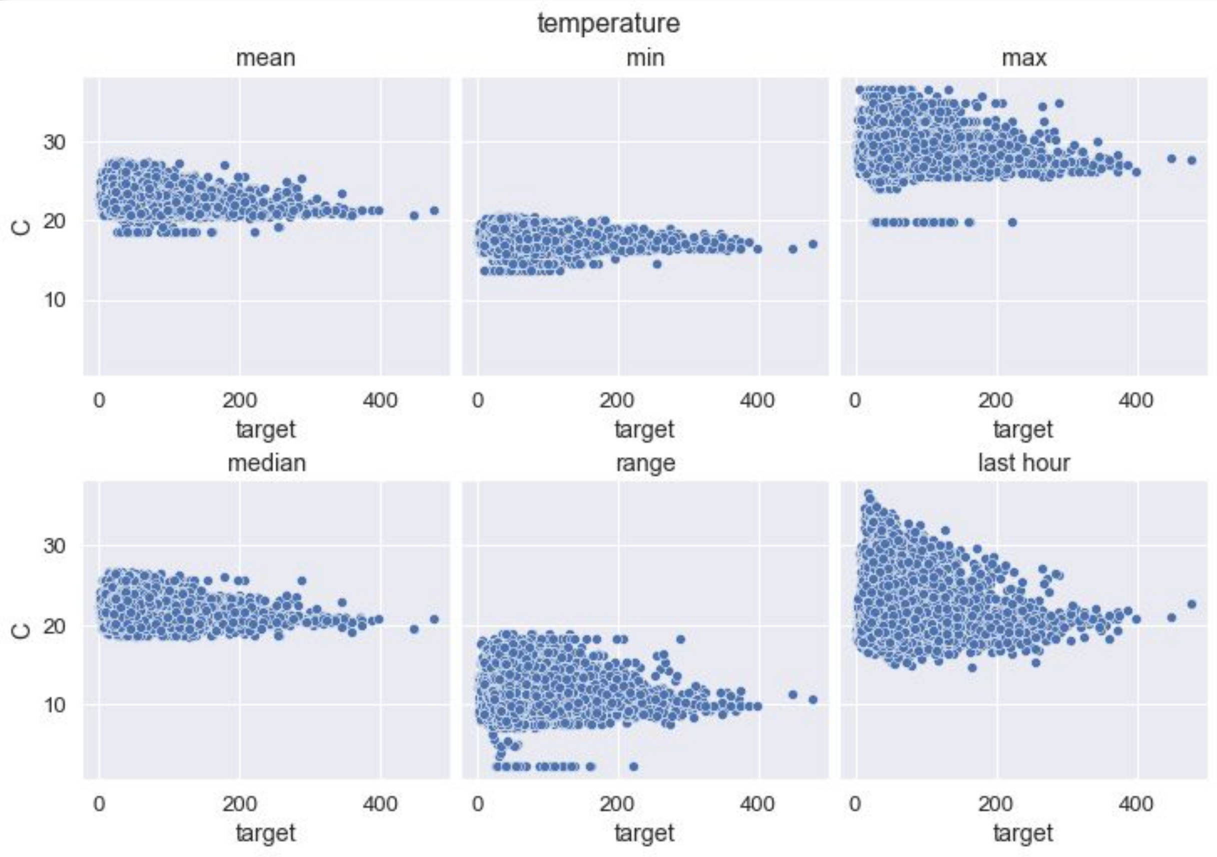
- Temperature
- Precipitation
- Humidity
- Wind direction
- Wind speed
- Atmospheric pressure



Engineered features

For the features, we calculate:

