



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

STATISTICAL LEARNING,
A.Y 2022/2023

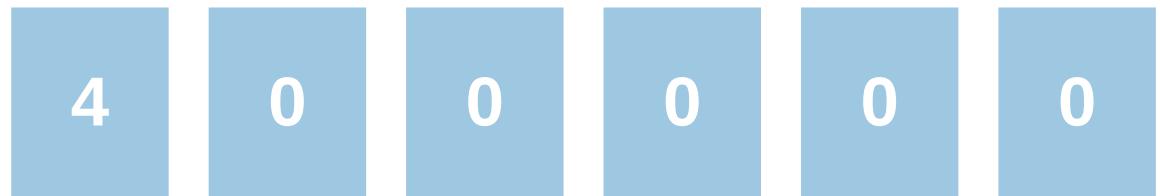
VENICE'S AIR QUALITY PREDICTION

LAURA LEGROTTAGLIE ID:2073222

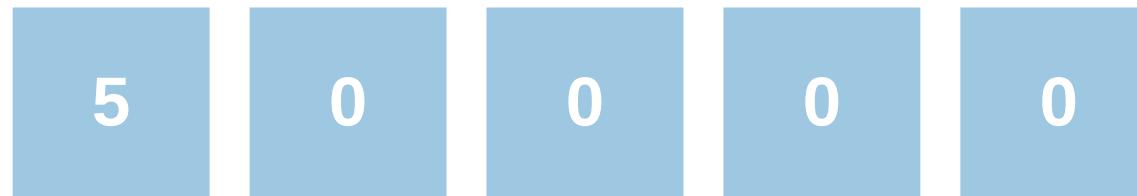
INTRODUCTION

The quality of the air that we breath is a crucial aspect of everyday life.

Every year air pollution causes:



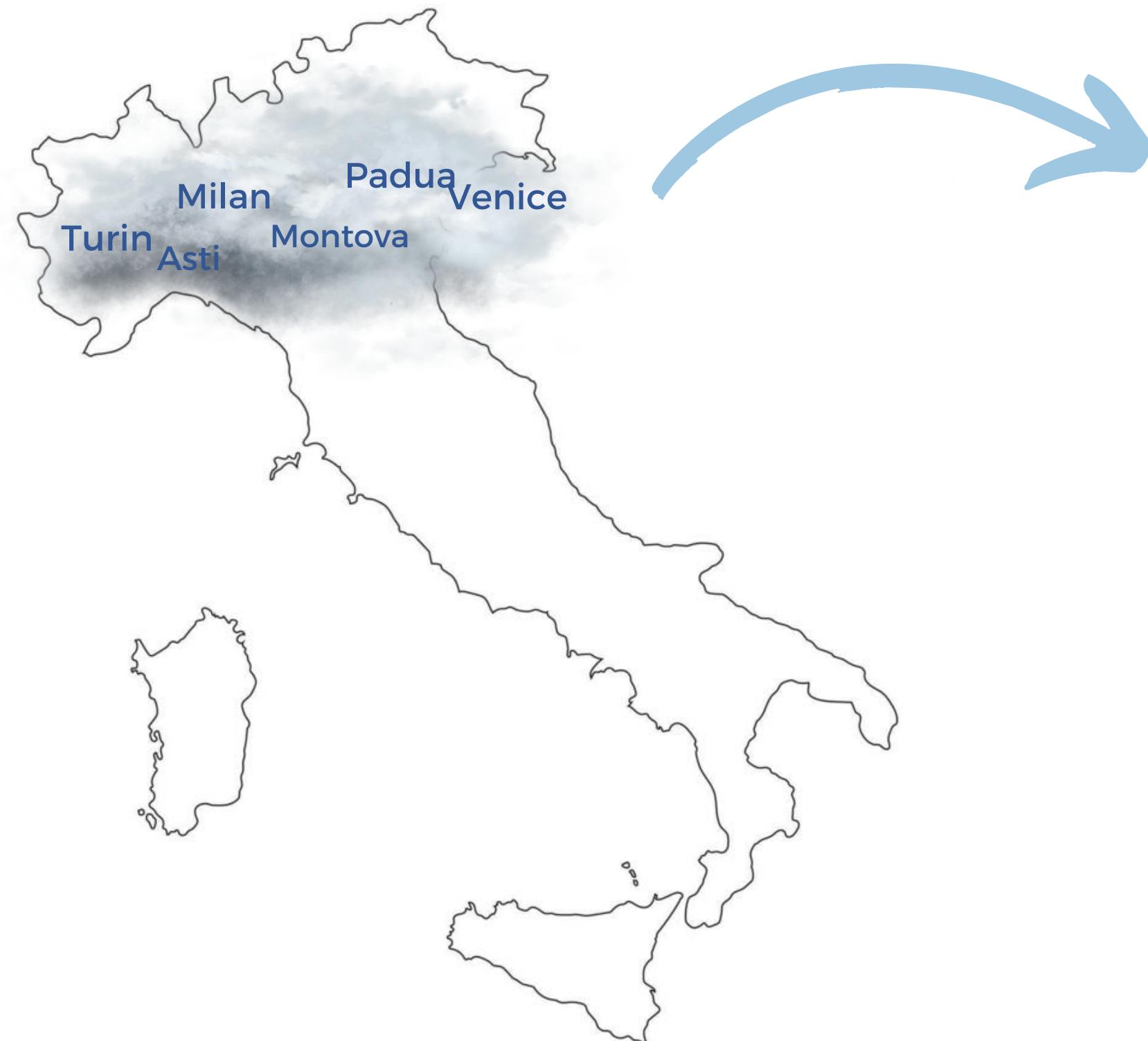
deaths in Europe



deaths in Italy



INTRODUCTION



Venice

PM10 AND AIR QUALITY

LIMIT

50 µg/m³

NUM/YEAR

35

TASK

Binary classification problem

TODAY

predict

TOMORROW

Meteorological conditions



Air quality

- PM10 < 50 µg/m³
- PM10 >= 50 µg/m³

WHAT ARE THE ADVANTAGES?

SOCIAL ASSISTANCE



Helping people with health conditions to detect critical days, avoid outdoor activities and take protective measures such as wearing masks

ECONOMIC



METEOROLOGICAL CONDITIONS

vs PM10

- easy to measure
- affordable
- can be measured everywhere
- specialized equipment
- expensive to maintain
- sparsely distributed

GOALS

- **INTERPRETABILITY**

1. Can meteorological data be used to predict air quality?
2. What are the weather predictors that are significant to the forecast and what is their influence?

- **PREDICTION**

1. How do models perform in making forecasts?
 - Comparing performances using different metrics

OVERVIEW

01

DATASET CONSTRUCTION

02

DATA PREPROCESSING

03

EXPLORATORY DATA ANALYSIS

04

MODELS

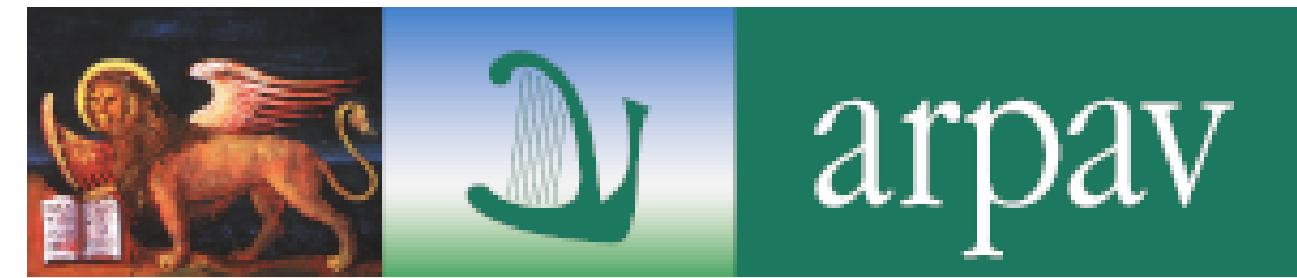
05

CONCLUSIONS

DATASET CONSTRUCTION



WEATHER DATA



Agenzia Regionale per la Prevenzione
e Protezione Ambientale del Veneto

PM10 DATA

FINAL DATASET

WEATHER DATASETS

- Data from years: 2018, 2019, 2021, 2022
- 15 variables



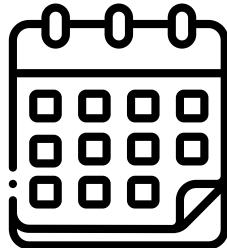
Location



Dew Point(°C)



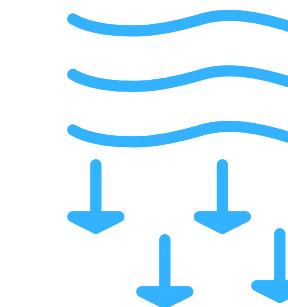
Wind_speed_med (km/h)
Wind_speed_max (km/h)
Gust(km/h)



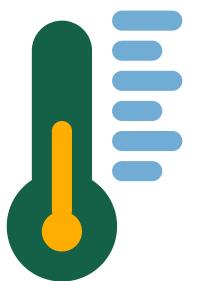
Date



Humidity



Pressure (mb)
Pressure_med (mb)



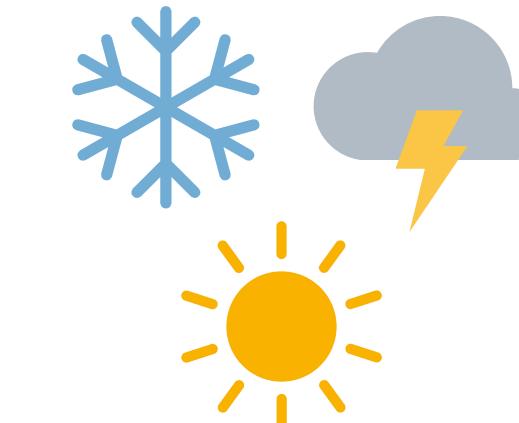
T_min(°C)
T_max(°C)
T_med(°C)



Visibility(km)



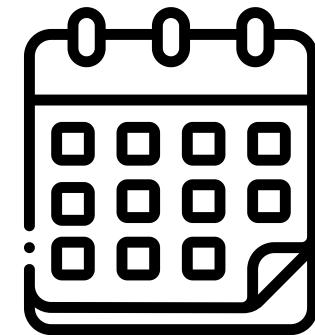
Rain(mm)



Phenomena
(fog, rain, snow,
thunderstorm...)

PM10 DATASETS

- 2 variables for years 2018 and 2022: Date and PM10
- For 2019 and 2021 datasets, from several stations, the selected ones are: Parco Bissuola (reference station) and Via Tagliamento.
- 3 variables for years 2019 and 2021



Date



PM10

- Parco Bissuola
- Via Tagliamento

DATA PREPROCESSING

1

HANDLE
MISSING
VALUES

WEATHER DATASETS

- Two days with missing values: 5/2/2019 and 14/1/2021
- Solution: median imputation considering the specific month and year

PM10 DATASETS

- The PM10 values of Parco Bissuola contain:
 - 1 NA value in 2018
 - 12 NA values in 2022
- Solution: consider the data that comes from the nearby station in Via Tagliamento

2

DATA CLEANING AND PREPARATION

```
##   Location          Date        T_med      T_max
## Length:1460    Length:1460  Min.   :-3.00  Min.   :-4.0
## Class  :character  Class  :character  1st Qu.: 7.00  1st Qu.: 4.0
## Mode   :character  Mode   :character  Median  :15.00  Median  :11.0
##                               Mean   :14.71  Mean   :10.8
##                               3rd Qu.:22.00 3rd Qu.:18.0
##                               Max.   :30.00  Max.   :27.0
##   T_min       Dew_point     Humidity     Visibility
## Min.   :-2.00  Min.   : 0.00  Min.   :35.00  Min.   : 0.00
## 1st Qu.:11.00 1st Qu.: 4.00  1st Qu.:65.00 1st Qu.:15.00
## Median :18.00  Median :10.00  Median :74.00  Median :19.00
## Mean   :18.46  Mean   :10.16  Mean   :73.94  Mean   :16.75
## 3rd Qu.:26.00 3rd Qu.:16.00 3rd Qu.:84.00 3rd Qu.:20.00
## Max.   :35.00  Max.   :24.00  Max.   :99.00  Max.   :24.00
## Wind_speed_med Wind_speed_max      Gust      Pressure      Pressure_med
## Min.   : 4.00  Min.   : 5.00  Min.   :0  Min.   :989  Min.   :0
## 1st Qu.: 8.00 1st Qu.:15.00  1st Qu.:0  1st Qu.:1011 1st Qu.:0
## Median :10.00  Median :17.00  Median :0  Median :1015  Median :0
## Mean   :10.76  Mean   :19.36  Mean   :0  Mean   :1016  Mean   :0
## 3rd Qu.:12.00 3rd Qu.:22.00  3rd Qu.:0  3rd Qu.:1020 3rd Qu.:0
## Max.   :39.00  Max.   :100.00  Max.   :0  Max.   :1037  Max.   :0
##   Rain   Phenomena
## Min.   :0  Length:1460
## 1st Qu.:0  Class  :character
## Median :0  Mode   :character
## Mean   :0
## 3rd Qu.:0
## Max.   :0
```

