CAC-2 MDS 272 : PROJECT REPORT

Title

Analyzing Urban Traffic in Bengaluru: Impact of Environmental and Situational Factors on Congestion and Mobility in Key Local Areas

Done by:

Kumar Yashu (2448062): **Q1**

Sreneedhi Konnor (2448055): **Q2**

Aaditya Kumar Dhaka (2448001): **Q3**

Sayandip Ghosh (24480048): **Q4**

Neelanjan Dutta (2448040): **Q5**

Submitted to:

Dr. Akshita Sharma

Department of Statistics and Data Science

Christ university, Bangalore

Problem Statement for the Study

" Understanding Traffic Dynamics in Bengaluru's Key Local Areas"

Bengaluru's rapid urban growth has intensified traffic congestion in areas like MG Road, Indiranagar, Koramangala, Jayanagar, and Electronic City. Environmental factors like weather and situational elements such as roadwork and traffic volume significantly affect mobility. **This study aims to explore these impacts to provide data-driven solutions for congestion management**.

Introduction for the Study

- The study focuses on understanding the interplay between environmental and situational factors and their effects on urban traffic dynamics.
- With the rapid urbanization and increased vehicle dependency, traffic congestion and its determinants have become pressing issues for modern cities.
- This project explores congestion levels, traffic volume, roadwork impacts, and public transport usage using comprehensive statistical analyses and visualizations. Leveraging real-world data and advanced statistical techniques, the study provides actionable insights into mitigating congestion and enhancing urban mobility.

Case Study Relevance

- Traffic congestion significantly affects urban living by increasing travel times, fuel consumption, and environmental pollution. This case study is critical as it addresses key urban challenges by analyzing the impact of factors like weather conditions, roadwork activities, and compliance with traffic regulations on congestion.
- By employing statistical methods such as ANOVA, T-tests, and Chi-square analyses, the project identifies patterns and relationships, guiding policymakers to design data-driven interventions.

Questions and Objectives

Question 1: Investigating the Effect of Area and Weather on Congestion Levels

Description:
This question examines whether the congestion levels differ significantly between specific areas and if weather conditions influence congestion. Addressing these factors helps identify localized congestion issues and the role of environmental factors, enabling targeted interventions to improve traffic flow.

Objectives:

1. Test if there is a significant difference in congestion levels between two areas (e.g., Indiranagar and Koramangala).
2. Investigate the impact of weather conditions on congestion levels using statistical methods and visualizations.
3. Evaluate if the observed distribution of weather conditions aligns with the expected distribution.

Question 2: Investigating the Effect of Weather Conditions on Traffic Volume

Description:
This question focuses on analyzing how various weather conditions affect traffic volume in urban areas. By understanding these effects, the study informs strategies to manage traffic during adverse weather, contributing to efficient urban mobility.

Objectives:

1. Compare traffic volumes across different weather conditions using the Kruskal-Wallis test.
2. Analyze road capacity utilization under contrasting weather conditions (e.g., Clear vs. Rain) using the Mann-Whitney U Test.
3. Determine the relationship between congestion levels and traffic signal compliance using correlation analysis.

Question 3: Impact of Roadwork and Construction Activity on Traffic Volume

Description:
This question evaluates the influence of roadwork and construction activities on traffic volume and patterns. Addressing these impacts aids in optimizing construction schedules and planning alternative routes to reduce congestion.

Objectives:

1. Assess differences in traffic volume between roads with and without roadwork/construction activity using the Mann-Whitney U Test.
2. Examine the relationship between roadwork activity and traffic volume categories using the Chi-square test for independence.
3. Conduct a two-way ANOVA to compare traffic volumes across different areas under the presence and absence of roadwork.

Question 4: Exploring the Effect of Weather Conditions on Public Transport Usage

Description:
This question analyzes how weather conditions influence public transport usage and explores potential relationships with environmental impact. Understanding these trends supports the development of resilient and weather-adaptive public transportation systems.

Objectives:

1. Perform one-way ANOVA to compare public transport usage across weather conditions.
2. Determine the correlation between environmental impact and public transport usage using Spearman rank correlation.
3. Fit regression models (Poisson and Negative Binomial) to predict public transport usage based on weather conditions.

Question 5: Analyzing the Relationship Between Congestion Level and Travel Time Index

Description:
This question explores the relationship between congestion levels and travel time indices to identify potential patterns or deviations. Insights from this analysis contribute to evaluating traffic efficiency and improving travel-time reliability in urban settings.

Objectives:

1. Test whether the median travel time index significantly deviates from a hypothesized value using the Wilcoxon Signed-Rank Test.
2. Analyze the monotonic relationship between congestion levels and travel time indices using Spearman's rank correlation.
3. Compare congestion levels across different ranges of travel time indices using the Kruskal-Wallis test.

Methodology/Codes:

```
# Load necessary libraries
library(ggplot2)

library(dplyr)

library(MASS)

library(BSDA)

# Load data
data <- read.csv("D:/cac2.csv")
```

Question 1: Investigating the Effect of Area and Weather on Congestion Levels
Objective 1: Test if there is a significant difference in congestion levels between two areas (e.g., Indiranagar and Koramangala)

Answer:

Let us set up the null hypothesis
H_0: There is no significant difference in the congestion levels between Indiranagar and Koramangala. i.e. $\mu_{\text{Indiranagar}} = \mu_{\text{Kormangala}}$

Against the alternative hypothesis

H_1: There is a significant difference in the congestion levels between Indiranagar and Koramangala. i.e. $\mu_{\text{Indiranagar}} \neq \mu_{\text{Kormangala}}$

Under H_0, the test statistic is given by:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

where:

