Statistics for Artificial Intelligence: Assignment 5

Marcel Aguilar Garcia, Id: 20235620

November 30, 2021

The aim of this report is to analyse winter time (May-August) weather and air pollution data from Christchurch to answer the questions of interest. The dataset that has been used contains weather related features and the target variable, PM_{10}, which is a measure of the daily average particulate matter concentrations below 10 micrometers.

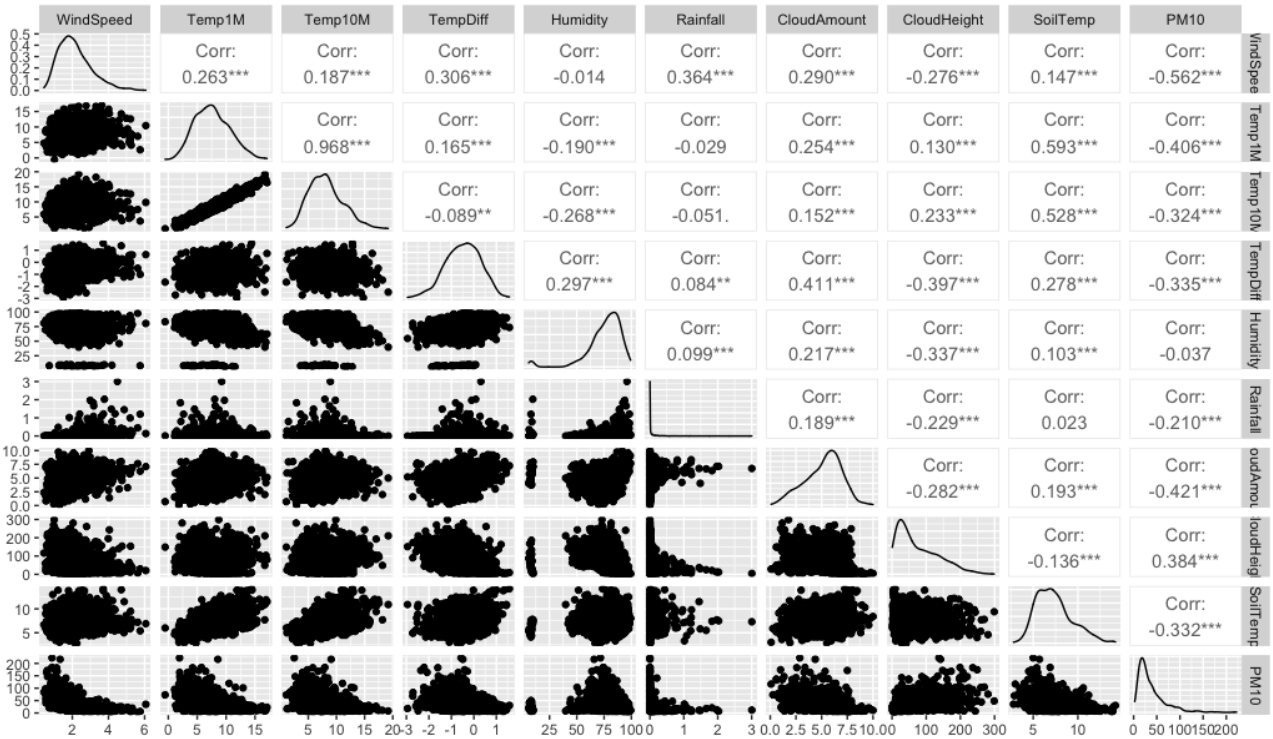
Question of Interest

In this report, I will answer to the two questions of interest which are:

1. Have the government intervention schemes been effective in reducing the pollution level?
2. How does the meteorology effect the concentration of PM_{10}?

Subjective Impressions or Exploratory Analysis

Let's start exploring the correlation between all features:



Correlation Features vs Target

As seen in the table above, there are many features that are correlated to the target. On one side, CloudHeight (0.384) is the only feature that has a positive linear association to the target PM_{10}. On

the other hand, WindSpeed (-0.562), CloudAmount (-0.421), and Temp1 (-0.406), are the top three features having a negative correlation to PM_{10}.