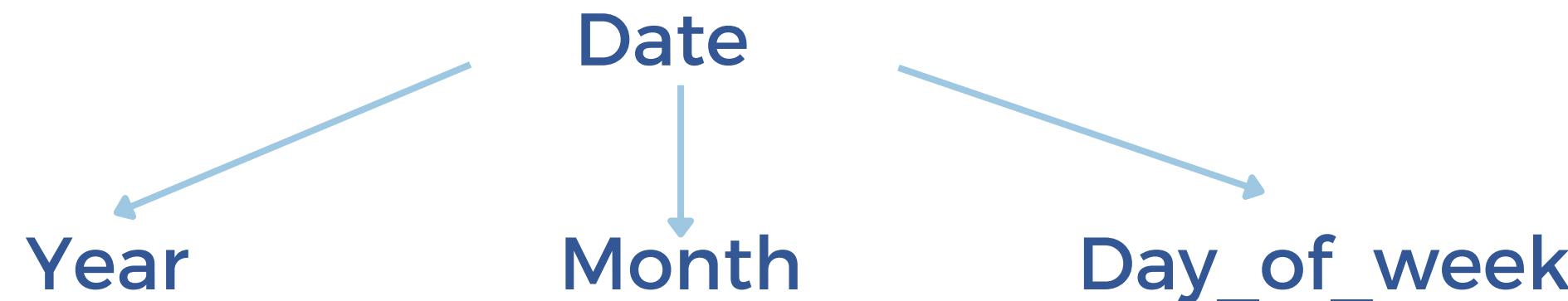
- Removal of unuseful covariates: Gust, Pressure_med, Rain, Location

2

DATA CLEANING AND PREPARATION

DATA PREPROCESSING

- Extraction of possible useful variables



- Encoding of categorical variables: Phenomena, Year, Month and Day_of_week

Initial levels Phenomena: 8	Final levels Phenomena: 4
nebbia	neve
821	3
pioggia	No event
293	824
pioggia temporale	Fog
126	156
pioggia temporale nebbia	Rain
2	352
	Storm
	128

3

CLASS
BALANCE
CHECK

DATA PREPROCESSING

- The response binary variable **BadAirQuality** is created from the PM10 values of Parco Bissuola station

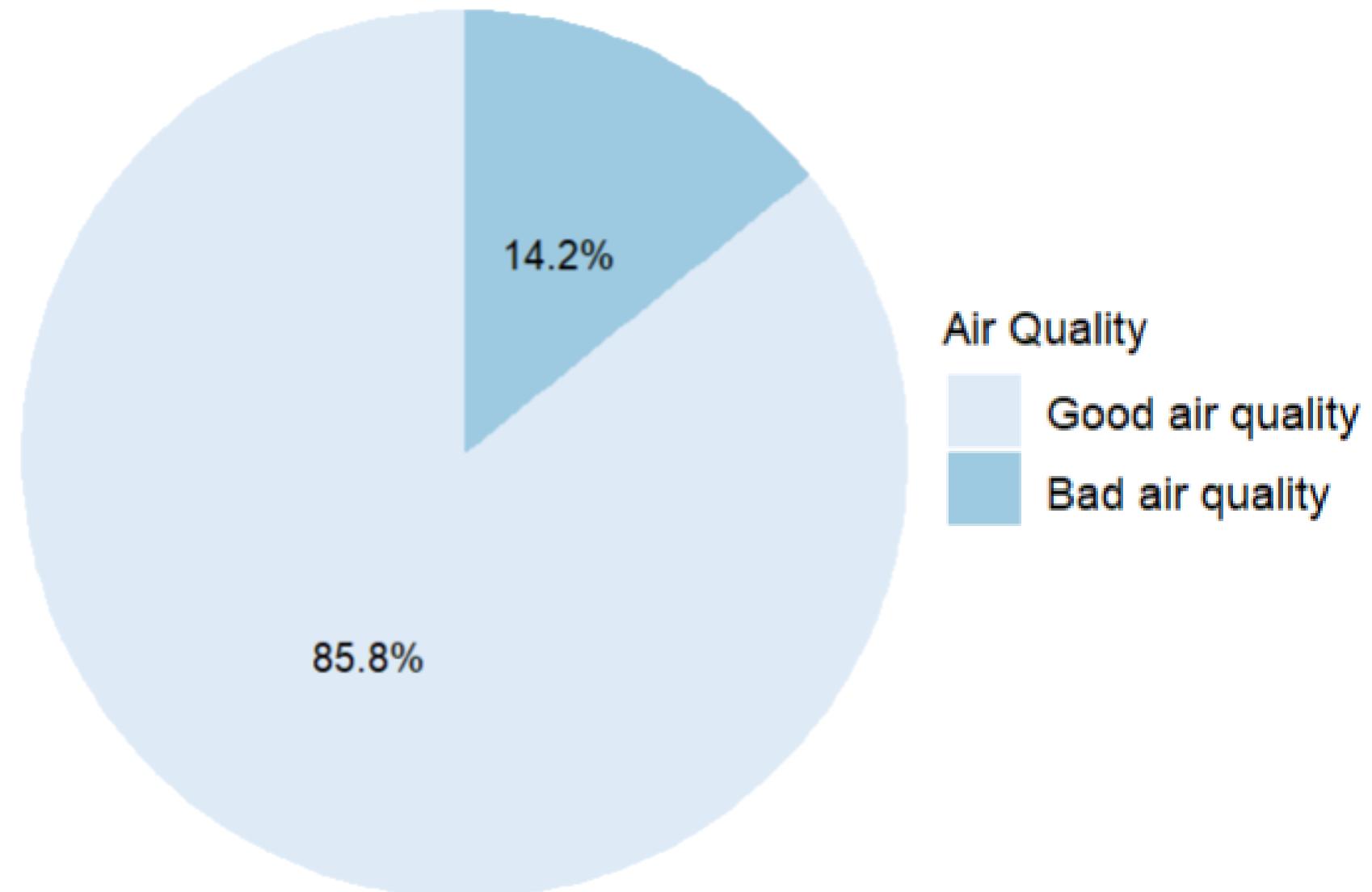
PM10 < 50 $\mu\text{g}/\text{m}^3$



PM10 $\geq 50 \mu\text{g}/\text{m}^3$



- Imbalanced problem

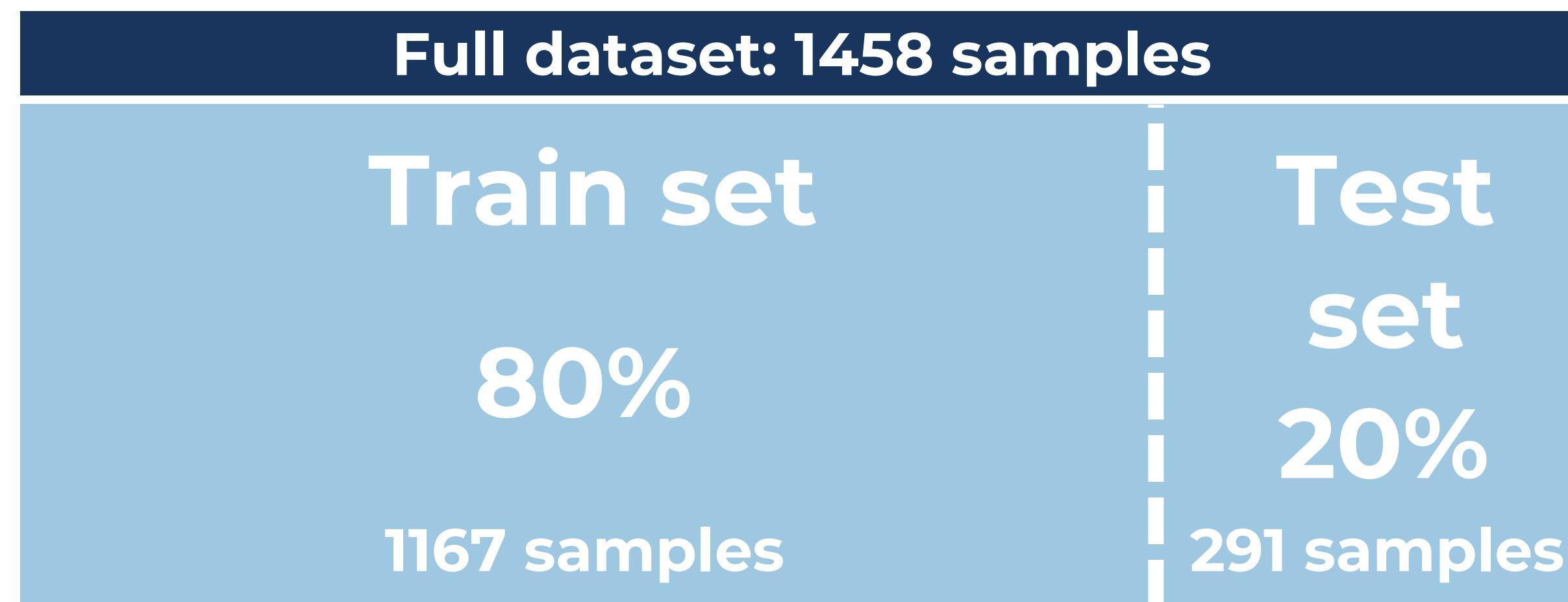


DATA PREPROCESSING

4

TRAIN-TEST
SPLIT

The dataset is divided into training (80%) and testing (20%) in a stratified manner to respect the proportion of the two classes

