# Capstone Project, Choose your own project (CYO), Air quality in PM10 levels after rain in Budapest, Hungary, from 2003 until 2012

Navarro H Daniel D

2024-Mar-13

# Contents

This is the second part of the Capstone project for the Data Science course by HarvardX, whose tasks is to prepare a Machine Learning project of my own choice, no major indications about the project to choose were given except the expected level of complexity going beyond a simple linear regression model, and the grading constraints.

As a guideline for the project selection two websites with ideas and data sets were provided, [UCI Machine Learning Repository] (https://archive.ics.uci.edu/ml/index.php) and Kaggle, however I was concerned about the substance of the project in terms of real application, immediate and direct impact and geographical location.

After a long search for a good theme and data disponibility combination, I was sure that I preferred climate related projects due to the importance and impact of the subject in everyday life, I ended working with information from the capital of Hungary, city of Budapest.

It would have been better to work with more local data, but there is not major data about the small town where I live, however the capital is 10Km away and the best reference obtained.

Some information was provided by the governmental meteorological agency www.met.hu and air quality information obtained from the European Environment Agency.

I used the information of the station north of the city https://odp.met.hu/climate/observations_hungary/daily_rain/historical/HABP_1RD_34413_20020101_20231231_hist.zip.

Still after collecting information with direct impact for me, matching the formats and time frame of both data needed a bit of data wrangling that resulted in a short time frame of 10 years of data collected in a daily basis from 2003 January 1st until 2012 December 31st.

Motivation

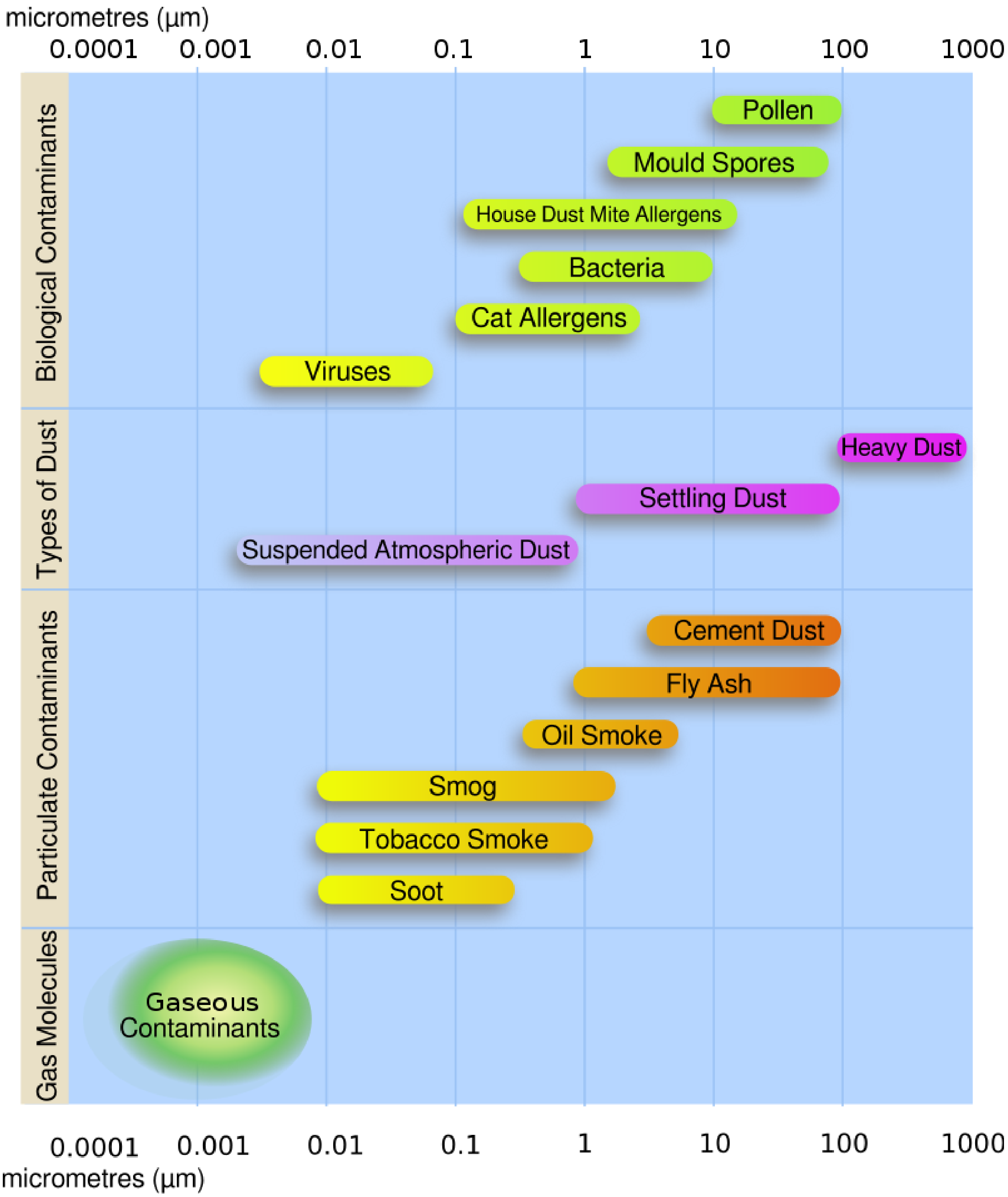
Air pollution is a variable of global concern nowadays as the world's weather system is getting unstable and new/still dynamically changing atmospheric conditions are affecting all life on earth. There is a systematic self feeding link between pollution and weather as air pollution contributes to the greenhouse effect who at the same time helps decrease the quality of the air.

