STA2005S - Regression Assignment

Jing Yeh
yhxjin001@myuct.ac.za

Saurav Sathnarayan
sthsau001@myuct.ac.za

2024-10-14

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

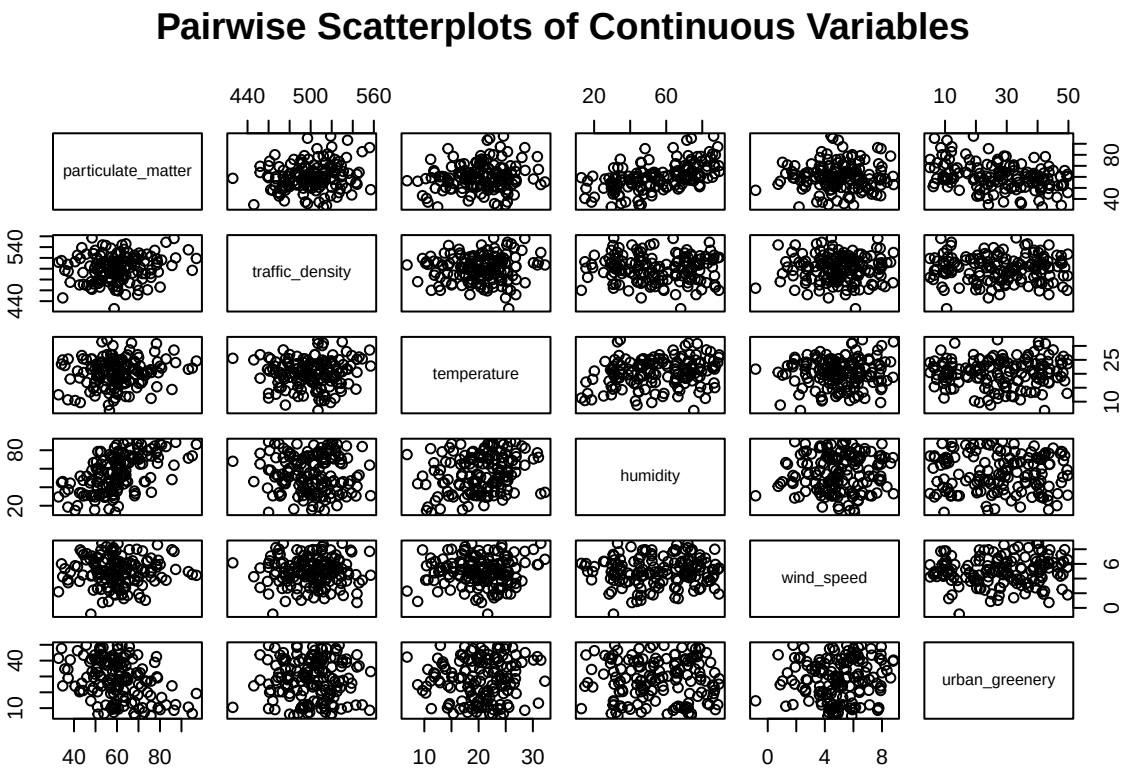
The goal of this analysis is to explore the relationships between PM levels and these explanatory variables. By identifying the most influential factors, we aim to inform urban planning and public health strategies that address air pollution and improve the quality of life in South African cities.