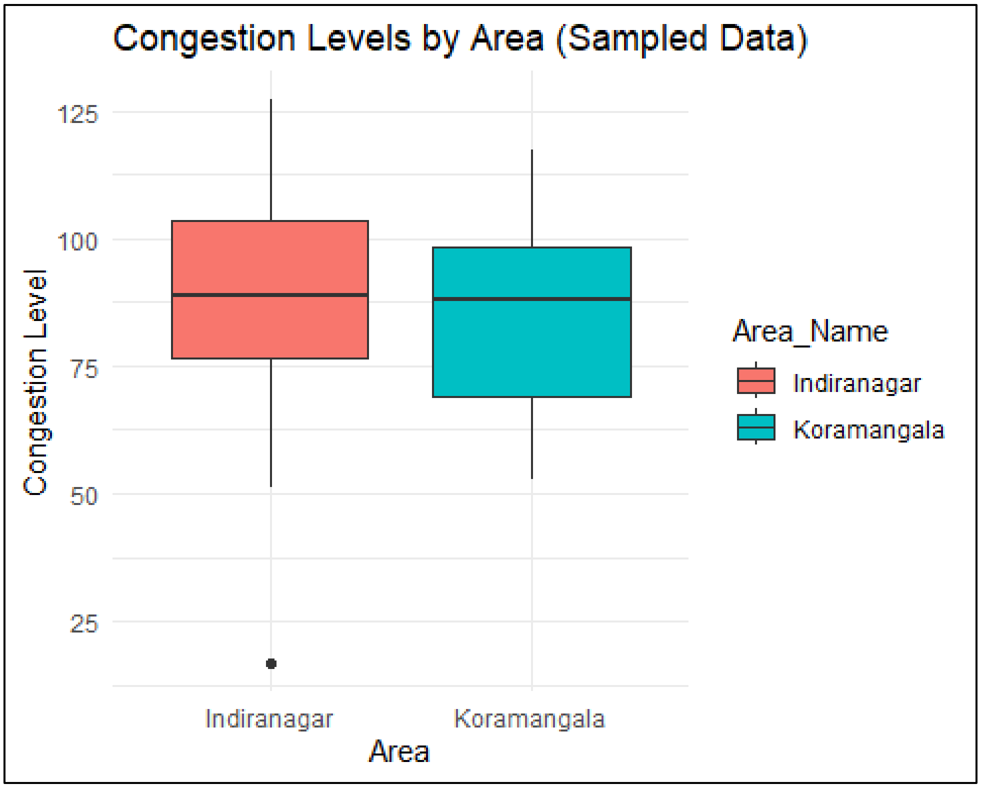
- $\bar{X}_1$ and $\bar{X}_2$ are the sample means for Indiranagar and Koramangala, respectively.
- s^2 is the pooled variance:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- n_1 and n_2 are the sample sizes for Indiranagar and Koramangala.
- s_1^2 and s_2^2 are the sample variances for Indiranagar and Koramangala.

```
# Step 1: Subset data for the two areas (Indiranagar and Koramangala)
area1 <- "Indiranagar"
area2 <- "Koramangala"
data_subset <- data %>% filter(Area_Name %in% c(area1, area2))

# Step 2: Randomly sample 25 observations from each area
set.seed(123) # Set seed for reproducibility
sampled_data <- data_subset %>%
  group_by(Area_Name) %>%
  slice_sample(n = 25) %>%
  ungroup()

# Step 3: Check for normality using Shapiro-Wilk test
shapiro1 <- shapiro.test(sampled_data$Congestion_Level[sampled_data$Area_Name == area1])
shapiro2 <- shapiro.test(sampled_data$Congestion_Level[sampled_data$Area_Name == area2])

cat("Shapiro-Wilk Test for Indiranagar: W =", shapiro1$statistic, "p-value =", shapiro1$p.value, "\n")

## Shapiro-Wilk Test for Indiranagar: W = 0.950033 p-value = 0.2511548

cat("Shapiro-Wilk Test for Koramangala: W =", shapiro2$statistic, "p-value =", shapiro2$p.value, "\n")

## Shapiro-Wilk Test for Koramangala: W = 0.9514907 p-value = 0.2708202

# Step 4: Test for equality of variances using F-test
variance_test <- var.test(Congestion_Level ~ Area_Name, data = sampled
```

```
_data)
cat("F-test for equality of variances: F =", variance_test$statistic,
"p-value =", variance_test$p.value, "\n")

## F-test for equality of variances: F = 1.961774 p-value = 0.1055383

# Step 5: Independent t-test (assuming normality and equal variances)
t_test_result <- t.test(Congestion_Level ~ Area_Name, data = sampled_d
ata, var.equal = TRUE)
cat("T-test result: t =", t_test_result$statistic, "p-value =", t_test
_result$p.value, "\n")

## T-test result: t = -0.1378291 p-value = 0.890952

# Step 6: Interpretation based on p-value
if (t_test_result$p.value < 0.05) {
  cat("Result: There is a significant difference in congestion levels
between Indiranagar and Koramangala.\n")
} else {
  cat("Result: No significant difference in congestion levels between
Indiranagar and Koramangala.\n")
}

## Result: No significant difference in congestion levels between Indi
ranagar and Koramangala.

# Step 7: Boxplot visualization
library(ggplot2)
ggplot(sampled_data, aes(x = Area_Name, y = Congestion_Level, fill = A
rea_Name)) +
  geom_boxplot() +
  labs(title = "Congestion Levels by Area (Sampled Data)", x = "Area",
y = "Congestion Level") +
  theme_minimal()
```



Interpretation

The results of the independent t-test indicate no significant difference in the mean congestion levels between Indiranagar and Koramangala. The test statistic ($t=-0.138$) and the corresponding p-value ($p=0.891$) suggest that the observed difference in means is likely due to random chance rather than a true difference. The Shapiro-Wilk test confirms that the data in both groups are approximately normally distributed ($p>0.05$), and the F-test indicates equal variances ($p=0.106$), validating the assumptions for the t-test.

Conclusion

There is no evidence to suggest a statistically significant difference in congestion levels between Indiranagar and Koramangala. The null hypothesis cannot be rejected, indicating that congestion levels in the two areas are similar based on the sampled data.

Graphical Interpretation:

The boxplot compares congestion levels in Indiranagar and Koramangala, showing similar medians, suggesting no significant difference in typical congestion levels between the two areas. The interquartile ranges (IQRs) indicate comparable variability, with overlapping ranges further supporting the statistical conclusion that the difference is not significant. Potential outliers, represented as dots, highlight occasional deviations but do not meaningfully affect the overall comparison. The visual evidence aligns with the t-test results, reinforcing that the observed differences in congestion levels are likely due to random variation rather than a true difference.

Objective 2: Investigate if weather conditions impact congestion levels

Answer:

Let us set up the null hypothesis

H_0: Weather conditions do not impact congestion levels

i.e. $\mu_{Clear} = \mu_{Rainy} = \mu_{Foggy} = \mu_{Snowy} = \mu_{Cloudy}$

Against the alternative hypothesis

H_1: At least one weather condition has a significant impact on congestion levels.

i.e. At least one μ is different.