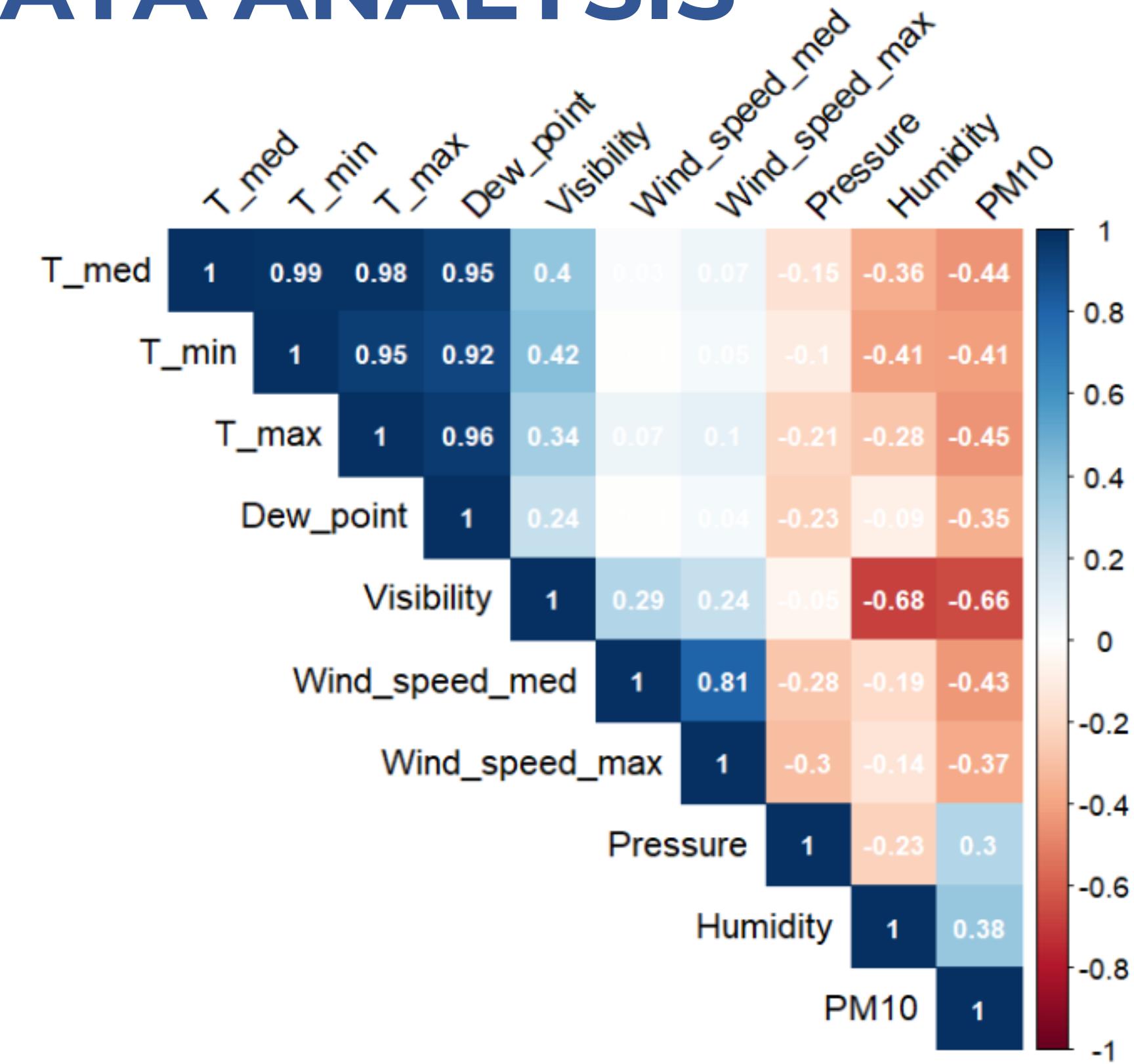


EXPLANATORY DATA ANALYSIS

CORRELATION MATRIX

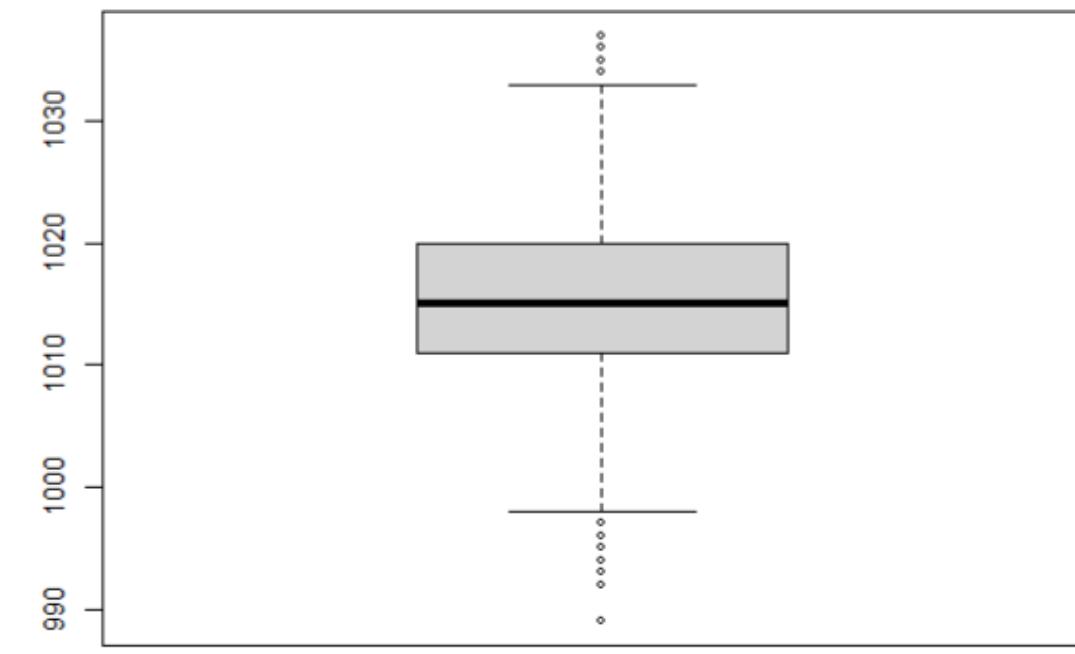
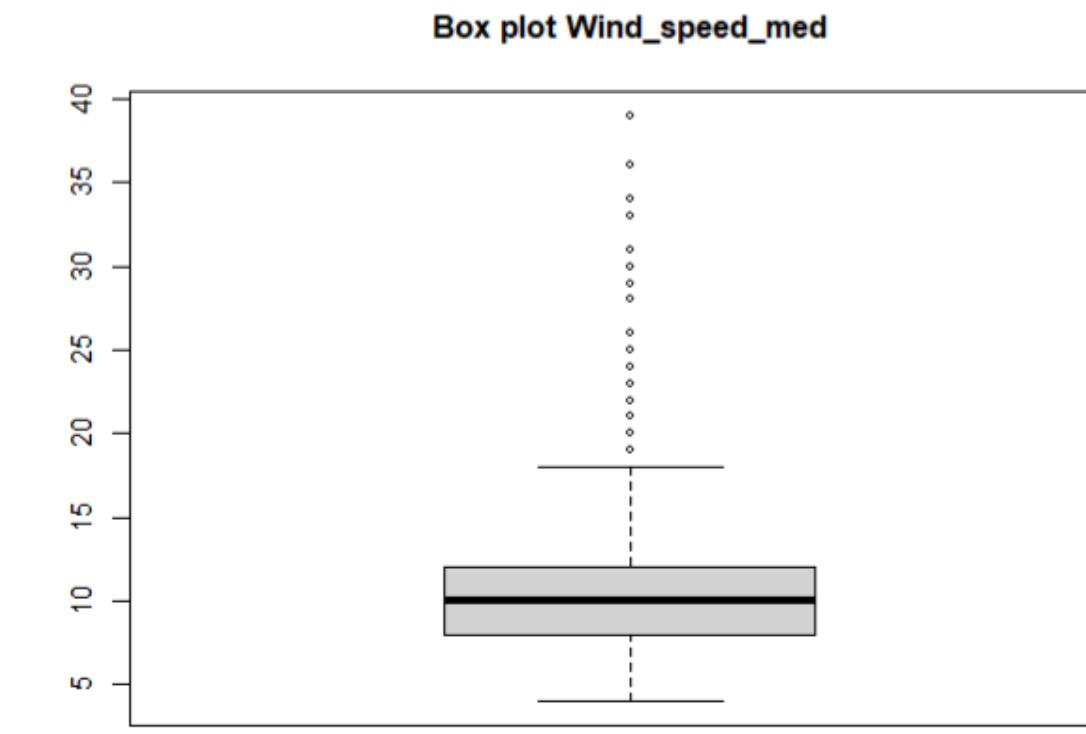
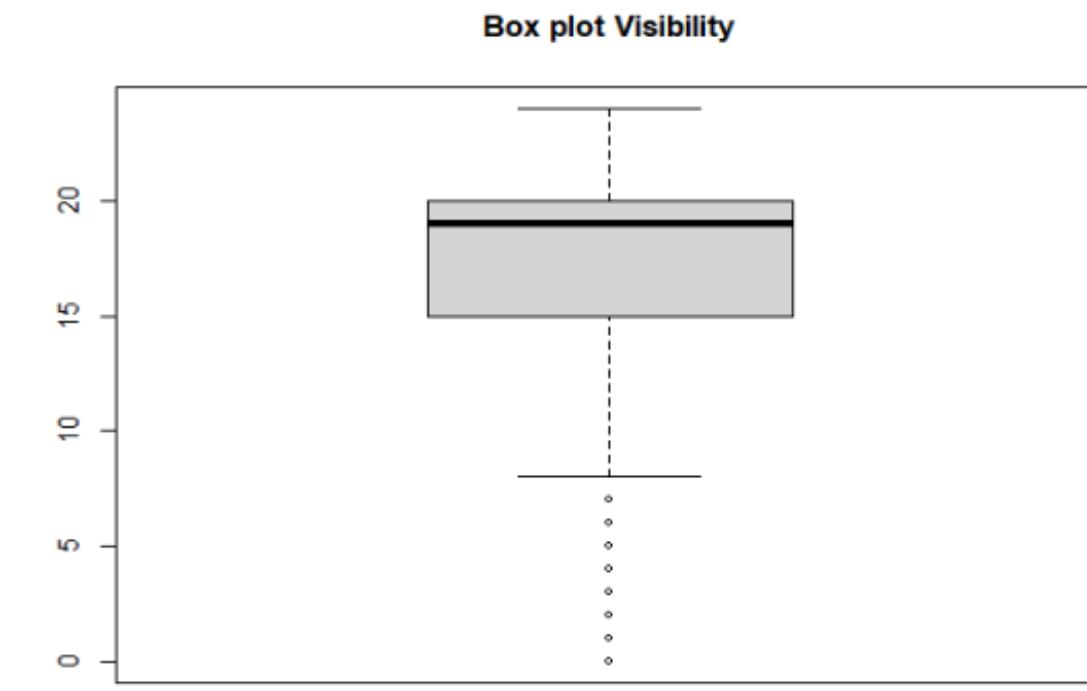
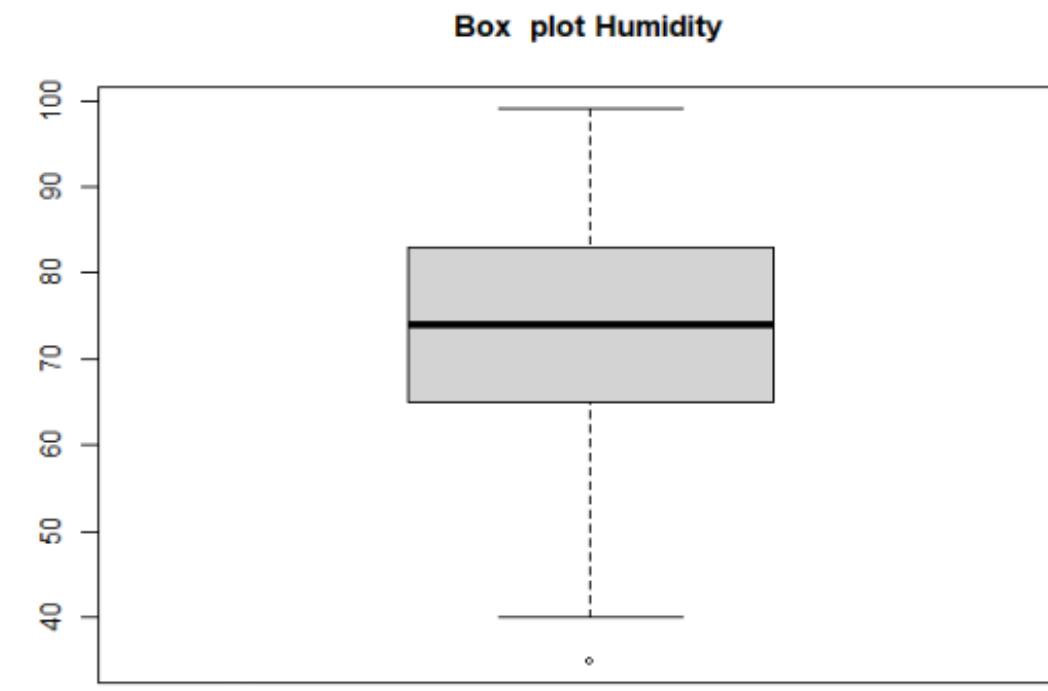
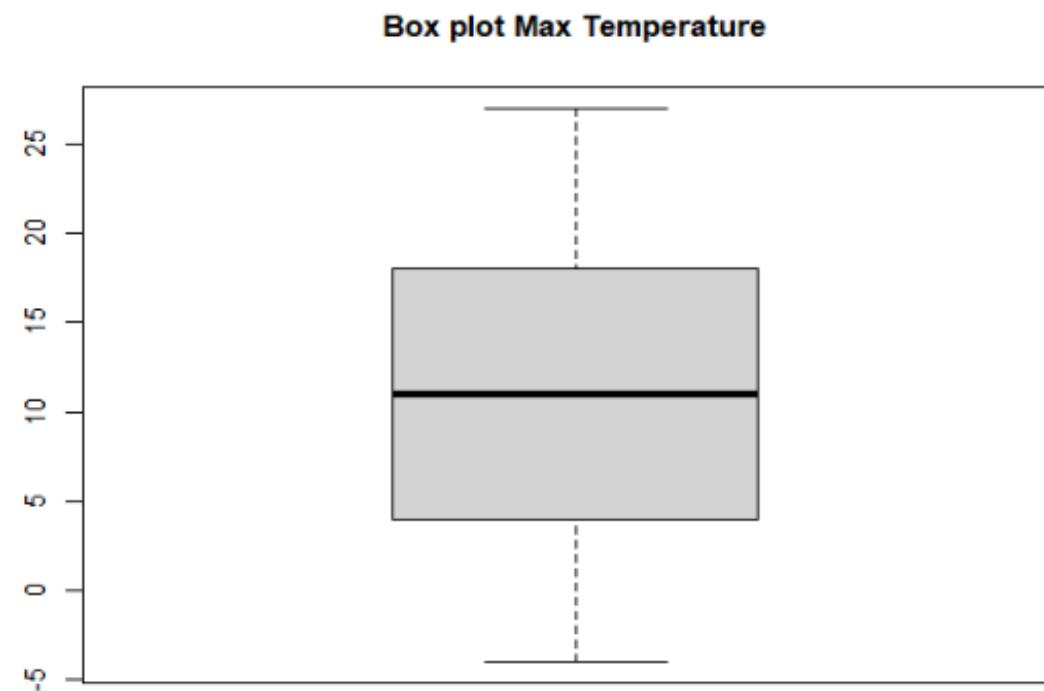
Removal of highly correlated variables based on the correlation with the numerical response PM10.