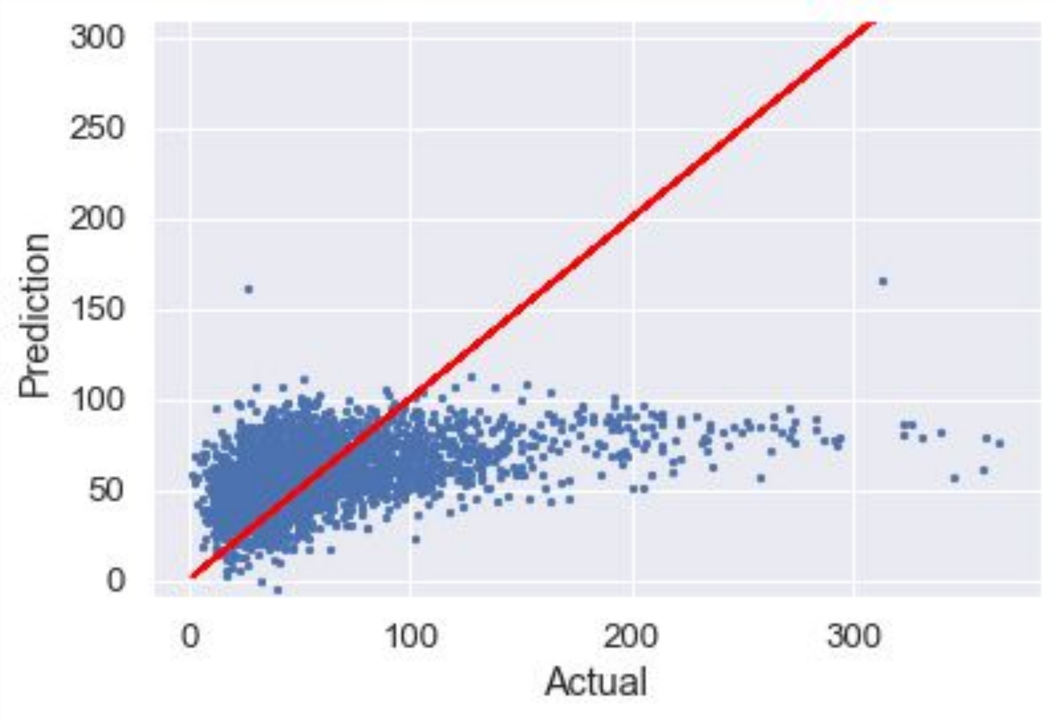
- Min, Max, Range
- Mean, Median
- Variance
- Value of the last recorded hour
- Mean Value of the last day



