

# Lab 5: Comprehensive EDA Assignment

Noah Gallego

14 September 2024

## Part I: Data Cleaning and Preperation

### Import Packages

```
library(readxl)
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(tidyr)
library(ggplot2)
```

### Read-In Data

```
df = read.csv("Airpollutants.csv")
head(df)
```

	County	Ozone	Pm2.5	CO	NO2	white.percentage	median.income
## 1	alameda	0.106	9.5	0.21	0.014	31.1	92,574
## 2	butte	0.082	11.4	0.24	0.017	71.6	48,443
## 3	contra costa	0.096	10.0	NA	0.020	43.2	93,712
## 4	fresno	0.074	12.5	0.32	0.016	29	51,261
## 5	humboldt	NA	7.4	0.19	0.015	white	45,528
## 6	imperial	0.111	10.2	0.26	NA	10.4	45,834
##	below.poverty.level						
## 1		9.4					
## 2		16.6					
## 3		8.8					
## 4		19.4					
## 5		19.4					
## 6		17.3					

## Check & Deal w/ Missing Values

```
# Apply to each missing value in DataFrame
print(paste("The number of missing values in the df are:", sum(is.na(df))))

## [1] "The number of missing values in the df are: 4"

numeric_columns = sapply(df, is.numeric)

for(col in names(df)[numeric_columns]) {
  df[[col]][is.na(df[[col]])] = mean(df[[col]], na.rm = TRUE)
}

df$Ozone[is.na(df$Ozone)] = mean(df$Ozone, na.rm = TRUE)
print(paste("The number of missing values in the IMPUTED df are:", sum(is.na(df))))

## [1] "The number of missing values in the IMPUTED df are: 0"

df
```

	County	Ozone	Pm2.5	CO	NO2	white.percentage
## 1	alameda	0.1060000	9.500000	0.2100000	0.01400000	31.1
## 2	butte	0.0820000	11.400000	0.2400000	0.01700000	71.6
## 3	contra costa	0.0960000	10.000000	0.2498889	0.02000000	43.2
## 4	fresno	0.0740000	12.500000	0.3200000	0.01600000	29
## 5	humboldt	0.1027222	7.400000	0.1900000	0.01500000	white
## 6	imperial	0.1110000	10.200000	0.2600000	0.01752778	10.4
## 7	inyo	0.0790000	6.400000	0.1700000	0.01300000	61.8
## 8	kern	0.0750000	13.200000	0.2900000	0.01600000	33.5
## 9	los angeles	0.0700000	11.000000	0.3460000	0.01900000	26.1
## 10	marin	0.0820000	7.500000	0.2200000	0.01700000	71.5
## 11	monterey	0.0740000	7.800000	0.2400000	0.01900000	29.5
## 12	orange	0.0680000	11.000000	0.3100000	0.01500000	40.1
## 13	riverside	0.1390000	9.483333	0.3300000	0.01800000	34.7
## 14	sacramento	0.0650000	8.700000	0.2600000	0.01500000	44.2
## 15	san bernardino	0.0850000	11.500000	0.3300000	0.01700000	27.9
## 16	san diego	0.0670000	7.000000	0.2900000	0.01400000	45.2
## 17	san francisco	0.7400000	8.500000	0.2700000	0.01300000	40.3
## 18	san joaquin	0.0720000	12.100000	0.2400000	0.02400000	31
## 19	san mateo	0.0830000	7.600000	0.2700000	0.01200000	38.9
## 20	santa barbara	0.0820000	7.600000	0.2700000	0.01800000	44.1
## 21	santa clara	0.1070000	9.000000	0.2300000	0.02000000	31
## 22	solano	0.0920000	9.700000	0.2500000	0.01500000	37.6
## 23	sonoma	0.0680000	7.300000	0.2100000	0.01600000	63.1
## 24	stanilaus	0.1040000	10.500000	0.3000000	0.01500000	41.1
## 25	ventura	0.0640000	7.900000	0.2800000	0.01900000	45
## 26	santa cruz	0.0700000	7.200000	0.2100000	0.02100000	56.9
## 27	tulare	0.1140000	14.900000	0.2800000	0.01700000	28.1
## 28	San Luis Obispo	0.0800000	8.000000	0.2300000	0.02300000	68.6
## 29	san benito	0.0790000	6.500000	0.1900000	0.01900000	33.5
## 30	nevada	0.1020000	8.800000	0.2000000	0.01600000	84.9
## 31	merced	0.0960000	12.300000	0.2600000	0.02200000	27.1
## 32	mendocino	0.0650000	9.200000	0.2200000	0.01800000	64.7
## 33	lake	0.0630000	6.300000	0.1800000	0.01500000	69.7
## 34	kings	0.0930000	15.900000	0.2500000	0.02500000	31.8
## 35	colusa	0.0720000	10.200000	0.2100000	0.02000000	34.3

```
## 36      calaveras 0.0990000  9.000000 0.2000000 0.01900000      80.7
## 37      yolo 0.0800000  7.800000 0.2400000 0.01900000      46.3
##      median.income below.poverty.level
## 1      92,574      9.4
## 2      48,443     16.6
## 3      93,712      8.8
## 4      51,261     19.4
## 5      45,528     19.4
## 6      45,834     17.3
## 7      52,874     12.8
## 8      52,479     18.5
## 9      64,251     14.1
## 10     110,217      7.8
## 11     66,676     12.1
## 12     85,398      9.9
## 13     63,948     11.6
## 14     63,902     13.0
## 15     60,164     13.2
## 16     74,855     10.7
## 17     104,552     11.4
## 18     61,145     12.3
## 19     113,776      6.8
## 20     71,657     15.2
## 21     116,178      6.9
## 22     77,609     10.0
## 23     76,753      9.1
## 24     57,387     14.1
## 25     84,017      8.9
## 26     78,041     10.6
## 27     47,518     18.7
## 28     70,699     13.1
## 29     81,977      8.9
## 30     63,240     11.8
## 31     50,129     21.9
## 32     49,233     16.1
## 33     42,475     16.5
## 34     53,865     17.7
## 35     56,704     11.4
## 36     58,151     13.5
## 37     65,923     14.8
```

### Ensure Correct DTypes throughout the DataFrame

```
# Check DTypes
str(df)
```

```
## 'data.frame':   37 obs. of  8 variables:
## $ County      : chr  "alameda " "butte" "contra costa" "fresno" ...
## $ Ozone       : num  0.106 0.082 0.096 0.074 0.103 ...
## $ Pm2.5       : num  9.5 11.4 10 12.5 7.4 10.2 6.4 13.2 11 7.5 ...
## $ CO          : num  0.21 0.24 0.25 0.32 0.19 ...
## $ NO2         : num  0.014 0.017 0.02 0.016 0.015 ...
## $ white.percentage : chr  "31.1" "71.6" "43.2" "29" ...
## $ median.income : chr  "92,574" "48,443" "93,712" "51,261" ...
```

```
## $ below.poverty.level: num  9.4 16.6 8.8 19.4 19.4 17.3 12.8 18.5 14.1 7.8 ...
# white.percentage & median.income are affected columns
df$white.percentage[df$white.percentage == "white"] = NA
df$white.percentage = as.numeric(df$white.percentage)
df$white.percentage[is.na(df$white.percentage)] = mean(df$white.percentage, na.rm = TRUE)

df$median.income = gsub(',', '', df$median.income)
df$median.income = as.numeric(df$median.income)

str(df)
```