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")
```

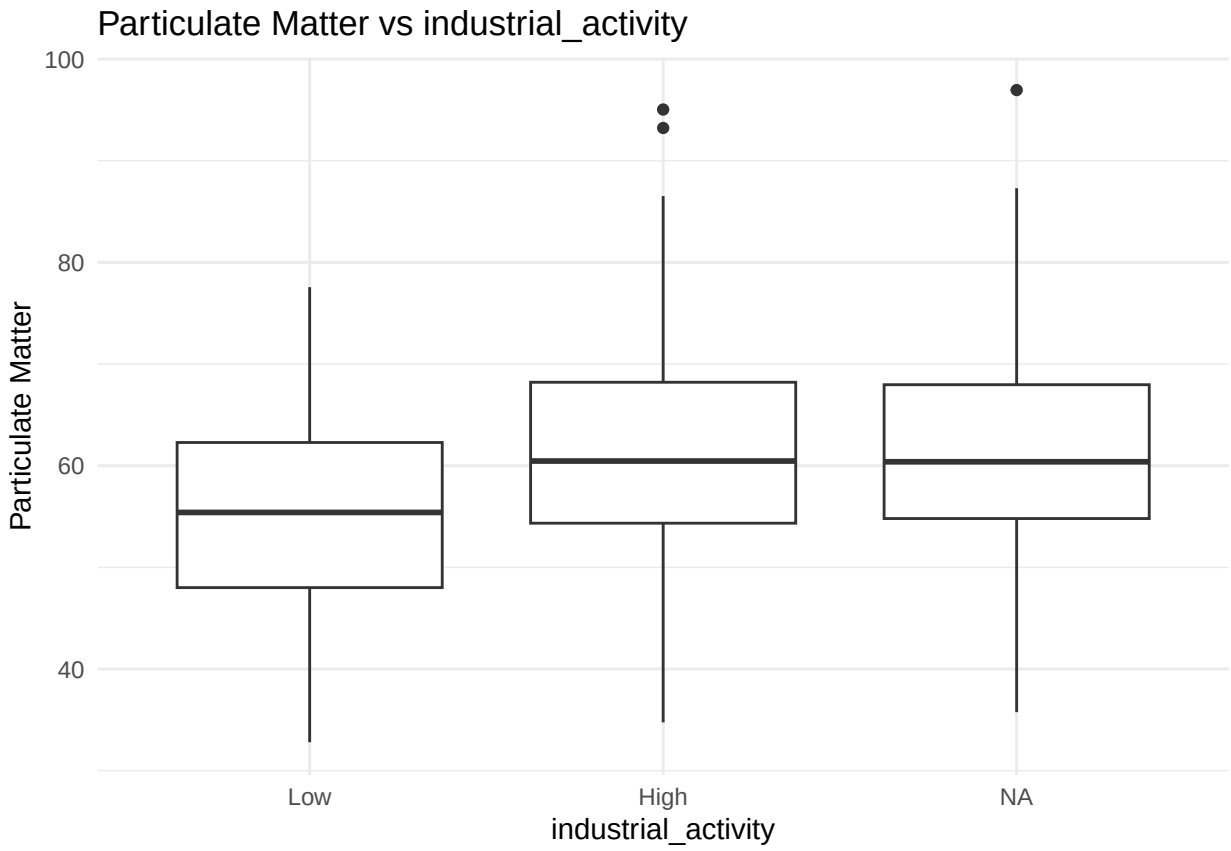


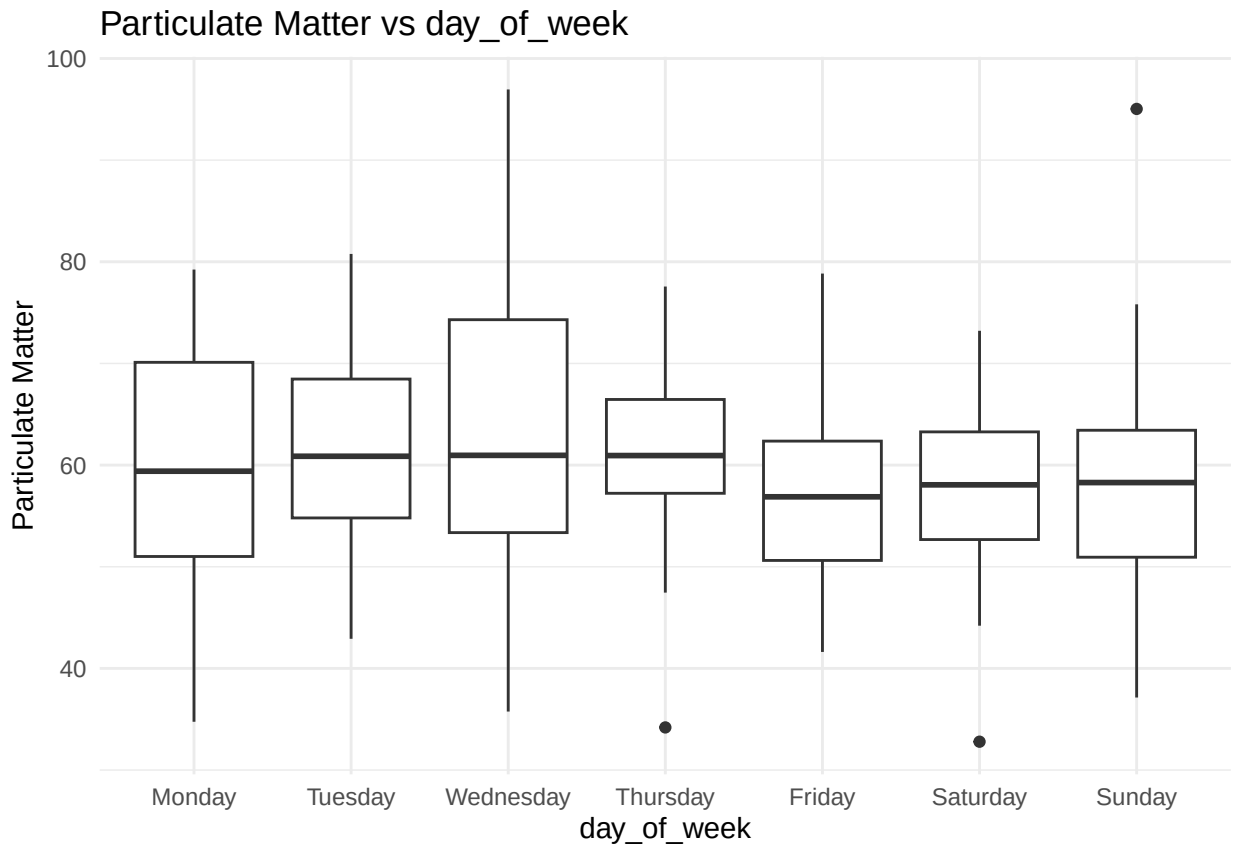
categorial variable plots

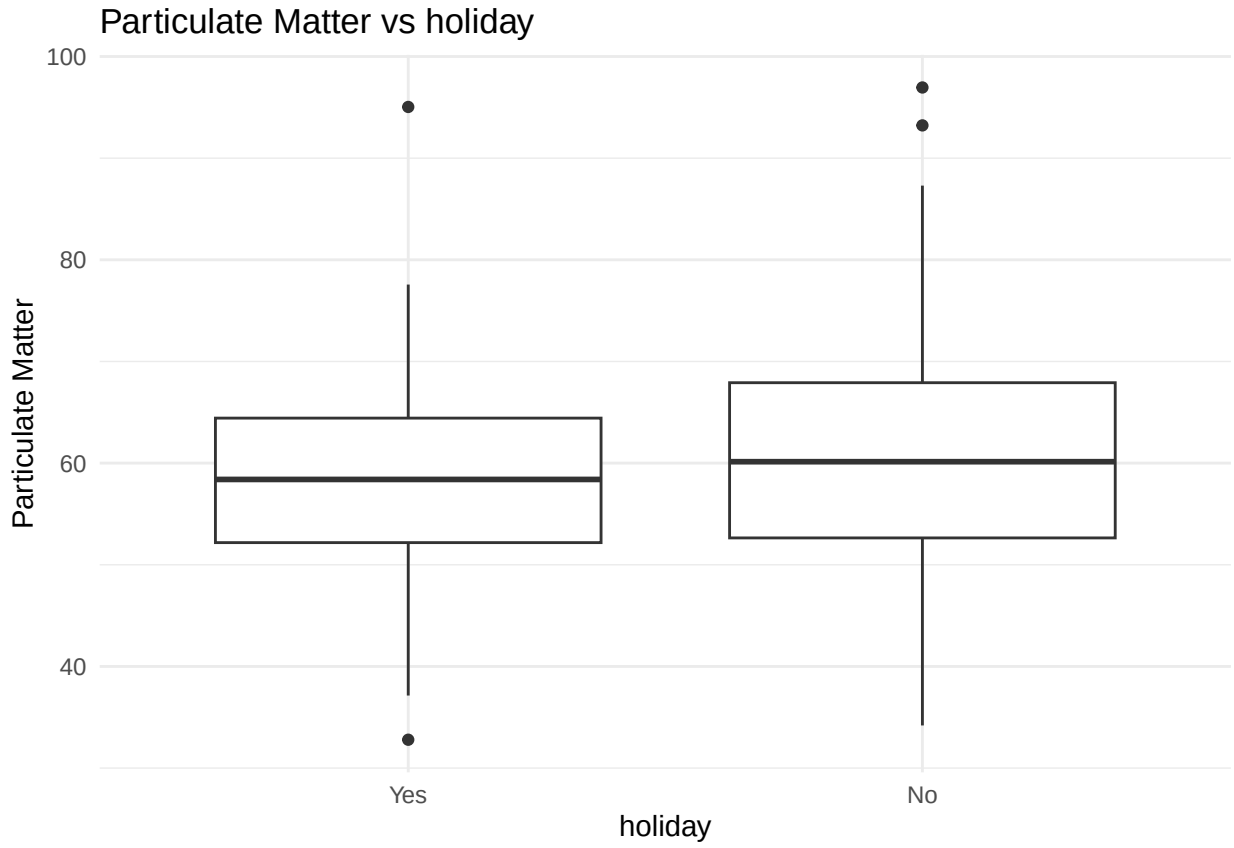
```r
data_tidy_air_quality$industrial_activity <- factor(data_tidy_air_quality$industrial_ac