The two features with the lowest correlation are Rainfall (-0.210) and Humidity (-0.037). While they do not show a linear association to PM_{10}, by looking at the scatterplots, we can see a visual pattern. Humidity seems to have have a quadratic association to the target and, in general, rainy days seem to have lower values of PM_{10}.

Overall, WindSpeed is the feature that has the strongest correlation with PM_{10}. For this reason, we can expect this feature to be relevant during the Formal Analysis.

Correlation Between Features

When interpreting the results from the model, we should take into account correlations between features. This is the case for Temp1M and Temp10M which have a strong linear correlation of 0.968. In a similar way, SoilTemp is correlated to Temp1M and Temp10M with correlations of 0.593 and 0.528, respectively.

Others

Note that PM_{10} appears to have a left-skewed distribution. For this reason, a logarithmic transformation will be used during the Formal Analysis.

Formal Analysis

Government intervention schemes and pollution level

In order to assess the government intervention schemes, I have fitted a simplistic straight line using the variable dateindex in order to see if PM_{10} reduces over time. The the linear equation representing this SLM is:

$$\hat{y} = 131.57 - 0.064 \cdot \text{dateindex} \tag{1}$$

- F-statistic of 70.72 and a p-value of 2.2e-16 indicates that there is a strong evidence for relationship between the dateindex variable and logPM_{10}.
- The p-value for dateindex is 2e-16 and so dateindex is deemed a useful predictor for logPM_{10}.
- $\mathcal{R}^2 = 0.0584$ which means that the explanatory variable dateindex explains only a 5.84% of the variability for logPM_{10}.

As this model has a significant p-value but has such a small $\mathcal{R}^2$, we can conclude that dataindex still provides information about the response variable even though data points fall further from the regression line. Additionally, the coefficient of dataindex is negative, which implies that PM_{10} decreases over time. The visualisation below shows logPM_{10} through time with a smooth line that helps to visualise the trend:

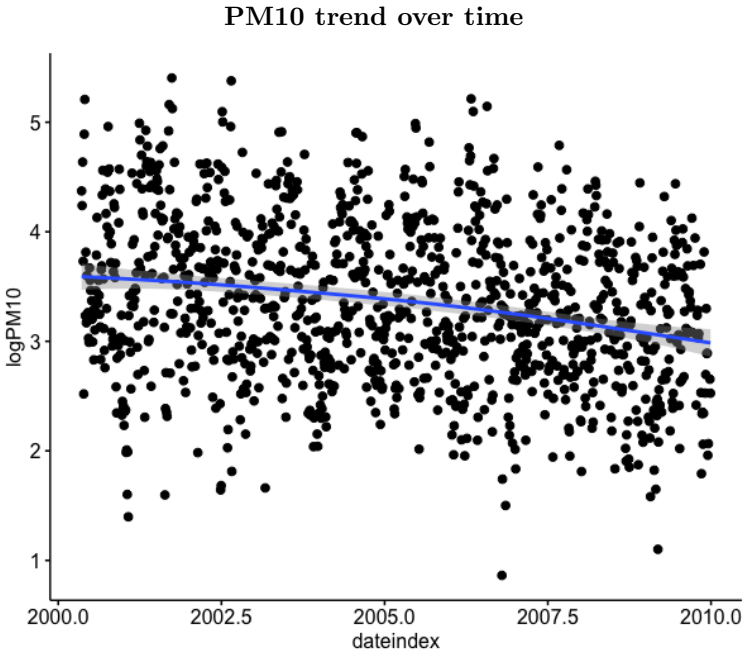


Figure 1: This plot shows the trend of logPM10 over time by using dateindex

Meteorology and pollution level

The weights of a Multiple Linear Regression model can be used to have a better understanding of the features and the target. For this reason, I have trained a multiple linear regression and I have analysed the results. Using all variables do not always lead to better results. So first, I have used different methods to help me assess which subset of features is more optimal.

