STA2005S - Regression Assignment

true true

2024-10-18

Abstract

In this report, we explored the efficiency of 6 programming languages through the approximation of π. We found that efficiency of various programming languages can vary widely, with C and C++ being the most efficient programming languages. We also presented evidence for compiled languages having better performance than interpreted languages. Our results suggest that programmers can benefit from taking the efficiency of various programming languages into account, rather than simply opting for simplicity in the syntax of these languages .

1 Part One : Analysis

1.1 Section 1: Introduction

Air pollution, particularly high levels of particulate matter (PM), is a major environmental and public health issue in South Africa's urban centers. Exposure to elevated PM levels is linked to respiratory diseases and other serious health conditions. Understanding the factors influencing PM concentrations is crucial for developing policies that improve air quality and protect public health. This analysis seeks to identify the key drivers of air pollution in South Africa's cities, focusing on how various urban, environmental, and socioeconomic factors affect particulate matter levels.

Unknown Factors to Investigate:

Traffic Density: How do varying levels of vehicle traffic contribute to PM levels in different areas?

Industrial Activity: What is the impact of industrial activity near monitoring stations on air quality?

Temperature & Humidity: How do changes in weather conditions, like temperature and humidity, influence PM concentrations?

Wind Speed: How does wind speed affect the dispersion or accumulation of particulate matter in urban areas?

Day of the Week & Public Holidays: Do patterns of human activity on weekdays, weekends, and holidays significantly influence pollution levels?

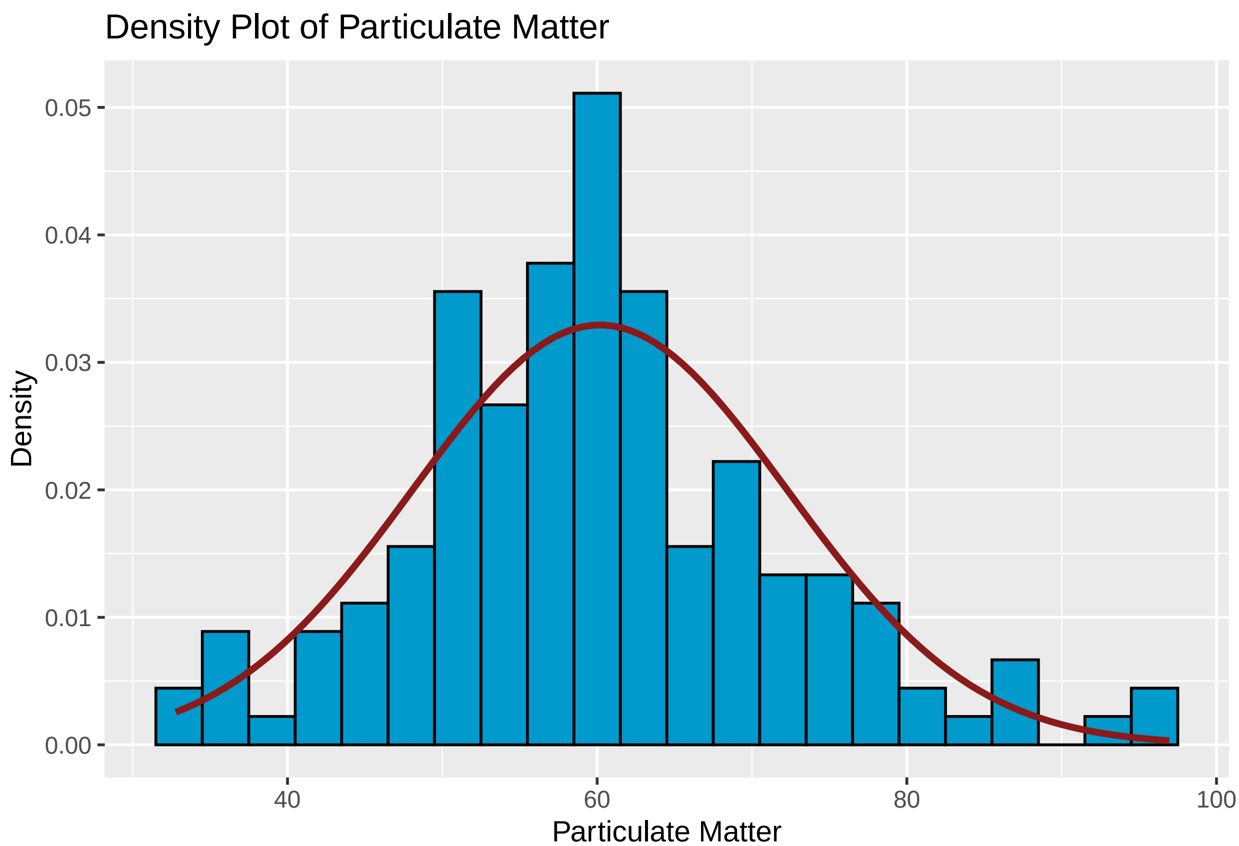
Urban Greenery: How effective are green spaces in reducing air pollution in densely populated areas?