                                    levels = c("Low", "Medium", "High"))  # Adjust the

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

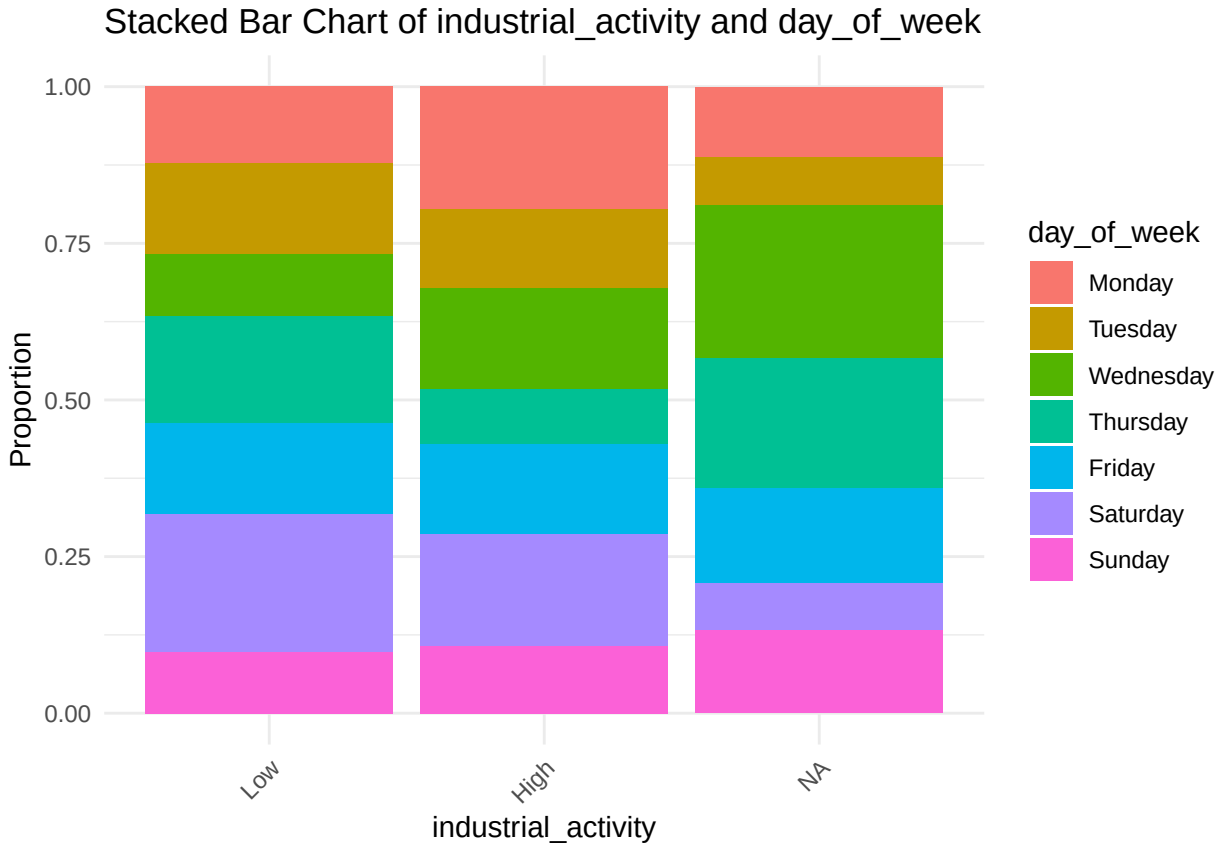
```r
for (i in 1:(length(categorical_vars)-1)) {
  for (j in (i+1):length(categorical_vars)) {