All features

In this method, I have fit a multiple linear regression model with all meteorological data. The equation of this model is given by:

$$\hat{y} = 4.77 - 0.35 \cdot \text{WindSpeed} - 0.040 \cdot \text{Temp1M} + 0.005 \cdot \text{Temp10M} + 0.014 \cdot \text{TempDiff} + 0.002 \cdot \text{Humidity} \\ -0.169 \cdot \text{Rainfall} - 0.0787 \cdot \text{CloudAmount} + 0.0031 \cdot \text{CloudHeight} - 0.043 \cdot \text{SoilTemp} \tag{2}$$

Model Summary:

- F-statistic of 187.2 and a p-value of 2.2e-16 indicates that there is a strong evidence for relationship between the meteorological variables and logPM_{10}.
- A $\mathcal{R}^2 = 0.5979$ tells us that the meteorological variables explain 59.79% of the variability of logPM_{10}.
- WindSpeed (2e-16), Humidity (0.0123), Rainfall (0.0195), CloudAmount (2e-16), CloudHeight (2e-16), and SoilTemp (2.67e-06) have got an acceptable p value (≤ 0.05). However Temp1M, Temp10M, and TempDiff are not particularly useful in this model.
- There seems to be evidence of multicollinearity:
    1. While F-statistic indicates that the model is useful, not all parameters have p-values that are significant.
    2. As seen in the exploratory analysis, some feature variables are strongly correlated between them (e.g. Temp1M and Temp10M have a correlation of 0.968).

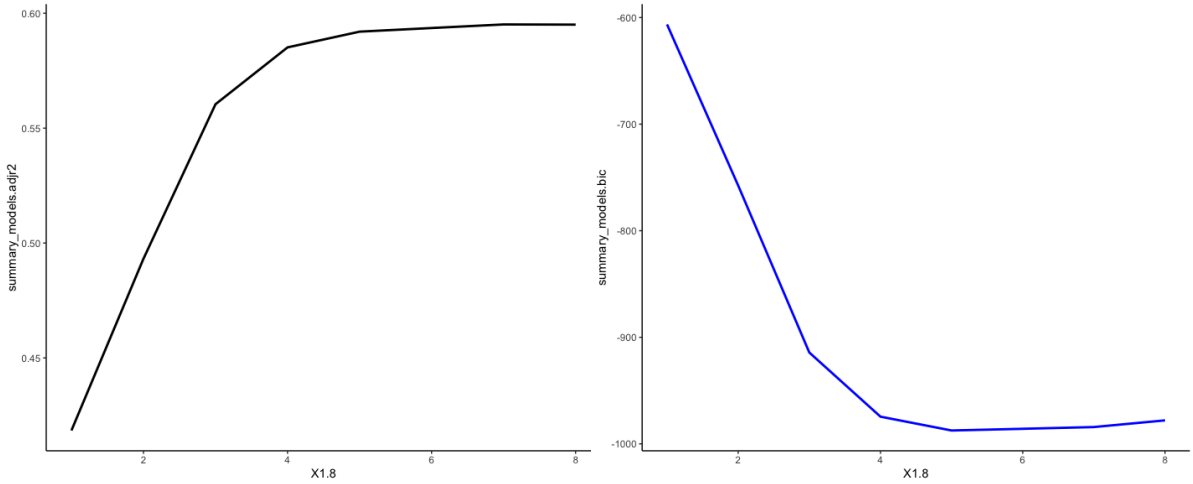
Because of possible multicollinearity issues, this model can not be used to interpret how all meteorological variables impact PM_{10}. For this reason, I have decide to test other methods that would help me to find a more relevant subset of features.

Best subsets regression

To start with, I have checked all possible combinations of subsets from 1 up to 8 elements (one less than the total number of variables). Additionally, I have used two different metrics, Adjusted R2 and BIC, to assess which subset had the best performance. As we would like to have a greater Adjusted R2 and a lower BIC, following the plots below, I have decided to use the best subset of four elements. This subset contains the following variables: **WindSpeed**, **Temp1**, **CloudAmount**, and **CloudHeight**.