Under H_0, the test statistic is given by:

$$F = \frac{\text{Between-group variability (Mean Square Between)}}{\text{Within-group variability (Mean Square Error)}}$$

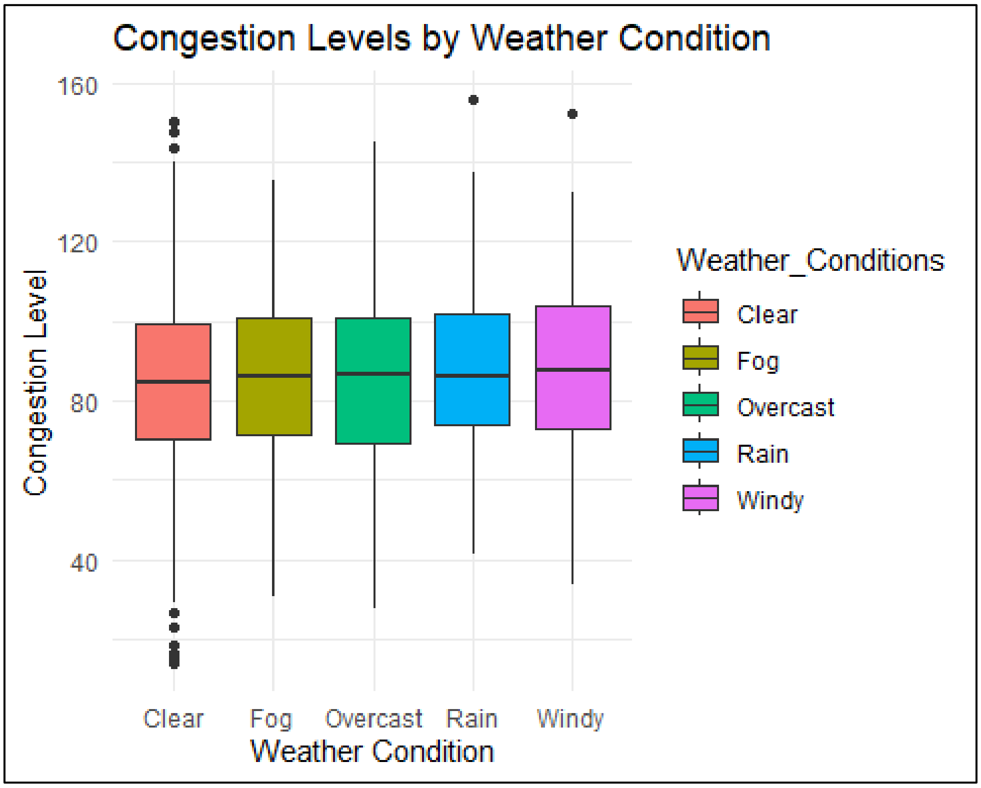
Where:

- $\text{Mean Square Between} = \frac{\text{Sum of Squares Between Groups (SSB)}}{\text{Degrees of Freedom Between Groups } (\text{df}_B)}$
- $\text{Mean Square Error} = \frac{\text{Sum of Squares Within Groups (SSW)}}{\text{Degrees of Freedom Within Groups } (\text{df}_W)}$

```
# Step 1: Assumption Check for Normality (Shapiro-Wilk Test)
shapiro_test <- shapiro.test(data$Congestion_Level)
cat("Shapiro-Wilk Test for Normality: W =", shapiro_test$statistic, "p
-value =", shapiro_test$p.value, "\n")

## Shapiro-Wilk Test for Normality: W = 0.9991248 p-value = 0.7274843

# Step 2: Assumption Check for Homogeneity of Variances (Bartlett's Te
st)
bartlett_test <- bartlett.test(Congestion_Level ~ Weather_Conditions,
data = data)
cat("Bartlett's Test for Homogeneity of Variances: K-squared =", bartl
ett_test$statistic,
    "p-value =", bartlett_test$p.value, "\n")

## Bartlett's Test for Homogeneity of Variances: K-squared = 2.562756
p-value = 0.6334347
```

```r
# Step 3: One-way ANOVA (if assumptions are met)
anova_result <- aov(Congestion_Level ~ Weather_Conditions, data = data
)
anova_summary <- summary(anova_result)
cat("ANOVA Summary: F =", anova_summary[[1]]["Weather_Conditions", "F
value"],
    "p-value =", anova_summary[[1]]["Weather_Conditions", "Pr(>F)"], "
\n")

## ANOVA Summary: F = 0.4077076 p-value = 0.8032037

# Step 4: Extract F-value and p-value for comparison with critical F-v
alue
f_value <- anova_summary[[1]]["Weather_Conditions", "F value"]
p_value <- anova_summary[[1]]["Weather_Conditions", "Pr(>F)"]
cat("F-value:", f_value, "p-value:", p_value, "\n")

## F-value: 0.4077076 p-value: 0.8032037

# Step 5: Compare F-value with critical F-value and decide on Tukey's
Test
f_critical <- qf(0.95, df1 = length(unique(data$Weather_Conditions)) -
1,
                 df2 = nrow(data) - length(unique(data$Weather_Conditi
ons)))
cat("Critical F-value:", f_critical, "\n")

## Critical F-value: 2.377986

if (f_value >= f_critical) {
  cat("F value is greater than or equal to the tabulated value. Procee
ding with Tukey's test.\n")
  tukey_result <- TukeyHSD(anova_result)
  print(tukey_result)
} else {
  cat("F value is less than the tabulated value. No Tukey test require
d.\n")
}

## F value is less than the tabulated value. No Tukey test required.

# Step 6: Visualization: Boxplot for Congestion Level by Weather Condi
tions
library(ggplot2)
ggplot(data, aes(x = Weather_Conditions, y = Congestion_Level, fill =
Weather_Conditions)) +
  geom_boxplot() +
  labs(title = "Congestion Levels by Weather Condition", x = "Weather
```

```
Condition", y = "Congestion Level") +
  theme_minimal()
```



Interpretation

The one-way ANOVA results indicate no statistically significant difference in congestion levels across different weather conditions. The test produced an F-value of 0.408 and a corresponding p-value of 0.803. Since the p-value is greater than the significance level (α=0.05), we fail to reject the null hypothesis. The Shapiro-Wilk test confirms that the data follows a normal distribution (p=0.727), and Bartlett's test suggests homogeneity of variances (p=0.633), meeting the assumptions for ANOVA. Additionally, the F-value is less than the critical F-value (2.378), so further post-hoc analysis (Tukey's test) is unnecessary.

Conclusion

There is no evidence to suggest that weather conditions have a significant impact on congestion levels. The null hypothesis cannot be rejected. This indicates that congestion levels remain consistent regardless of the prevailing weather conditions.

