STA2005S - Regression Assignment

Jing Yeh
yhxjin001@myuct.ac.za

Saurav Sathnarayan
sthsau001@myuct.ac.za

2024-10-16

0.1 Part One : Analysis

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

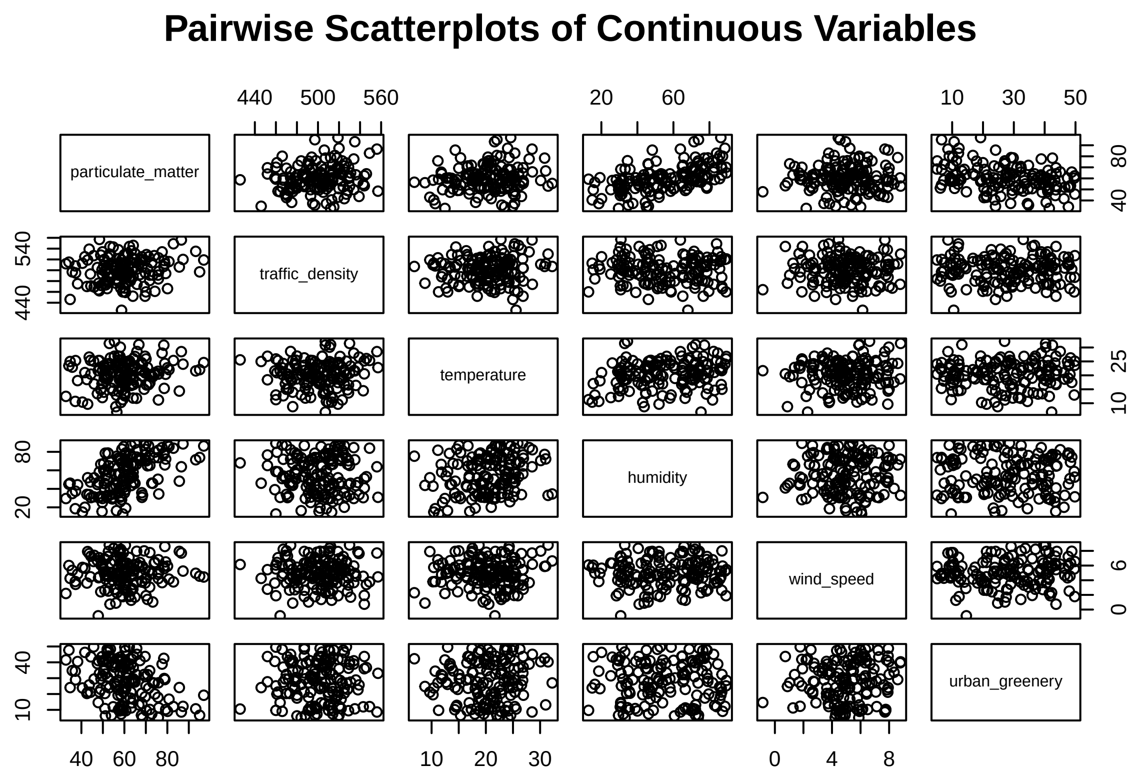
The goal of this analysis is to explore the relationships between PM levels and these explanatory variables. By identifying the most influential factors, we aim to inform urban planning and public health strategies that address air pollution and improve the quality of life in South African cities.