The equation of the MLR with this subset is:

$$\hat{y} = 4.74 - 0.36 \cdot \text{WindSpeed} - 0.064 \cdot \text{Temp1M} - 0.072 \cdot \text{CloudAmount} + 0.0033 \cdot \text{CloudHeight} \quad (3)$$



(a) Adjusted R2 **(b)** BIC

Model Summary:

- F-statistic is 403.7 with a p-value of 2.2e-16 indicating that there is a strong evidence for relationship between this subset and logPM_{10}
- A $\mathcal{R}^2 = 0.5866$ tells us that this subset explain 58.66% of the variability of logPM_{10}.
- All features have a significant p-value: WindSpeed (2e-16), Temp1M (2e-16), CloudAmount (2.87e-16), CloudHeight (2e-16).
- There are no signs of multicollinearity

Backward Selection

By using backward selection, we start with a multiple linear regression model that has all features and, in each step, we remove a feature that, by removing it, provides the best performance to the model. This algorithm stops when by removing any feature, the model has worse performance. The subset returned by using this method is **WindSpeed**, **Temp1M**, **Humidity**, **Rainfall**, **CloudAmount**, **CloudHeight**, and **SoilTemp**.

The equation of the MLR with this subset is:

$$\begin{aligned}\hat{y} = 4.73 - 0.35 \cdot \text{WindSpeed} - 0.045 \cdot \text{Temp1M} + +0.0026 \cdot \text{Humidity} - 0.175 \cdot \text{Rainfall} \\ -0.077 \cdot \text{CloudAmount} + 0.0031 \cdot \text{CloudHeight} - 0.042 \cdot \text{SoilTemp}\end{aligned} \quad (4)$$

Model Summary:

- F-statistic of 240.8 and a p-value of $\leq$ 2.2e-16 indicates that there is a strong evidence for relationship between the meteorological variables and logPM_{10}.

- A $\mathcal{R}^2 = 0.5976$ tells us that the meteorological variables explain 59.76% of the variability of $\log PM_{10}$.
- All features have a significant p-value: WindSpeed (2e-16), Temp1M (2e-16), Humidity (0.00663), Rainfall (0.01436), CloudAmount (2e-16), CloudHeight (2e-16), and SoilTemp (3.74e-06)
- There are no signs of multicollinearity

Stepwise Backwards

In this case, using *Backward* or *Stepwise Backwards* leads to the same subset of features. Therefore, the model is exactly the same.

Forward Selection

By using forward selection, we start with a multiple linear regression model that does not contain any features and, in each step, we add the feature that gives the best performance to the model. This algorithm stops when, by adding any feature, the model has worse performance. The subset returned by using this method is **WindSpeed**, **CloudAmount**, **SoilTemp**, **CloudHeight**, **Temp10M**, **Humidity**, **Rainfall**.

The equation of the MLR with this subset is:

$$\begin{aligned}\hat{y} = 4.82 - 0.36 \cdot \text{WindSpeed} - 0.043 \cdot \text{Temp10M} + +0.0023 \cdot \text{Humidity} - 0.158 \cdot \text{Rainfall} \\ -0.082 \cdot \text{CloudAmount} + 0.0032 \cdot \text{CloudHeight} - 0.045 \cdot \text{SoilTemp}\end{aligned} \tag{5}$$

Model Summary:

- F-statistic of 240.6 and a p-value of 2.2e-16 indicates that there is a strong evidence for relationship between the meteorological variables and $\log PM_{10}$.
- A $\mathcal{R}^2 = 0.5974$ tells us that the meteorological variables explain 59.74% of the variability of $\log PM_{10}$.
- All features have a significant p-value: WindSpeed (2e-16), CloudAmount (2e-16), SoilTemp (1.34e-07), CloudHeight (2e-16), Temp10M (2.10e-11), Humidity (0.0163), and Rainfall (0.0277)
- There are no signs of multicollinearity

Stepwise Forward

In this case, using *Forward* or *Stepwise Forward* leads to the same subset of features. Therefore, the model is exactly the same.

Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. When training a lasso regression with the default parameters, we get the following equation:

$$\begin{aligned}\hat{y} = 4.76 - 0.344 \cdot \text{WindSpeed} - 0.037 \cdot \text{Temp1M} - 0.00034 \cdot \text{Rainfall} \\ -0.062 \cdot \text{CloudAmount} + 0.0024 \cdot \text{CloudHeight} - 0.029 \cdot \text{SoilTemp}\end{aligned} \tag{6}$$

Model Comparation

A summary of the coefficients for each of the methods that we have analysed can be seen in the table below:

-	All Features	Best Subset	Backward Selection	Foward Selection	Lasso
Intercept	4.77	4.74	4.73	4.82	4.76
WindSpeed	-0.35	-0.36	-0.35	-0.36	-0.344
Temp1M	-0.04	-0.064	-0.045	-	-0.037
Temp10M	-0.004	-	-	-0.043	-
TempDiff	0.014	-	-	-	-
Humidity	0.0025	-	0.0026	0.0023	-
Rainfall	-0.16	-	-0.175	-0.158	-0.00034
CloudAmount	-0.078	-0.072	-0.077	-0.082	-0.062
CloudHeight	0.0031	0.0033	0.0031	0.0032	0.0024
SoilTemp	-0.043	-	-0.042	-0.045	-0.029

Comparation Summary:

- WindSpeed, CloudAmount, and CloudHeight are included in all models with similar coefficients.
- Temp1M has been included in all methods other than Forward Selection. However, in FS Temp10M has been used. As seen during exploratory analysis, Temp1M and Temp10M have a strong correlation and therefore, we can expect them to be behave similarly.
- While Best Subset method provides a more parsimonious solution, the final subset is missing Humidity, and Rainfall, which both seem to be relevant in other methods.

Final Model

In order to select a final model, we can take a look to the Bayesian Information Criterion (BIC). BIC considers the performance of the model while applying a penalty on the model complexity.

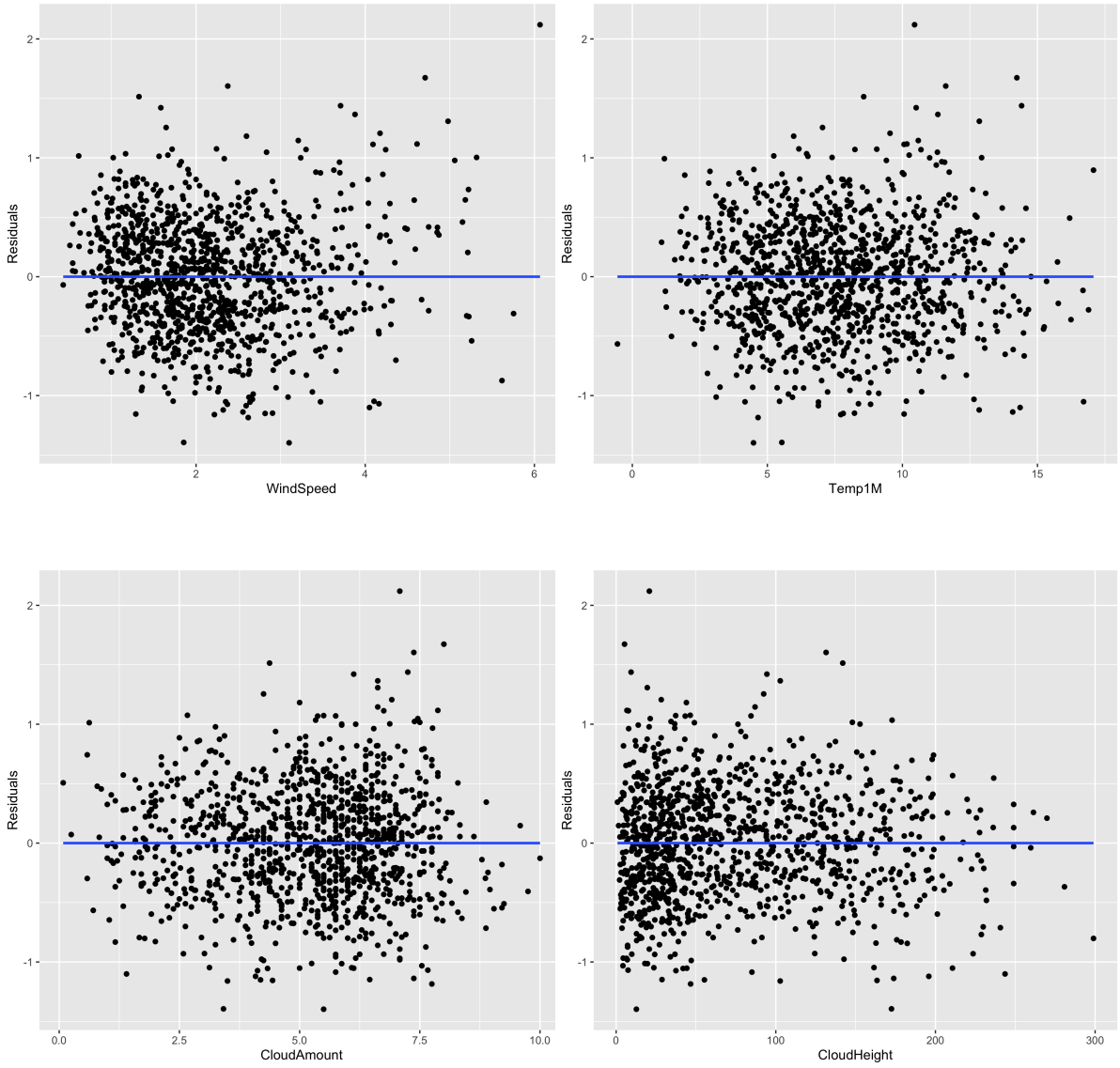
- MLR with features from backward and forward selection have the lowest BIC, 1552
- MLR with the Best Subset of features has a BIC of 1562
- MLR with all features has a BIC of 1565