2 Objective

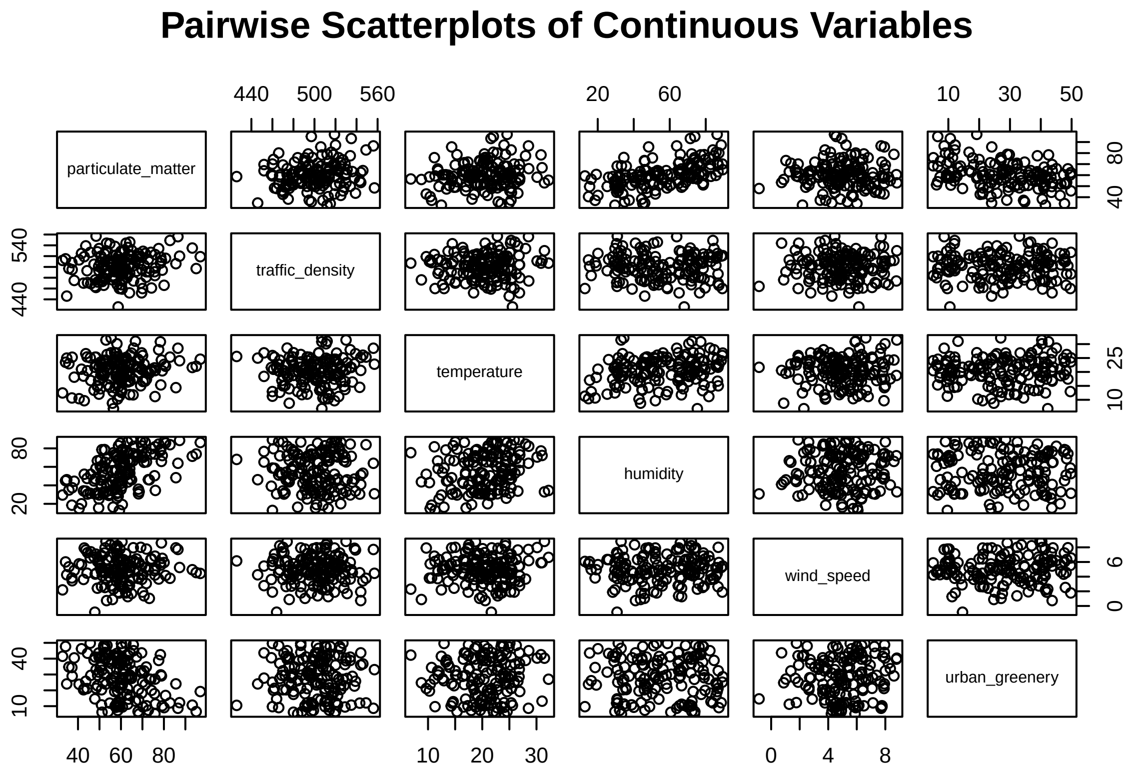
The goal of this analysis is to explore the relationships between PM levels and these explanatory variables. By identifying the most influential factors, we aim to inform urban planning and public health strategies that address air pollution and improve the quality of life in South African cities.