    cat("Contingency Table for", categorical_vars[i], "and", categorical_vars[j], "\n"
    print(table(data_tidy_air_quality[[categorical_vars[i]]], data_tidy_air_quality[[ca
    cat("\n")
  }
}
```

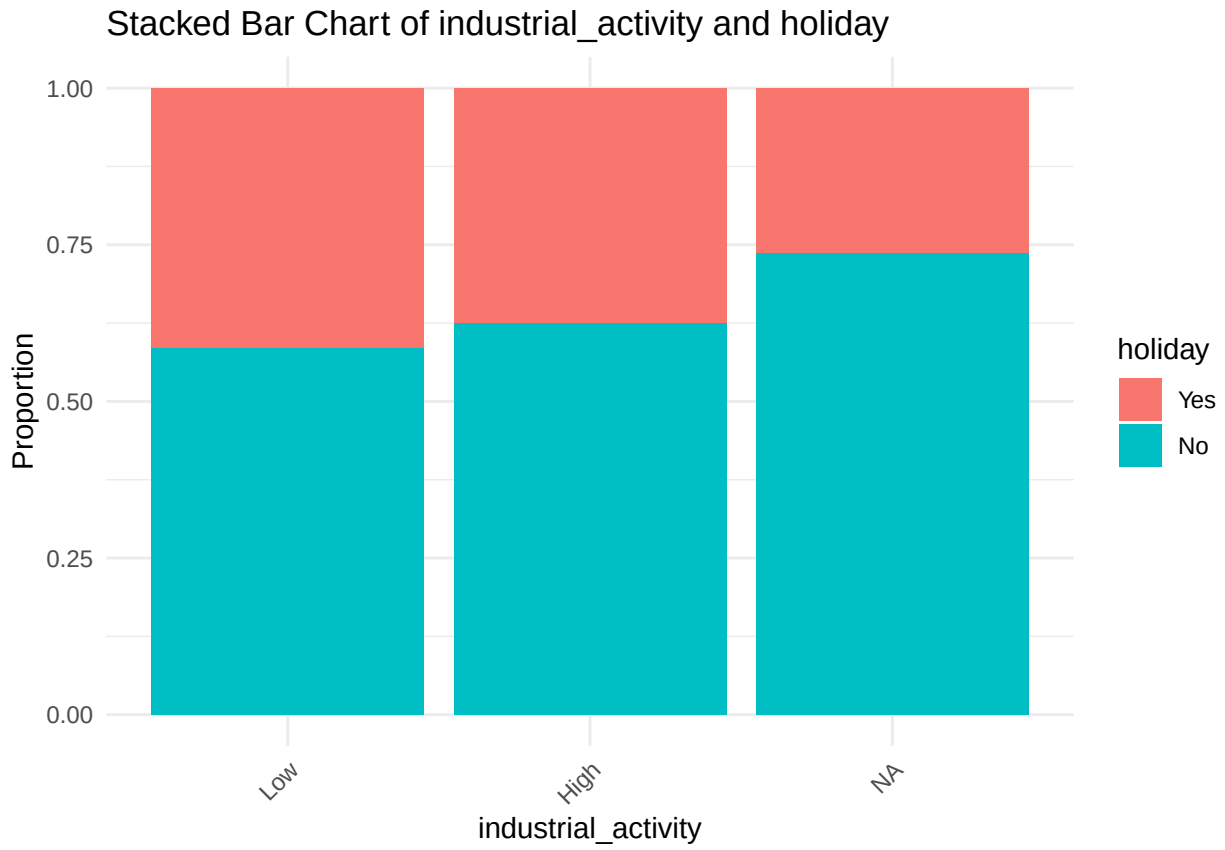
```
## Contingency Table for industrial_activity and day_of_week
##
##          Monday