- T_min
 - T_med
 - Dew_point
 - Wind_speed_max



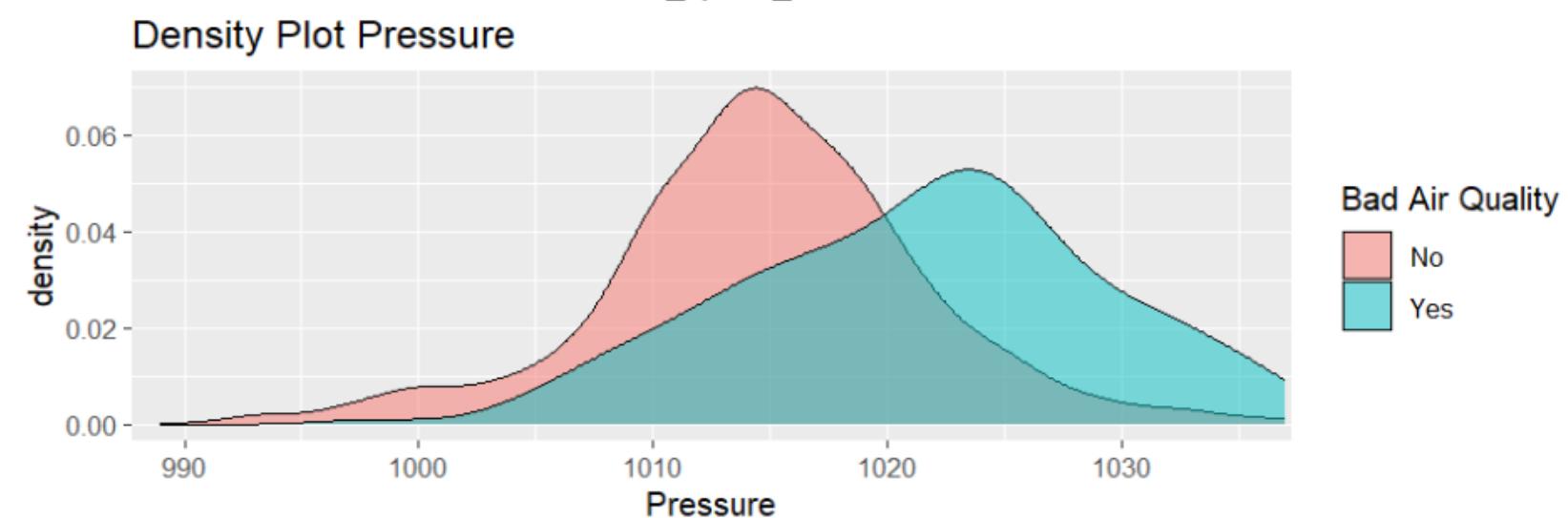
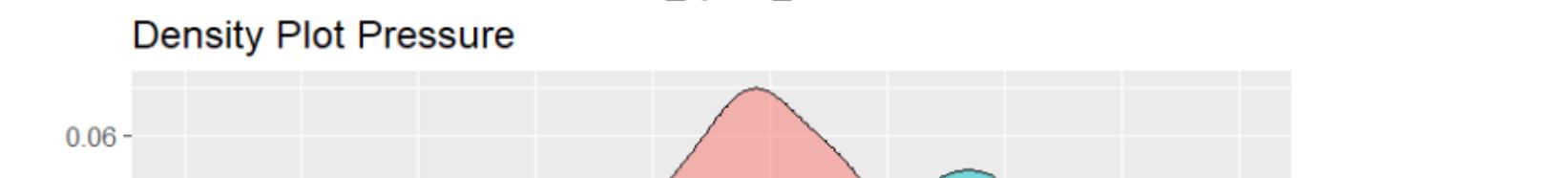
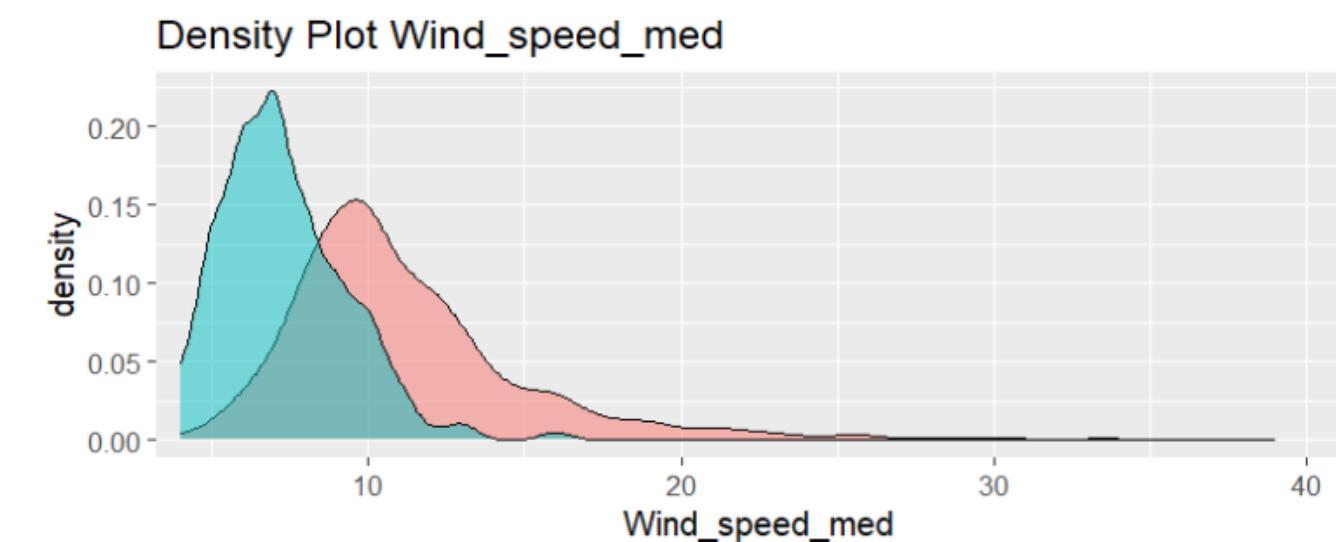
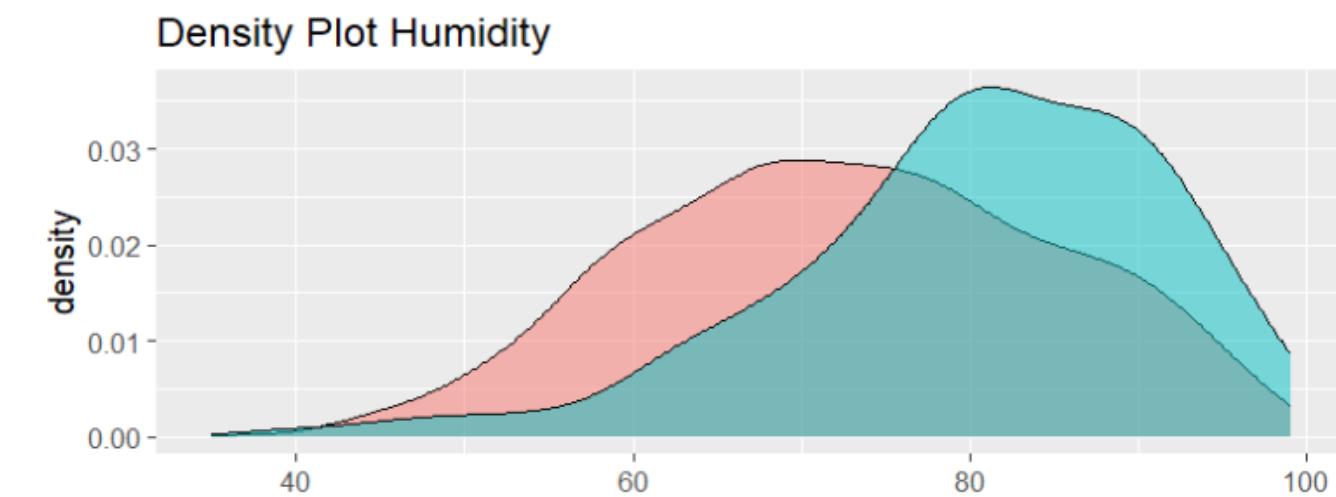
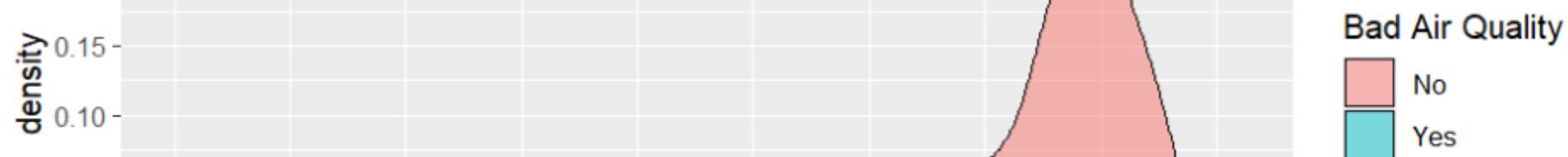
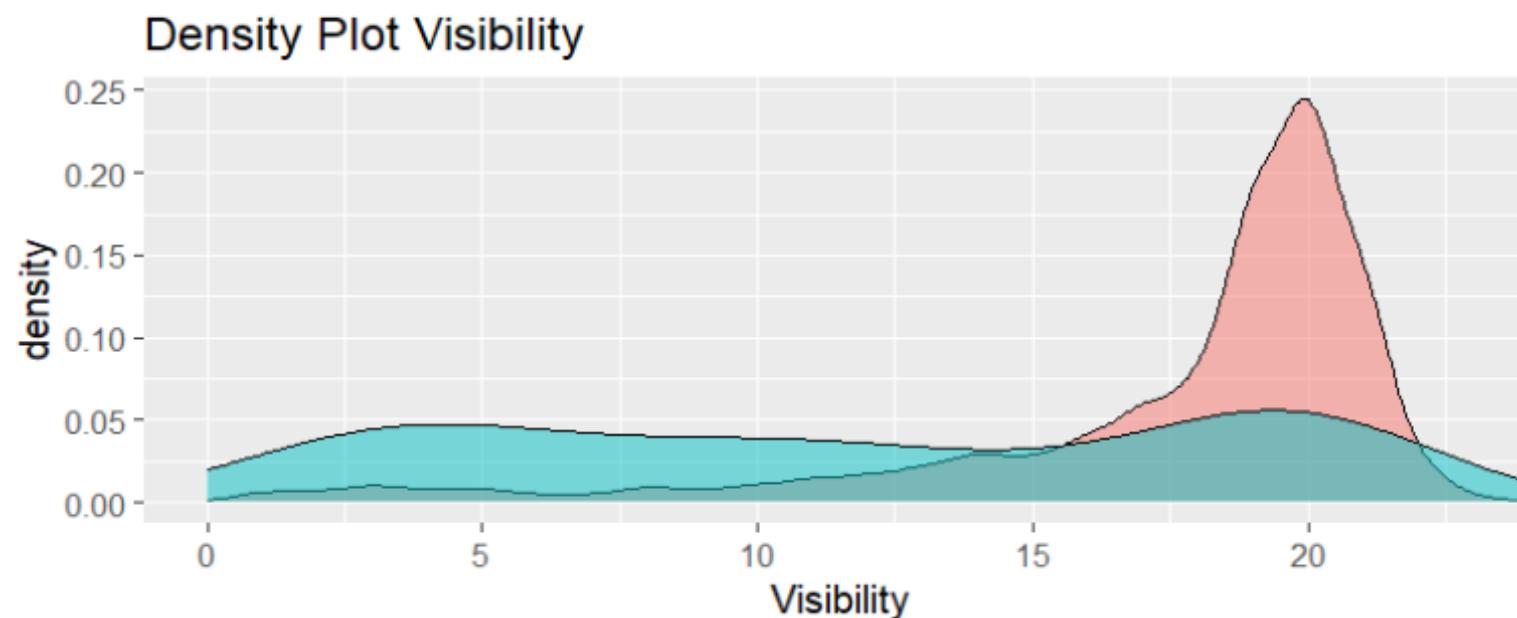
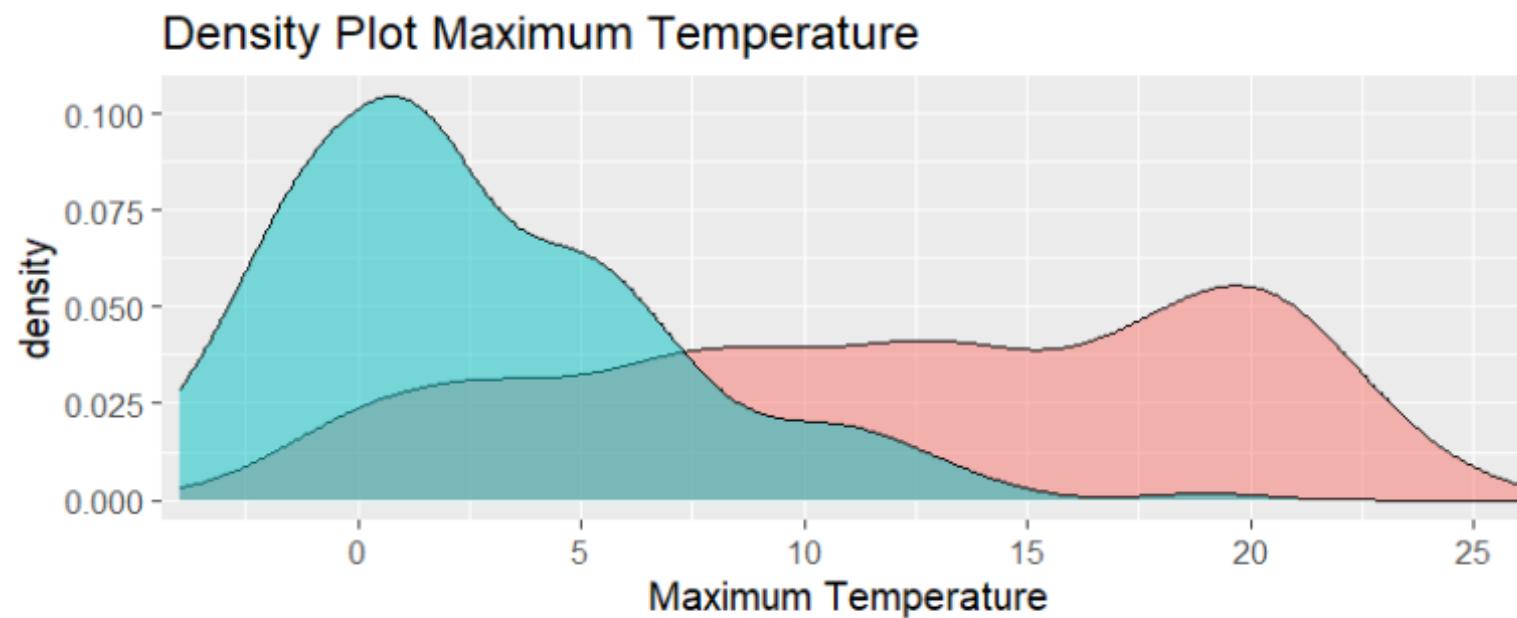
EXPLANATORY DATA ANALYSIS