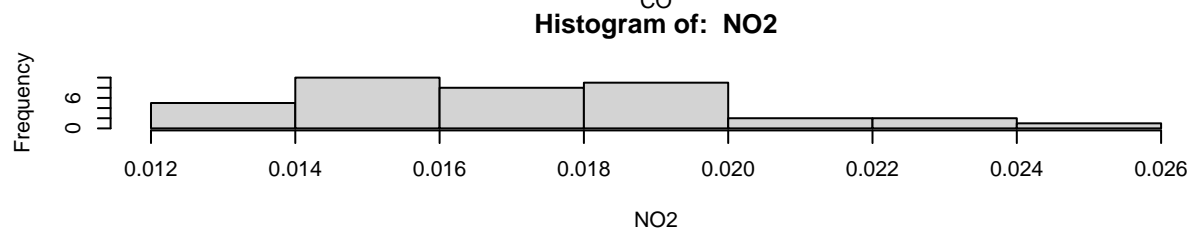
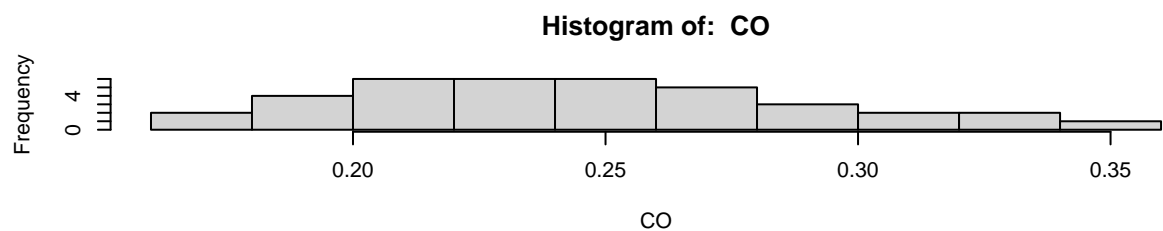
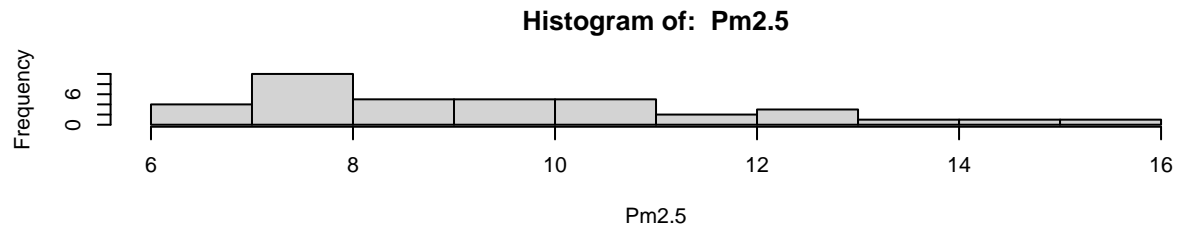
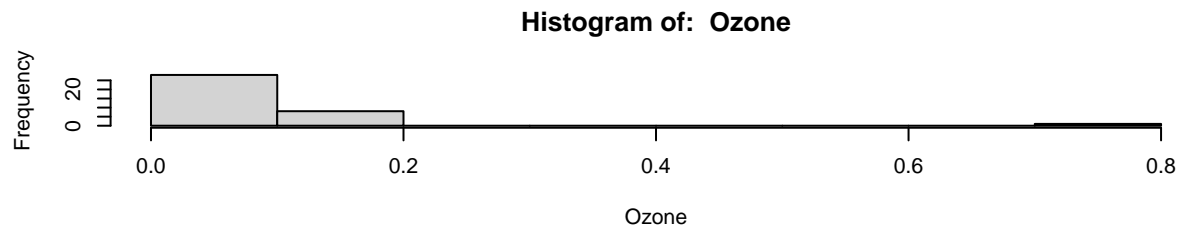
```
## 'data.frame':  37 obs. of  8 variables:
## $ County      : chr  "alameda " "butte" "contra costa" "fresno" ...
## $ Ozone       : num  0.106 0.082 0.096 0.074 0.103 ...
## $ Pm2.5       : num  9.5 11.4 10 12.5 7.4 10.2 6.4 13.2 11 7.5 ...
## $ CO          : num  0.21 0.24 0.25 0.32 0.19 ...
## $ NO2         : num  0.014 0.017 0.02 0.016 0.015 ...
## $ white.percentage : num  31.1 71.6 43.2 29 44.4 ...
## $ median.income  : num  92574 48443 93712 51261 45528 ...
## $ below.poverty.level: num  9.4 16.6 8.8 19.4 19.4 17.3 12.8 18.5 14.1 7.8 ...
```

## View Data Distribution

```
view_dist = function(df) {
  par(mfrow=c(3, 1))

  for (col in names(df)[numeric_columns]) {
    hist(df[[col]], main = paste("Histogram of: ", col), xlab=col)
  }
}

view_dist(df)
```



## Handling Outliers

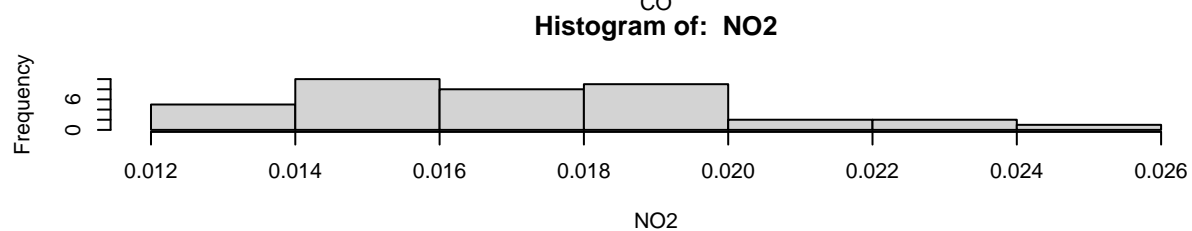
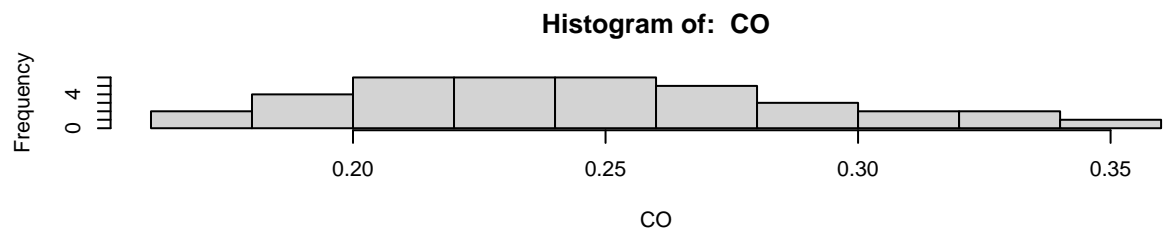
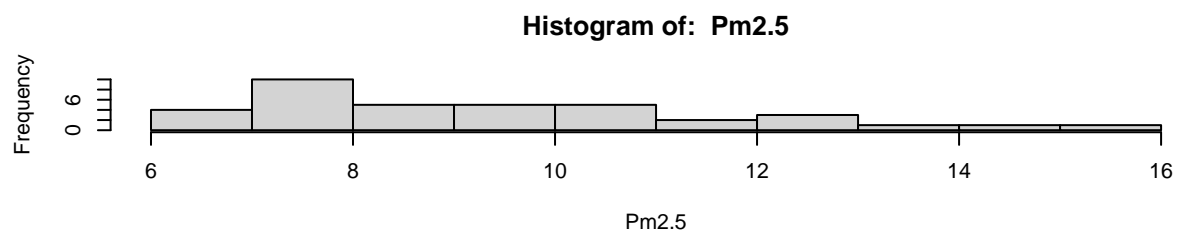
We will use the IQR method to detect outliers. To handle these outliers, we will impute them with the median since the dataset is relatively small and the data is slightly skewed.

```
for (col in names(df)[numeric_columns]) {
  q1 = quantile(df[[col]], 0.25, na.rm = TRUE) # Q1
  q3 = quantile(df[[col]], 0.75, na.rm = TRUE) # Q3
  iqr = IQR(df[[col]], na.rm = TRUE)

  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
}
```

```
df[[col]] [df[[col]] < lower_bound | df[[col]] > upper_bound] = mean(df[[col]])
}

view_dist(df)
```



## Part II: Numerical Summaries and Visualizations

### Numerical Summaries

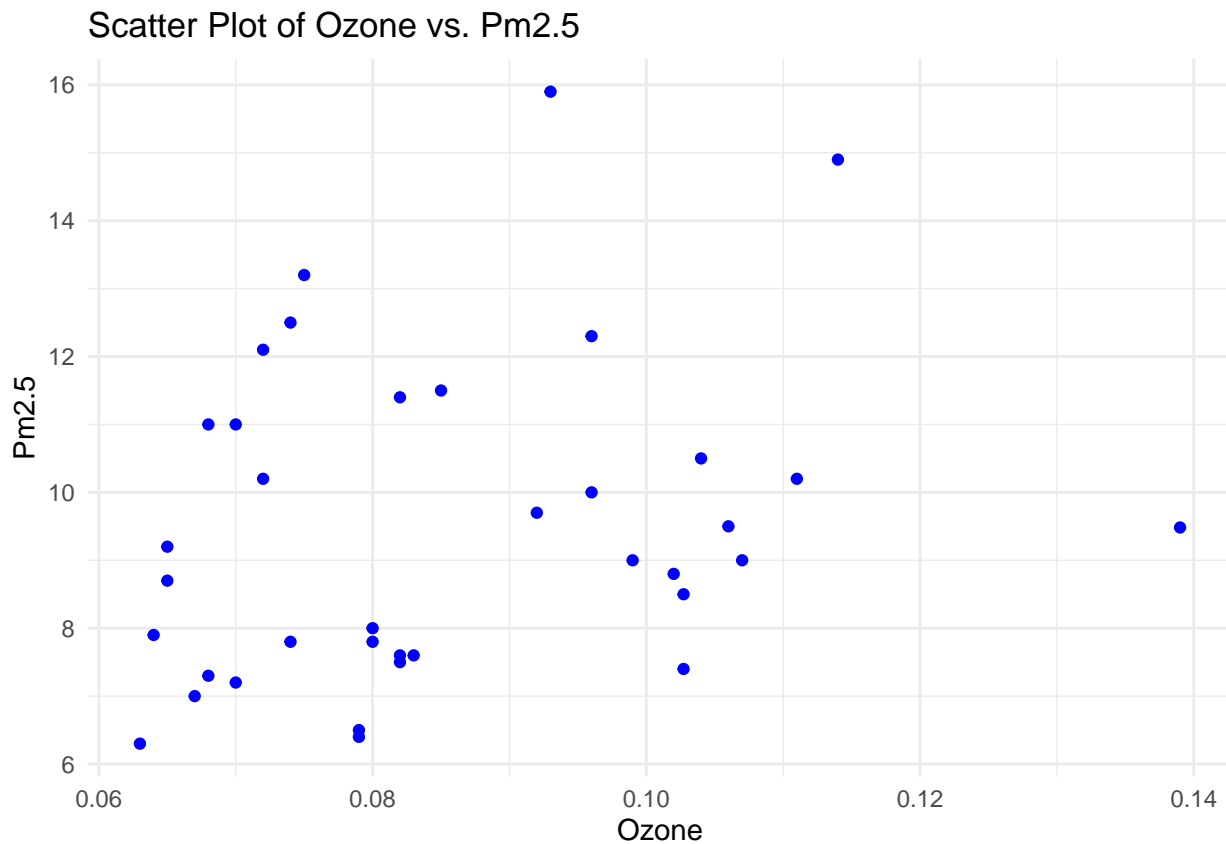
```
summary(df)
```

```
##      County      Ozone      Pm2.5      CO
## Length:37      Min.    :0.0630      Min.    : 6.300      Min.    :0.1700
## Class :character 1st Qu.:0.0720      1st Qu.: 7.600      1st Qu.:0.2100
## Mode  :character Median :0.0820      Median : 9.000      Median :0.2499
##                      Mean  :0.0855      Mean  : 9.483      Mean  :0.2499
##                      3rd Qu.:0.0990      3rd Qu.:11.000      3rd Qu.:0.2800
##                      Max.   :0.1390      Max.   :15.900      Max.   :0.3460
##      NO2      white.percentage median.income below.povery.level
## Min.    :0.01200      Min.    :10.4      Min.    : 42475      Min.    : 6.80
## 1st Qu.:0.01500      1st Qu.:31.1      1st Qu.: 52874      1st Qu.:10.00
## Median :0.01700      Median :40.3      Median : 63948      Median :12.80
## Mean   :0.01753      Mean   :44.4      Mean   : 69004      Mean   :13.09
## 3rd Qu.:0.01900      3rd Qu.:56.9      3rd Qu.: 78041      3rd Qu.:16.10
## Max.   :0.02500      Max.   :84.9      Max.   :116178      Max.   :21.90
```