The Baseline Model

- Linear Regression

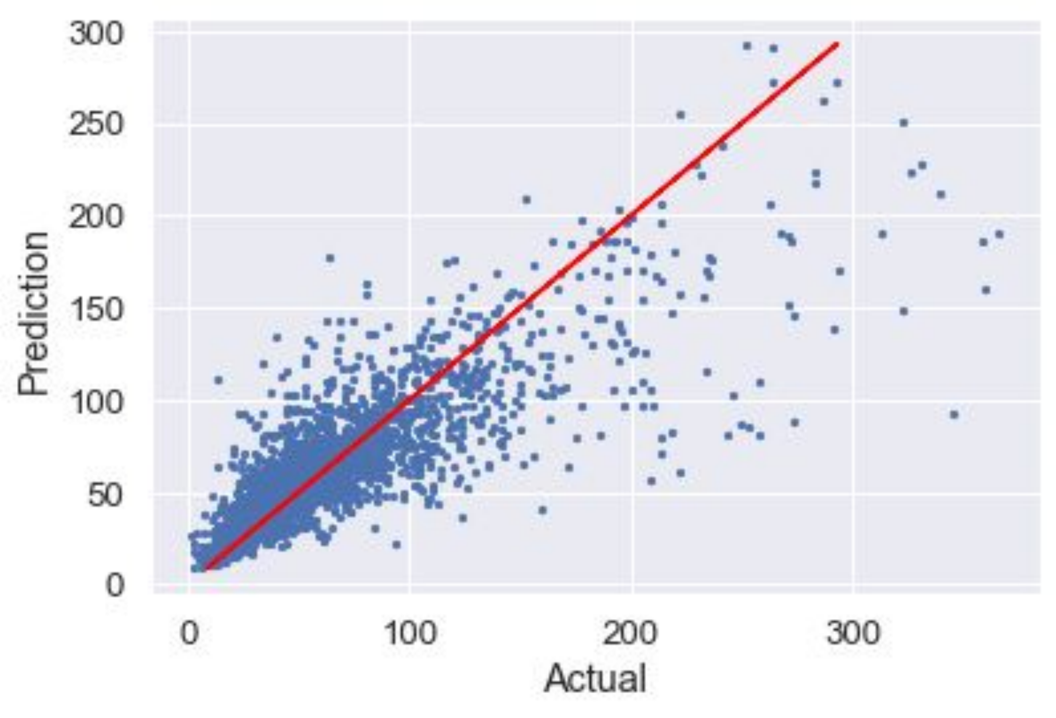
- RMSE: 39.9



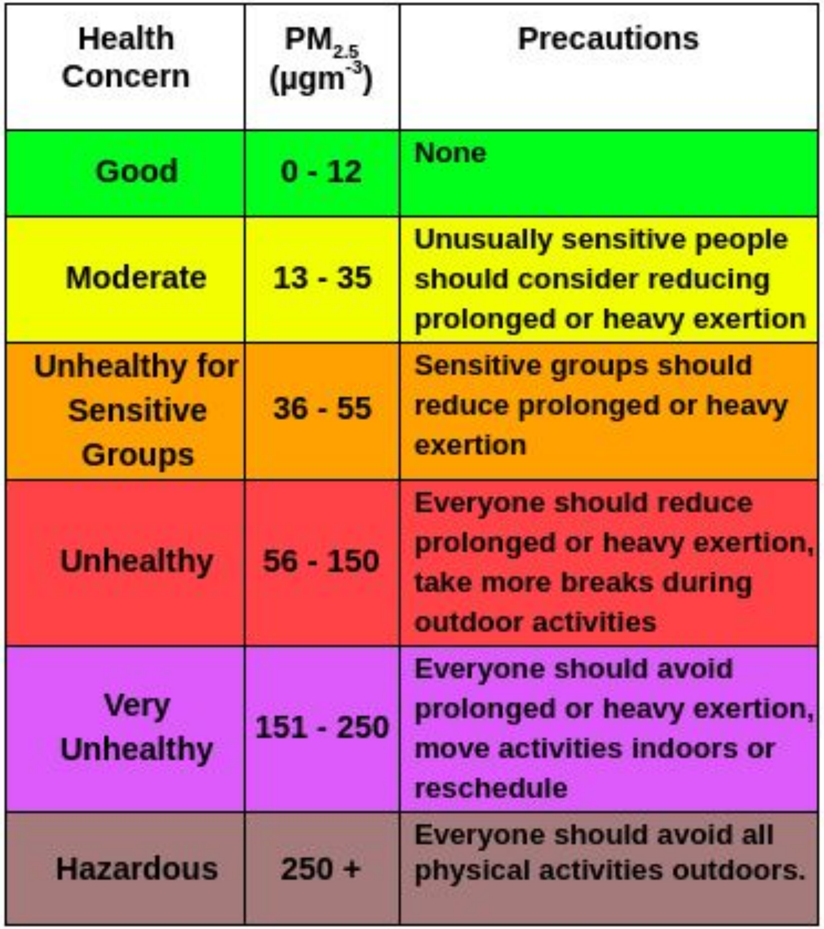
The Model

- ExtraTrees Regression

- RMSE: 24.5

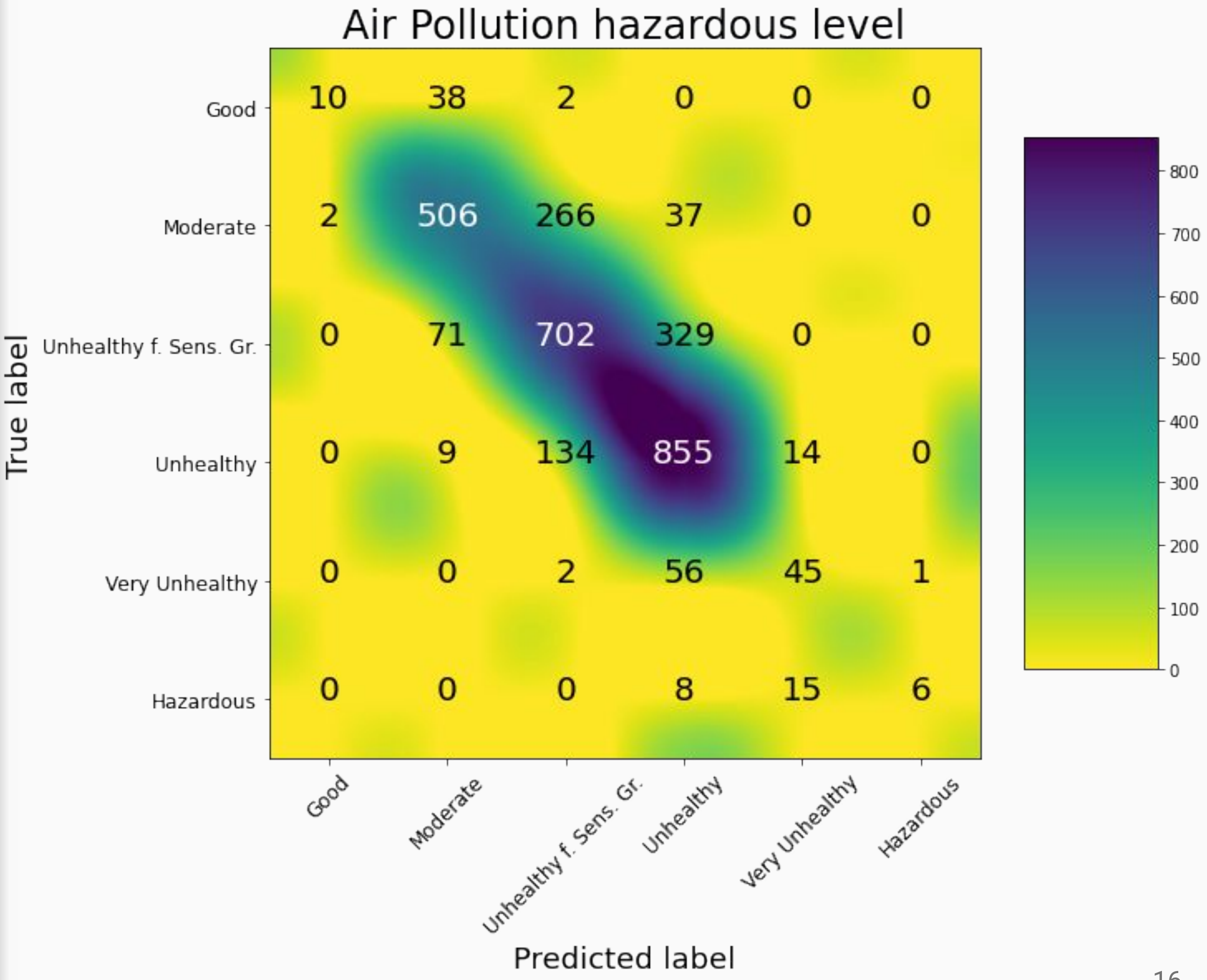


Evaluation



Health Concern	$PM_{2.5}$ (μgm^{-3})	Precautions
Good	0 - 12	None
Moderate	13 - 35	Unusually sensitive people should consider reducing prolonged or heavy exertion
Unhealthy for Sensitive Groups	36 - 55	Sensitive groups should reduce prolonged or heavy exertion
Unhealthy	56 - 150	Everyone should reduce prolonged or heavy exertion, take more breaks during outdoor activities
Very Unhealthy	151 - 250	Everyone should avoid prolonged or heavy exertion, move activities indoors or reschedule
Hazardous	250 +	Everyone should avoid all physical activities outdoors.