Graphical Interpretation:

The boxplot visualizes congestion levels across different weather conditions (Clear, Fog, Overcast, Rain, and Windy), showing that the median and interquartile ranges (IQRs) are similar across all categories, indicating no significant variation. The overlapping ranges across the conditions further support the ANOVA results, confirming that the differences in mean congestion levels are statistically insignificant. Outliers are present but are evenly distributed across weather conditions, suggesting they do not impact the overall conclusion. This graphical representation aligns with the statistical analysis, confirming that weather conditions do not have a significant effect on congestion levels.

 # Objective 3: Test if the observed distribution of weather conditions deviates from the expected distribution.

Answer:

Let us set up the null hypothesis

H_0: The observed frequencies of weather conditions match the expected frequencies based on the given proportions.

Against the alternative hypothesis

H_1: The observed frequencies of weather conditions do not match the expected frequencies.

Under H_0, the test statistic is given by:

The test statistic for a chi-squared test is calculated as:

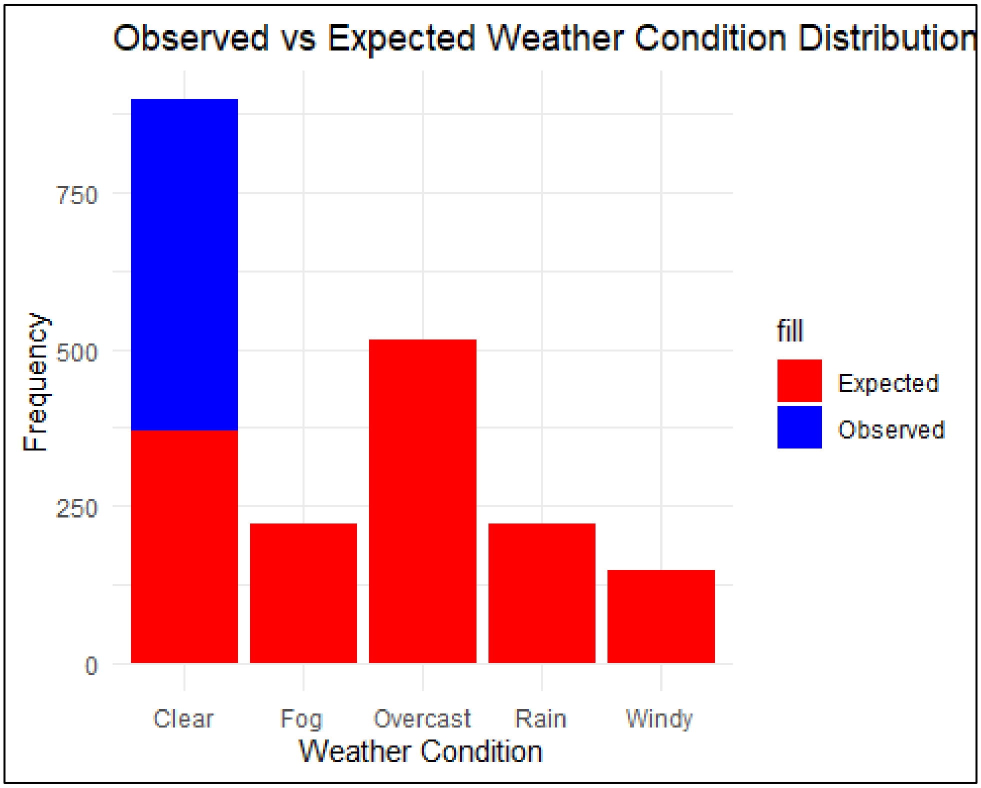
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency for each weather condition
- E_i = Expected frequency for each weather condition (calculated as $\text{expected proportion} \times \text{total observations}$)

```
# Step 1: Calculate observed frequencies
observed <- table(data$Weather_Conditions)  # Frequencies of each cond
ition
cat("Observed Frequencies:\n")

## Observed Frequencies:

print(observed)

##
##    Clear      Fog Overcast     Rain    Windy
##      897      180      208      141       48

# Step 2: Define custom expected proportions
expected_proportions <- c(0.25, 0.15, 0.35, 0.15, 0.10)  # Adjust prop
ortions to match your hypothesis
cat("Expected Proportions:\n")

## Expected Proportions:
```

```
print(expected_proportions)

## [1] 0.25 0.15 0.35 0.15 0.10

# Step 3: Ensure the length of observed and expected match
if (length(observed) == length(expected_proportions)) {
  # Step 4: Perform Chi-squared test
  chi_sq_result <- chisq.test(x = observed, p = expected_proportions)
  cat("Chi-squared Test Result:\n")
  print(chi_sq_result)

## Chi-squared Test Result:
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 1045.4, df = 4, p-value < 2.2e-16

  # Step 5: Visualization: Observed vs Expected Frequencies
  observed_df <- as.data.frame(observed)
  expected_df <- data.frame(Weather_Conditions = names(observed),
                            Frequency = sum(observed) * expected_propo
rtions)

  library(ggplot2)
  ggplot() +
    geom_bar(data = observed_df, aes(x = Var1, y = Freq, fill = "Obser
ved"),
             stat = "identity", position = "dodge") +
    geom_bar(data = expected_df, aes(x = Weather_Conditions, y = Frequ
ency, fill = "Expected"),
             stat = "identity", position = "dodge") +
    labs(title = "Observed vs Expected Weather Condition Distribution"
,
         x = "Weather Condition", y = "Frequency") +
    scale_fill_manual(values = c("Observed" = "blue", "Expected" = "re
d")) +
    theme_minimal()
} else {
  cat("Error: Length of observed categories does not match length of e
xpected proportions.\n")
}
```



Interpretation:

The Chi-squared test result shows a very small p-value (< 2.2e-16), which indicates that there is a significant difference between the observed and expected frequencies of weather conditions. Specifically, the observed data do not conform to the hypothesized proportions of weather conditions (Clear: 25%, Fog: 15%, Overcast: 35%, Rain: 15%, Windy: 10%). The large Chi-squared statistic (1045.4) suggests a substantial deviation, meaning that the weather conditions in the dataset are not distributed according to the expected proportions.

Conclusion:

Given the extremely small p-value, we reject the null hypothesis that the observed weather conditions follow the expected distribution. This suggests that the actual distribution of weather conditions is significantly different from the expected one. Therefore, further investigation is needed to understand the factors driving the observed distribution, and the current hypothesis about the weather conditions may need to be adjusted.

Graphical Interpretation:

The bar chart compares the observed and expected distributions of weather conditions, with blue bars representing expected frequencies and red bars representing observed frequencies. Significant differences are noticeable, such as "Clear" having much higher observed frequencies than expected, while categories like "Fog" and "Windy" have lower observed frequencies than expected.

This visual disparity aligns with the statistical test results, showing that the observed weather conditions deviate significantly from the expected distribution.