2.1 Section 2 : Data Exploration

density plot

pairwsie plots

```r
continuous_vars <- data_tidy_air_quality[, sapply(data_tidy_air_quality, is.numeric)]
pairs(continuous_vars, main = "Pairwise Scatterplots of Continuous Variables")
```

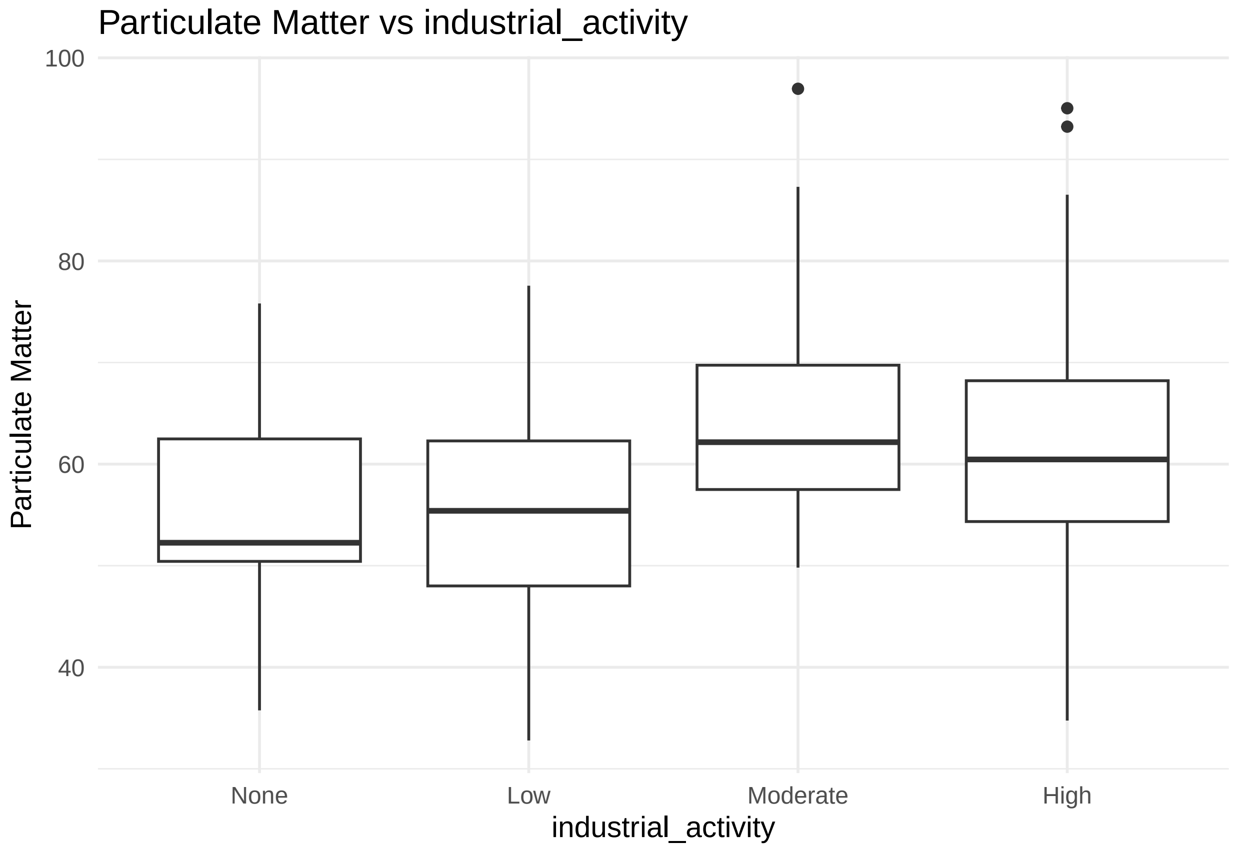


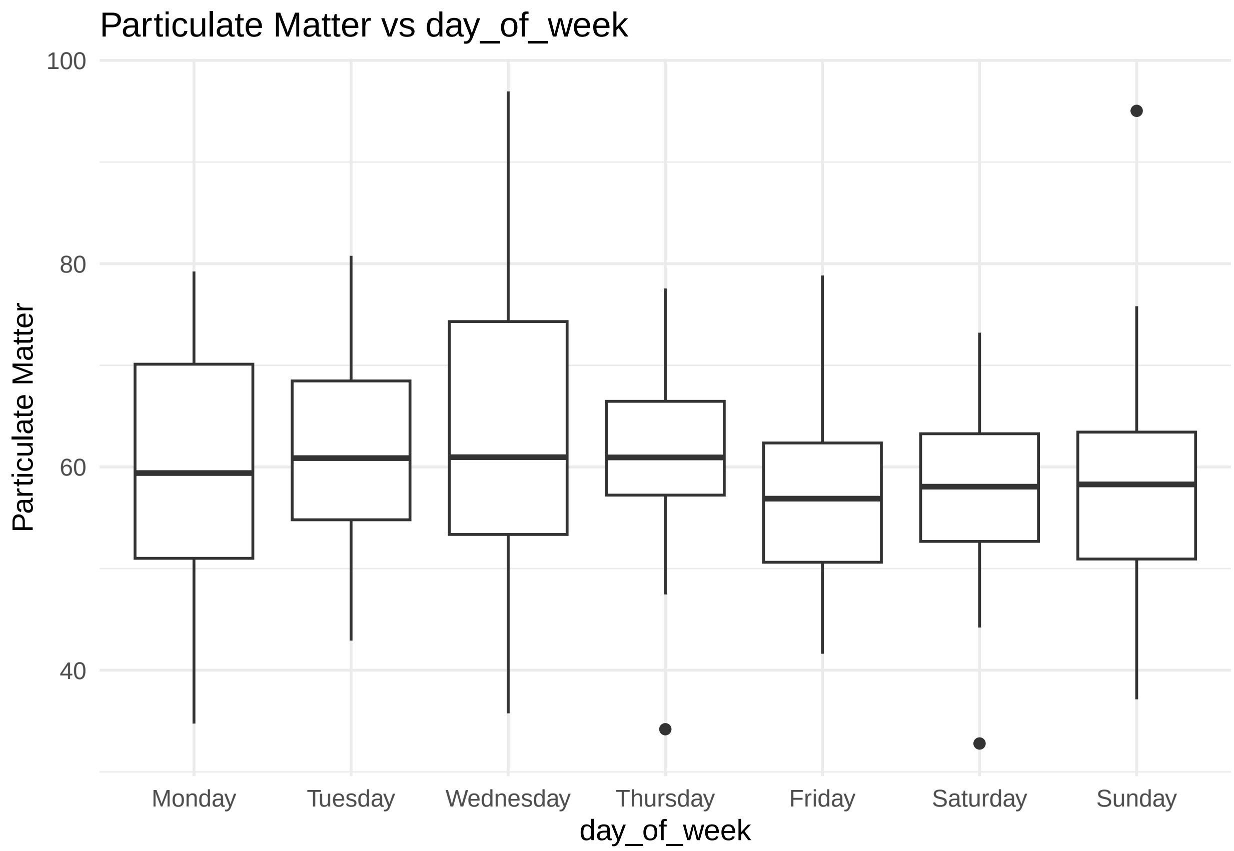
categorial variable plots

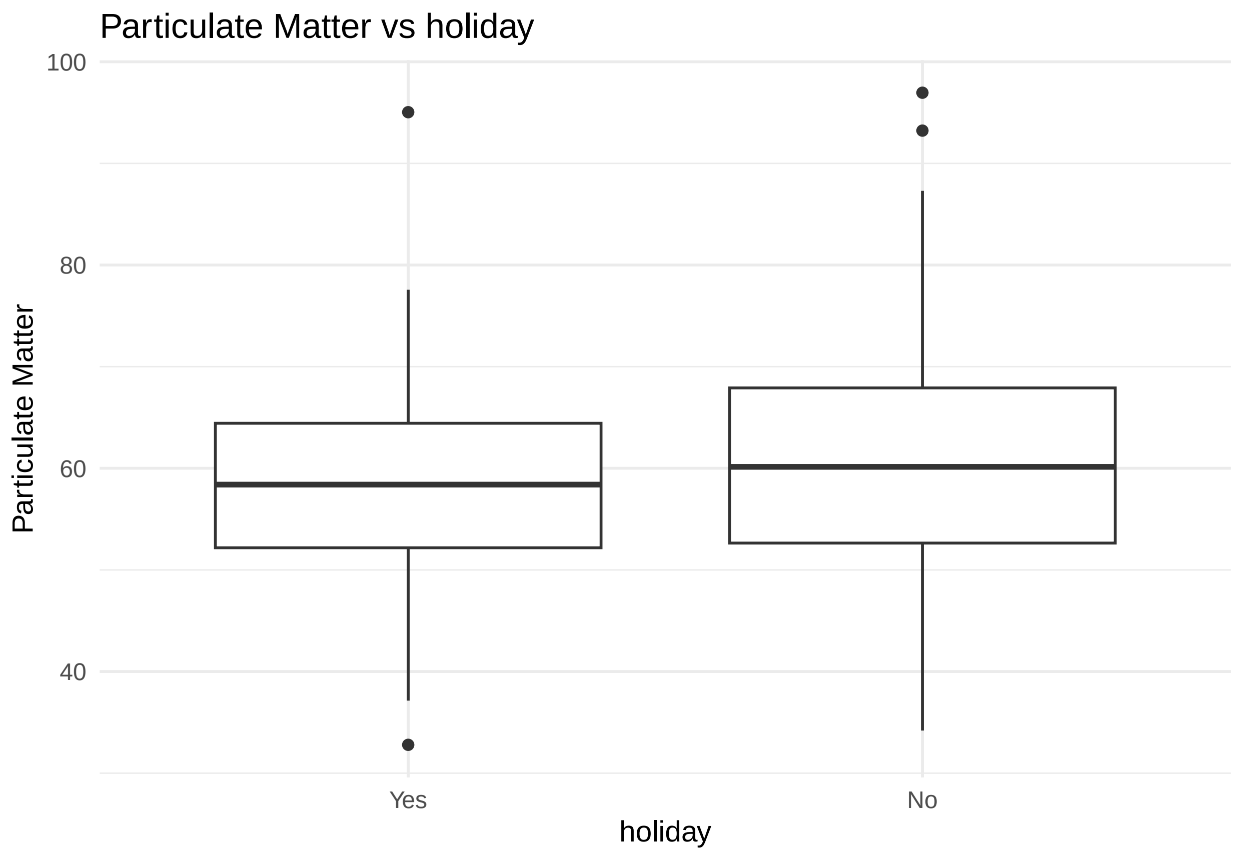
```r
data_tidy_air_quality$industrial_activity <- factor(data_tidy_air_quality$industrial_ac