Contaminated air is causing millions of death around the world, specially for most susceptible population (Classified as Sensitive Population) like children, weak people and elderly, the situation is similar around the world but specially impacting some populated cities in Asia.

The quality of the air of Budapest, with more than 1 500 000 inhabitants, ranks as 256 of 375 cities controlled in Europe with air quality qualified as moderate quality, 13.0 ug/m3 according to measures between 2022 and 2023. More information at https://www.eea.europa.eu/themes/air/urban-air-quality/european-city-air-quality-viewer.

Moderate air quality qualification is in the range of 10 and to 15 $\mu g/m^3$, the European Union set an annual limit of 25 $\mu g/m^3$.

As it can be seeing in the following image in the PM10 micrometers category are included several common pollutants, thus we kept it as a good indicator of the air quality.



Measures of concentration of PM10 in the air is one of the important indicators of air quality due to the capacity of such small particulates to enter into the human body, a PM10 particle is 1/5 until 1/7 part of a human hair diameter, is reduced to the level of 1/9 of a beach sand grain, such small elements can enter into the body pores and pass to the blood circulation.

There are more dangerous contaminants and smaller particle than PM10, however for the scope of this project it was decided to work with PM10 as a central indicator of quality of the air following level of rain.

More about the subject in Wikipedia.

For the population group sensitive to bad air quality due to physical health conditions it is often advised to avoid outside activities during bad air quality days.

This is why it is important to be able to forecast pollution level rather than just submit warnings after the pollution levels are high.

Rain is known to clean the air from pollutants, "As a raindrop falls through the atmosphere, it can attract tens to hundreds of tiny aerosol particles to its surface before hitting the ground. The process by which droplets and aerosols attract is coagulation, a natural phenomenon that can act to clear the air of pollutants like soot, sulfates, and organic particles." Source

With actual technology rain can be predicted with somehow good accuracy and some days of anticipation, this is why the intention is to predict a non-predictable variable, air pollution, after a predictable variable, rain.

## Goal

To generate a model that predicts the expected air quality level in terms of PM10 particulates concentration in the air based on the given rain level in mm.

A machine learning algorithm that, starting from the information about raining prediction provided by the meteorological services, helps predict the levels of PM10 pollutants concentration expected in the air.