Question 2: Investigating the Effect of Weather Conditions on Traffic Volume

Objective 1: Compare Traffic Volumes Across Different Weather Conditions Using Kruskal-Wallis Test

Answer:

Let us set up the null hypothesis

H_0: The median traffic volumes are the same across all weather conditions.

Against the alternative hypothesis

H_1: At least one weather condition has a different median traffic volume.

Under H_0, the test statistic is given by

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Where:

- N = Total number of observations across all groups.
- k = Number of groups (in this case, weather conditions).
- R_i = Sum of ranks for group i.
- n_i = Number of observations in group i.

```
# 1. Test for Normality (Optional, for understanding; not required for
Kruskal-Wallis)
shapiro_test <- shapiro.test(data$Traffic_Volume)
print(shapiro_test)

##
##  Shapiro-Wilk normality test
##
## data:  data$Traffic_Volume
## W = 0.99879, p-value = 0.4167

# 2. Homogeneity of Variances: Bartlett test (Optional check, not requ
ired for Kruskal-Wallis)
bartlett_test <- bartlett.test(Traffic_Volume ~ Weather_Conditions, da
ta = data)
print(bartlett_test)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Traffic_Volume by Weather_Conditions
## Bartlett's K-squared = 11.18, df = 4, p-value = 0.02462
```

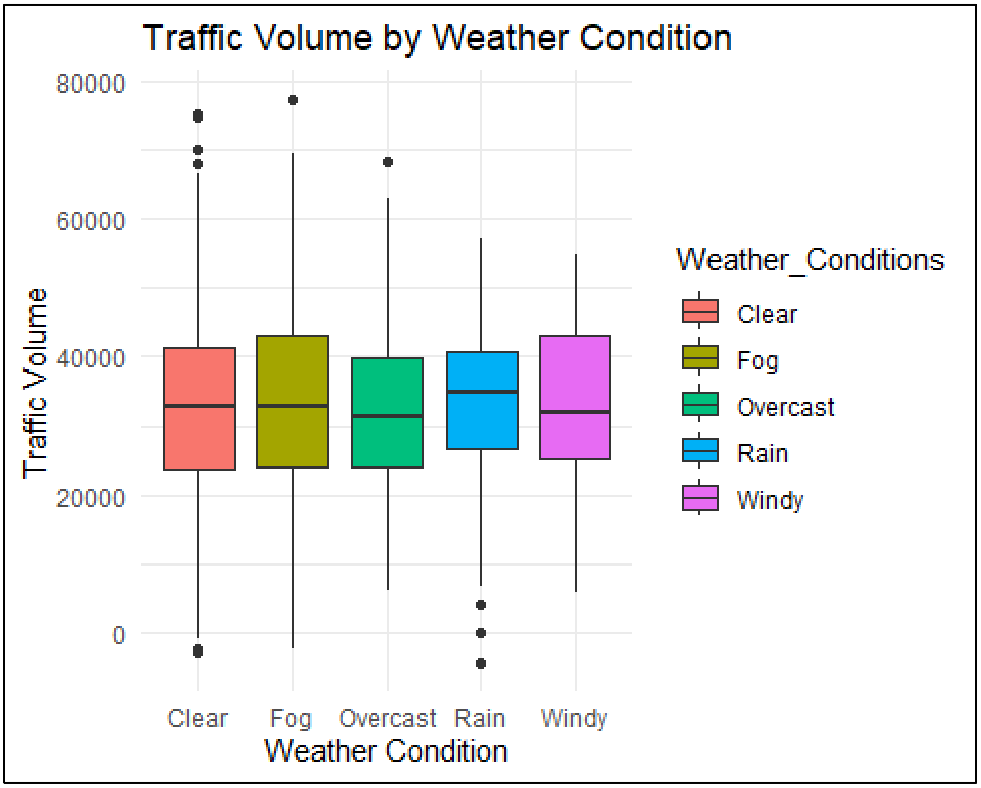
```
# Since Bartlett test shows significant p-value, Kruskal-Wallis is the
better choice.
# Main Test: Kruskal-Wallis Test
kruskal_result <- kruskal.test(Traffic_Volume ~ Weather_Conditions, da
ta = data)

# Display results
cat("Kruskal-Wallis Test Statistic:", kruskal_result$statistic, "\nP-v
alue:", kruskal_result$p.value, "\n")
```

```
## Kruskal-Wallis Test Statistic: 1.842464
## P-value: 0.7647049
```

```
# Check if Kruskal-Wallis test result is significant
if (kruskal_result$p.value < 0.05) {
  cat("There is a significant difference in traffic volumes across wea
ther conditions.\n")
} else {
  cat("There is no significant difference in traffic volumes across we
ather conditions.\n")
}
```

```
## There is no significant difference in traffic volumes across weathe
r conditions.
```

```
# Visualization: Boxplot for Traffic Volumes Across Weather Conditions
library(ggplot2)

ggplot(data, aes(x = Weather_Conditions, y = Traffic_Volume, fill = We
ather_Conditions)) +
  geom_boxplot() +
  labs(title = "Traffic Volume by Weather Condition", x = "Weather Con
dition", y = "Traffic Volume") +
  theme_minimal()
```



Interpretation:

The results of the Kruskal-Wallis test show a **p-value of 0.7647**, which is much greater than the typical significance threshold of 0.05. This suggests that there is no statistically significant difference in traffic volumes across the different weather conditions (Clear, Fog, Overcast, Rain, Windy). Additionally, although the Bartlett test showed a significant p-value (0.0246) indicating a potential violation of the assumption of equal variances, the Kruskal-Wallis test is robust to such violations and was therefore used as the appropriate non-parametric alternative.

Conclusion:

Since the Kruskal-Wallis test did not show a significant result (p-value = 0.7647), we conclude that there is **no significant difference** in traffic volumes across different weather conditions in the dataset. This implies that weather conditions may not have a strong influence on traffic volumes in this particular dataset, and other factors might be more important in explaining the variations in traffic.

Graphical Interpretation:
The boxplot visualizes traffic volume across different weather conditions (Clear, Fog, Overcast, Rain, and Windy). Each box represents the spread of traffic volume for a specific condition, with overlapping ranges indicating similarity across categories.

The visual pattern supports the Kruskal-Wallis test results, which show no statistically significant difference (p-value = 0.7647) in traffic volumes among weather conditions. This suggests that weather conditions do not strongly influence traffic volume.