                                                    levels = c("None","Low", "Moderate", "High"))   # A

data_tidy_air_quality$day_of_week <- factor(data_tidy_air_quality$day_of_week,
                                            levels = c("Monday", "Tuesday", "Wednesday",
                                                       "Thursday", "Friday", "Saturday", "Sunday"))

data_tidy_air_quality$holiday <- factor(data_tidy_air_quality$holiday,
                                        levels = c("Yes", "No"))

categorical_vars <- names(data_tidy_air_quality)[sapply(data_tidy_air_quality, is.facto


for (var in categorical_vars) {
  plt<- ggplot(data_tidy_air_quality, aes_string(x = var, y = "particulate_matter")) +
    geom_boxplot() +
```

```
    labs(title = paste("Particulate Matter vs", var),
         x = var,
         y = "Particulate Matter") +
    theme_minimal()


  print(plt)  # Print the plot
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```







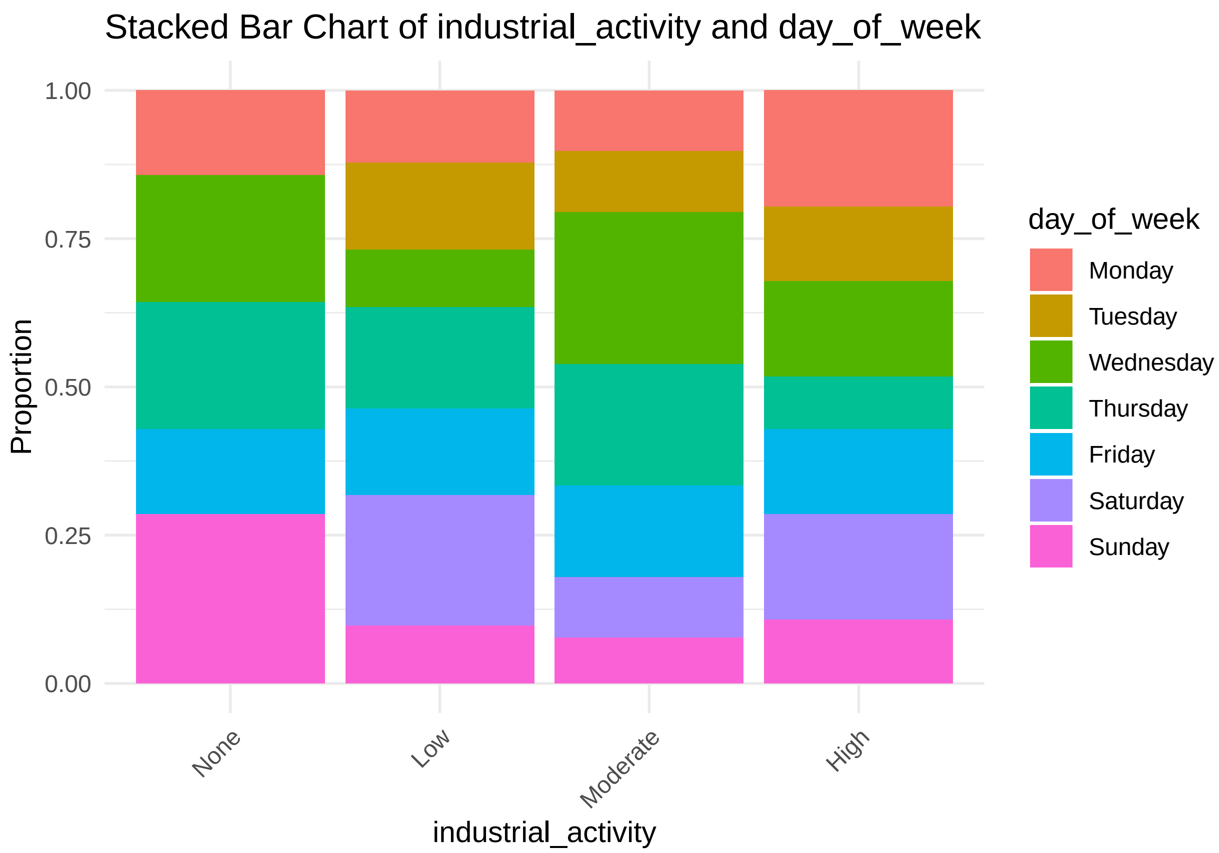
tabular representation of relationship between categorial variables

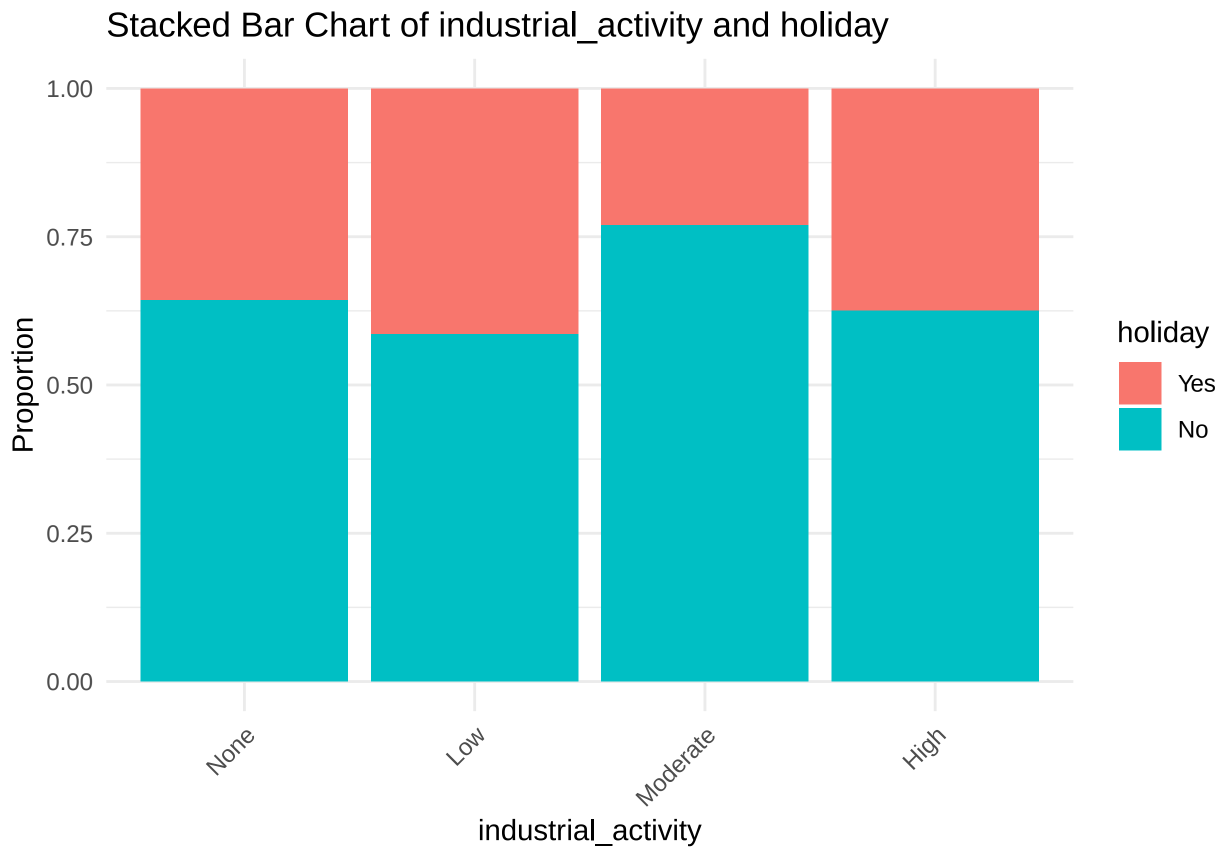
```
for (i in 1:(length(categorical_vars)-1)) {
  for (j in (i+1):length(categorical_vars)) {
    cat("Contingency Table for", categorical_vars[i], "and", categorical_vars[j], "\n'
    print(table(data_tidy_air_quality[[categorical_vars[i]]], data_tidy_air_quality[[ca
    cat("\n")
  }
}
```