OUTLIERS



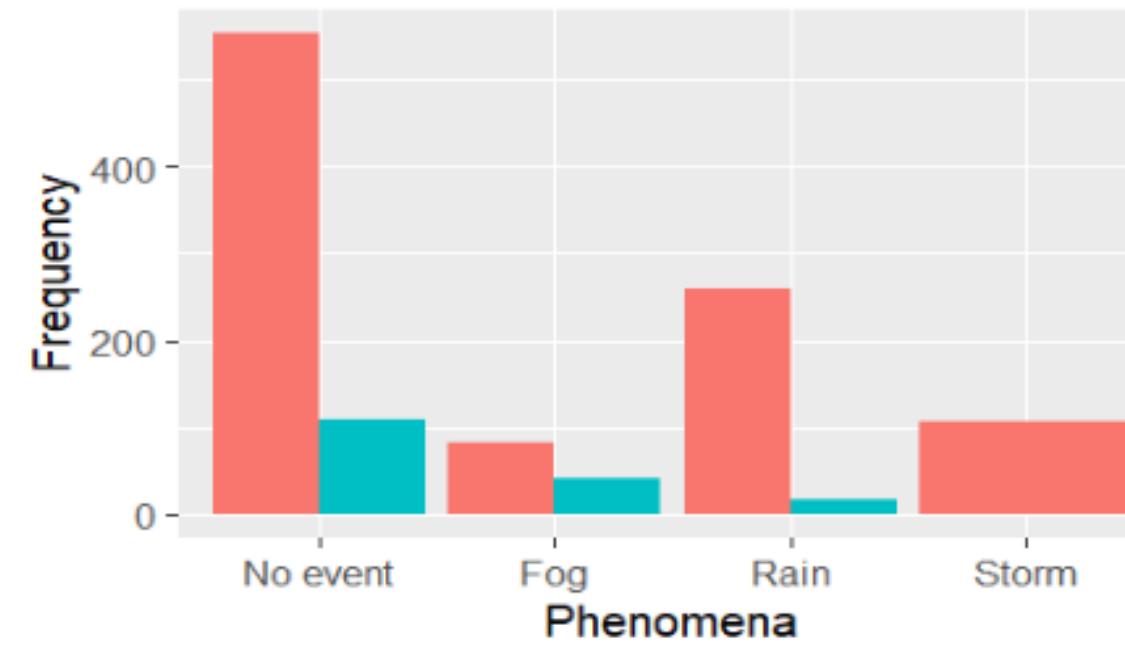
BIVARIATE DENSITY ANALYSIS

CONTINUOUS FEATURES

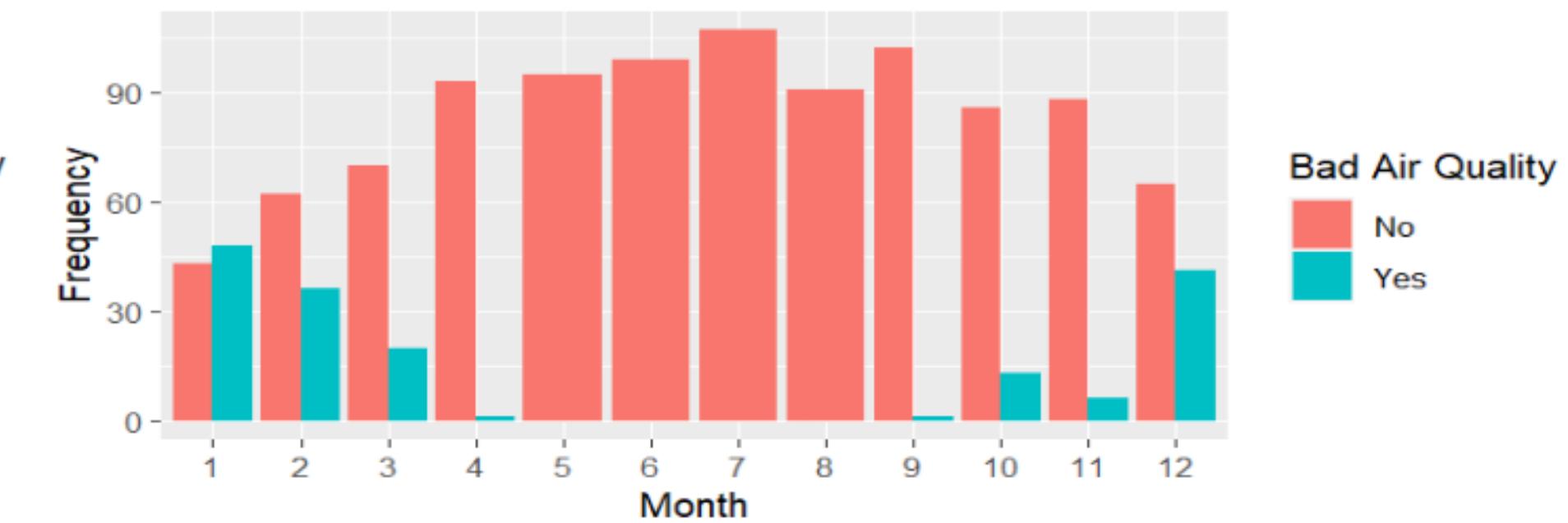


BIVARIATE CATEGORICAL ANALYSIS

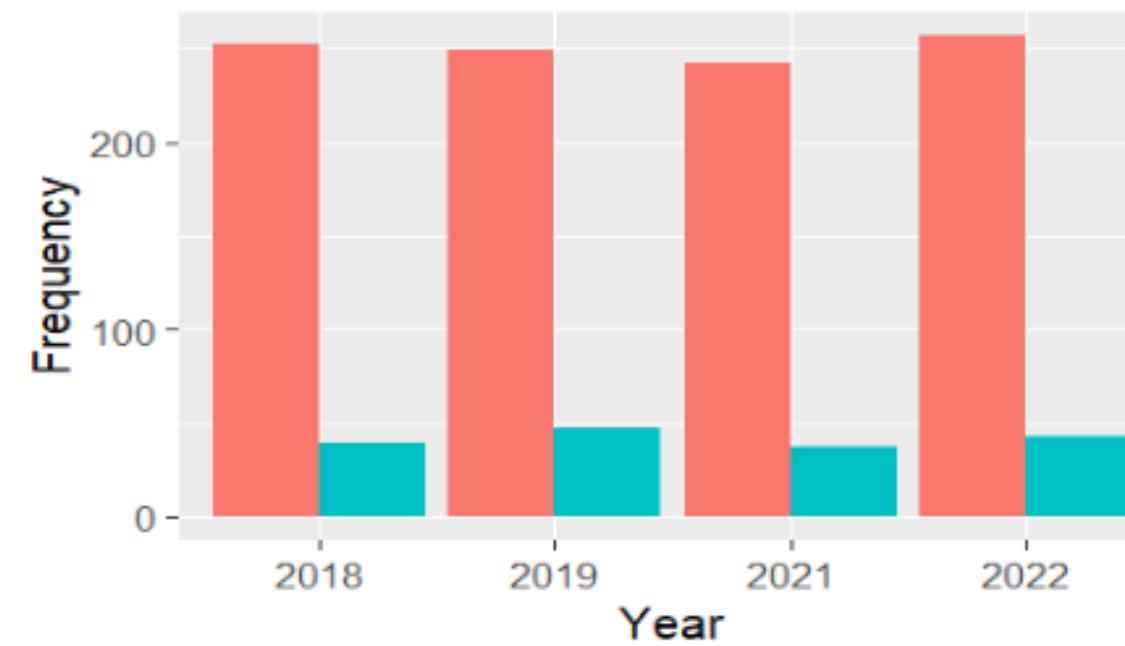
Bar Plot Phenomena



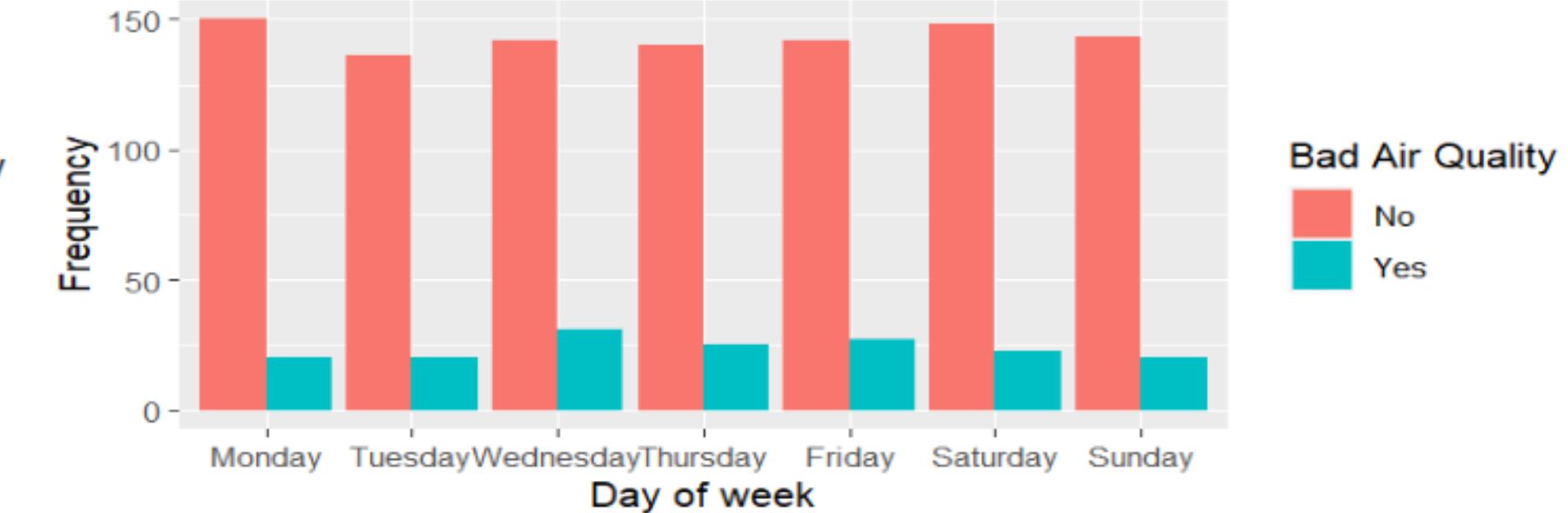
Bar Plot Month



Bar Plot Year

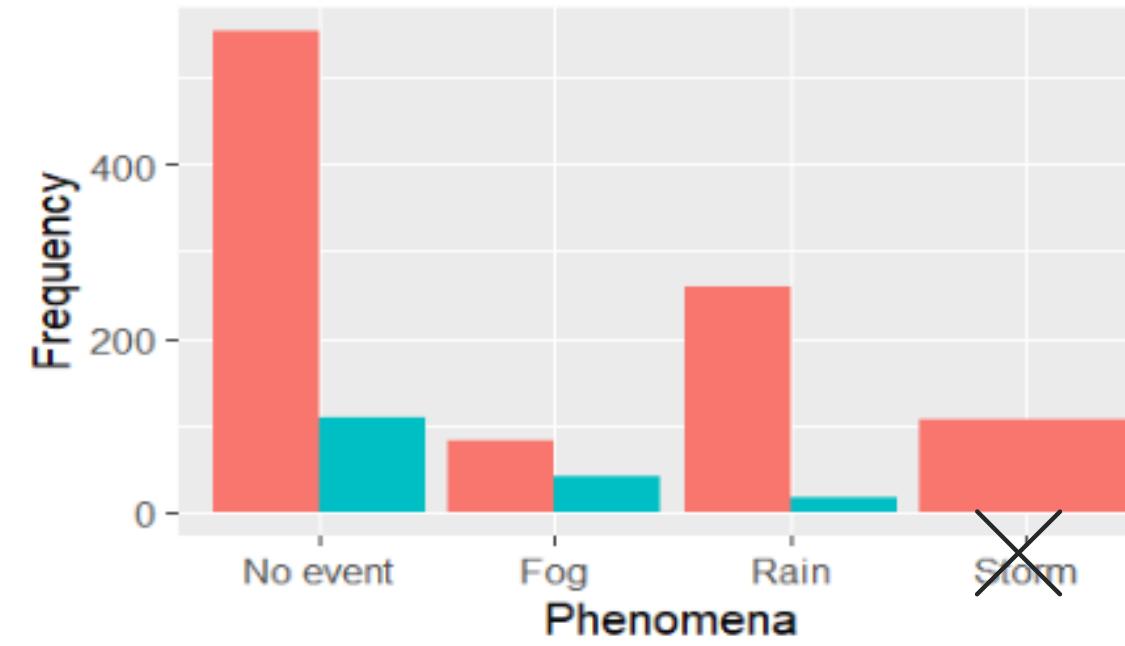


Bar Plot Day of week

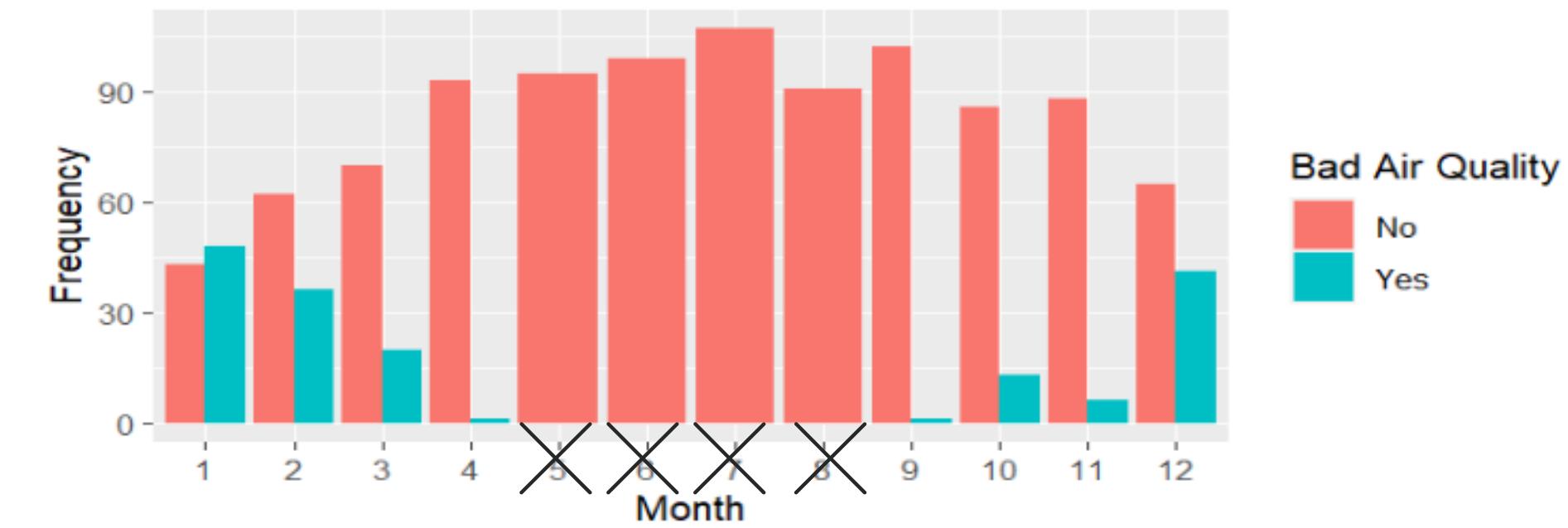


BIVARIATE CATEGORICAL ANALYSIS

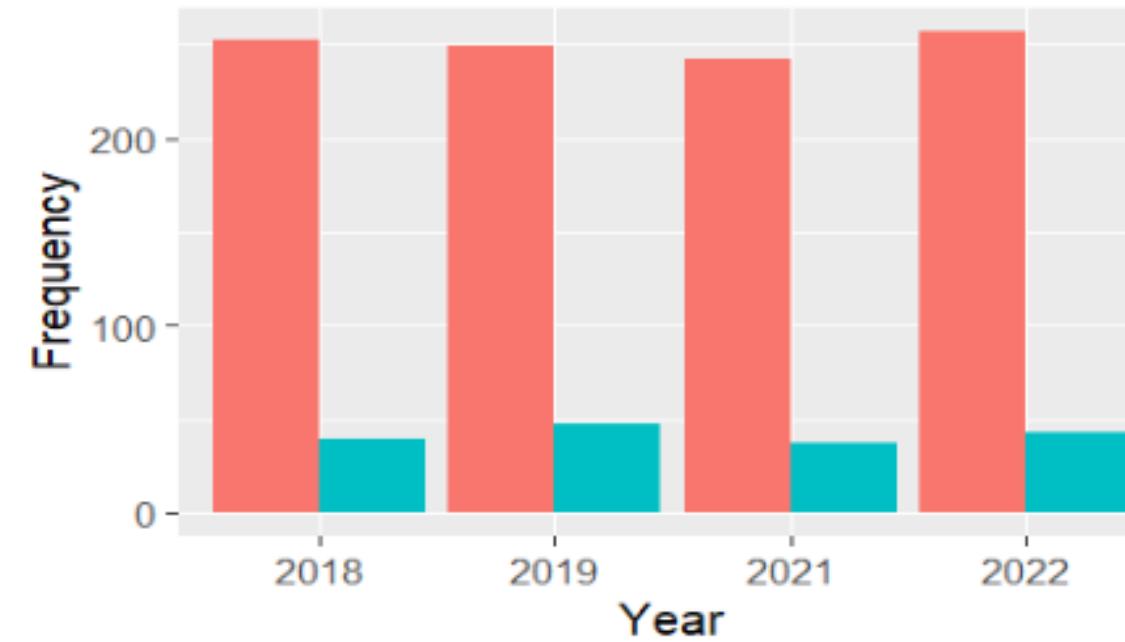
Bar Plot Phenomena



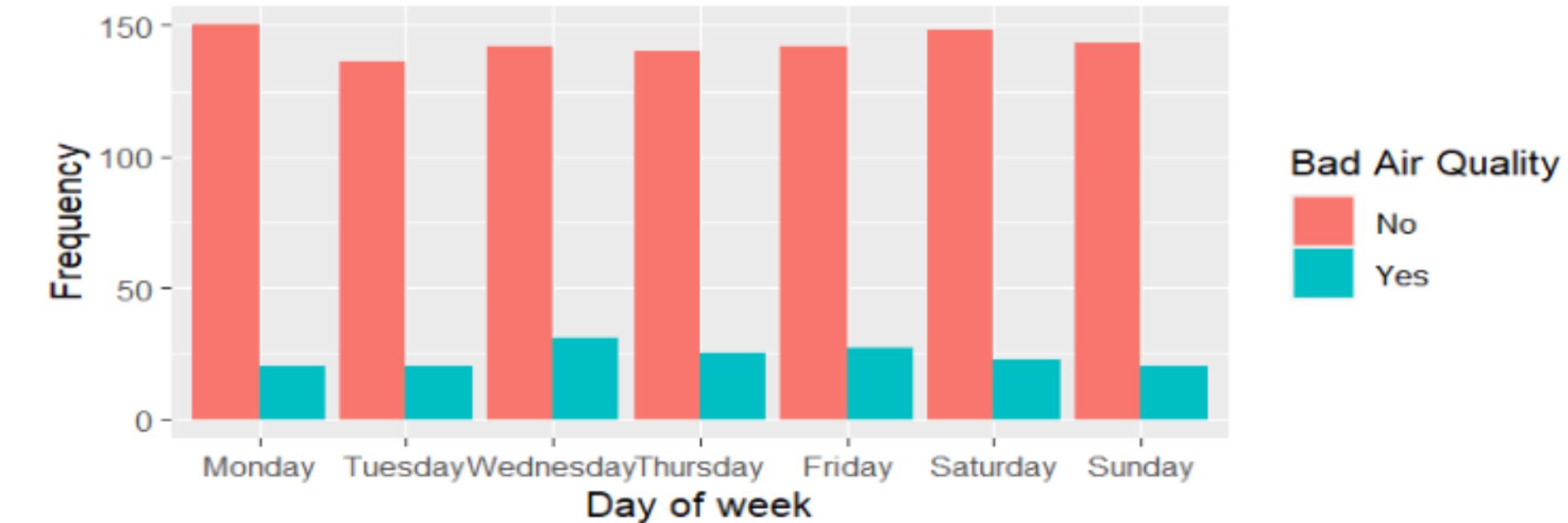
Bar Plot Month



Bar Plot Year



Bar Plot Day of week



FINAL DATASET AND UNDERSAMPLING

The **final training dataset** is composed of 743 occurrences and 10 features

9 PREDICTORS

CONTINUOUS

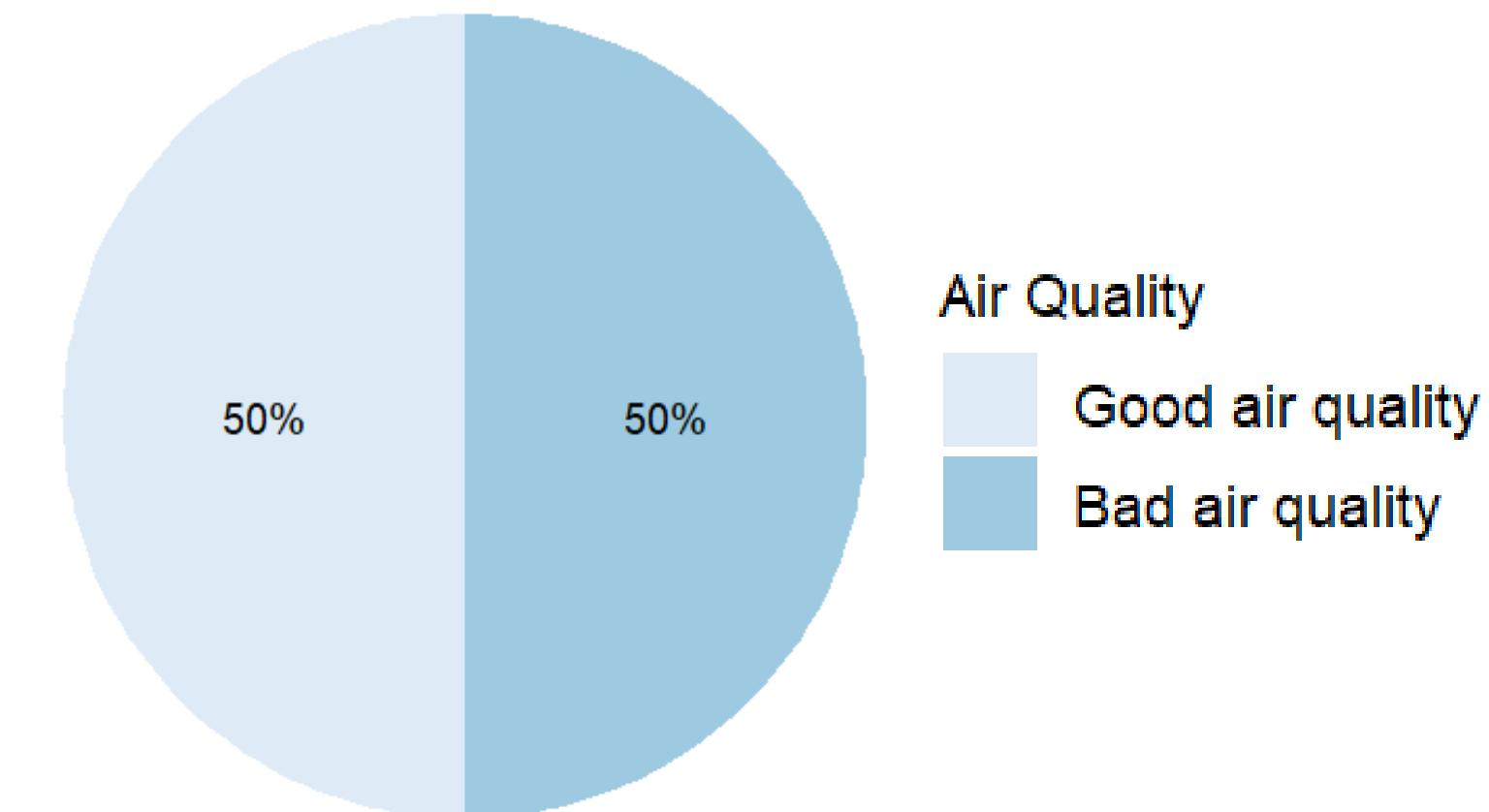
T_max	🌡️
Pressure	⬇️
Humidity	%
Visibility	👁️
Wind_speed_med	💨

CATEGORICAL

Year	📅
Month	
Day_of_week	
Phenomena	☀️ ☁️

RESPONSE: BadAirQuality

To balance the classes, **undersampling** technique is applied. The balanced training set contains 332 observations



MODELS

METRICS

- **Accuracy**, used to measure the overall correctness of the model
- **Precision**, which evaluates the proportion of true positives among all positive predictions
- **Sensitivity**. This metrics shows the extent to which the model is able to detect the days with high PM10 values
- **F1-Score**, harmonic mean of Precision and Sensitivity

Logistic Regression on the original dataset

- All the predictors are initially included
- Since the dataset contains categorical predictors, the presence of multicollinearity is checked by the means of Generalized Variance Inflation Factor (GVIF)

	GVIF	df	GVIF^(1/(2*df))
T_max	3.290769	1	1.814048
Humidity	3.579425	1	1.891937
Visibility	3.266605	1	1.807375
Wind_speed_med	1.555637	1	1.247252
Pressure	1.472572	1	1.213496
Phenomena	2.623416	2	1.272673
Year	1.364436	3	1.053155
Month	5.168010	7	1.124480
Day_of_week	1.200242	6	1.015327

Logistic Regression: original dataset

		TRUE	
		0	1
PRED	0	137	17
	1	11	23

- Variable selection: **Backward Stepwise selection** based on the Bayesian Information Criterion (BIC)
- Initial threshold: 0.5

Coefficients:

