Air-Quality – Ozone Analysis

Sandip Banerjee

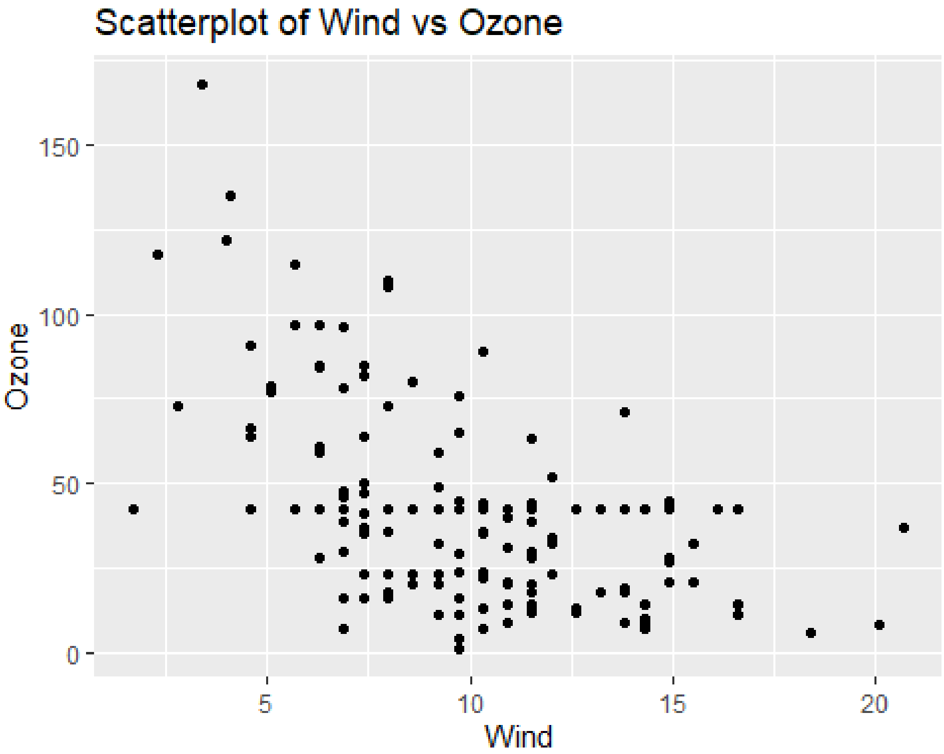
2025-03-20

```
library(ggplot2)
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(caret)

## Warning: package 'caret' was built under R version 4.4.3

## Loading required package: lattice

library(Metrics)

## Warning: package 'Metrics' was built under R version 4.4.3

## 
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
## 
##     precision, recall

data("airquality")

summary(airquality)

##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

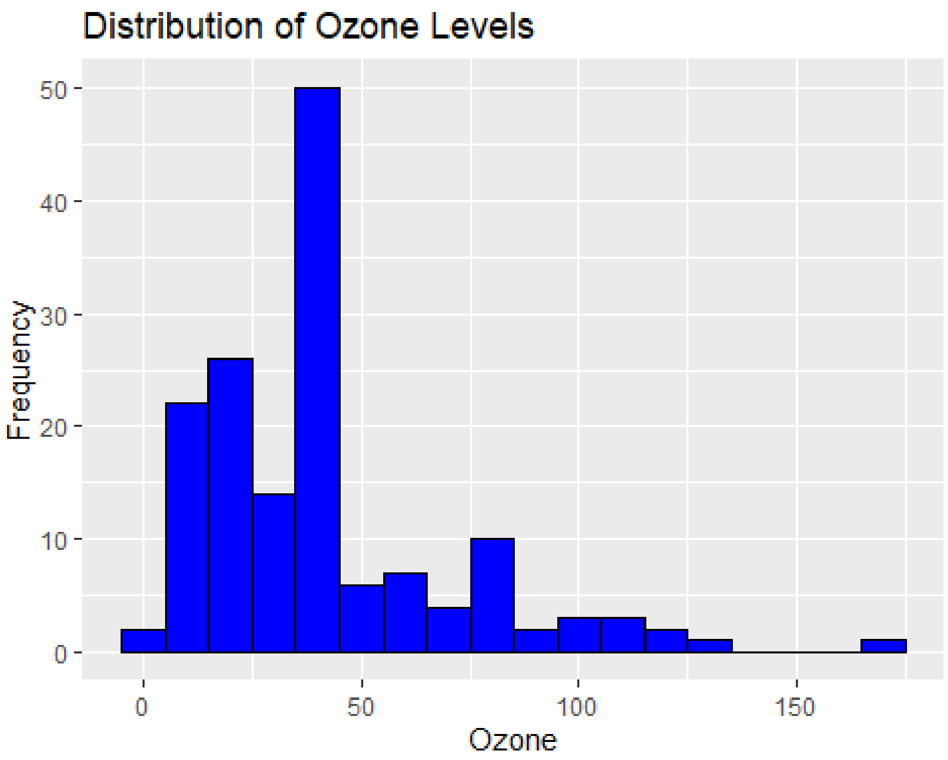
```
missing_values <- colSums(is.na(airquality))
print(paste("Missing values per column:", missing_values))
```