With satellite systems, weather information and mathematical prediction models, meteorological services predict and publish the rain forecast, however contamination cannot be predicted this way and is only informed in a historical basis after registers have been collected in several stations. If we use rain predictions in a accurate way we can also have a prediction of the pollution level as well, information that is of importance for the citizens when taking care or preventing exposition to the contamination is important.

# 1 Data preparation

A match between a significant subject and proper data took several days of internet search and selection, some valuable data could not be retrieved from websites in enough quantity and quality to produce a report, other data was not valuable enough.

Due to local limitations and lack of relevance of the country in the most known databases I have to work with two separate data sets, weather historical registry from the national meteorological service in Hungary MET and pollution historical registry from the EEA, even this two data sets were not a perfect match and required more data wrangling.

The meteorological data was extensive, well collected with a frequency of daily register from the years 2002 until 2023, you can see more at https://odp.met.hu/climate/observations_hungary/daily_rain/historical/.

The pollution data from the EEA is somewhat disparate, split into different periods and formats, more difficult to follow and not really well detailed, with a download form that can be found at https://eeadmz1-downloads-webapp.azurewebsites.net/. The data comes extremely split in several files and not very well identified.

Data is also provided in .parquet format for which the "arrow" package is needed in this project to open the data.

As the time frame of the pollution data is 2003-01-02 until 2013-01-01 in a daily basis the study period was limited to these years, the meteorological data was good enough for this small period and it is still a good base for the project to provide insights into the study and further possible analysis. I also removed extra weather stations of the data set leaving only the capital city, Budapest, into consideration.

The final data went from 1 465 600 rows to 3 654 observations that covered daily registers in the period of 2003 until 2013.

## 1.2 Linear regression and other models

# 2 Initial exploration

Minimum values for rain registered by the meteorological authority between 2003 and 2012 is 1.79 with maximum of 82 with standard deviation of 4.94.

Minimum values for PM10 particulate in the air registered between 2003 and 2012 is 31.39 with maximum of 236.1 with standard deviation of 23.45.

## 2.1 Checking the quality of the data

Final table is a 3 653 observations data frame with 3 numerical variables, a discrete: date, and two continuous: rain in millimeters and pollution PM10 particles in micro grams per cubic meter. Date is formatted as yyyymmdd; year, month day.

```r
dim(workingData)
```

```
## [1] 3653    3
```

```r
str(workingData)
```

```
## 'data.frame':    3653 obs. of  3 variables:
##  $ Date   : num  2e+07 2e+07 2e+07 2e+07 2e+07 ...
##  $ Rain_mm: num  0.1 6.7 0 1.4 0.8 1.8 15.3 0 2.6 3.8 ...
##  $ PM10   : num  -1 -1 -1 -1 -1 ...
```

After reducing the data to the matching number of observations between the two data sets I check the quality of the data.

```
sum(!complete.cases(workingData))
```

```
## [1] 0
```

```
colSums(is.na(workingData))
```

```
##    Date Rain_mm    PM10 
##       0       0       0
```

```
sum(is.na(workingData))
```

```
## [1] 0
```

The resulting data set of the weather reports has not empty nor invalid values.

```
head(workingData)
```

```
##       Date Rain_mm PM10
## 1 20030101     0.1   -1
## 2 20030102     6.7   -1
## 3 20030103     0.0   -1
## 4 20030104     1.4   -1
## 5 20030105     0.8   -1
## 6 20030106     1.8   -1
```

Rain level in mm is acts as the predictor and PM10 is the outcome or response that we want to predict for the future (extrapolation).

## 2.2 Maximum registered rain volume in the period

```
max(as.numeric(workingData$Rain_mm))
```

```
## [1] 82
```

The maximum registered rain volume was 82 mm registered in 2010-May-15.

```
##          Date Rain_mm PM10
## 2692 20100515      82   11
```

And a high level of PM10 was reached that day, 32 $\mu g/m^3$.