2.1 Section 2 : Data Exploration

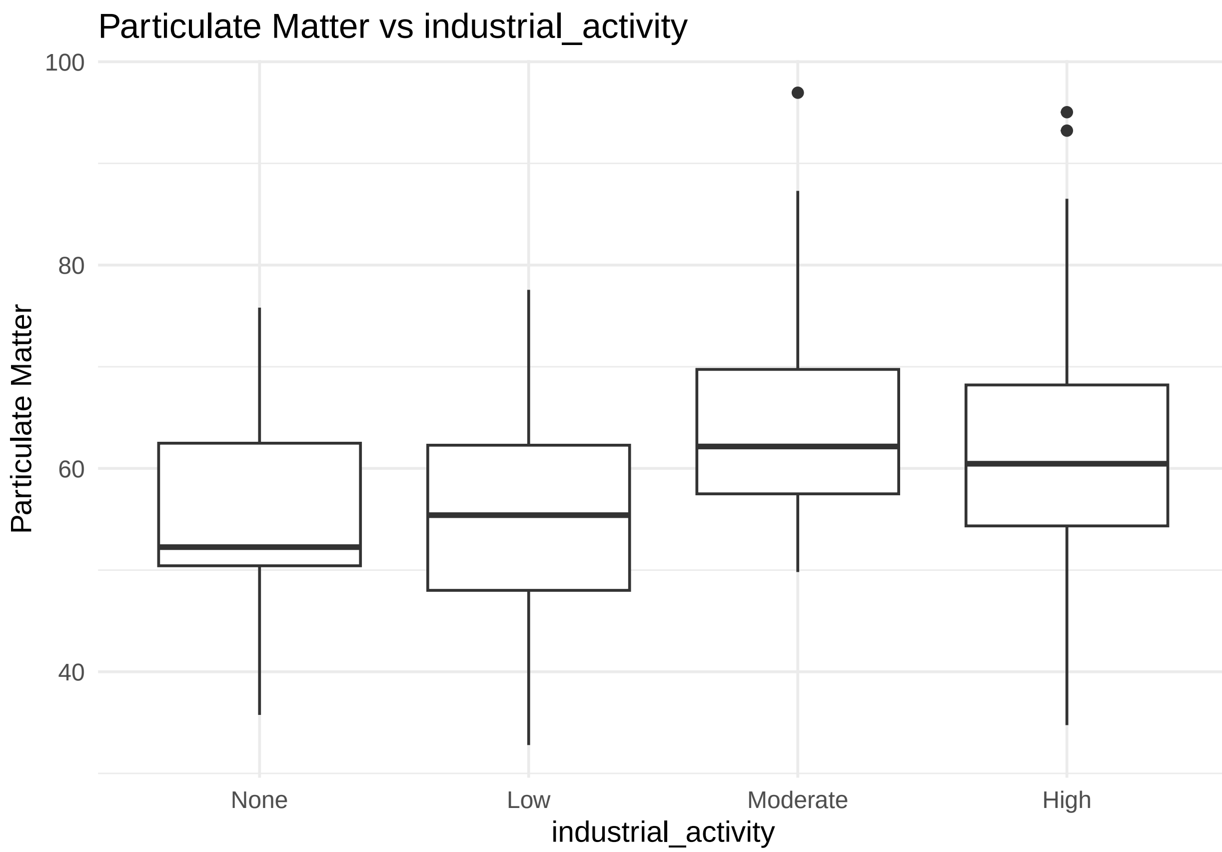
2.1.1 Density plot

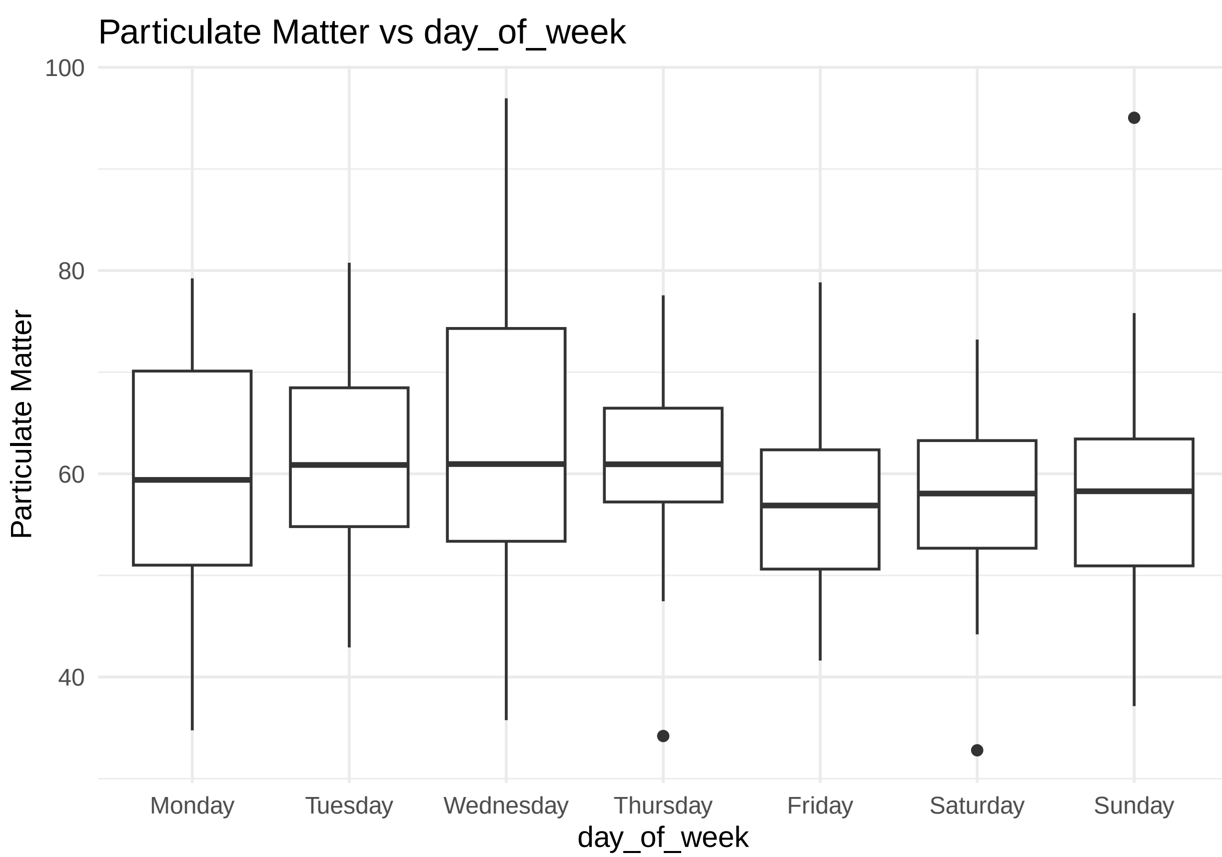


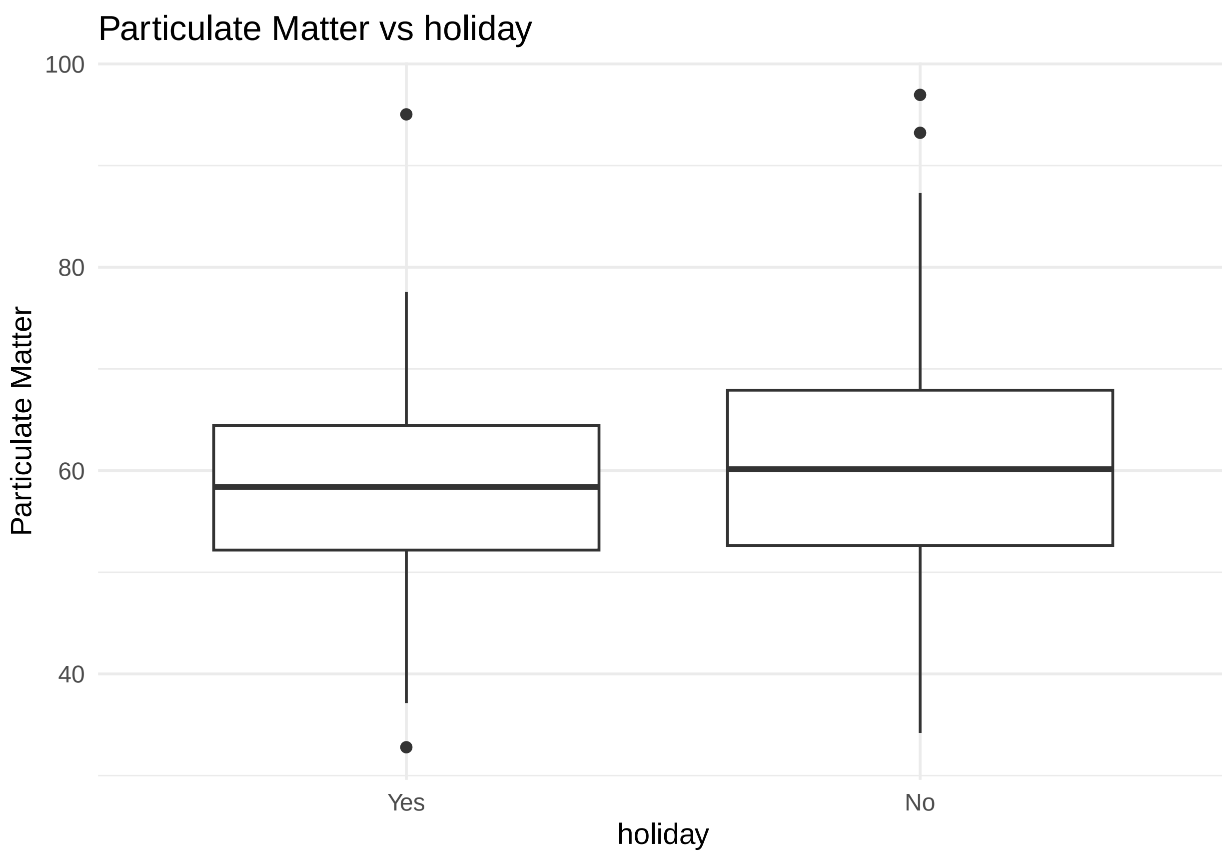
2.1.2 Pairwsie Plots



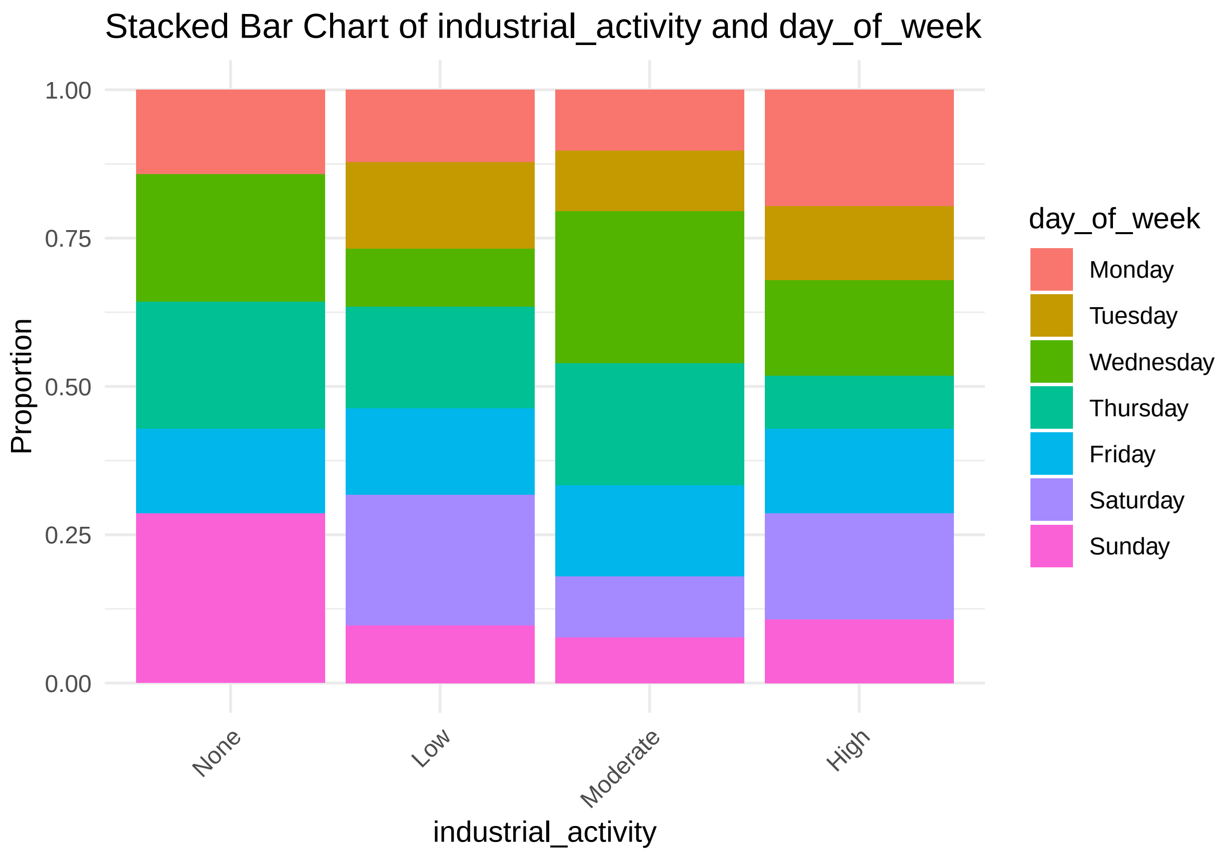
2.1.3 Categorial Variable Plots

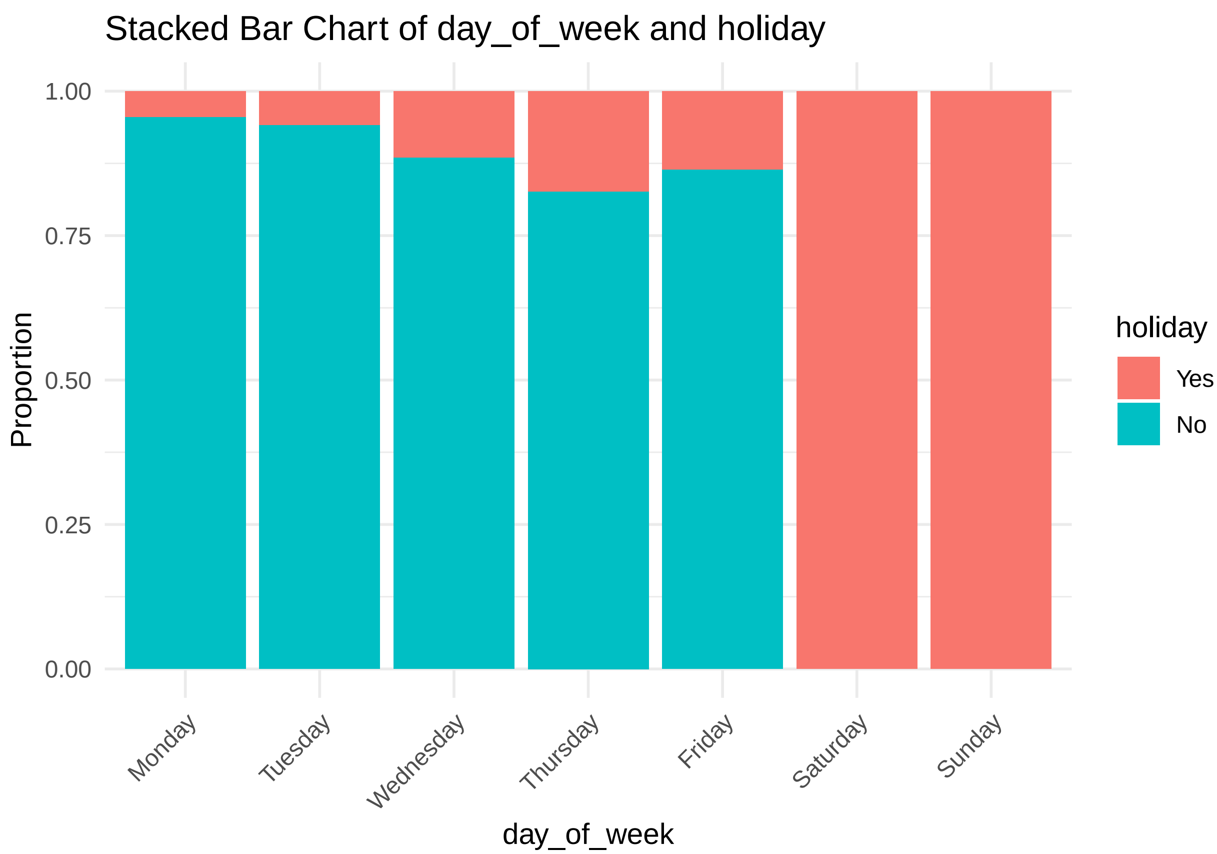





Particulate Matter vs holiday
100
80
60
40
Particulate Matter
Yes
No
holiday

2.1.4 Visual Representation of Relationship between Categorical Variables





2.1.5 Comments

Distribution characterisitcs:

The distribution of particulate matter levels is generally right-skewed, indicating that a small number of observations have significantly high levels of particulate matter while most observations are clustered at lower levels. The presence of outliers suggests variations in local conditions affecting air quality.

Observed Relationships

1. Traffic Density: A positive correlation exists between particulate matter levels and traffic density, suggesting that areas with higher vehicle traffic tend to experience elevated levels of particulate matter.