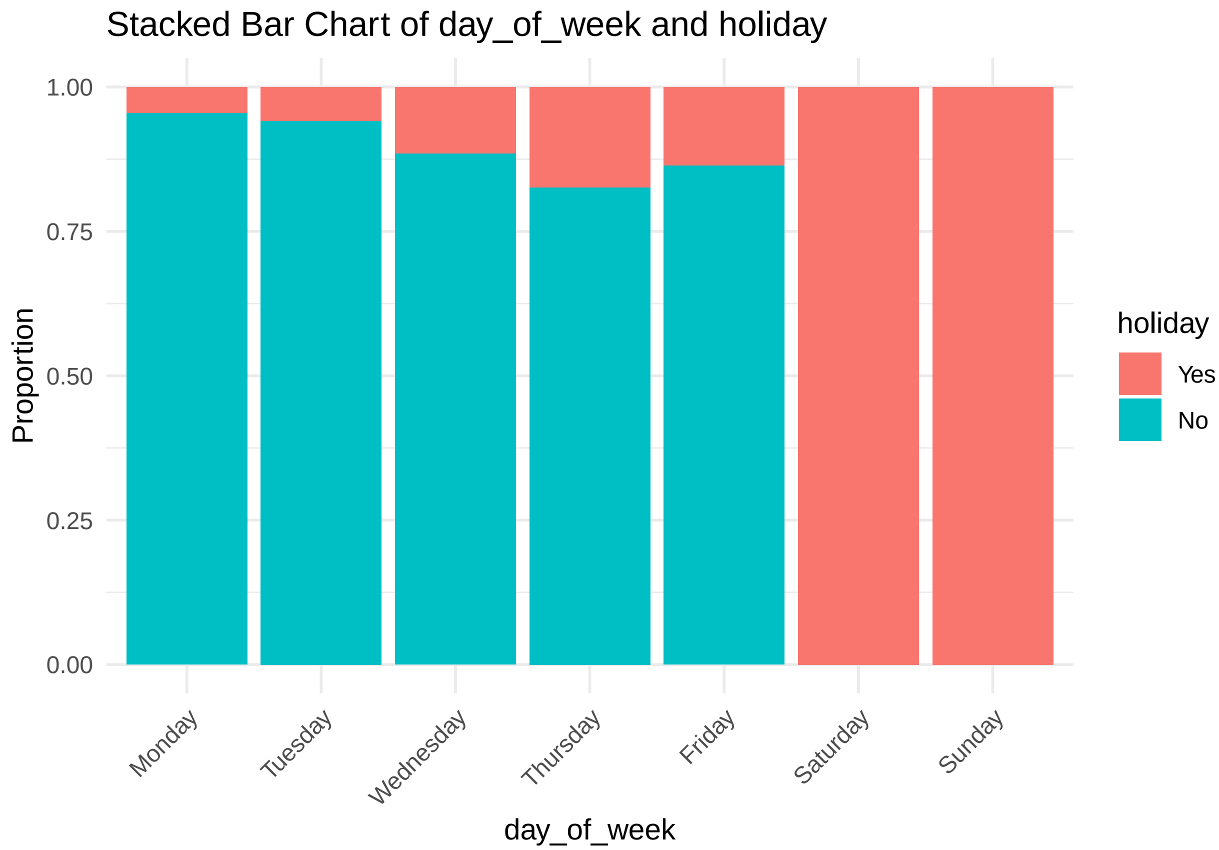
```
## Contingency Table for industrial_activity and day_of_week
##
##              Monday Tuesday Wednesday Thursday Friday Saturday Sunday
##    None           2       0         3        3      2        0      4
##    Low            5       6         4        7      6        9      4
##    Moderate       4       4        10        8      6        4      3
##    High          11       7         9        5      8       10      6
##
## Contingency Table for industrial_activity and holiday
##
##              Yes No
##    None        5  9
##    Low        17 24
##    Moderate    9 30
##    High       21 35
##
## Contingency Table for day_of_week and holiday
##
##               Yes No
##    Monday       1 21
##    Tuesday      1 16
##    Wednesday    3 23
##    Thursday     4 19
##    Friday       3 19
##    Saturday    23  0
##    Sunday      17  0
```

visual representation of relationship between categorial variables