### 2.3 Maximum concentration of PM10 per cubic meter in the period

```
max(as.numeric(workingData$PM10))
```

```
## [1] 236.1
```

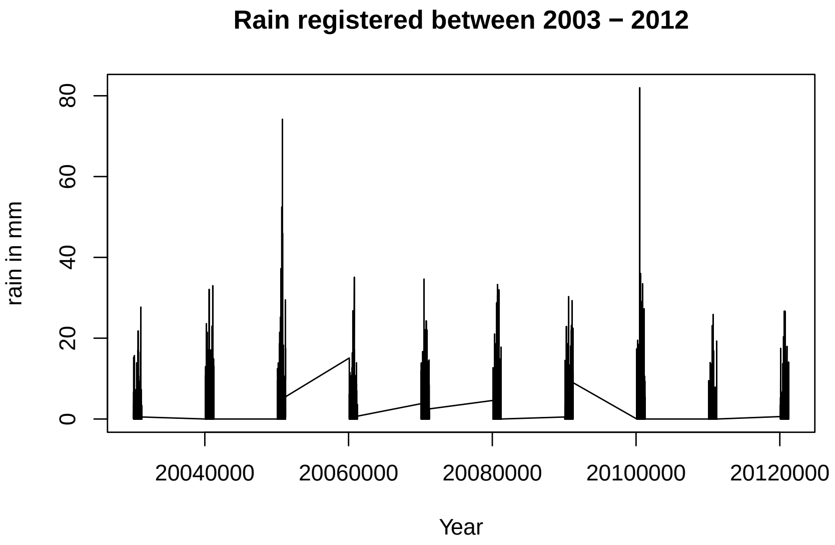
The maximum registered concentration of PM10 in the study period is 236.1 $\mu$, way above the set level of the European union of 25 $\mu g/m^3$.

```
##         Date Rain_mm  PM10
## 773 20050211     2.5 236.1
```

# 3 Data visualization

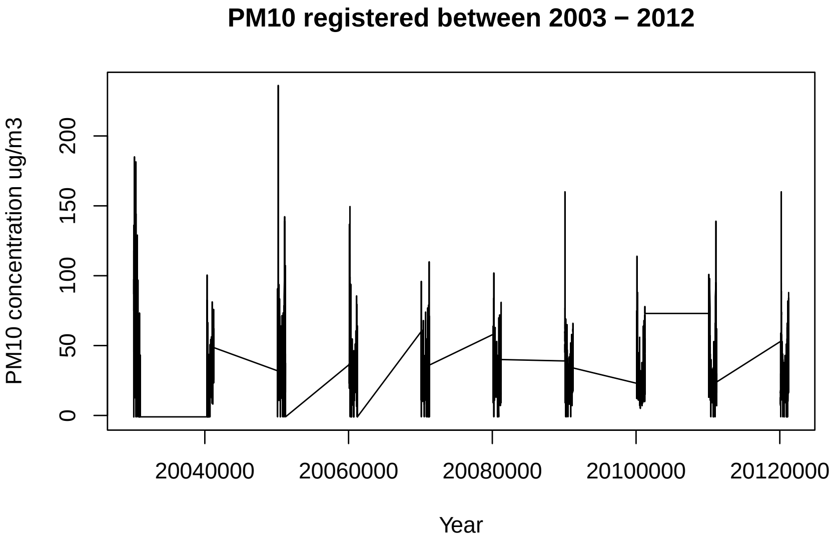
We can see rain levels increasing moderately every year from 2003 until 2012.

```
plot(workingData$Date,workingData$Rain_mm, main = "Rain registered between 2003 - 2012", xlab = "Year",
```