```
## [1] "Missing values per column: 37" "Missing values per column: 7" 
## [3] "Missing values per column: 0"  "Missing values per column: 0" 
## [5] "Missing values per column: 0"  "Missing values per column: 0"
```

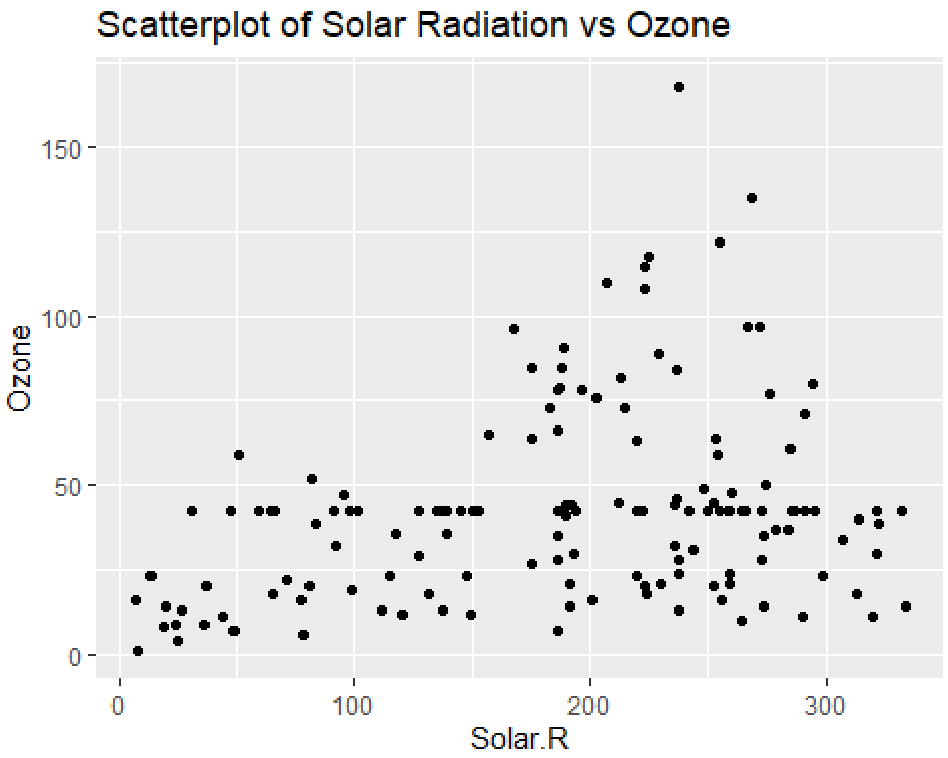
```
airquality_clean <- airquality %>%
  mutate(across(where(is.numeric), ~ifelse(is.na(.), mean(., na.rm = TRUE), .
)))

# 1. Exploratory Data Analysis (EDA) -

ggplot(airquality_clean, aes(x = Wind, y = Ozone)) +
  geom_point() + labs(title = "Scatterplot of Wind vs Ozone")
```



```
ggplot(airquality_clean, aes(x = Ozone)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") + labs(title
= "Distribution of Ozone Levels", x = "Ozone", y = "Frequency")
```



```r
ggplot(airquality_clean, aes(x = Solar.R, y = Ozone)) +
  geom_point() + labs(title = "Scatterplot of Solar Radiation vs Ozone")
```



```
correlation_matrix <- cor(airquality_clean[, sapply(airquality_clean, is.nume
ric)])
print("Correlation Matrix:")
```