MLR with the subset of features from Backward and Forward Selection have the lowest BIC. On top of that, they have most of features in common and with very similar coefficients, therefore, we could use any of them as our final model.

MLR Residual Diagnostic Plots

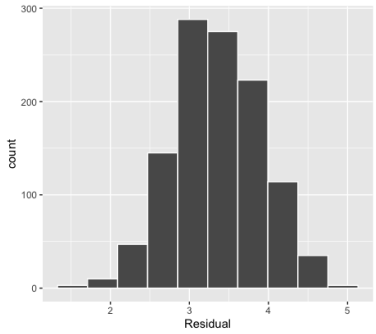
The final model will be used to help us have a better understanding on how the meteorological values impact PM_{10}. While its main used may not be inference, I have finished this section providing a quick overview to Residual Diagnostic Plots. For the purpose of this exercise, I have used the features provided by the best subset method and the MLR provided by these features. In a more complete report, we should be plotting the residuals against all the feature variables and not just the subset that we have used.

Analysis of residuals vs fitted values per feature



From these plots we can see that the residuals are equally spread around zero for all features. For this reason, if we were to use this model for inference, we could say that the predictions are robust to the changing variance and the uncertainty estimates would be reliable. Additionally, the histogram below shows that the error follows a normal distribution:

Histogram Residuals



Conclusion and Translation

Have the government intervention schemes been effective in reducing the pollution level?

In the Formal Analysis, we have used a Simple Linear Regression to see if the levels of PM_{10} were decreasing over time. We have proved the usefulness of the model with the only caveat that the data points fall further from the regression line. As the coefficient of dataindex in the model is negative, we can conclude that PM_{10} decreases over time and therefore, that the government intervention schemes have been effective in reducing the pollution level.

How does the meteorology effect the concentration of PM_{10}?

In order to assess the impact of the meterological effect on PM_{10}, I have trained five different Multiple Regression Models with different subsets of features. On top of that, I have proved the usefulness of all models and we have seen that many of the features have similar coefficients in all of them. By looking at the coefficients of the common features, we can say that:

- **WindSpeed** is the feature that decreases the concentration of PM_{10} the most.
- **Temp1M/Temp10M** and **CloudAmount** decrease the concentration of PM_{10} in a similar way.
- **CloudHeight** is the only meteorology effect that raises the levels of PM_{10}.

On top of that, when using Backward and Forward Selection method, we have been able to see that **Rainfall** decreases substantially the levels of PM_{10} and that **Humidity** increases them slightly. While **SoilTemp** seems to be an important feature in some of the MLR, it's important to note that, as seen in the exploratory analysis, this feature is correlated to Temp1 and Temp10M.