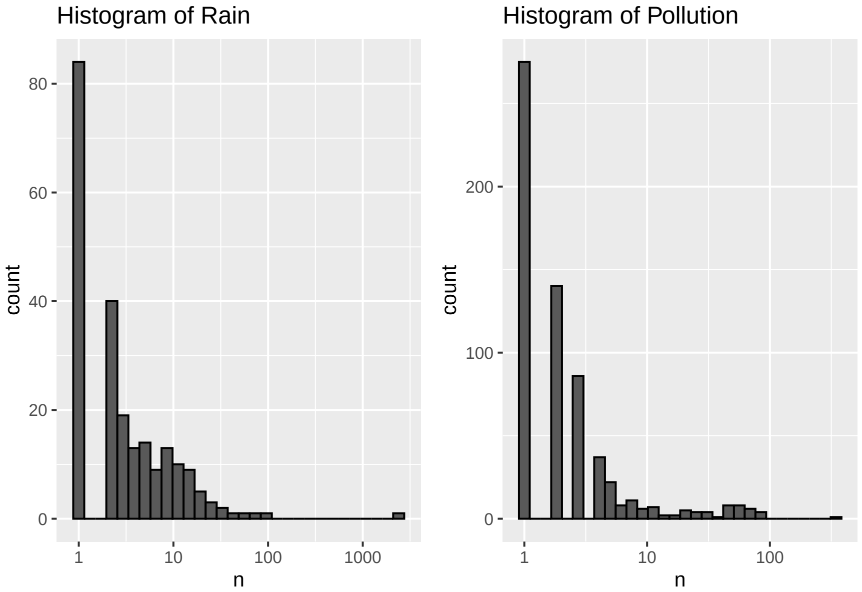


PM10 pollution has, on the other hand, varied from year to year.

```
plot(workingData$Date,workingData$PM10,main = "PM10 registered between 2003 - 2012", xlab = "Year", ylal
```

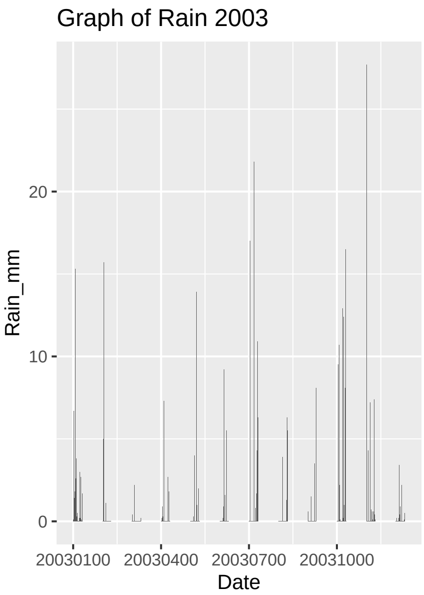


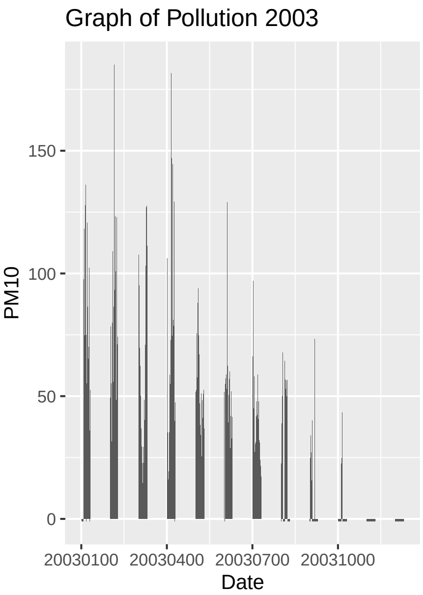
For a better calibration of our intuition of the relation between rain level and pollution in the air we shall match both parameters in a graph



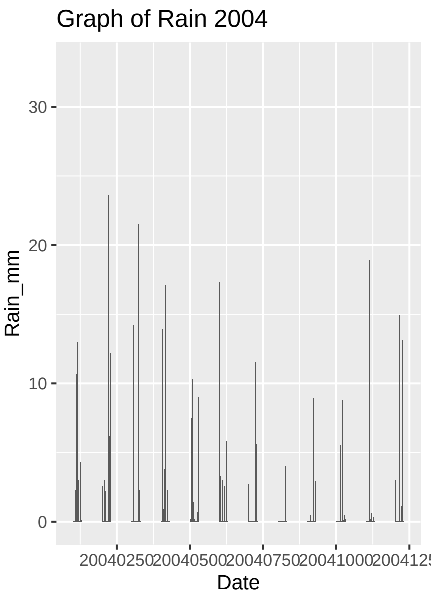
However this frequency graph is not depicting the relation of the variables.