 # Objective 2: Compare Road Capacity Utilization in Different Weather Conditions Using Mann-Whitney U Test

Answer:

Let us set up the null hypothesis

H_0: The median road capacity utilization for Clear weather is equal to the median for Rain weather.
Against the alternative hypothesis
H_1: The median road capacity utilization for Clear weather is not equal to the median for Rain weather

Under H_0, the test statistic is given by
The test statistic for the Mann-Whitney U test is denoted by W. The U statistic represents the sum of ranks for one group (or the difference between the number of observations in each group), and it is used to test whether the two independent samples come from the same distribution.

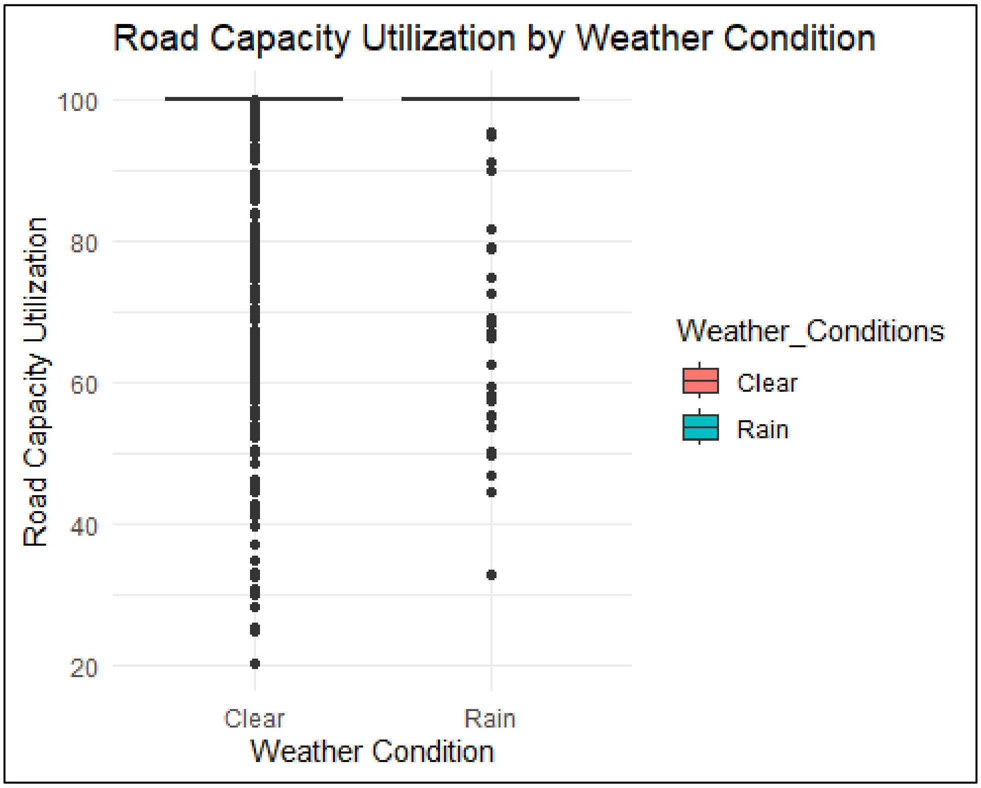
```
# Subset data for two weather conditions
weather_condition1 <- "Clear"
weather_condition2 <- "Rain"
data_subset <- data %>% filter(Weather_Conditions %in% c(weather_condition1, weather_condition2))

# Separate data for each weather condition
clear_weather_data <- data_subset$Road_Capacity_Utilization[data_subset$Weather_Conditions == weather_condition1]
rain_weather_data <- data_subset$Road_Capacity_Utilization[data_subset$Weather_Conditions == weather_condition2]

# 1. Mann-Whitney U Test (Wilcoxon rank-sum test)
mann_whitney_result <- wilcox.test(clear_weather_data, rain_weather_data, alternative = "two.sided", conf.level = 0.95)

# Print Mann-Whitney U test results
cat("\nMann-Whitney U Test Results:\n")
```

```
##
## Mann-Whitney U Test Results:
```

```
print(mann_whitney_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clear_weather_data and rain_weather_data
## W = 62751, p-value = 0.8368
## alternative hypothesis: true location shift is not equal to 0
```

```
# Visualization: Boxplot for road capacity utilization in the two weat
her conditions
library(ggplot2)

ggplot(data_subset, aes(x = Weather_Conditions, y = Road_Capacity_Util
ization, fill = Weather_Conditions)) +
  geom_boxplot() +
  labs(title = "Road Capacity Utilization by Weather Condition", x = "
Weather Condition", y = "Road Capacity Utilization") +
  theme_minimal()
```



Interpretation:

The Mann-Whitney U test compares the distributions of road capacity utilization between two weather conditions, in this case, "Clear" and "Rain." The **p-value of 0.8368** is much larger than the typical significance threshold of 0.05, indicating that there is **no significant difference** in road capacity utilization between the two weather conditions. This suggests that, in this dataset, weather conditions (Clear vs. Rain) do not have a substantial effect on road capacity utilization.

Conclusion:

Since the p-value (0.8368) is greater than 0.05, we fail to reject the null hypothesis. This means that there is **no significant difference** in road capacity utilization between Clear and Rain weather conditions. The data suggests that road capacity utilization is similar in both weather conditions, and weather does not appear to influence road capacity utilization in this dataset.

Graphical Interpretation:

The boxplot visualizes road capacity utilization for the two weather conditions ("Clear" and "Rain"). Both groups show similar distributions, with medians near 100 and minimal visible differences in spread or central tendency. This supports the conclusion that road capacity utilization does not vary significantly between the two weather conditions.

Objective 3: Investigate whether there's a relationship between congestion level and traffic signal compliance using a correlation test.
Answer

Let us set up the null hypothesis

H_0: There is no correlation between congestion level and traffic signal compliance, i.e. $\rho = 0$

Against the alternative hypothesis

H_1: There is a correlation between congestion level and traffic signal compliance ($\rho \neq 0$).