### Scatterplots & Pairwise Analysis

```
par(mfrow=c(4, 1))

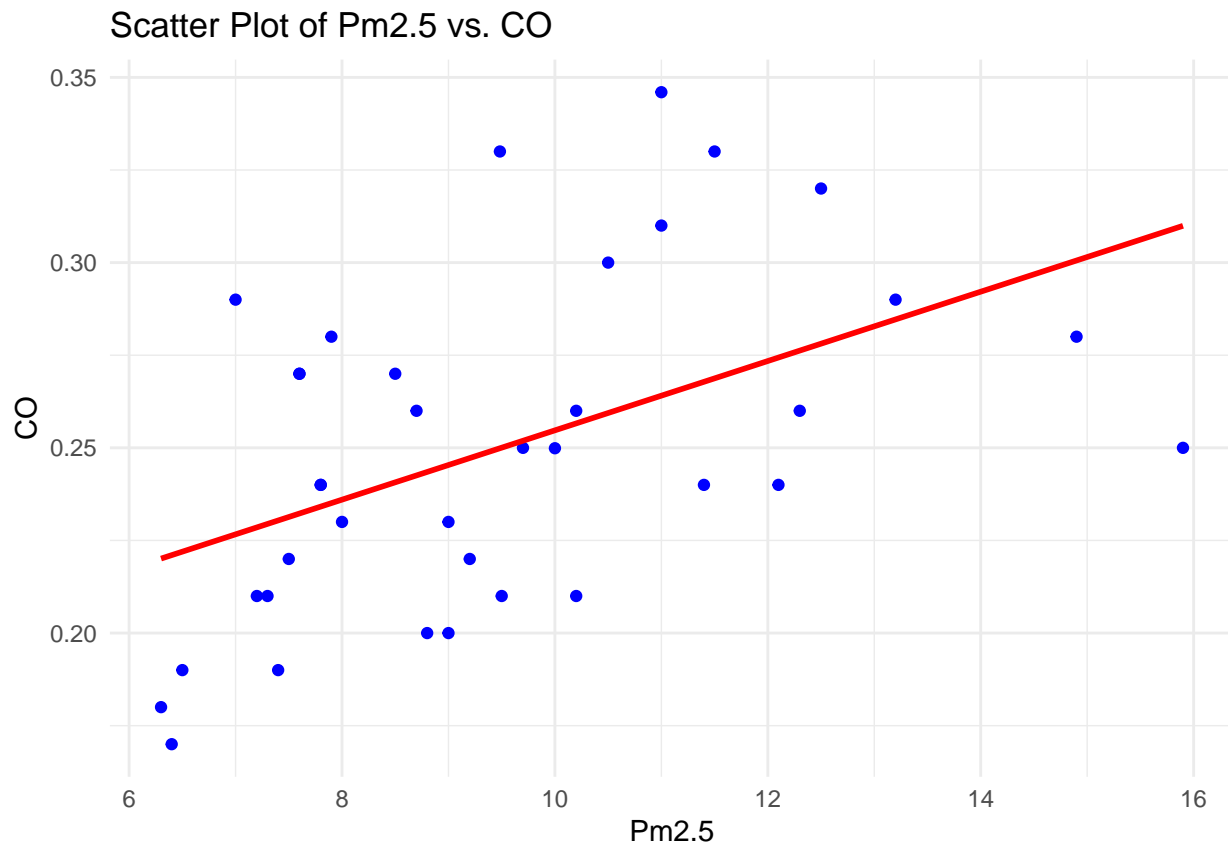
# Ozone V. Pm2.5
ggplot(df, aes(x = Ozone, y = Pm2.5)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot of Ozone vs. Pm2.5", x = "Ozone", y = "Pm2.5") +
  theme_minimal()
```



```
# Pm2.5 vs. CO
ggplot(df, aes(x = Pm2.5, y = CO)) +
```

```
geom_point(color = "blue") +
geom_smooth(method = "lm", se = FALSE, color = "red") +
labs(title = "Scatter Plot of Pm2.5 vs. CO", x = "Pm2.5", y = "CO") +
theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

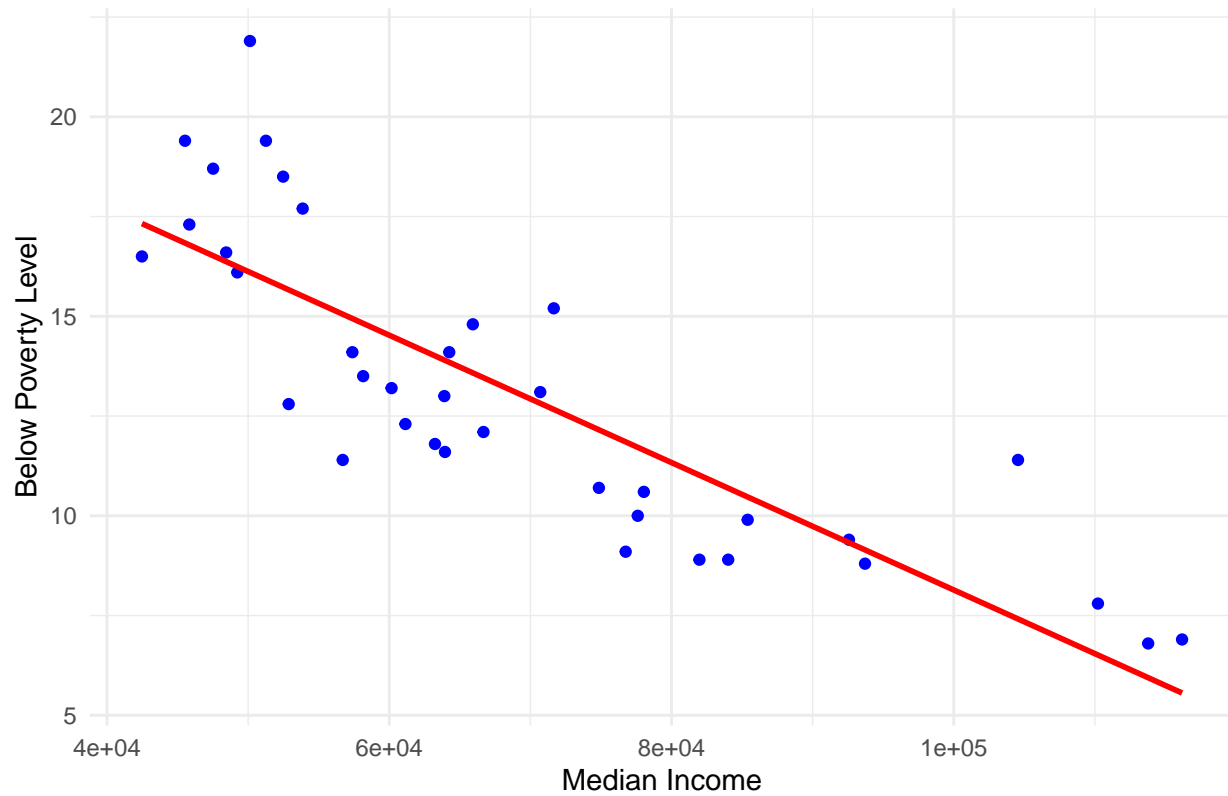