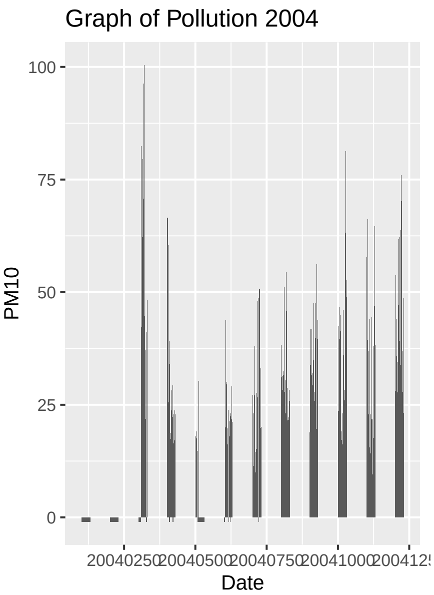
When comparing side by side only one year levels we can see the relation is rather poor as pollution goes down during November and December, however some patter can be observe during spring and summer where rain is low and pollution concentration high.





Preparing the same comparison for 2004 we can see:





Pollution in January and February 2004, following the previous observed November and December 2003 is also minimal, was is a one off event? Is there a problem with the data? Modeling will need to work with this possible mismatch.

# 4 Modeling the algorithm

## 4.1 Linear Regression model

The first approach is to use the simple linear regression model of considering that the expected pollution is the mean of the pollution, in order to have a idea of the model possibility and complexity.

```
## # A tibble: 1 x 2
##   method        RMSE
##   <chr>        <dbl>
## 1 Simple Model  23.5
```

The simple regression model give a high Root-Mean-Standard-Error, meaning the difference between an expected value and the real value in average is extreme.

This result is due to the high variability in PM10 pollutants concentration between high and low days, making it difficult to predict a level of pollution by just the average contamination level.

This machine learning model presents a challenge in the sense that the prediction is a continuous variable and we are using only one predictor.

### 4.2 More complex algorithms with caret

Other algorithms were tested as well, however some data quality defect made impossible to run the confusion matrix thus ignoring the accuracy of the models, to avoid models using the Date as one of the predictors it was removed from the training and test sets used.

According to Max Kuhn and Kjell Johnson, Applied Predictive Modeling, 2016:

> There are potential advantages to removing predictors prior to modeling. First, fewer predictors means decreased computational time and complexity. Second, if two predictors are highly correlated, this implies that they are44 3 Data Pre-processing measuring the same underlying information. Removing one should not compromise the performance of the model and might lead to a more parsimonious and interpretable model. Third, some models can be crippled by predictors with degenerate distributions. In these cases, there can be a significant improvement in model performance and/or stability without the problematic variables.

```
## # A tibble: 4 x 2
##   method        RMSE
##   <chr>        <dbl>
## 1 Simple Model  23.5
## 2 kknn          35.7
## 3 rpart         23.1
## 4 bagEarth      22.9
```