```
## [1] "Correlation Matrix:"
```

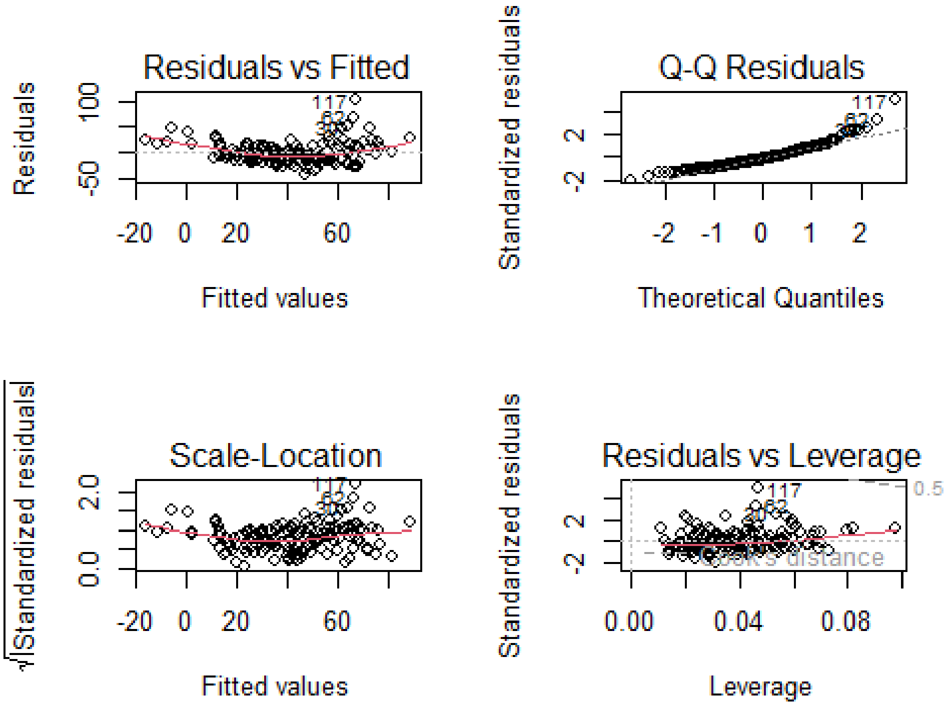
```
print(correlation_matrix)
```

```
##                Ozone     Solar.R        Wind       Temp        Month
## Ozone    1.00000000  0.30296951 -0.53093584  0.6087420  0.149081301
## Solar.R  0.30296951  1.00000000 -0.05524488  0.2625689 -0.072904429
## Wind    -0.53093584 -0.05524488  1.00000000 -0.4579879 -0.178292579
## Temp     0.60874201  0.26256886 -0.45798788  1.0000000  0.420947252
## Month    0.14908130 -0.07290443 -0.17829258  0.4209473  1.000000000
## Day     -0.01135537 -0.14562113  0.02718090 -0.1305932 -0.007961763
##                  Day
## Ozone   -0.011355366
## Solar.R -0.145621130
## Wind     0.027180903
## Temp    -0.130593175
## Month   -0.007961763
## Day      1.000000000
```

```
# 2. Model Building -

model <- lm(Ozone ~ ., data = airquality_clean)
summary(model)
```

```
## 
## Call:
## lm(formula = Ozone ~ ., data = airquality_clean)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -40.263 -13.699  -3.592  10.596 100.549 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -42.16840   19.61843  -2.149   0.0332 *  
## Solar.R       0.05492    0.02049   2.680   0.0082 ** 
## Wind         -2.67022    0.54032  -4.942 2.08e-06 ***
## Temp          1.40550    0.23116   6.080 9.87e-09 ***
## Month        -1.85576    1.34301  -1.382   0.1691    
## Day           0.26508    0.19326   1.372   0.1723    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.78 on 147 degrees of freedom
## Multiple R-squared:  0.4927, Adjusted R-squared:  0.4755 
## F-statistic: 28.56 on 5 and 147 DF,  p-value: < 2.2e-16
```

```
# 3. Model Evaluation -

par(mfrow = c(2, 2))

plot(model)
```



```
# 4. Performance Metrics -

predictions <- predict(model, airquality_clean)


rmse_value <- rmse(airquality_clean$Ozone, predictions)
print(paste("RMSE: ", rmse_value))

## [1] "RMSE:  20.3693869698654"
```

```
# 5. Model Selection -

print("Model Coefficients:")

## [1] "Model Coefficients:"

print(coef(model))