```
# Median Income V. Below Poverty Level
ggplot(df, aes(x = median.income, y = below.poverty.level)) +
geom_point(color = "blue") +
geom_smooth(method = "lm", se = FALSE, color = "red") +
labs(title = "Scatter Plot of Median Income vs. Below Poverty Level", x = "Median Income", y = "Below
theme_minimal()
```

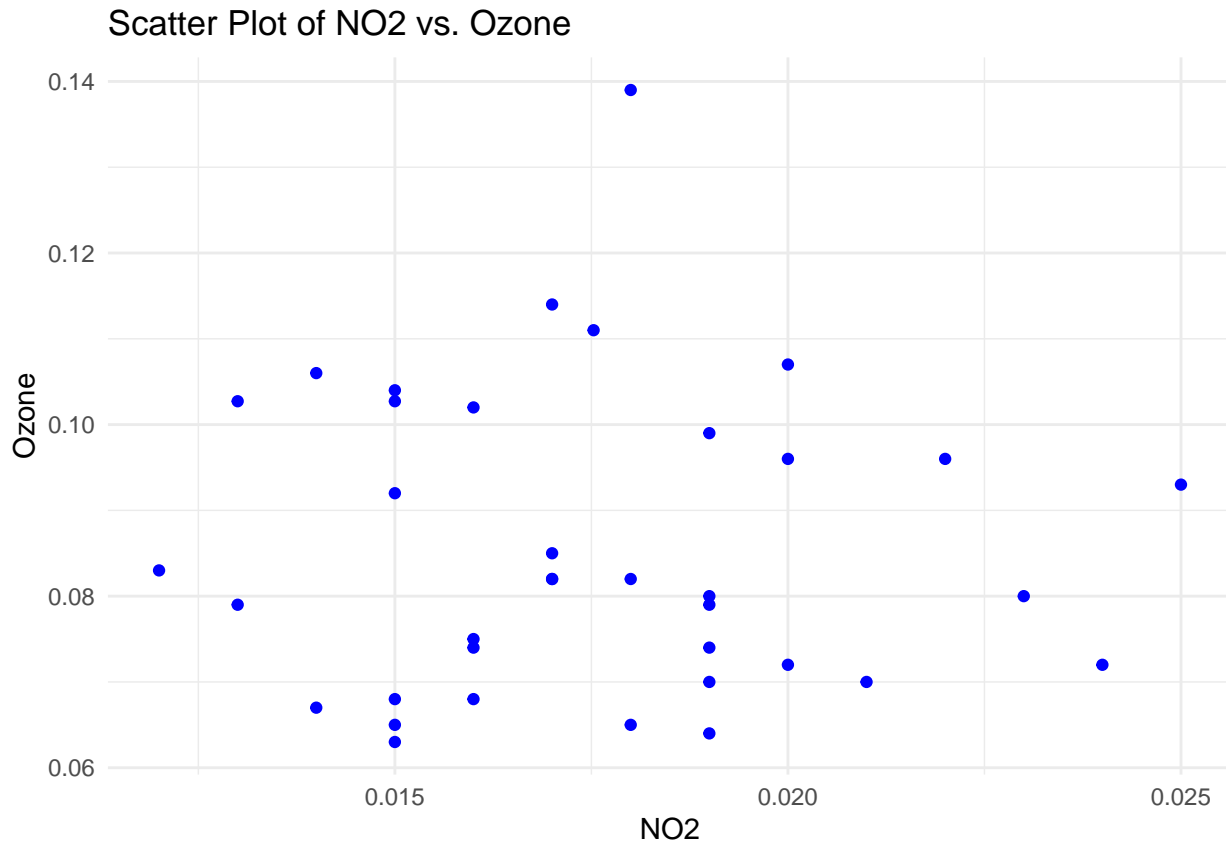
```
## `geom_smooth()` using formula = 'y ~ x'
```



Scatter Plot of Median Income vs. Below Poverty Level



```
# NO2 V. Ozone
ggplot(df, aes(x = NO2, y = Ozone)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot of NO2 vs. Ozone", x = "NO2", y = "Ozone") +
  theme_minimal()
```



### Correlation

```
# Calculate Pearson's Correlation
corr_Pm2_Ozone = cor(df$Ozone, df$Pm2.5, method = 'pearson')
corr_Pm2_CO = cor(df$Pm2.5, df$CO, method = 'pearson')
corr_Income_Poverty = cor(df$median.income, df$below.poverty.level, method = 'pearson')
corr_NO2_Ozone = cor(df$NO2, df$Ozone, method = 'pearson')

cat("Correlation between Ozone and Pm2.5:", corr_Pm2_Ozone, "\n")

## Correlation between Ozone and Pm2.5: 0.2444828
cat("Correlation between Pm2.5 and CO:", corr_Pm2_CO, "\n")

## Correlation between Pm2.5 and CO: 0.486746
cat("Correlation between Median Income and Below Poverty Level:", corr_Income_Poverty, "\n")

## Correlation between Median Income and Below Poverty Level: -0.8323263
cat("Correlation between NO2 and Ozone:", corr_NO2_Ozone, "\n")

## Correlation between NO2 and Ozone: -0.02552856
```

**1. Correlation between Ozone and Pm2.5: 0.244:** - The correlation is weakly positive. This suggests that there is a slight tendency for higher Ozone levels to be associated with higher Pm2.5 levels, but the relationship is not strong. Factors other than Pm2.5 are likely influencing Ozone levels.

**2. Correlation between Pm2.5 and CO: 0.487:** - This is a moderate positive correlation. It indicates that as Pm2.5 levels increase, CO levels also tend to increase. This suggests that these two pollutants may

be related, potentially coming from similar sources such as vehicle emissions or industrial activities.

**3. Correlation between Median Income and Below Poverty Level: -0.832:** - This is a strong negative correlation, which is expected. As median income increases, the percentage of people below the poverty level decreases significantly. This suggests a strong inverse relationship, where areas with higher income levels have fewer people living below the poverty line.

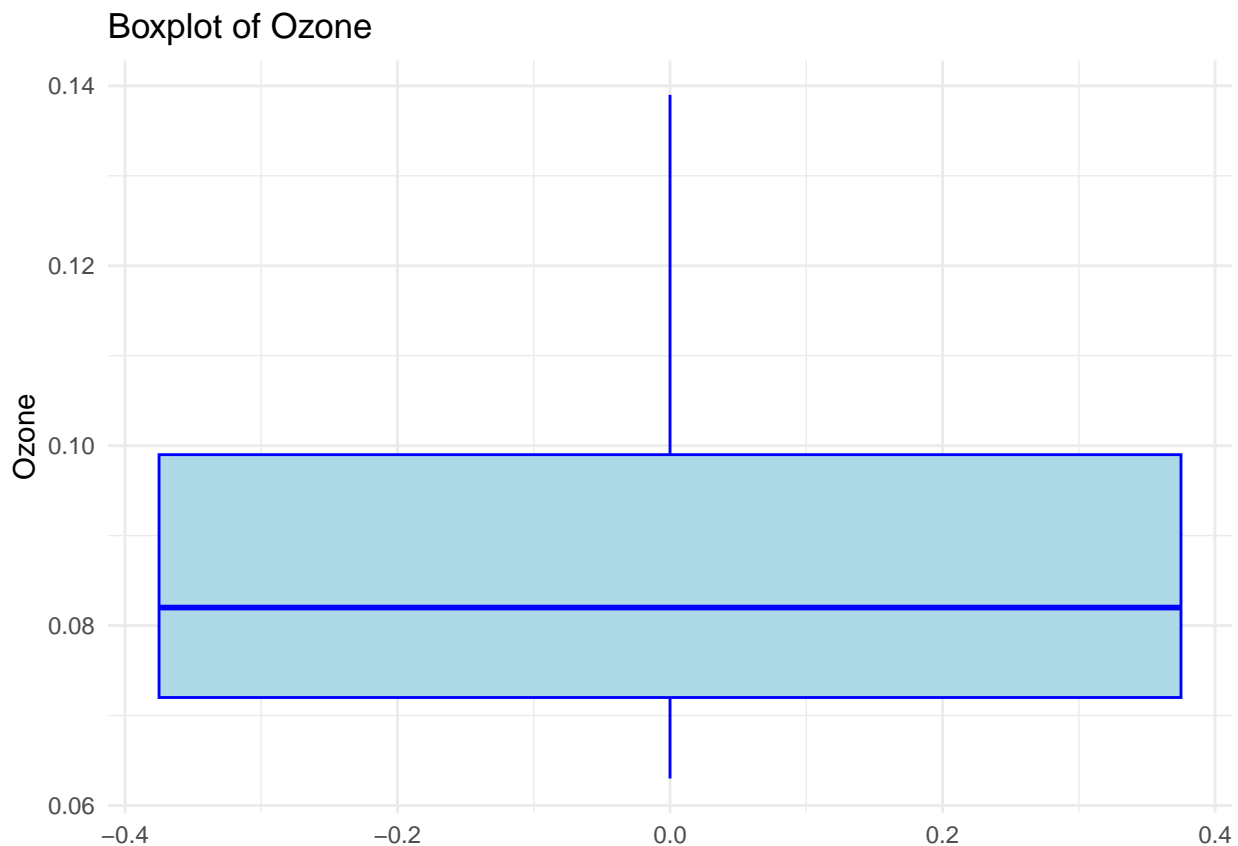
**4. Correlation between NO2 and Ozone: -0.026:** - This is a very weak negative correlation. The near-zero value indicates that there is no significant linear relationship between NO2 and Ozone in the dataset. These two pollutants do not seem to be directly related, or their relationship is influenced by other factors not captured in this simple correlation.

## Histograms & Boxplots