Under H_0 the test statistic is given by:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

```r
# Assumption: Normality of data for both variables
shapiro_congestion <- shapiro.test(data$Congestion_Level)
shapiro_signal_compliance <- shapiro.test(data$Traffic_Signal_Compliance)

# Print Shapiro-Wilk test results
print(shapiro_congestion)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Congestion_Level
## W = 0.99912, p-value = 0.7275
```

```
print(shapiro_signal_compliance)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Traffic_Signal_Compliance
## W = 0.95655, p-value < 2.2e-16
```

```
# Assumption check: Normality
if (shapiro_congestion$p.value > 0.05 && shapiro_signal_compliance$p.v
alue > 0.05) {
  cat("Both variables are normally distributed. Proceeding with Pearso
n correlation.\n")

  # Main Test: Pearson Correlation
  correlation_test <- cor.test(data$Congestion_Level, data$Traffic_Sig
nal_Compliance, method = "pearson")
} else {
  cat("Normality violated for one or both variables. Proceeding with S
pearman correlation.\n")

  # Main Test: Spearman Correlation
  correlation_test <- cor.test(data$Congestion_Level, data$Traffic_Sig
nal_Compliance, method = "spearman")
}
```

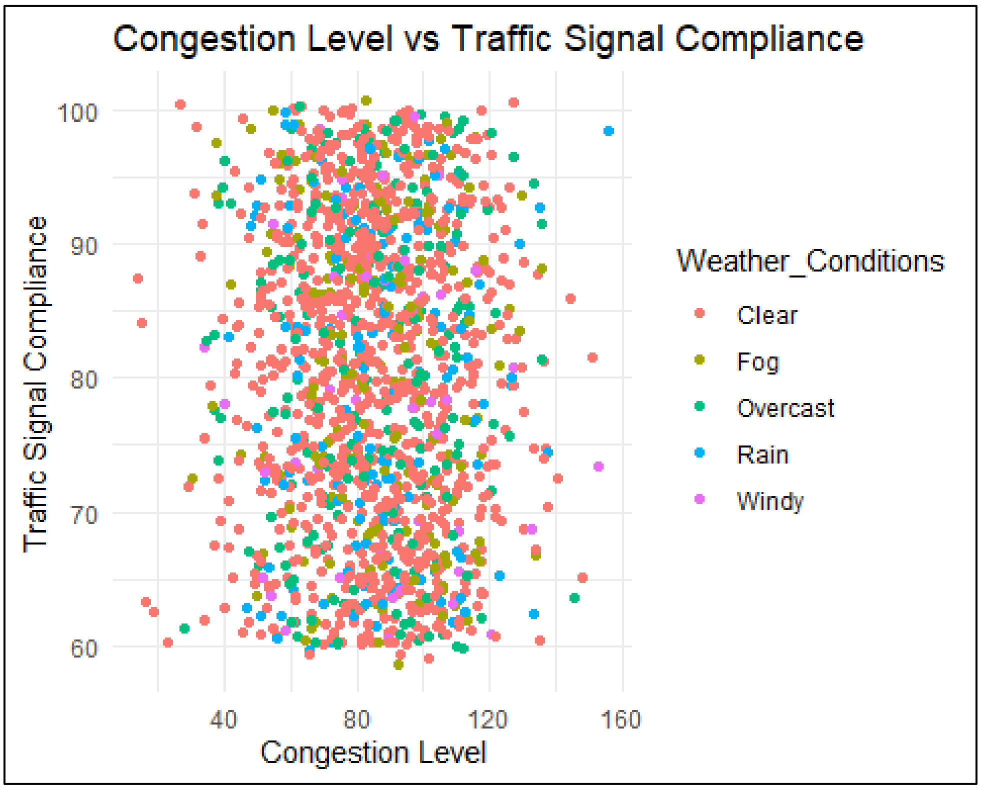
```
## Normality violated for one or both variables. Proceeding with Spear
man correlation.
```

```
## Warning in cor.test.default(data$Congestion_Level,
## data$Traffic_Signal_Compliance, : Cannot compute exact p-value with
ties
```

```
# Print correlation test results
print(correlation_test)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Congestion_Level and data$Traffic_Signal_Compliance
## S = 548492723, p-value = 0.2894
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##         rho
## -0.02761366

# Visualization: Scatter plot for congestion level and traffic signal compliance
library(ggplot2)
ggplot(data, aes(x = Congestion_Level, y = Traffic_Signal_Compliance))
+
  geom_point(aes(color = Weather_Conditions)) +
  labs(title = "Congestion Level vs Traffic Signal Compliance",
       x = "Congestion Level", y = "Traffic Signal Compliance") +
  theme_minimal()
```



Interpretation:

The analysis proceeds with the **Spearman's rank correlation** test because the normality assumption was violated for **Traffic Signal Compliance**, as indicated by the **Shapiro-Wilk test** (p-value < 2.2e-16). The **Congestion Level** data did not show a significant deviation from normality (p-value = 0.7275), but since at least one variable violates the normality assumption, the **Spearman correlation** is used as a non-parametric alternative.

The **Spearman's rank correlation** coefficient ρ\rhoρ is calculated as **-0.0276**, and the corresponding p-value is **0.2894**.

Conclusion:

Given the **p-value = 0.2894**, which is greater than the typical significance threshold of 0.05, we **fail to reject the null hypothesis**. This means there is **no statistically significant monotonic relationship** between **Congestion Level** and **Traffic Signal Compliance** in the dataset.

In other words, based on the Spearman correlation, the data suggests that **Congestion Level** and **Traffic Signal Compliance** do not have a strong monotonic relationship, and changes in one do not necessarily correlate with changes in the other.

Graphical Interpretation

The scatterplot displays congestion levels against traffic signal compliance across different weather conditions. The points appear randomly scattered without a discernible trend, which aligns with the test results indicating no significant correlation.

 # Question 3: Impact of Roadwork and Construction Activity on Traffic Volume
Objective 1: Mann-Whitney U test for difference in Traffic Volume between roads with and without Roadwork/Construction Activity.

Answer:
Let us set up the null hypothesis

H_0: The distributions of Traffic Volume for "Yes" and "No" Roadwork and Construction Activity are the same.
Against the alternative hypothesis
H_1: The distributions of Traffic Volume for "Yes" and "No" Roadwork and Construction Activity are different.