##  (Intercept)      Solar.R          Wind          Temp         Month
Day
```

```
## -42.16840125   0.05492231  -2.67022091   1.40549971  -1.85575798   0.26507
622
```

Report: Analysis of Air Quality Data

1. Introduction

The objective of this analysis is to examine the relationship between various meteorological factors and ozone levels in the air using the `air-quality` dataset, which contains daily air quality measurements from May to September in New York. The analysis covers exploratory data analysis (EDA), model building, and evaluation of a linear regression model for predicting ozone levels.

2. Exploratory Data Analysis *(EDA)*

2.1 Summary Statistics

The **air-quality** dataset contains the following variables:

- **Ozone**: Ozone levels in the air (in ppb).
- **Solar**: Solar radiation (in Langley).
- **Wind**: Wind speed (in mph).
- **Temp**: Temperature (in Fahrenheit).
- **Month**: Month of the year (numeric).
- **Day**: Day of the month (numeric).

The dataset has some missing values, specifically:

- 37 missing values for Ozone.
- 7 missing values for Solar Radiation.
- No missing values for the other variables.

2.2 Missing Values Handling

To address the missing values, we imputed the missing values in the numeric columns (Ozone, Solar Radiation, Wind, and Temp) with the mean value of the respective columns.

2.3 Visualizations

- **Scatterplot of Wind vs. Ozone**: This scatterplot indicates a negative relationship between Wind and Ozone levels, meaning that higher wind speeds are generally associated with lower ozone levels.
- **Histogram of Ozone Levels**: The distribution of Ozone levels shows a right-skewed pattern, indicating that most of the data points fall under lower ozone levels.

- **Scatterplot of Solar Radiation vs. Ozone**: This plot indicates a slight positive relationship, suggesting that higher solar radiation correlates with higher ozone levels.

2.4 Correlation Matrix

The correlation matrix shows the following significant relationships between the variables:

- **Ozone and Temp**: A moderate positive correlation (0.61), indicating that as temperature increases, ozone levels also tend to increase.
- **Ozone and Wind**: A moderate negative correlation (-0.53), suggesting that higher wind speeds may lower ozone levels.
- **Ozone and Solar Radiation**: A weak positive correlation (0.30), indicating a mild positive relationship between solar radiation and ozone levels.

3. Model Building

We built a multiple linear regression model to predict Ozone levels based on the other meteorological factors (Solar Radiation, Wind, Temp, Month, and Day).

The summary of the regression model shows the following:

- **Intercept**: -42.17
- **Solar**: 0.055 (significant at the 0.01 level)
- **Wind**: -2.67 (significant at the 0.001 level)
- **Temp**: 1.41 (significant at the 0.001 level)
- **Month**: -1.86 (not significant at the 0.05 level)
- **Day**: 0.27 (not significant at the 0.05 level)

The **Multiple R-squared** value is 0.4927, indicating that approximately 49.27% of the variance in ozone levels is explained by the model. The **Adjusted R-squared** is 0.4755, which accounts for the number of predictors in the model.

4. Model Evaluation

We evaluated the model using diagnostic plots and performance metrics:

- **Residuals Plot**: The residuals appear fairly evenly spread, suggesting that the model is a reasonable fit.
- **RMSE (Root Mean Square Error)**: The RMSE value is 20.37, indicating that the model's predictions deviate from the actual ozone levels by approximately 20.37 units on average.

5. Model Coefficients

The model coefficients are as follows:

- Intercept: -42.17
- Solar: 0.055
- Wind: -2.67
- Temp: 1.41
- Month: -1.86
- Day: 0.27

These coefficients suggest the following:

- Solar radiation has a positive impact on ozone levels.
- Wind has a negative impact on ozone levels.
- Temperature has a positive impact on ozone levels.
- Month and Day do not have a significant impact on ozone levels.

---

1. COMPLETE ANALYSIS OF THE GIVEN DATA USING APPROPRIATE STATISTICAL TECHNIQUES

Exploratory Data Analysis (EDA): Various visualizations were created to understand the relationships between variables:

- A scatter plot was generated to visualize the relationship between **Wind** and **Ozone**, showing a negative correlation (higher wind speeds tend to be associated with lower Ozone levels).
- A histogram of **Ozone** values was plotted to assess the distribution of the target variable, revealing a slightly right-skewed distribution.
- Another scatter plot between **Solar Radiation** and **Ozone** indicated a moderate positive relationship between the two.

Correlation Analysis: A correlation matrix was calculated for the numeric variables, revealing several key insights:

- **Wind** and **Ozone** have a negative correlation of -0.53, indicating that higher wind speeds are generally associated with lower ozone levels.
- **Temp** and **Ozone** have a moderate positive correlation of 0.61, suggesting that higher temperatures tend to coincide with higher ozone levels.