2. Urban Greenery: A negative trend is observed, where higher urban greenery correlates with lower particulate matter, indicating that vegetation may help mitigate air pollution.
3. Temperature and Wind Speed: No strong relationship was identified between particulate matter and temperature. However, there is a slight negative correlation with wind speed, indicating that higher wind speeds may help disperse particulate matter.

Potential Collinearity

Some potential collinearity is observed among the explanatory variables, particularly between traffic density and urban greenery. High traffic areas often have less vegetation, leading to a relationship that may confound the analysis. Additionally, temperature and wind speed may also exhibit collinearity, as changes in one could affect the other.

3 Section 3

3.1 Simple linear regression

```
X <- cbind(1,data_tidy_air_quality$traffic_density)

Y <-data_tidy_air_quality$particulate_matter
bhat <- solve(t(X) %*% X) %*% t(X) %*% Y

Cmat <- solve(t(X) %*% X)

k <- ncol(X)
rss <- t(Y - X %*% bhat) %*% (Y - X %*% bhat)
# Calculate s2 = RSS/(n-k)
s2 <- as.numeric((rss)/148)
s2
```

```
## [1] 143.5745
```

```
c_ii <- diag(Cmat)

std.error <- sqrt(s2 * c_ii)
std.error
```

```
## [1] 20.37801682  0.04065266
```

3.2 Hypothesis Test

```
# Calculate F-statistic and p-value manually
group_means <- tapply(data_tidy_air_quality$particulate_matter,
                      data_tidy_air_quality$industrial_activity, mean)
overall_mean <- mean(data_tidy_air_quality$particulate_matter)

# Calculate SST
SST <- sum((data_tidy_air_quality$particulate_matter - overall_mean)^2)

# Calculate SStreatment
n <- table(data_tidy_air_quality$industrial_activity)
SStreatment <- sum(n * (group_means - overall_mean)^2)

# Calculate SSerror
group_means_vector <- unlist(tapply(data_tidy_air_quality$particulate_matter, data_tidy_air_quality$indu
[data_tidy_air_quality$industrial_activity])
SSerror <- sum((data_tidy_air_quality$particulate_matter - group_means_vector)^2)

# Calculate degrees of freedom
k <- length(unique(data_tidy_air_quality$industrial_activity))
N <- nrow(data)
DFtreatment <- k - 1
```

```r
DFerror <- 150 - k

# Calculate Mean Squares
MStreatment <- SStreatment / DFtreatment
MSerror <- SSerror / DFerror