Evaluation

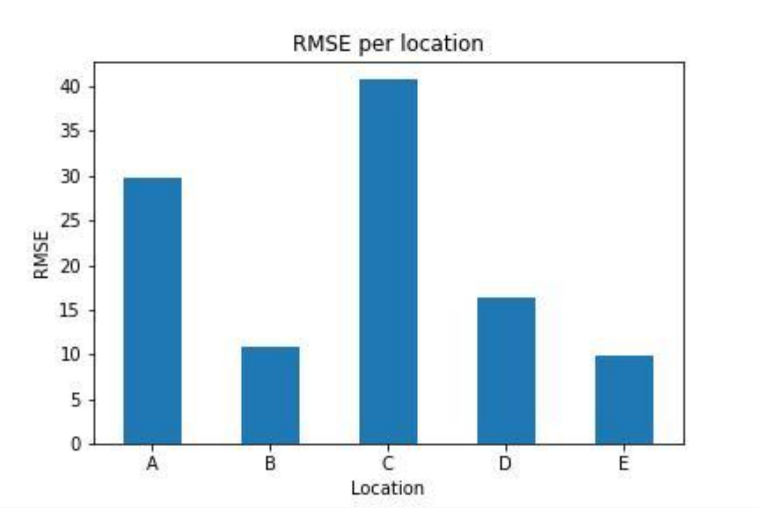


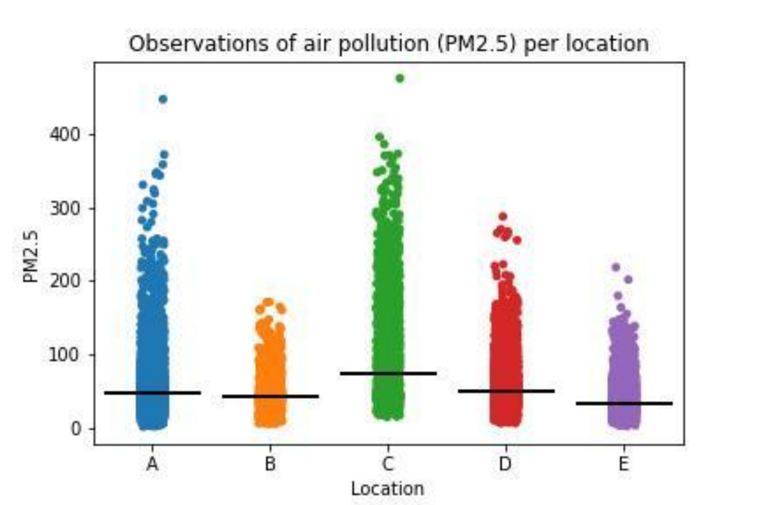
Air Pollution hazardous level

True label \ Predicted label	Good	Moderate	Unhealthy f. Sens. Gr.	Unhealthy	Very Unhealthy	Hazardous
Good	10	38	2	0	0	0
Moderate	2	506	266	37	0	0
Unhealthy f. Sens. Gr.	0	71	702	329	0	0
Unhealthy	0	9	134	855	14	0
Very Unhealthy	0	0	2	56	45	1
Hazardous	0	0	0	8	15	6