```r
for (i in 1:(length(categorical_vars) - 1)) {
  for (j in (i + 1):length(categorical_vars)) {
    # Create the plot
    p <- ggplot(data_tidy_air_quality, aes_string(x = categorical_vars[i], fill = cate
      geom_bar(position = "fill") +  # Use "fill" to make it a stacked bar chart (pr
      labs(title = paste("Stacked Bar Chart of", categorical_vars[i], "and", categoric
           x = categorical_vars[i],
           y = "Proportion") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))

    # Print the plot
    print(p)
  }
}
```







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

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
```

```r
c_ii <- diag(Cmat)

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

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

```r
mod1<-lm(data_tidy_air_quality$particulate_matter ~ data_tidy_air_quality$traffic_densi

summary(mod1)
```

```
##
## Call:
## lm(formula = data_tidy_air_quality$particulate_matter ~ data_tidy_air_quality$traf
##     data = data_tidy_air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.332  -7.561  -1.050   6.110  35.243
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           18.11537   20.37802   0.889   0.3755
## data_tidy_air_quality$traffic_density  0.08400    0.04065   2.066   0.0406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.98 on 148 degrees of freedom
## Multiple R-squared:  0.02804,    Adjusted R-squared:  0.02147
## F-statistic: 4.269 on 1 and 148 DF,  p-value: 0.04055
```

hypthesis test

```r
# Summary of ANOVA results
summary(aov(particulate_matter ~ industrial_activity, data = data_tidy_air_quality))
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## industrial_activity   3   2182   727.3   5.396 0.0015 **
## Residuals           146  19680   134.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

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

# Calculate SSW
group_means_vector <- unlist(tapply(data_tidy_air_quality$particulate_matter, data_tidy
SSerror <- sum((data_tidy_air_quality$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
DFerror <- 150 - k

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



# Calculate F-statistic
F_statistic <- MStreatment/MSerror

# Output F-statistic
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 Coefficients

	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

4.0.1 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

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

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_unrestricted, model_restricted)
```

```
## Analysis of Variance Table
##
## Model 1: particulate_matter ~ traffic_density + industrial_activity +
##     temperature + humidity + wind_speed + day_of_week + holiday +
##     urban_greenery + temperature:humidity
## Model 2: particulate_matter ~ traffic_density + industrial_activity +
##     humidity + wind_speed + day_of_week + holiday + urban_greenery
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    133 11032
## 2    135 11096 -2   -63.801 0.3846 0.6815
```

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

We now test for

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

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

```
## Analysis of Variance Table
##
```

```
## Model 1: particulate_matter ~ traffic_density + industrial_activity +
##     temperature + humidity + wind_speed + day_of_week + holiday +
##     urban_greenery + temperature:humidity
## Model 2: particulate_matter ~ traffic_density + industrial_activity +
##     temperature + wind_speed + day_of_week + holiday + urban_greenery
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    133 11032
## 2    135 16101 -2   -5069.2 30.556 1.204e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Repeating the same procedure, we obtain a P value less than 0.00001, indicating that humidity has a significant effect on the concentration of particular matter

4.0.2 Categorical Variables

For the categorical levels with more than 2 levels, we can read directly from the summary table that only Moderate Industrial Activity yields a P value less than 0.05 (P value = 0.0308). We can thus only reject the null hypothesis at 0.05 α level.

```
summary_df <- as.data.frame(summary(multi_model)$coefficients)
summary_categorical_df <- summary_df[c(3:5, 9:14),]
summary_table <- kable(summary_categorical_df, digits=4)|>
  kable_styling(font_size = 12)
summary_table
```

	Estimate	Std. Error	t value	Pr(¿|t|)
industrial_activityLow	2.6589	2.9480	0.9019	0.3687
industrial_activityModerate	6.4545	2.9575	2.1824	0.0308
industrial_activityHigh	5.3652	2.8390	1.8898	0.0610
day_of_weekTuesday	0.0133	3.0339	0.0044	0.9965
day_of_weekWednesday	0.1565	2.7840	0.0562	0.9553
day_of_weekThursday	0.1662	2.8832	0.0576	0.9541
day_of_weekFriday	-2.4221	2.8505	-0.8497	0.3970
day_of_weekSaturday	-4.4832	3.9825	-1.1257	0.2623
day_of_weekSunday	-2.0885	4.1094	-0.5082	0.6121