# Calculate F-statistic
F_statistic <- MStreatment/MSerror
F_statistic
```

```
## [1] 5.395959
```

```r
# Calculate p-value
p_value <- pf(F_statistic, DFtreatment, DFerror, lower.tail = FALSE)
p_value
```

```
## [1] 0.001502236
```

4 Question 4

Table 1: Confidence Interval for each Coefficient

	2.5 %	Estimate	97.5 %
Intercept			
(Intercept)	-21.0568	13.7937	48.6442
Traffic Density			
traffic_density	0.0155	0.0799	0.1444
Industrial Activity			
industrial_activityLow	-3.1721	2.6589	8.4900
industrial_activityModerate	0.6047	6.4545	12.3043
industrial_activityHigh	-0.2503	5.3652	10.9806
Natural Factors			
temperature	-1.1521	-0.2815	0.5891
humidity	-0.1111	0.1926	0.4962
wind_speed	-0.8040	0.0193	0.8426
temperature:humidity	-0.0088	0.0061	0.0209
Day of Week			
day_of_weekTuesday	-5.9877	0.0133	6.0142
day_of_weekWednesday	-5.3501	0.1565	5.6630
day_of_weekThursday	-5.5367	0.1662	5.8690
day_of_weekFriday	-8.0602	-2.4221	3.2161
day_of_weekSaturday	-12.3605	-4.4832	3.3940
day_of_weekSunday	-10.2167	-2.0885	6.0396
Holiday			
holidayNo	-6.7151	-0.9961	4.7228
Urban Greenery			
urban_greenery	-0.4142	-0.2954	-0.1766