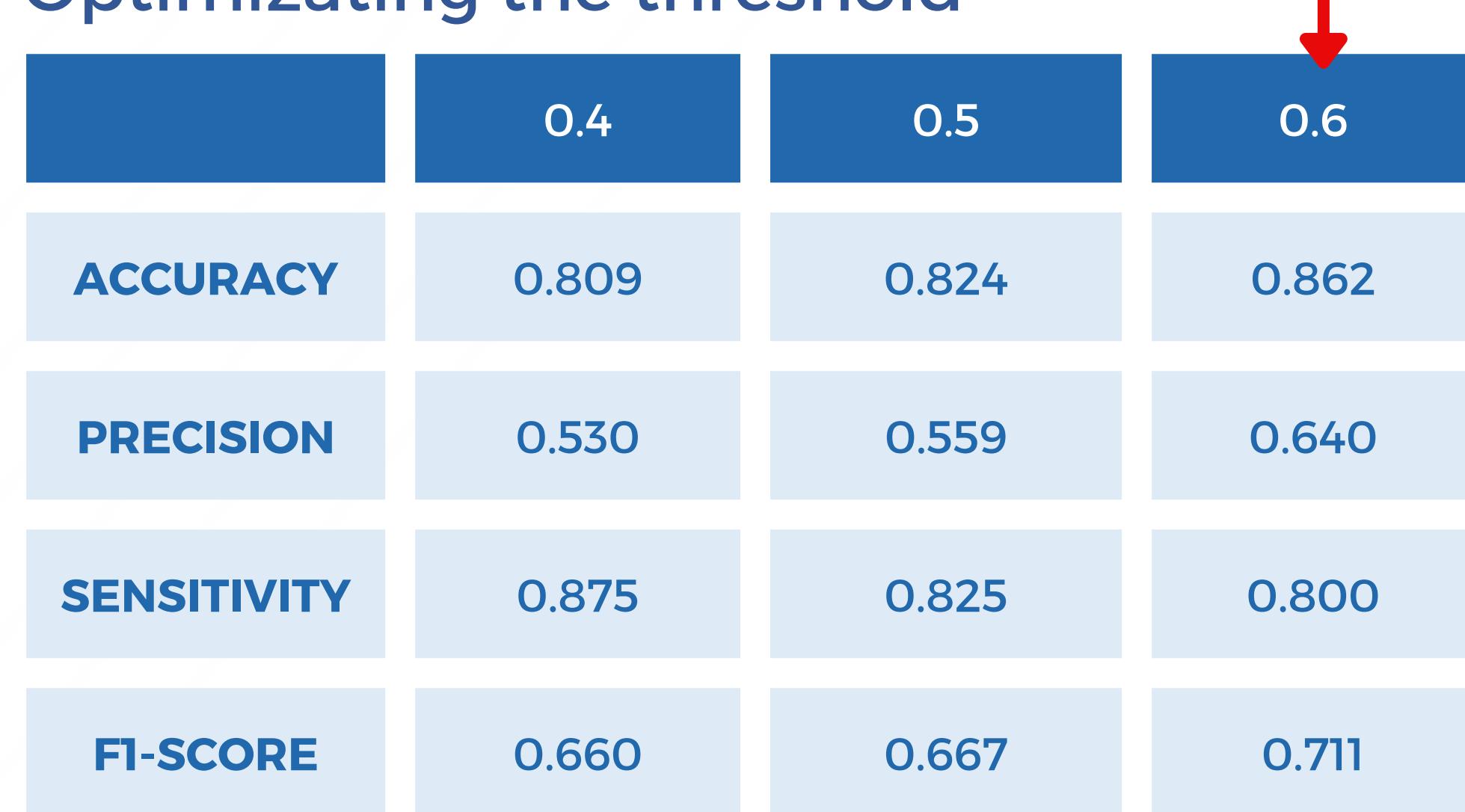
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-41.93796	15.88063	-2.641	0.008270	**
T_max	-0.21235	0.02662	-7.977	1.50e-15	***
Visibility	-0.13698	0.02598	-5.273	1.34e-07	***
Wind_speed_med	-0.36897	0.05822	-6.337	2.34e-10	***
Pressure	0.04662	0.01551	3.005	0.002654	**
PhenomenaFog	-1.43224	0.38665	-3.704	0.000212	***
PhenomenaRain	-1.66530	0.44178	-3.770	0.000164	***



Logistic Regression: undersampled dataset

		TRUE	
		0	1
		0	130
PRED	0	130	8
1	18	32	

- The results obtained in terms of Recall on the original dataset are poor
- Fitting the model on the undersampled dataset with Backward Stepwise Selection successfully decreases the rate of FN
- Optimizing the threshold



Logistic Regression with interactions: undersampled dataset

ACCURACY	0.819
PRECISION	0.545
SENSITIVITY	0.900
F1-SCORE	0.679

- Main effects: features selected by the Backward Stepwise procedure
- Interaction effects: second order interactions between the variables kept, third and fourth order interactions between numerical variables
- After fitting the model, Backward Stepwise selection is again performed with AIC
- **Best threshold:** 0.4

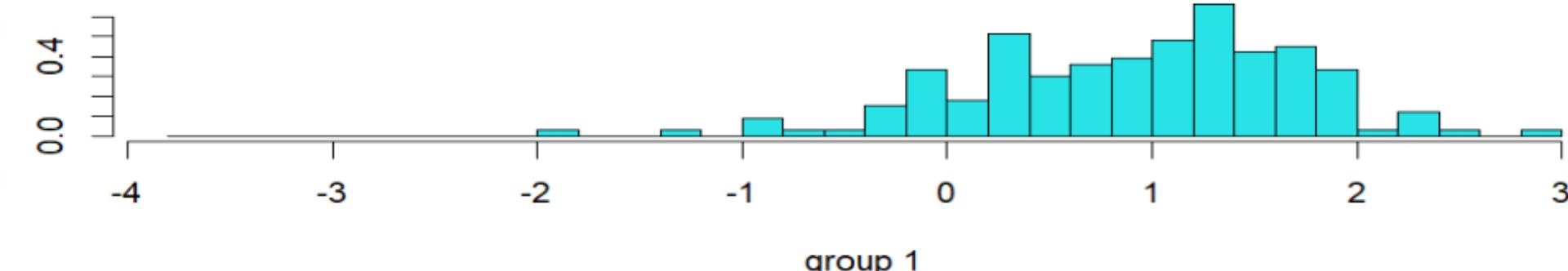
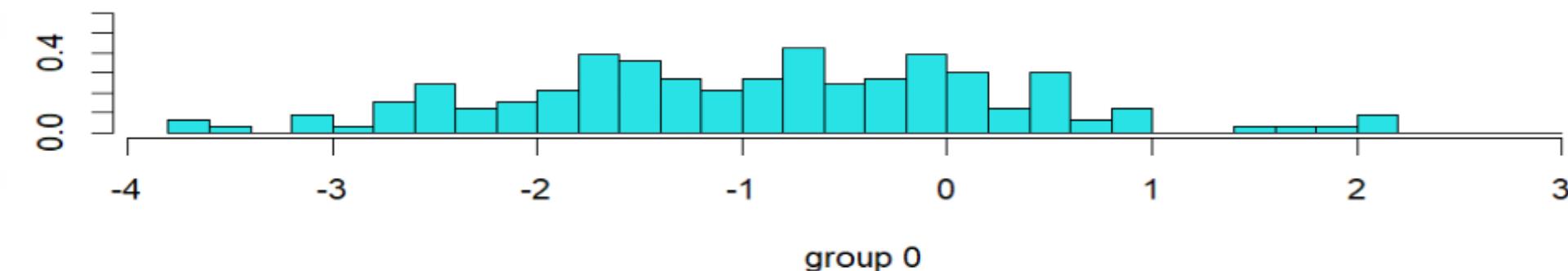
Coefficients :

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	236.852331	176.762778	1.340	0.18026
Visibility	-25.793239	11.344209	-2.274	0.02298 *
Pressure	-0.228198	0.173591	-1.315	0.18865
Wind_speed_med	-27.334825	21.607728	-1.265	0.20585
T_max	-9.369122	6.136833	-1.527	0.12683
PhenomenaFog	-1.352649	0.525970	-2.572	0.01012 *
PhenomenaRain	-1.559590	0.563537	-2.768	0.00565 **
Visibility:Pressure	0.025306	0.011144	2.271	0.02316 *
Visibility:Wind_speed_med	2.482656	1.307560	1.899	0.05760 .
Pressure:Wind_speed_med	0.026679	0.021275	1.254	0.20984
Visibility:T_max	-0.016235	0.006100	-2.661	0.00778 **
Pressure:T_max	0.009204	0.006030	1.526	0.12696
Visibility:Pressure:Wind_speed_med	-0.002447	0.001287	-1.901	0.05729 .

Linear Discriminant Analysis (LDA)

ACCURACY	0.830
PRECISION	0.565
SENSITIVITY	0.875
F1-SCORE	0.686

- All the following models are fitted on the undersampled dataset and on the features selected by the backward stepwise procedure since the achieved performances are better
- Normality assumptions do not hold
- **Best threshold:** 0.3



Quadratic Discriminant Analysis (QDA)

ACCURACY	0.835
PRECISION	0.576
SENSITIVITY	0.850
F1-SCORE	0.687

- Assumption: each class has its own covariance matrix (less restrictive than LDA)
- Quadratic decision boundary
- **Best threshold:** 0.5

		TRUE	
		0	1
PRED	0	123	6
	1	25	34

Naive Bayes

ACCURACY	0.835
PRECISION	0.579
SENSITIVITY	0.825
F1-SCORE	0.680

- Independence assumption of the covariates in each class fails
- For the quantitative variables, it's like applying QDA with the additionally assumption of independence
- For categorical variables, the distribution of each predictor in the specific class can be computed considering the specific proportion
- **Best threshold:** 0.5

		TRUE	
		0	1
PRED	0	123	6
	1	25	34

Ridge regression

ACCURACY

0.835

PRECISION

0.576

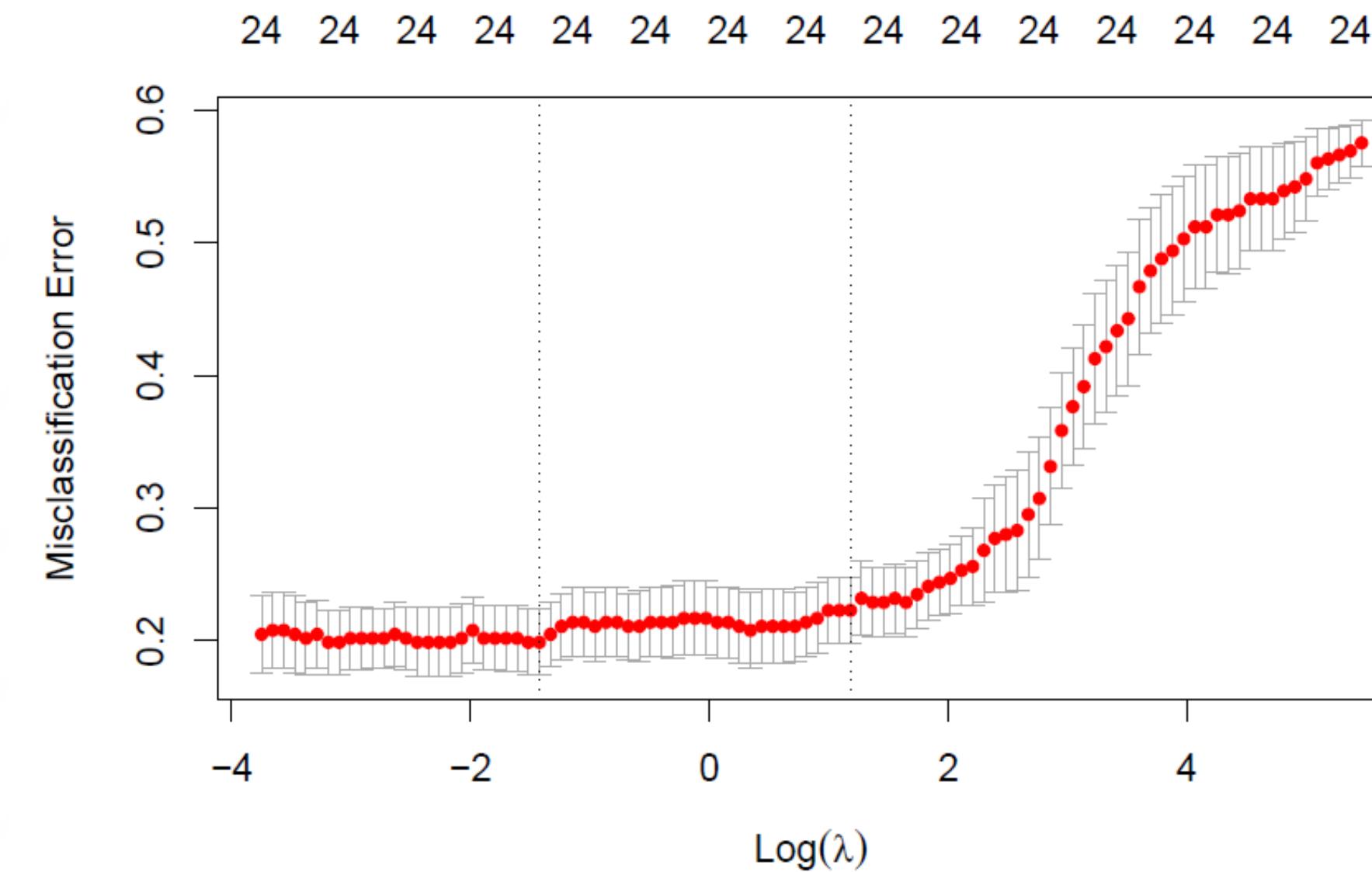
SENSITIVITY

0.850

F1-SCORE

0.687

- All features are kept
- Shrinking of the coefficients towards 0
- L2 penalization
- Best λ : 0.241