By checking the Bayesian Information Criterion for each MLR, we have concluded that the MLR with the subset from Backward Selection or Forward Selection are the best candidates for the final model. However, in some cases, it may be desirable to find the simplest model and, in this way, the best subset method provides the most parsimonious solution.

For the purpose of this report, I thought that the models used were good enough, however I could have added additional improvements by considering interactions and quadratic relationships with the target.

Appendix: R Code

```
library(table1)
library(dplyr)
library(ggplot2)
options(rgl.useNULL = TRUE)
library(rgl)
library(tolerance)
library(infer)
library(ggridges)
library(viridis)
library(GGally)
library(tidyverse)
library(lubridate)
library(ggpubr)
library(tibble)
library(moderndive)
library(leaps)
library(glmnet)
library(Rcpp)
require(methods)
library(gridExtra)


#Code from Assignment
chch = read.csv("chchpollution.csv")
names(chch) # column names
head(chch)
summary(chch)
# Make date column
chch = chch %>% mutate(date = dmy(paste(Day, Month, Year, sep = "/")))
ggplot(chch, aes(x = date, y = PM10)) + geom_point()
# Data wrangling:
# Extract the winter data and from year 2000, as government interventions started in 2002
# There are 123 days of winter, so create date index over winters days (not calendar days)
# To make your analysis easier, the days with a missing PM10 measurement are also ignored
chchwinter = chch %>%
  filter(Month %in% 5:8, Year >= 2000) %>%
  mutate(dateindex = 2000 + (row_number() - 1)/123) %>%
  filter(!is.na(PM10))
chchwinter = na.omit(chchwinter)
ggplot(chchwinter, aes(x = dateindex, y = PM10)) + geom_point() +
  xlab("Winter Days") + ylab("Daily Average PM10 Concentration") +
  labs(title = "Time Series of Daily Average PM10 Concentration Each Winter")


#Adding natural logarithm for PM10
chchwinter <- chchwinter %>%
              mutate(logPM10 = log(PM10))

#Plotting correlations between all features (target inclusive)
#chchwinter %>%
#
    select(WindSpeed,Temp1M,Temp10M,TempDiff,Humidity,Rainfall,CloudAmount,CloudHeight,SoilTemp,PM10)
    %>%
#  ggpairs(chchwinter,progress=FALSE)

ggsave("correlation_matrix_plot.png")


#Training and analysing results of Simple Linear Model with dateindex
```

```r
simple.model <- lm(logPM10 ~ dateindex,data=chchwinter)
summary(simple.model)

#Plotting logPM10 through time with trend line
chchwinter %>%
  ggscatter(x = "dateindex", y = "logPM10") +
  geom_smooth() +
  labs(x = "dateindex", y = "logPM10")

ggsave("plot_logPM10_dateindex.png")


#Training and analysing results of Multiple Linear Regression METHOD: ALL FEATURES
all_features.model =
    lm(logPM10~WindSpeed+Temp1M+Temp10M+TempDiff+Humidity+Rainfall+CloudAmount+CloudHeight+SoilTemp,data=chchwin
summary(all_features.model)

# Training and analysing results of Multiple Linear Regression METHOD: BEST SUBSET
model_best_subset<-
    regsubsets(logPM10~WindSpeed+Temp1M+Temp10M+TempDiff+Humidity+Rainfall+CloudAmount+CloudHeight+SoilTemp,data
summary_model_best_subset <- summary(model_best_subset)

# Plotting Adjusted R2 and BIC to select number of features
data_best_subset<-data.frame(1:8,summary_model_best_subset$adjr2,summary_model_best_subset$bic)
data_best_subset
ggplot(data_best_subset,aes(x= X1.8,y=summary_model_best_subset.adjr2)) +
  geom_line(colour='black',size=1)+
  theme_classic()

ggsave("best_subset_adjr2.png")

ggplot(data_best_subset,aes(x= X1.8,y=summary_model_best_subset.bic)) +
  geom_line(colour='blue',size=1)+
  theme_classic()

ggsave("best_subset_bic.png")

#Analysing model with optimal features (4)
model_optimal_4_features =
    lm(logPM10~WindSpeed+Temp1M+CloudAmount+CloudHeight,data=chchwinter)