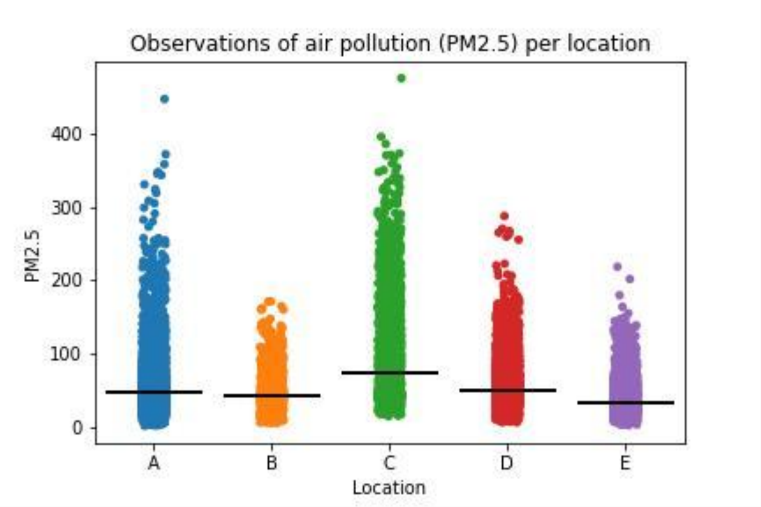
Accuracy Score
68.3%

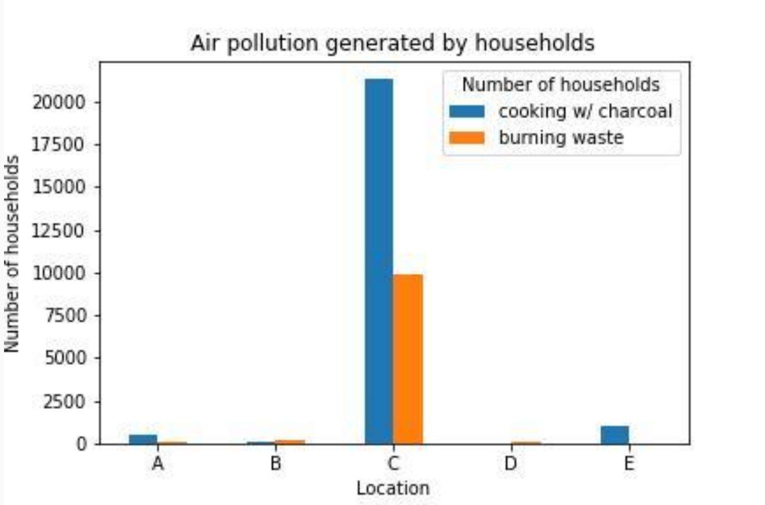
Error Analysis





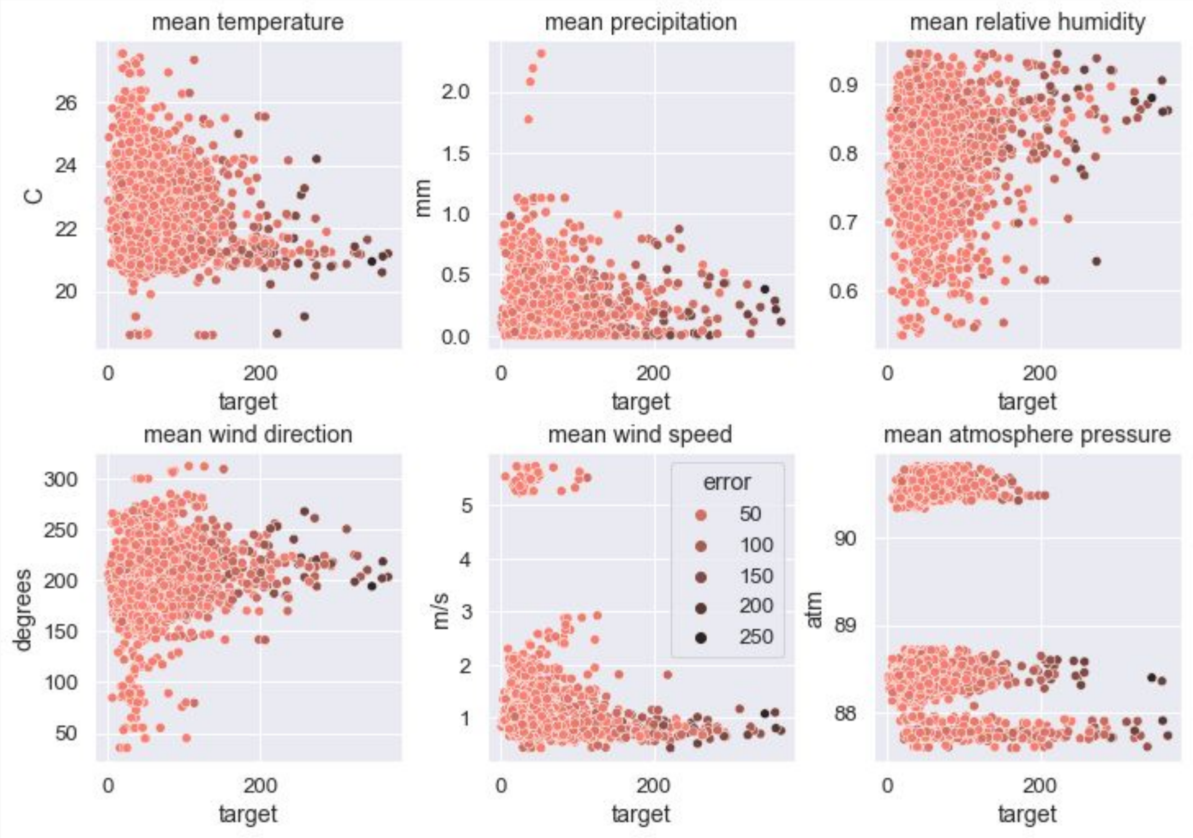
Error Analysis





Error Analysis

The error is the highest for high targets.



Error Analysis

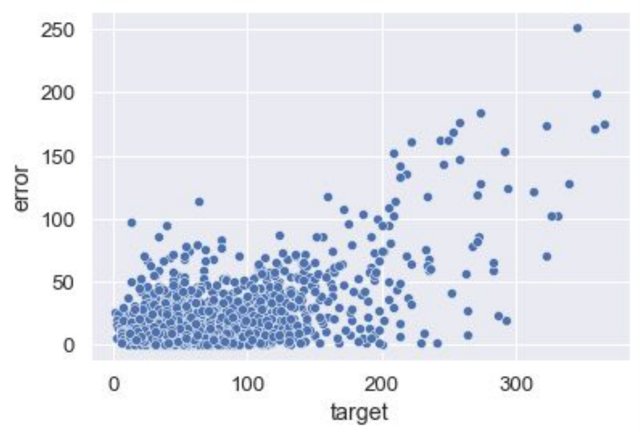
The error is the highest for high targets.

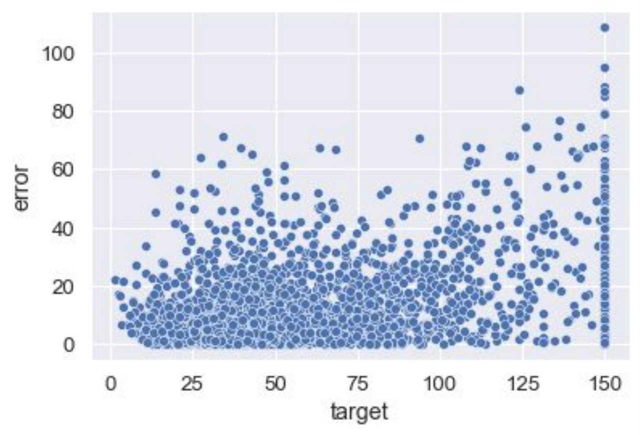
PM2.5 values above 150 means pollution level is very unhealthy.

We don’t gain crucial information in this range.

If we cut off target values above 150 and label them with 150, we can reduce the error.