Lasso regression

ACCURACY

0.814

PRECISION

0.541

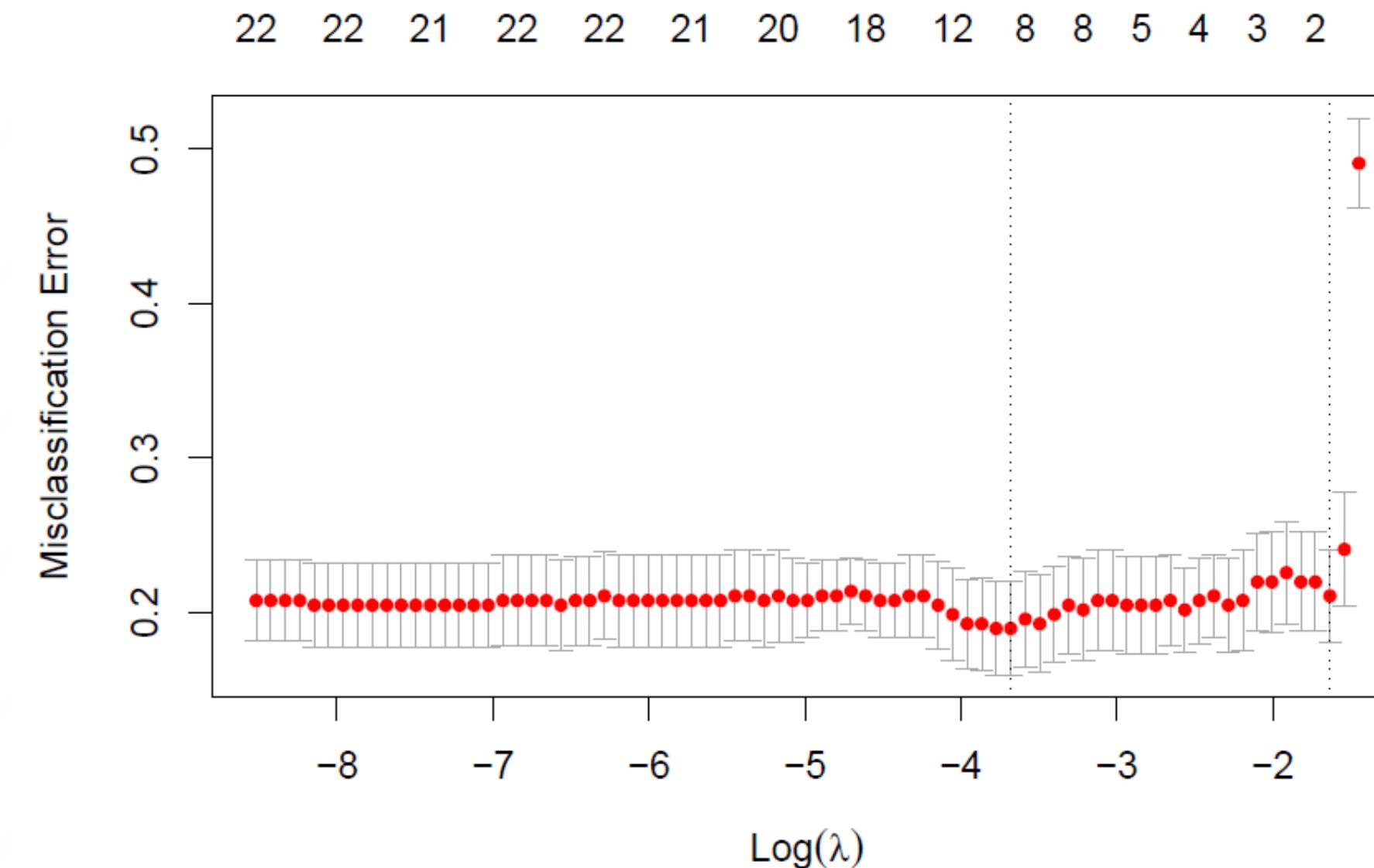
SENSITIVITY

0.825

F1-SCORE

0.653

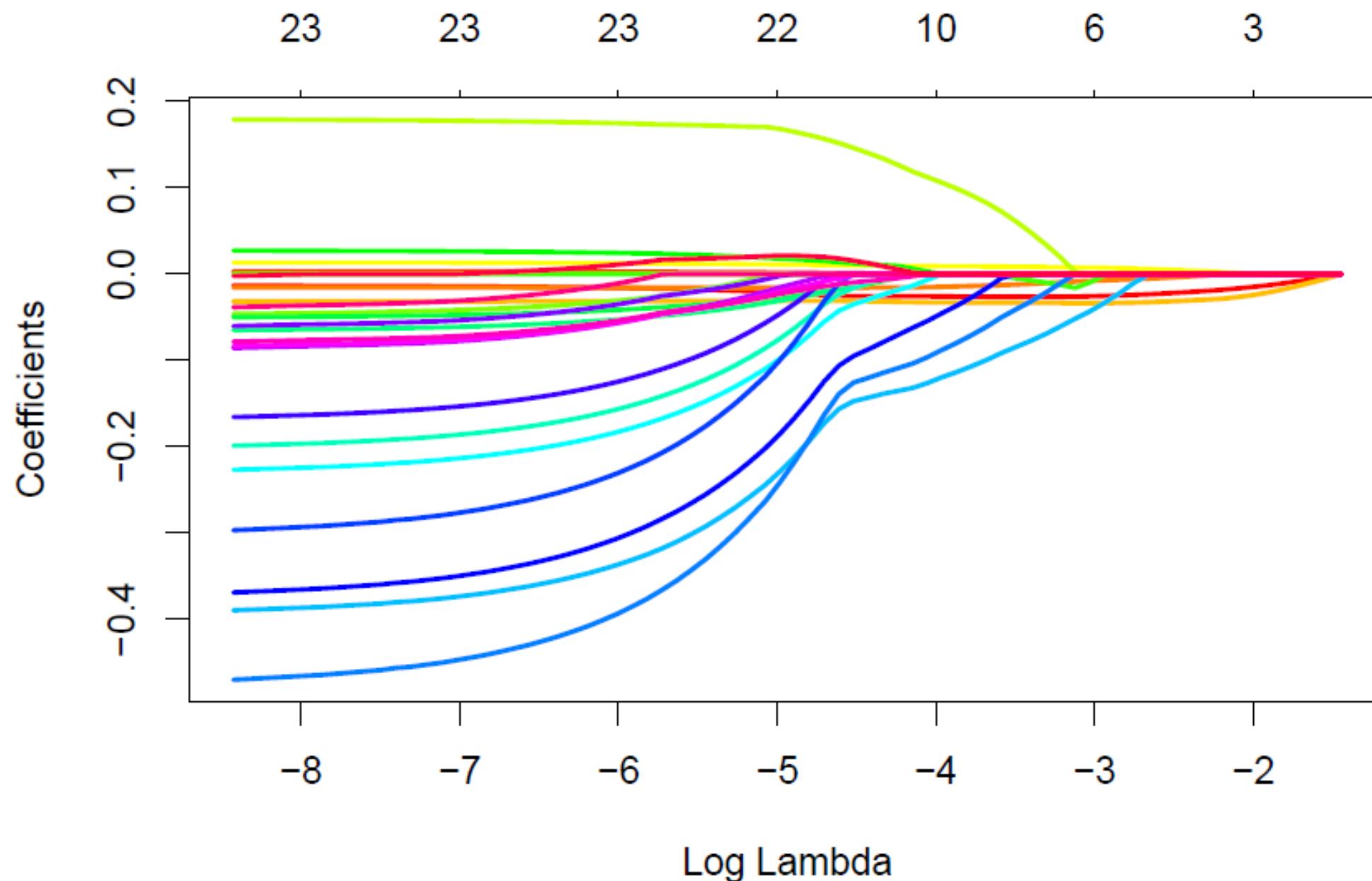
- It perform feature selection: non significant features are set exactly equal to 0
- L1 penalization
- Best λ : 0.025



Lasso regression



The most significant predictors are similar to the one kept by the backward stepwise process:
T_max, Visibility, Wind_speed_med, Pressure, PhenomenaRain, Month4, Month9, Month11.



K-Nearest Neighbors (KNN)

ACCURACY	0.809
PRECISION	0.533
SENSITIVITY	0.800
F1-SCORE	0.640

- Non-parametric model
- Since the dataset contain numerical and categorical features, traditional distance metrics are not suitable.
- Solution: **Gower distance**

$$\text{Gower}(X, Y) = \frac{\sum_{i=1}^n w_i \cdot d(x_i, y_i)}{\sum_{i=1}^n w_i}$$

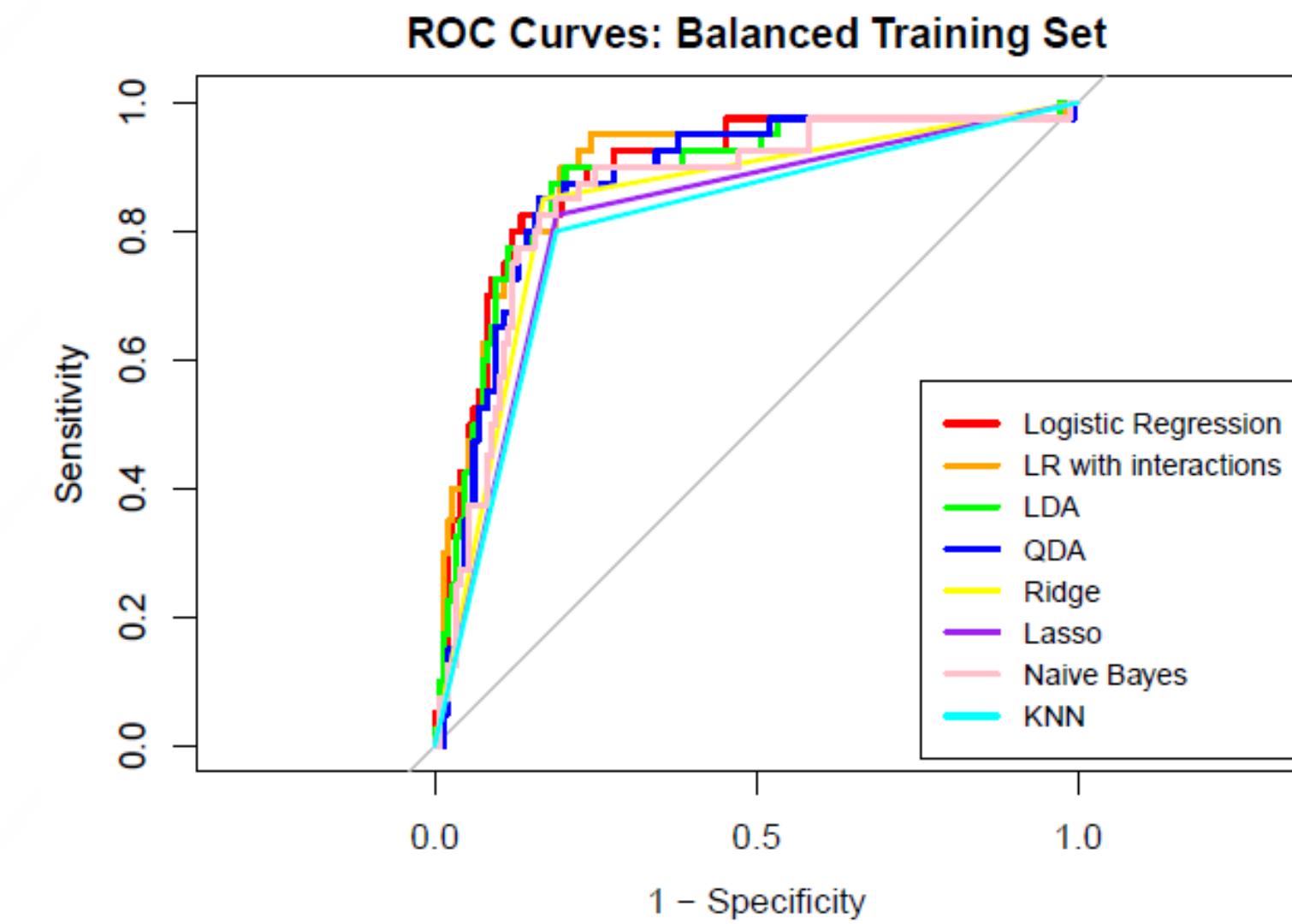
- For numeric variables, Gower distance uses a normalized L1-norm
- For categorical covariates, it uses a simple matching coefficient (0/1): same level implies 0 distance, 1 otherwise
- **Number of neighbors K = 3**

MODEL COMPARISON

UNDERSAMPLED DATASET

- All the models have achieved good performances, each one with their strengths and weaknesses
- Best overall model:
 - Logistic Regression with 0.6 as threshold
- Models with highest Sensitivity and AUC:
 - Logistic Regression with interactions
 - LDA

Model	Accuracy	Recall	Precision	F1_score	AUC
Logistic Regression	0.862	0.800	0.640	0.711	0.888
Logistic with interactions	0.819	0.900	0.545	0.679	0.892
LDA	0.830	0.875	0.565	0.686	0.879
QDA	0.835	0.850	0.576	0.687	0.872
Ridge	0.835	0.850	0.576	0.687	0.841
Lasso	0.814	0.825	0.541	0.653	0.818
Naive Bayes	0.835	0.825	0.579	0.680	0.856
KNN	0.809	0.800	0.533	0.640	0.805



CONCLUSIONS

Meteorological conditions can accurately predict whether the PM10 particulate levels in the air will be high or low in the following day

The best predictors are:

- **Phenomena:** the presence of rain and fog tends to lower the likelihood of high PM10
- **Visibility, T_max and Wind_speed_med.** High values exhibit a negative influence on this probability.
- **Pressure.** High pressure systems increases the probability of having bad air quality the next day.