```
# Histograms Already Completed in Prior Step **

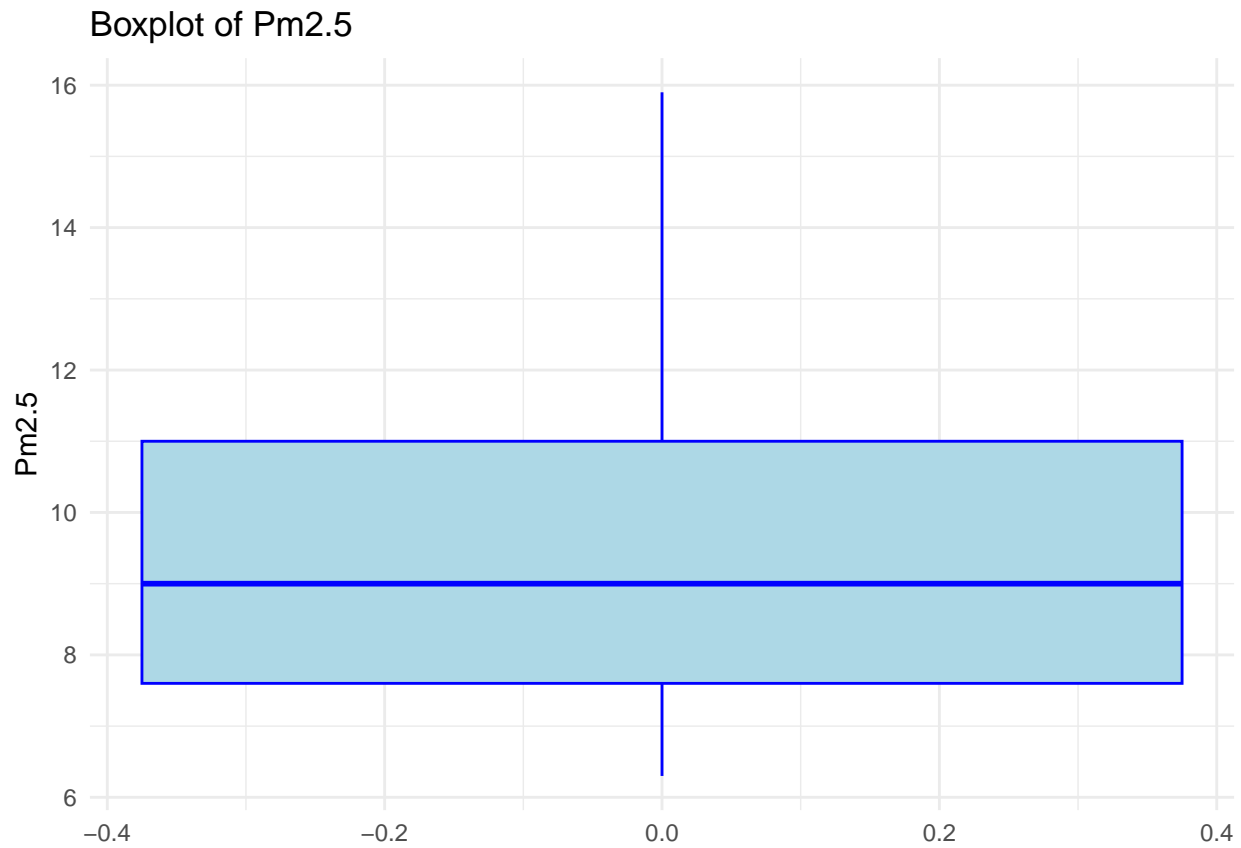
par(mfrow=c(3, 1))

# Create Box Plots
# Ozone
ggplot(df, aes(y = Ozone)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  labs(title = "Boxplot of Ozone", y = "Ozone") +
  theme_minimal()
```

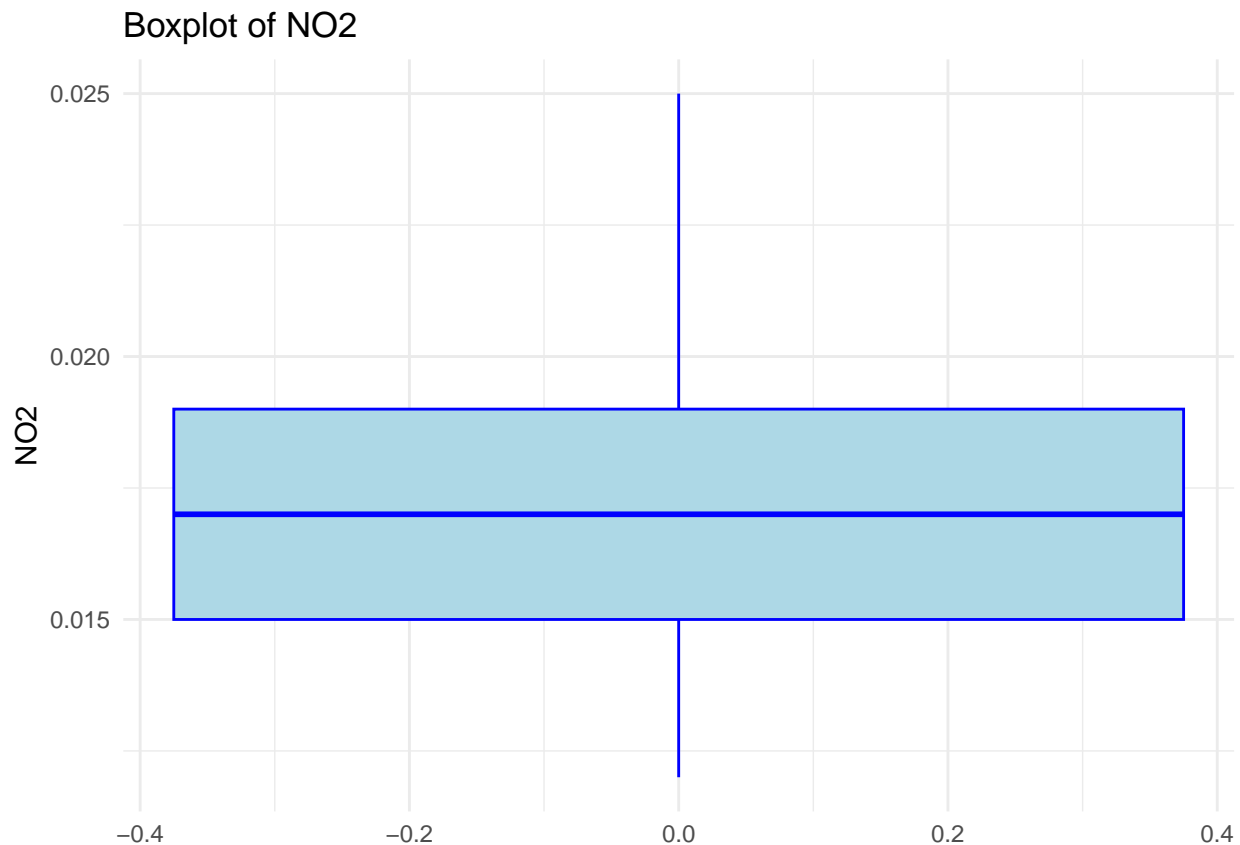


```
# Pm2.5
ggplot(df, aes(y = Pm2.5)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
```

```
labs(title = "Boxplot of Pm2.5", y = "Pm2.5") +  
theme_minimal()
```



```
# NO2  
ggplot(df, aes(y = NO2)) +  
  geom_boxplot(fill = "lightblue", color = "blue") +  
  labs(title = "Boxplot of NO2", y = "NO2") +  
  theme_minimal()
```



## Part III: Creating and Analyzing New Categorical Variables

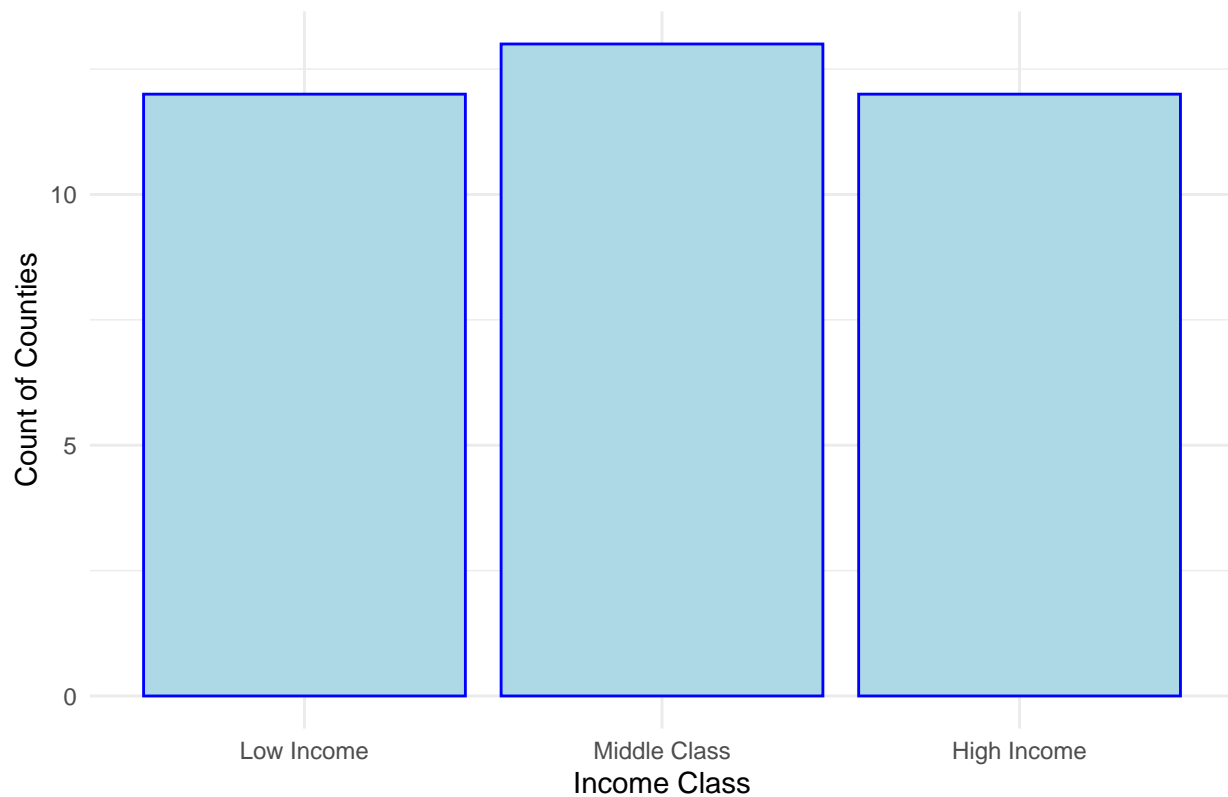
### Creating Income Classes