- **Solar.R** has a weak positive correlation with **Ozone** (0.30), suggesting some relationship but not a strong one.

---

2. EVALUATION OF THE FINAL MODEL FOR ESTIMATION OF RESPONSE VARIABLE (OZONE)

Linear Regression Model: A multiple linear regression model was built to predict the **Ozone** levels based on all other available variables.

The summary of the model indicates the following:

- **Intercept**: -42.17, with a significant p-value (0.033), meaning that the intercept is statistically significant.
- **Solar.R**: Coefficient of 0.0549, significant with a p-value of 0.0082, suggesting that solar radiation has a positive impact on ozone levels.
- **Wind**: Coefficient of -2.67, highly significant (p-value < 0.001), suggesting a negative impact on ozone levels.
- **Temp**: Coefficient of 1.41, significant (p-value < 0.001), showing that temperature positively affects ozone levels.
- **Month** and **Day**: Both have high p-values (0.1691 and 0.1723), indicating they are not statistically significant predictors of **Ozone** levels.

Model Performance:

- **Multiple R-squared**: 0.4927, which means the model explains approximately 49.27% of the variance in ozone levels. This suggests that while the model explains a reasonable portion of the variance, there is still a significant amount of unexplained variance.
- **Adjusted R-squared**: 0.4755, accounting for the number of predictors in the model.
- **RMSE (Root Mean Square Error)**: 20.37, which provides an estimate of how much the predictions deviate from the actual values. This value is relatively large, indicating room for improvement in prediction accuracy.

Conclusion on Model: The final model is not the best for estimation of the response variable, **Ozone**, as it only explains about 49% of the variance, and the RMSE value indicates substantial prediction errors. Though the model captures key predictors like **Wind**, **Temp**, and **Solar.R**, it may require further refinement to improve accuracy.

---

3. DRAWBACKS OF THE MODEL

Several drawbacks are apparent in the current model:

- **Low R-squared Value**: The model explains only about 49% of the variance, suggesting that many factors influencing ozone levels are not captured. The model could potentially benefit from more sophisticated techniques.

- **Insignificant Predictors**: Variables like **Month** and **Day** did not show statistically significant coefficients. Including these variables in the model may add unnecessary complexity without contributing meaningful predictive power.
- **Linear Model Assumptions**: The linear regression model assumes a linear relationship between predictors and the target variable. However, in the case of air quality data, the relationships might be more complex (e.g., non-linear or interactive effects between variables), which the linear model may not adequately capture.
- **Overfitting**: While not conclusively proven, the high number of predictors for a relatively moderate R-squared value raises concerns about potential overfitting. This suggests that the model may fit the training data well but struggle to generalize to new data.

4. Suggestions for Model Improvement

To improve the model and enhance its accuracy:

- **Use of Non-linear Models**: Exploring non-linear regression techniques or machine learning models such as **Random Forest** or **Gradient Boosting** could help capture more complex relationships between predictors and ozone levels.
- **Feature Engineering**: Creating new features or interactions between existing features could improve the model. For example, interactions between **Solar Radiation** and **Temperature** could be considered, as these might jointly influence ozone levels.
- **Addressing Multicollinearity**: Correlations between predictors, such as between **Temp** and **Solar Radiation**, could lead to multicollinearity, which can inflate standard errors of regression coefficients. Addressing multicollinearity (e.g., by using **Principal Component Analysis**) might improve model stability.
- **Data Augmentation**: If more data is available, training the model on a larger dataset could improve its robustness and predictive power.

Practical Usefulness of the Report in its Domain: The report offers valuable insights into how various factors such as wind speed, solar radiation, and temperature influence ozone levels. In the context of air quality monitoring, understanding these relationships can help in designing better environmental policies. Additionally, the model's evaluation and improvement suggestions can guide future data-driven studies to provide more accurate predictions for ozone and other environmental indicators.

Conclusion

The current linear regression model provides a basic understanding of the factors affecting ozone levels, but its performance could be improved. The analysis of missing data, correlations, and the initial model evaluation has revealed areas where more complex modeling approaches and better feature selection could lead to more accurate predictions. By addressing the drawbacks and applying suggestions for improvement, the model can be refined to be more effective in air quality estimation.