4.1 Hypothesis Testing

We'd like to perform hypothesis tests on the following variables: Temperature, Humidity, Industrial Levels, and Day of Week.

We'll start by examining whether Temperature has an effect on the concentration of Particulate Matter. We'll use the following set of hypothesis:

$$H_0 : \beta_{temp} = \beta_{hum:temp} = 0$$
$$H_A : \beta_{temp} \neq 0 \text{ and } \beta_{hum:temp} \neq 0$$

This can be done by comparing the restricted and un restricted model:

$$Y_R = \beta_0 + \beta_{traffic} X +$$

```
model_unrestricted <- lm(particulate_matter ~ . +
                         temperature:humidity,
                         data=data_tidy_air_quality)
model_restricted <- update(model_unrestricted, .~.
                           - temperature
                           - temperature:humidity)
anova(model_restricted, model_restricted)
```

Using the anova function in R, we compare the two models with F test. The F test yields a P value 0.6815, suggesting that temperature doesn't have a significant effect on the concentration of particular matter.

We now test for the effect of humidity. Repeating the same procedure, we obtain a P value < 0.00001. Suggesting that it's likely that humidity has an effect on the concentration of particulate matters.

$$H_0 : \beta_{hum} = \beta_{hum:temp} = 0$$
$$H_A : \beta_{hum} \neq 0 \text{ and } \beta_{hum:temp} \neq 0$$

```
model_unrestricted <- lm(particulate_matter ~ . +
                         temperature:humidity,
                         data=data_tidy_air_quality)
model_restricted <- update(model_unrestricted, .~.
                           - humidity
                           - temperature:humidity)
anova(model_restricted, model_unrestricted)
```

4.1.1 Categorical Variables

For the day of week, we take Monday as the reference category and test for the following set of hypothesis, using the same procedure.

$$H_0 : \beta_{Tuesday} = \beta_{Wednesday} = \beta_{Thursday} = \beta_{Friday} = \beta_{Saturday} = \beta_{Sunday} = 0$$
$$H_A : \beta_{Tuesday} \neq 0 \text{ and } \beta_{Wednesday} \neq 0 \text{ and } \beta_{Thursday} \neq 0 \text{ and } \beta_{Friday} \neq 0 \text{ and } \beta_{Saturday} \neq 0 \text{ and } \beta_{Sunday} \neq 0$$

```
data_tidy_air_quality$day_of_week <-
  relevel(factor(data_tidy_air_quality$day_of_week), ref="Monday")
model_unrestricted <- lm(particulate_matter ~ .,
                         data=data_tidy_air_quality)
model_restricted <- update(model_unrestricted, .~.
                           - day_of_week)
anova(model_restricted, model_unrestricted)
```

The P value is 0.7735, indicating that there is no evidence that supports rejecting the null hypothesis.

We do the same for industrial activity, taking No Activity as the reference category to test for the set of hypothesis:

$$H_0 : \beta_{Low} = \beta_{Moderate} = \beta_{High} = 0$$
$$H_A : \beta_{Low} \neq 0 \text{ and } \beta_{Moderate} \neq 0 \text{ and } \beta_{High} \neq 0$$

```
data_tidy_air_quality$industrial_activity <-
  relevel(factor(data_tidy_air_quality$industrial_activity), ref="None")
model_unrestricted <- lm(particulate_matter ~ .,
                         data_tidy_air_quality)
```

```r
model_restricted <- lm(particulate_matter ~ . - industrial_activity,
                        data_tidy_air_quality)
anova(model_restricted, model_unrestricted)
```

We obtain a P-value of 0.0707, suggesting that there's no evidence that supports rejecting the null hypothesis at 5% significance level.

4.2 Intepretation

From the summary output and the F test for Natural Factors, we get that only Moderate Industrial Activity (p-value: 0.03), Traffic Density (p-value: 0.0155), Urban Greenery (p-value: < 0.0001), and Humidity (p-value: < 0.0001) have a significant effect (p value less than or equal to 0.05) on the concentration of particulate matters.