```
income_33rd <- quantile(df$median.income, 0.33)
income_67th <- quantile(df$median.income, 0.67)

df$income_class = cut(df$median.income,
  breaks = c(-Inf, income_33rd, income_67th, Inf),
  labels = c("Low Income", "Middle Class", "High Income")
)

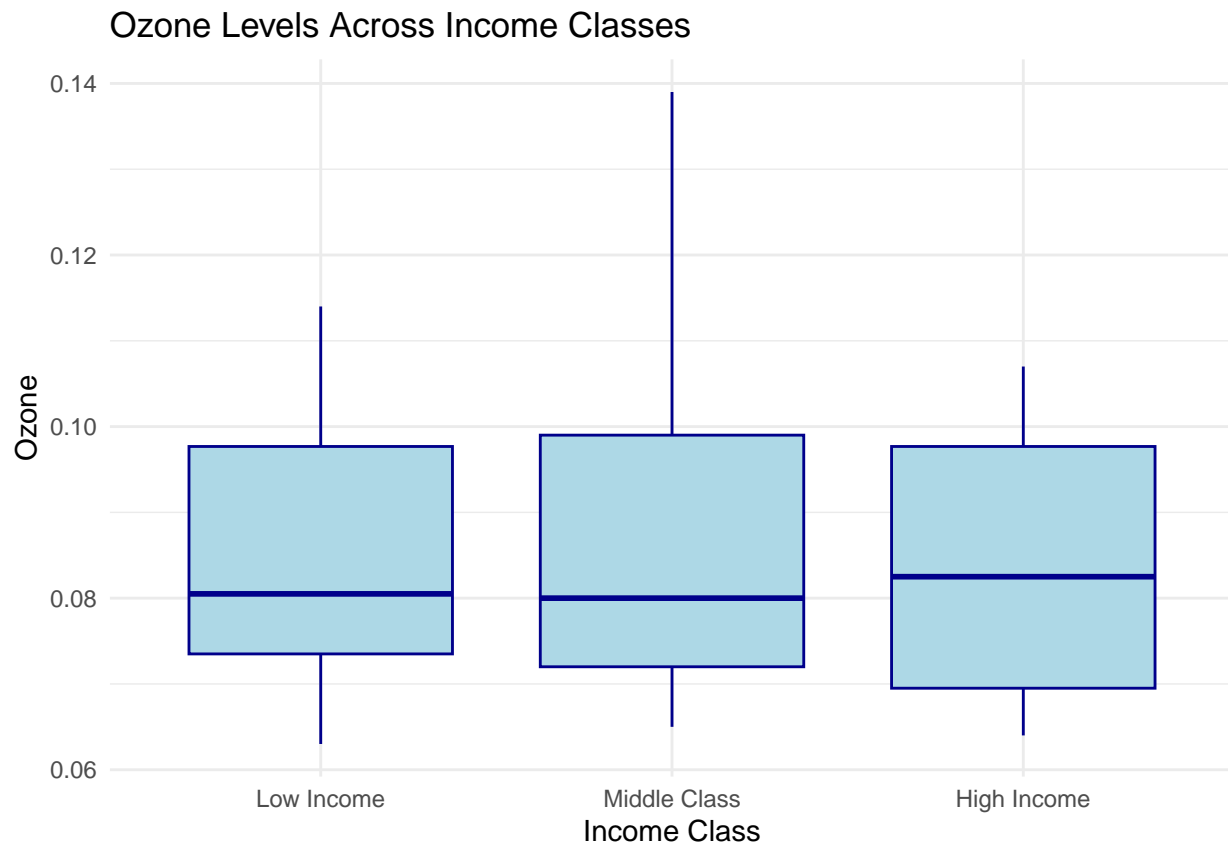
ggplot(df, aes(x = income_class)) +
  geom_bar(fill = "lightblue", color = "blue") +
  labs(title = "Distribution of Income Classes", x = "Income Class", y = "Count of Counties") +
  theme_minimal()
```

Distribution of Income Classes

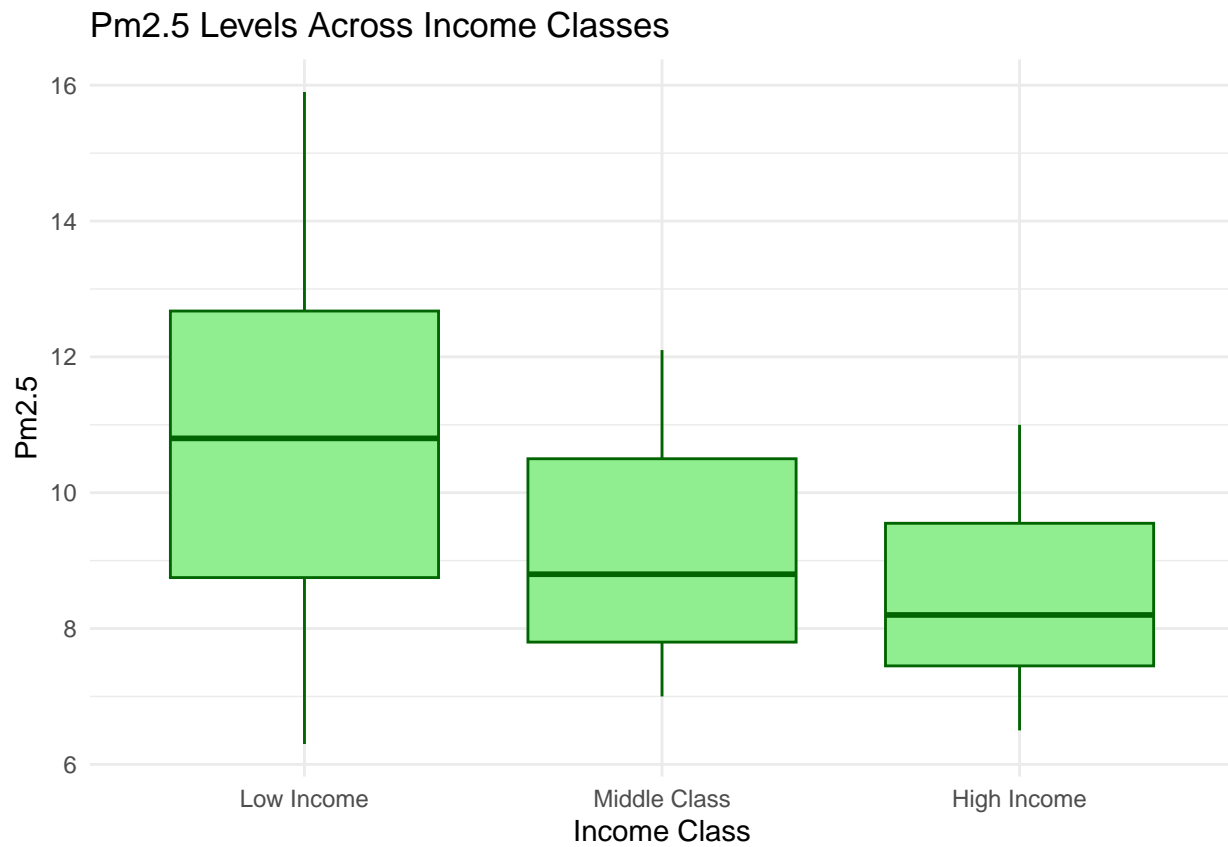


Analysis By Income Classes

```
# Ozone boxplot by income class  
ggplot(df, aes(x = income_class, y = Ozone)) +  
  geom_boxplot(fill = "lightblue", color = "darkblue") +  
  labs(title = "Ozone Levels Across Income Classes", x = "Income Class", y = "Ozone") +  
  theme_minimal()
```

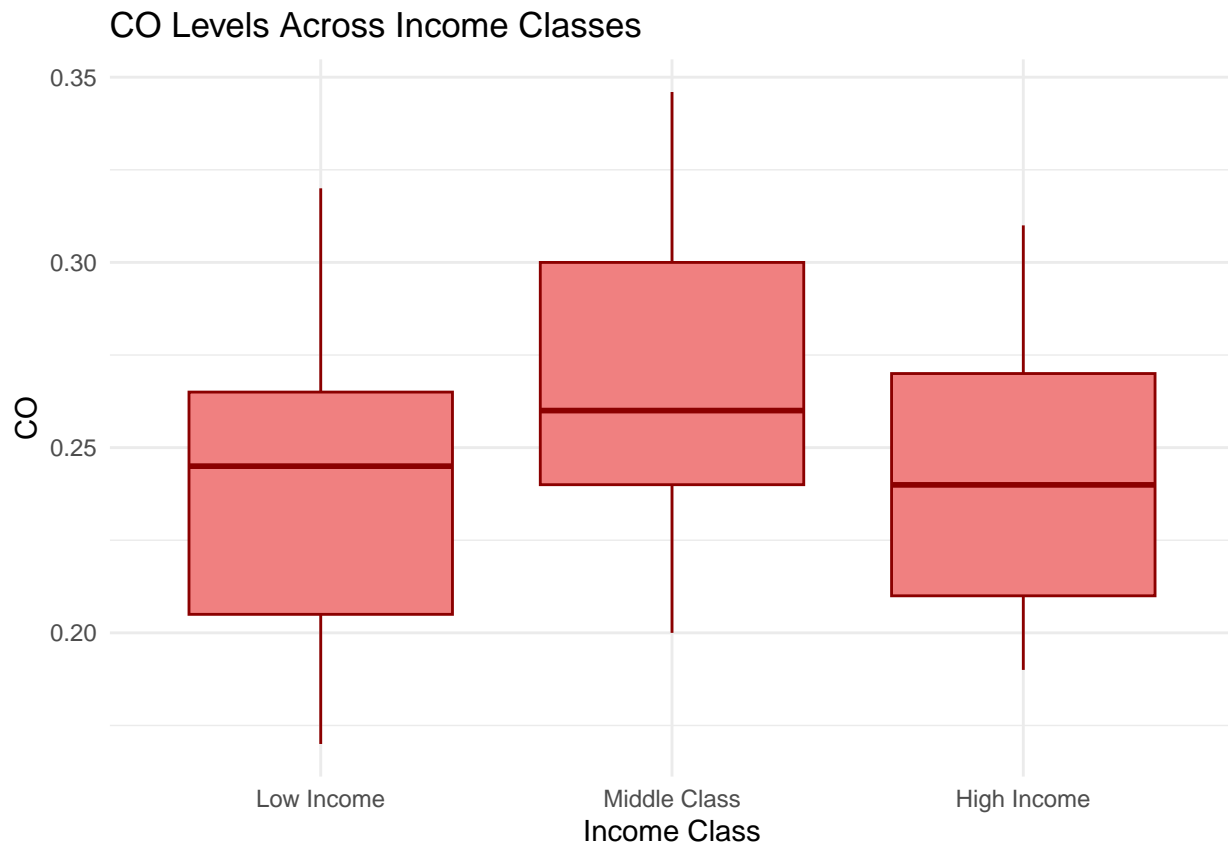