Tuesday Wednesday Thursday Friday Saturday Sunday
##   Low         5       6         4        7      6        9      4
##   Medium      0       0         0        0      0        0      0
##   High       11       7         9        5      8       10      6
##
## Contingency Table for industrial_activity and holiday
##
##          Yes No
##   Low     17 24
##   Medium   0  0
##   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
```

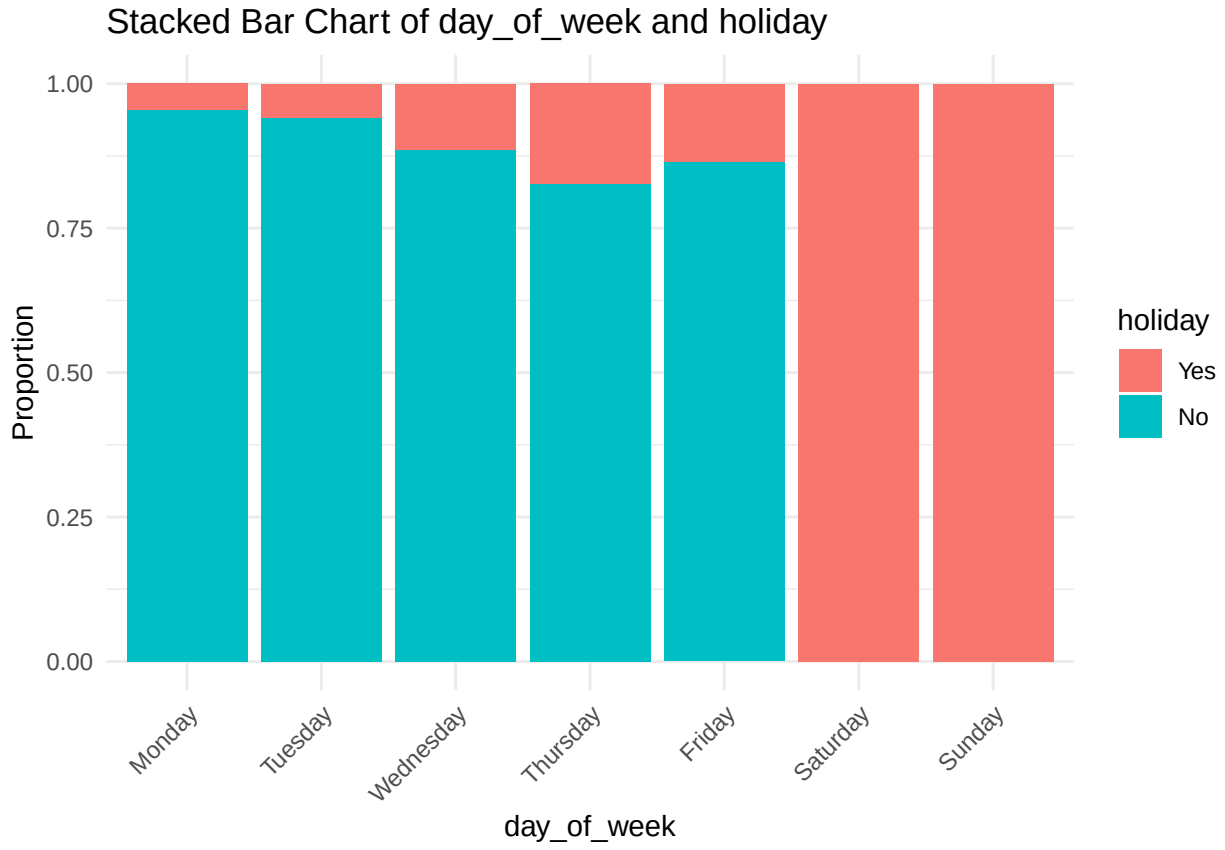
visual representation of relationship between categorial variables

```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.