summary(model_optimal_4_features)

#Training and analysing results of Multiple Linear Regression METHOD: BACKWARD SELECTION
step(all_features.model,direction = "backward")

#Analysing model with backward features
model_optimal_backward_features=
    lm(logPM10~WindSpeed+Temp1M+Humidity+Rainfall+CloudAmount+CloudHeight+SoilTemp,data=chchwinter)
summary(model_optimal_backward_features)


#Training and analysing results of Multiple Linear Regression METHOD: STEPWISE BACKWARD
step(all_features.model,direction = "both")


#Training and analysing results of Multiple Linear Regression METHOD: FORWARD SELECTION
min.model <- lm(logPM10~1, data=chchwinter)
step(min.model, direction = "forward",
    scope = list(lower = ~1,
                 upper = ~
                     WindSpeed+Temp1M+Temp10M+TempDiff+Humidity+Rainfall+CloudAmount+CloudHeight+SoilTemp))
```

```r
#Analysing model with forward features
model_optimal_forward_features= lm(formula = logPM10 ~ WindSpeed + CloudAmount + SoilTemp +
    CloudHeight + Temp10M + Humidity + Rainfall, data = chchwinter)
summary(model_optimal_forward_features)

#Training and analysing results of Multiple Linear Regression METHOD: STEPWISE FORWARD
min.model <- lm(logPM10~1, data=chchwinter)
step(min.model, direction = "both",
     scope = list(lower = ~1,
                  upper = ~
                      WindSpeed+Temp1M+Temp10M+TempDiff+Humidity+Rainfall+CloudAmount+CloudHeight+SoilTemp))

#Training and analysing results of Multiple Linear Regression. MLR Asumptions
get_regression_points(model_optimal_4_features)

## Residuals vs Fits
get_regression_points(model_optimal_4_features) %>%
  ggplot(aes(x = logPM10_hat, y = residual)) +
  geom_point() +
  labs(x = "Fitted Values", y = "Residual", title = "Residuals vs Fits")

## histogram of residuals
get_regression_points(model_optimal_4_features) %>%
  ggplot(aes(x = logPM10_hat)) +
  geom_histogram(color = "white", bins = 10) +
  labs(x = "Residual")

#Plotting residuals against all feature variables
chchwinter_best_subset <- chchwinter %>%
                        select(WindSpeed, Temp1M, CloudAmount, CloudHeight)

chchwinter_best_subset.res = cbind(chchwinter_best_subset,
                                resids = residuals(model_optimal_4_features))

## Residuals vs Fits
plot1 = ggplot(chchwinter_best_subset.res, aes(x = WindSpeed, y = resids)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(x = "WindSpeed", y = "Residuals")

  ggsave('windspeed_res.png')

plot2 = ggplot(chchwinter_best_subset.res, aes(x = Temp1M, y = resids)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Temp1M", y = "Residuals")

  ggsave('temp1_res.png')

plot3 = ggplot(chchwinter_best_subset.res, aes(x = CloudAmount, y = resids)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(x = "CloudAmount", y = "Residuals")

  ggsave('CloudAmount_res.png')

plot4 = ggplot(chchwinter_best_subset.res, aes(x = CloudHeight, y = resids)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(x = "CloudHeight", y = "Residuals")

  ggsave('CloudHeight_res.png')


#Training Lasso and analysing results
```

```
X <- chchwinter[, !names(chchwinter) %in%
    c("PM10","logPM10","date","dateindex","Day","Month","Year","WindDirection")]
y <- chchwinter[, "logPM10"]
fit.lasso = glmnet(X, y, alpha = 1, lambda = 10^seq(-3, 6, 0.1)) # alpha = 1 is lasso
    penalty, see help()
plot(fit.lasso, xvar = "lambda", label = TRUE)
fit.cvlasso = cv.glmnet(data.matrix(X), y, alpha = 1, lambda = 10^seq(-3, 6, 0.1))
plot(fit.cvlasso)
coefficients(fit.cvlasso)
coefficients(model_optimal_4_features)

#Comparing BIC for models
broom::glance(all_features.model)
broom::glance(model_optimal_4_features)
broom::glance(model_optimal_backward_features)
broom::glance(model_optimal_forward_features)
```