→ RSME: 18.2





Future Work

- Use full range of data information instead of using aggregations: emphasize chronological dimension of the data.

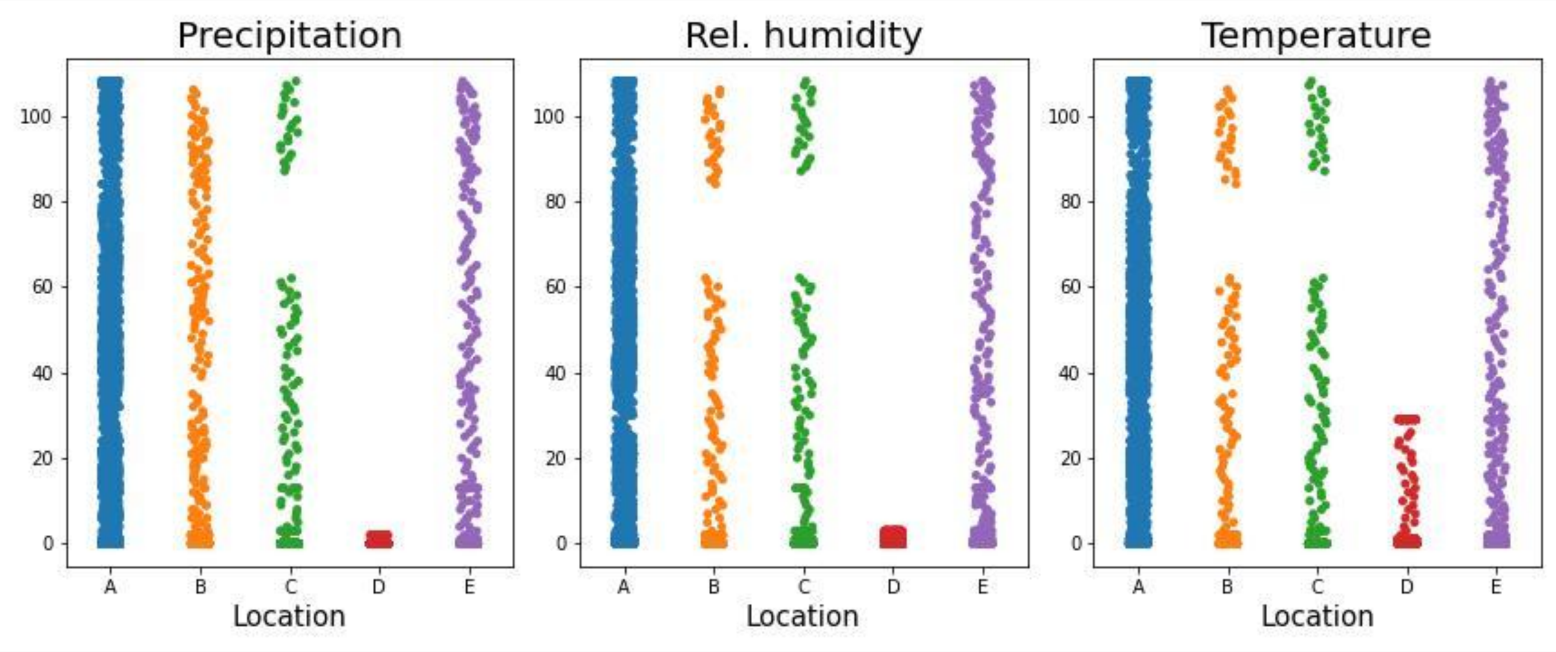
- Advanced feature engineering with domain knowledge (esp. wind).

Future Work

- Use full range of data information instead of using aggregations: emphasize chronological dimension of the data.
- Advanced feature engineering with domain knowledge (esp. wind).

Thank you for your attention!

Data Quality – Missing values


Precipitation
Rel. humidity
Temperature
A
B
C
D
E
Location