```
# Pm2.5 boxplot by income class
ggplot(df, aes(x = income_class, y = Pm2.5)) +
  geom_boxplot(fill = "lightgreen", color = "darkgreen") +
  labs(title = "Pm2.5 Levels Across Income Classes", x = "Income Class", y = "Pm2.5") +
  theme_minimal()
```

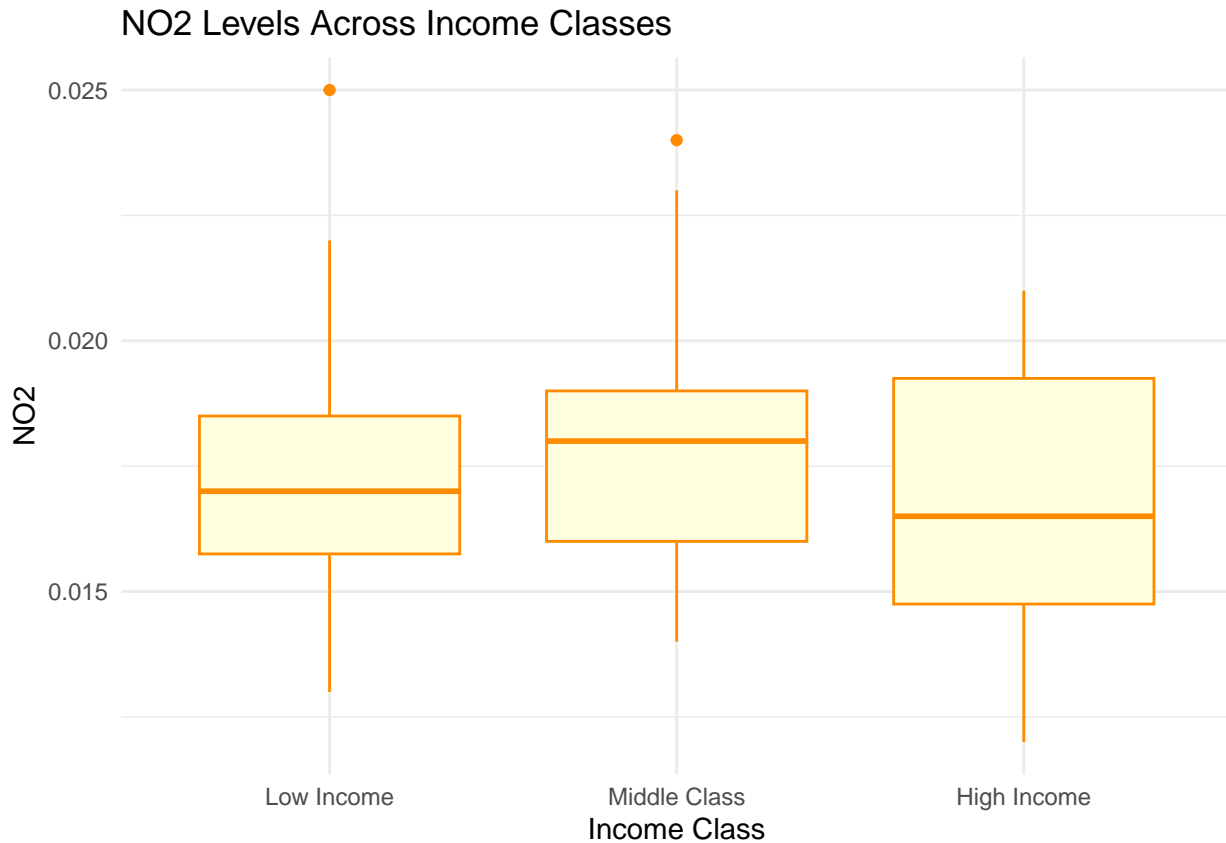


```
# CO boxplot by income class
ggplot(df, aes(x = income_class, y = CO)) +
  geom_boxplot(fill = "lightcoral", color = "darkred") +
  labs(title = "CO Levels Across Income Classes", x = "Income Class", y = "CO") +
  theme_minimal()
```





```
# NO2 boxplot by income class  
ggplot(df, aes(x = income_class, y = NO2)) +  
  geom_boxplot(fill = "lightyellow", color = "darkorange") +  
  labs(title = "NO2 Levels Across Income Classes", x = "Income Class", y = "NO2") +  
  theme_minimal()
```



**Ozone:** No significant differences in Ozone levels across income classes; median and IQR are similar for all groups.

**Pm2.5:** Higher levels and greater variability in low-income counties, suggesting particulate pollution is more prevalent in these areas. High-income counties have the lowest Pm2.5 levels.

**CO:** Slightly higher CO levels in middle-income counties, but overall CO distributions are similar across all income classes.

**NO2:** Minimal differences across income classes, with consistent distributions and a couple of outliers in middle and high-income groups.

In general, Pm2.5 shows the clearest disparity, with higher levels in lower-income counties, while Ozone, CO, and NO2 remain relatively stable across income classes.

### Analyzing By Poverty Levels

```
df$poverty_class <- cut(df$below.poverty.level,
                        breaks = c(-Inf, 15, 30, Inf),
                        labels = c("Low Poverty", "Moderate Poverty", "High Poverty"))