The average increase in concentration of particulate matter by Moderate Industrial Activity is estimated to be 6.4545 $\frac{\mu g}{m^3}$(CI95%: 0.6047, 12.3043), indicating a significant positive association between the two factors. The average increase in concentration of particulate matter caused by Traffic Density, that is one extra vehicle per hour, is estimated to be 0.0799 $\frac{\mu g}{m^3}$ (CI95%: 0.0155, 0.1444), indicating a positive association. The average decrease in concentration of particulate matter caused by a unit in the percentage of area covered by Urban Greenery is estimated to be 0.2954 $\frac{\mu g}{m^3}$ (CI95%:-0.4142, -0.1766), indicating a negative association between the two. The effect of humidity can interact with temperature, and is thus not independent of other factors. Using the anova model, we can only deduce that there is likely an effect, but we can't be certain about the confidence interval.

5 part 2

5.1 Scenario A

5.1.1 Methodology

Under the null hypothesis, the response variable is not related to temperature. We have that $\beta_0 = 0$, which means that the response variable $Y = 30 + \epsilon$ where ϵ is some random noises. We violated the model assumption by making $\epsilon \sim Unif(-17.32, 17.32)$ where $Var(\epsilon) = 100$, as opposed to being normally distributed.

5.2 Scenario A:

5.2.1 Methodology and discussion of results for Scenario A

Simulation under the null hypothesis (β_1= 0):

We simulate the data assuming $\beta_0 = 30$, β_1= 0, and errors are uniformly distributed. The errors will be sampled from a uniform distribution where $e \sim U(a, b)$ with the constraint that $Var(e) = 100$

For a uniform distribution $e \sim U(a, b)$ with a variance $\sigma^2 = 100$, $Var(e) = \frac{(b-a)^2}{12}$. Solving for a and b we get $a = -17.32$ and $b = 17.32$

To simulate our data, we created a function 'run_simulation' which runs a simulation once. In a simulation, we used 'runif(length(temperature), min = a, max = b)' to generate random errors. Next calculated $Y_i = 30 + e$. We then used the lm function to fit our regression model. We extracted p values from our model, and ran ifelse statement to check whether our null hypothesis was rejected or not. Then we used the ‘textit{replicate()’ function to repeat each procedure 1000 times. Finally, we counted the number of null hypothesis rejected and we observed that our type 1 error rate for this scenario is 0.043

Type I error is the probability of incorrectly rejecting the null hypothesis when it is true (false positive). Under the null hypothesis, the expected Type I error rate should be equal to the chosen significance level, typically 0.05.

Under a uniform distribution, the error terms tend to be more tightly spread compared to the tails of a normal distribution (which has more extreme values due to its longer tails). This can result in:

Underestimated variability in the model, leading to more frequent rejections of the null hypothesis when it should not be rejected. Inflated Type I error rate: The test may incorrectly reject the null hypothesis more often than expected under the nominal significance level.

5.2.2 Results

We obtain a type 1 error rate of 0.054, which is not too far from 0.05. Indicating that this model assumption does not significantly affect the type 1 error rate. This can be explained by central limit theorem, with the sampling distribution of e converges to a normal distribution. (For interest’s sake, if we change the simulation count to 10000, the type 1 error rate becomes 0.0501, even closer to 0.05)

5.3 Scenario B:

5.3.1 Methodology and discussion of results for Scenario B

Again we created a single function to run a single simulation. In each simulation, We simulate the error variances using the normal distribution with mean of 100 and variance of 50. We then ensure that the variances are positive by taking the absolute value of the simulated variances. Then we simulated error terms using a normal distribution with a mean of 0 and a variance of our error_variance calculated before. This ensures that our errors have a non constant variance. We then repeated the same steps as above to obtain our type 1 error rate. Our observated type 1 error rate is 0.047.

Impact of Heteroscedasticity: Effect on Type I Error Rate: Heteroscedasticity violates the assumption of constant variance in regression models. This leads to incorrect estimates of standard errors and, consequently, incorrect p-values. As a result, the Type I error rate will likely increase beyond the nominal 0.05 level.

Our error rate is lower than 0,05. There could be a few reasons for this:

Conservative Hypothesis Test: Our test may be too conservative. This means the test is less likely to reject the null hypothesis even when it might be warranted. It could happen if: The variability of the errors isn’t large enough to make the standard errors inaccurate in a way that increases the Type I error.

The model has adjusted in such a way that it becomes harder to reject the null hypothesis, potentially overcorrecting for the heteroscedasticity.

5.4 Scenario C:

5.4.1 Methodology and discussion of results for Scenario C:

Using our function again, we replicate 1000 simulations,but this time we test the dependence of errors in our model. To generate Correlated errrors in our model we, use the provided function given to us. As above we repeated the same steps in order to obtain the proportion of the number of times our null hypothesis was incorrectly rejected. OUr type 1 errot rate was 0.053.

Autocorrelation violates the assumption of independent errors in linear regression. This can lead to incorrect standard errors for the regression coefficients, making hypothesis tests less reliable.

Typically, autocorrelated errors inflate Type I errors if unaccounted for because the model underestimates the true error variability. However, depending on the specific pattern of autocorrelation and sample size, this effect can vary.