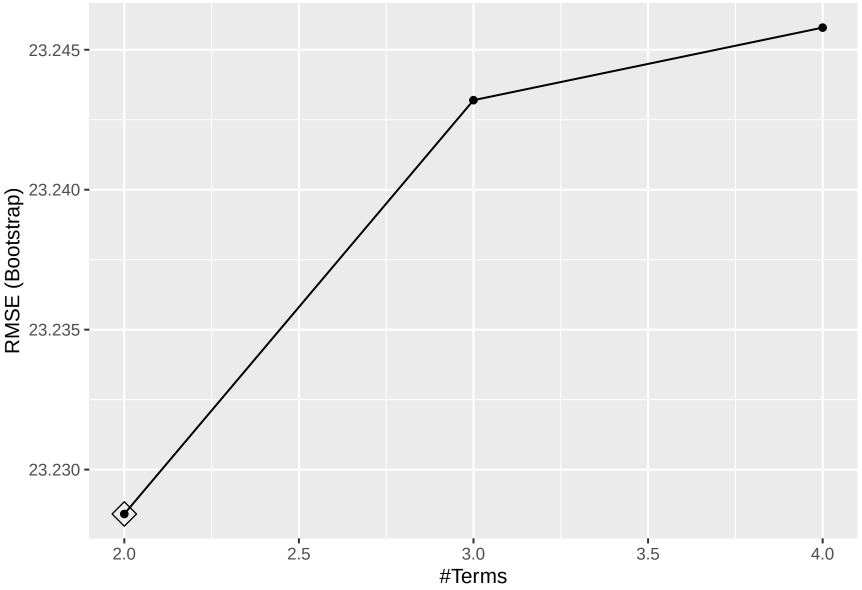
After testing 5 different algorithms with not noticeable difference in their results: glm, kknn, rpart, bagEarth and glmboost, the more accurate were kept: kknn, rpart and bagEarth.

From the last two we kept bagEarth with RMSE of 23.06.

```
train_bagEarth$finalModel
```

```
##
## Call:
## (function (x, y, weights = NULL, B = 50, summary = mean, keepX = TRUE,     ...) {    requireNamespac
##
## Data:
##    # variables:   1
##    # samples:     2920
## case weights used
##
## B: 50
```

```r
ggplot(train_bagEarth, highlight = TRUE)
```



Some other models were tested with errors, by example Loess, however to avoid an Error Type III by loosing sight of the goal of the project and focusing in the technicalities, we keep the previously mentioned.

# 5 Results

Working with real world data that comes from different collections determine the possibilities of the algorithm. In the actual project, although after data wrangling the format and time coverage was stable, having a short period of 10 years and only one predictor impacted the algorithm accuracy negatively.

The accuracy measures in term of RMSE or error from the prediction of complex algorithm of caret package did not really improve much over the simple mean of the outcome, leaving clear that pattern between rain and pollution was not enough, pollution is too variable and more data is needed.

One consideration that was out of the scope of this essay is the time delay between rains and pollution levels, worth to be further investigated in search of possible patterns that cannot be seeing when checking one by one rain and pollution in the same dates.

Still intuition indicates rain and clean air have a relation that, might not be direct but through any other ignored elements that link both the pretended predictor and outcome. Although rain is a very commonly used weather indicator, there is plenty of many other indicators like wind, humidity, atmospheric pressure that can be added to the model for future investigations.

Still this project never pretended to state that low pollution was caused by rain levels but, that my be a causality relation between both events.

Various techniques were used for the present project that can be resumed in the following:

```r
rmse_results|>  knitr::kable()
```

| method | RMSE |
|---|---|
| Simple Model | 23.46386 |
| kknn | 35.69290 |
| rpart | 23.12566 |
| bagEarth | 22.88759 |

The accuracy in terms of the error from the prediction of the algorithm to the real value is not satisfactory, thus it was necessary to try several models in search of a better result, internet search of guidance in this respect did not really helped.

Thus this project leads to the following conclusions.

# 6 Conclusion

Weather related prediction has long proved to be a difficult subject due to various factors, mainly the complexity of weather as a system of different elements and their relations, including human activity.

The main goal in this project was to predict a pollution variable, PM10 concentration in the air, starting from the prediction of a weather variable, rain, in the city of Budapest, Hungary.

The availability of data was a challenge and having a final subset of data required some work.

However the project gives light to other paths into more investigation and different approach to the same goal given its impact con everyday life on earth. The living, economical and health effect of being able to predict pollution and establish sound prevention alarms is worth the investigation.

New approaches can include, between others, a check of the time delay between rain and pollution levels, the inclusion of other predictors into the model, the test of the evaluation of other pollutants different from PM10 and the test of other algorithm that might be more suitable for weather related projects.