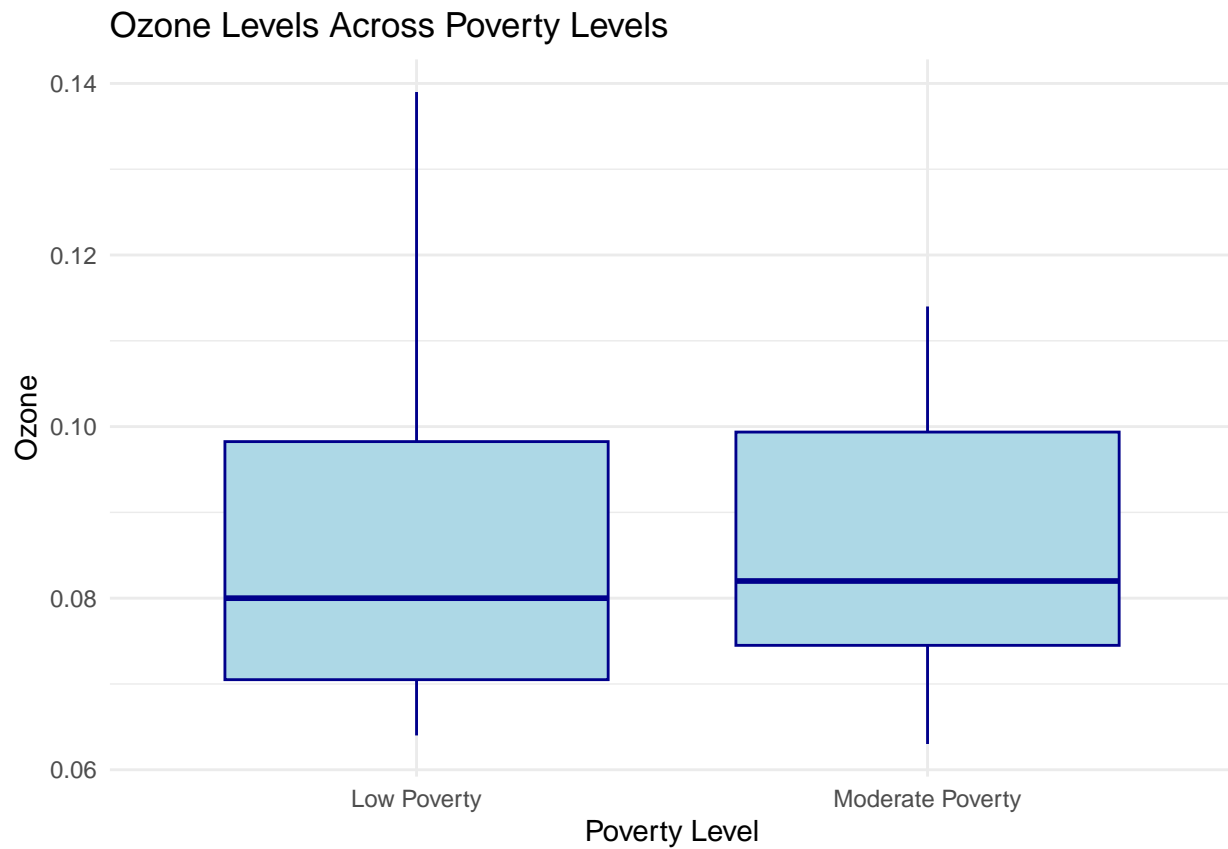
ggplot(df, aes(x = poverty_class)) +
  geom_bar(fill = "skyblue", color = "blue") +
  labs(title = "Distribution of Counties Across Poverty Levels", x = "Poverty Level", y = "Count of Counties") +
  theme_minimal()
```

Distribution of Counties Across Poverty Levels



Analysis By Poverty Levels

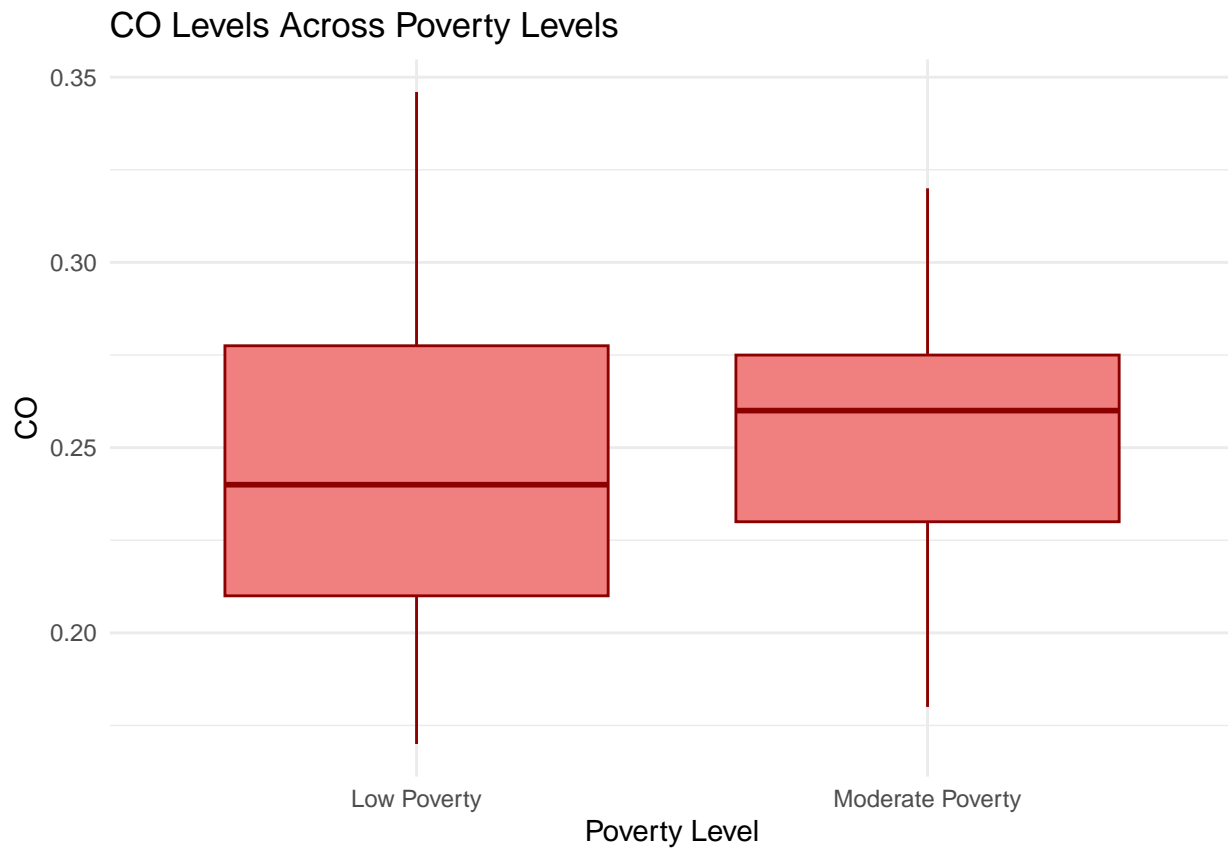
```
# Ozone boxplot by poverty class
ggplot(df, aes(x = poverty_class, y = Ozone)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Ozone Levels Across Poverty Levels", x = "Poverty Level", y = "Ozone") +
  theme_minimal()
```



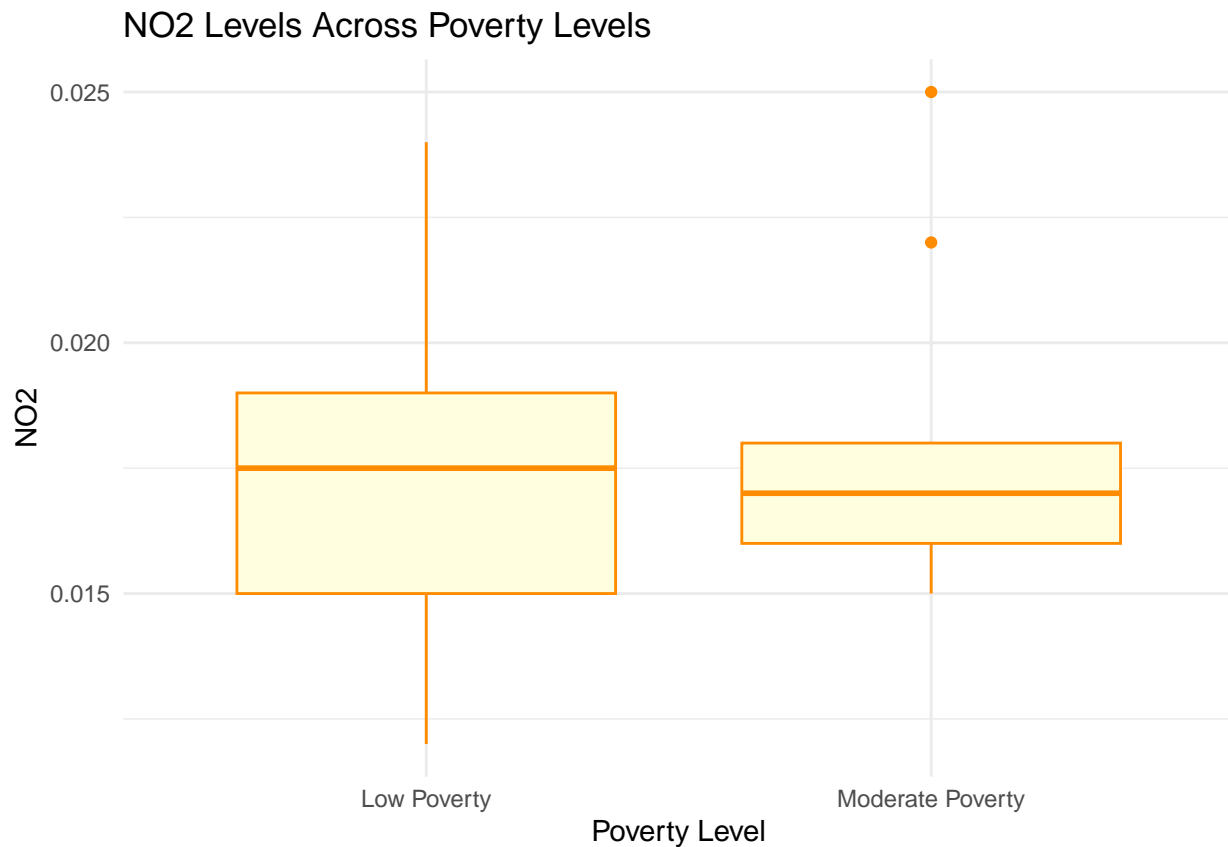
```
# Pm2.5 boxplot by poverty class
ggplot(df, aes(x = poverty_class, y = Pm2.5)) +
  geom_boxplot(fill = "lightgreen", color = "darkgreen") +
  labs(title = "Pm2.5 Levels Across Poverty Levels", x = "Poverty Level", y = "Pm2.5") +
  theme_minimal()
```



```
# CO boxplot by poverty class
ggplot(df, aes(x = poverty_class, y = CO)) +
  geom_boxplot(fill = "lightcoral", color = "darkred") +
  labs(title = "CO Levels Across Poverty Levels", x = "Poverty Level", y = "CO") +
  theme_minimal()
```



```
# NO2 boxplot by poverty class
ggplot(df, aes(x = poverty_class, y = NO2)) +
  geom_boxplot(fill = "lightyellow", color = "darkorange") +
  labs(title = "NO2 Levels Across Poverty Levels", x = "Poverty Level", y = "NO2") +
  theme_minimal()
```



Metrics that we measured did not indicate much change between Low and Moderate Poverty classes. *However:*

- \* Moderate poverty counties appear to have higher Pm2.5 levels and more variability compared to low-poverty counties. This suggests that moderate poverty areas may experience worse air quality in terms of particulate matter pollution.
- \* Low poverty counties show lower and more consistent Pm2.5 levels, indicating better air quality in terms of particulate pollution.

*# Check the unique values and the distribution of poverty levels*

```
summary(df$below.poverty.level)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      6.80  10.00   12.80   13.09  16.10   21.90
```

```
table(df$poverty_class)
```

```
##
##      Low Poverty Moderate Poverty    High Poverty
##             26             11             0
```