Under H_0, the test statistic is given by

The test statistic for the **Mann-Whitney U test** is denoted by **W**. The U statistic represents the sum of ranks for one group (or the difference between the number of observations in each group), and it is used to test whether the two independent samples come from the same distribution.

```
# Subset data based on Roadwork and Construction Activity ("Yes" vs "No")
data_subset <- data %>% filter(Roadwork_and_Construction_Activity %in% c("Yes", "No"))

# Assumptions for Mann-Whitney U Test:
```

```r
# 1. The two groups are independent.
# 2. The data does not need to be normally distributed.

# Main Test: Mann-Whitney U Test (Wilcoxon rank-sum test)
mann_whitney_result <- wilcox.test(Traffic_Volume ~ Roadwork_and_Const
ruction_Activity, data = data_subset)
cat("\nMann-Whitney U Test Result:\n")
```

```
##
## Mann-Whitney U Test Result:
```

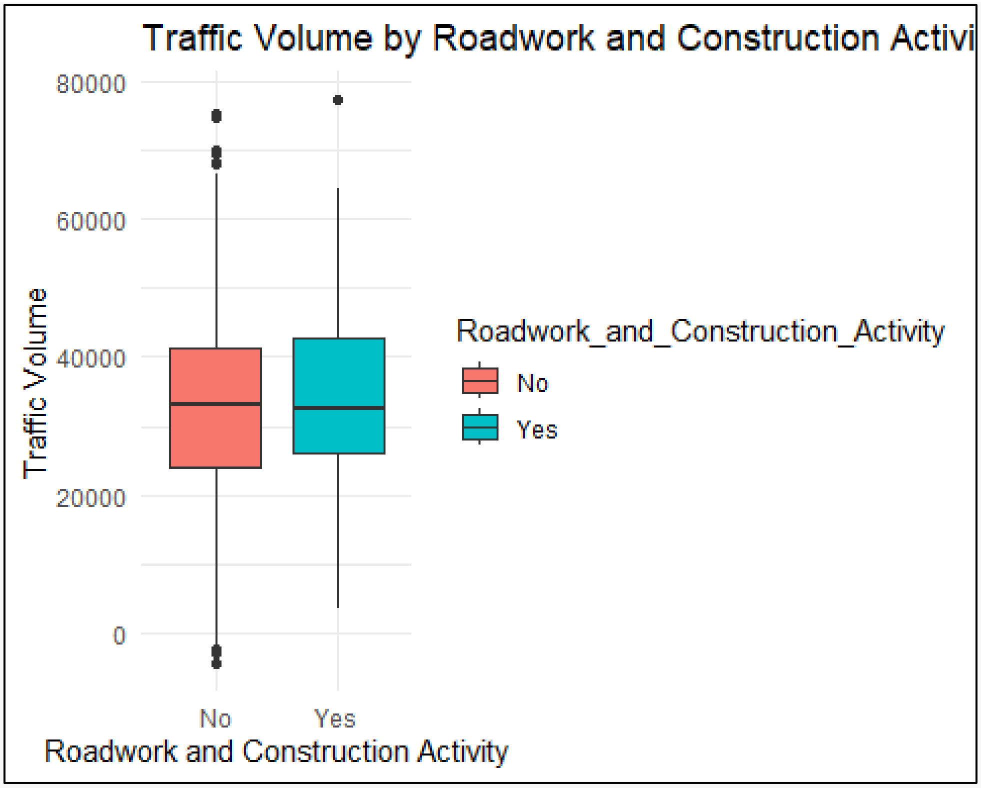
```r
print(mann_whitney_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Traffic_Volume by Roadwork_and_Construction_Activity
## W = 92315, p-value = 0.4777
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Results Interpretation:
if (mann_whitney_result$p.value < 0.05) {
  cat("There is a significant difference in Traffic Volume between the
two groups.\n")
} else {
  cat("No significant difference in Traffic Volume between the two gro
ups.\n")
}
```

```
## No significant difference in Traffic Volume between the two groups.
```

```r
# Visualization: Boxplot for Traffic Volume based on Roadwork and Cons
truction Activity
library(ggplot2)
ggplot(data_subset, aes(x = Roadwork_and_Construction_Activity, y = Tr
affic_Volume, fill = Roadwork_and_Construction_Activity)) +
  geom_boxplot() +
  labs(title = "Traffic Volume by Roadwork and Construction Activity",
       x = "Roadwork and Construction Activity",
       y = "Traffic Volume") +
  theme_minimal()
```



Interpretation:

The **p-value** associated with the Mann-Whitney U test is **0.4777**, which is greater than th e commonly used significance threshold of **0.05**. This means that there is no statistically s ignificant difference in **Traffic Volume** between the two groups.

The relatively large p-value indicates that the observed differences in traffic volumes bet ween the two groups could likely be due to random variation, rather than being a result of roadwork and construction activity.

Conclusion:

Since the **p-value (0.4777)** is greater than **0.05**, we **fail to reject the null hypothesis**. Thi s means that there is **no significant difference** in the **Traffic Volume** between the groups with and without roadwork and construction activity. Therefore, based on this test, it cann ot be concluded that roadwork and construction activity significantly affect traffic volum e in the dataset.

Graphical Interpretation:

The boxplot compares traffic volumes for cases with and without roadwork and construct ion activity, showing similar distributions with overlapping ranges, comparable medians, and interquartile ranges. While the "Yes" group (with roadwork) has a slightly higher me dian, the presence of outliers in both groups suggests variability. The Mann-Whitney U te st (p-value = 0.4777) confirms that the observed differences are not statistically significant, meaning any variation in traffic volume between the two groups is likely due to random chance.

 # Objective 2: Chi-Squared Test for Independence (Roadwork and Construction Activity vs Traffic Volume Category)

 # Categorize Traffic Volume into categories (Low, Medium, High, Very High)

Answer:

Let us set up the null hypothesis

H_0: Roadwork and Construction Activity is independent of Traffic Volume Category.
Against the alternative hypothesis
H_1: Roadwork and Construction Activity is dependent on Traffic Volume Category.

Under H_0, the test statistic is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency in each cell of the contingency table.
- E_i = Expected frequency in each cell of the contingency table, which is calculated under the assumption that the two variables are independent.

```r
data$Traffic_Volume_Category <- cut(
  data$Traffic_Volume,
  breaks = c(0, 10000, 30000, 50000, Inf),
  labels = c("Low", "Medium", "High", "Very High")
)

# Remove rows with NA values in Traffic_Volume_Category or Roadwork_and_Construction_Activity
data_clean <- na.omit(data[, c("Traffic_Volume_Category", "Roadwork_and_Construction_Activity")])

# Create a contingency table
contingency_table <- table(data_clean$Roadwork_and_Construction_Activity, data_clean$Traffic_Volume_Category)

# Assumptions for Chi-squared test:
```

```r
# Check if all expected frequencies are > 5
expected_frequencies <- chisq.test(contingency_table)$expected
if (all(expected_frequencies > 5)) {
  cat("All expected frequencies are greater than 5. Proceeding with th
e Chi-squared test.\n")
} else {
  cat("Warning: Some expected frequencies are less than 5. Chi-squared
test may not be valid.\n")
}
```

```
## All expected frequencies are greater than 5. Proceeding with the Ch
i-squared test.
```

```r
# Perform Chi-squared test
chi_sq_test <- chisq.test(contingency_table)
cat("\nChi-squared Test for Independence:\n")
```

```
##
## Chi-squared Test for Independence:
```

```r
print(chi_sq_test)
```

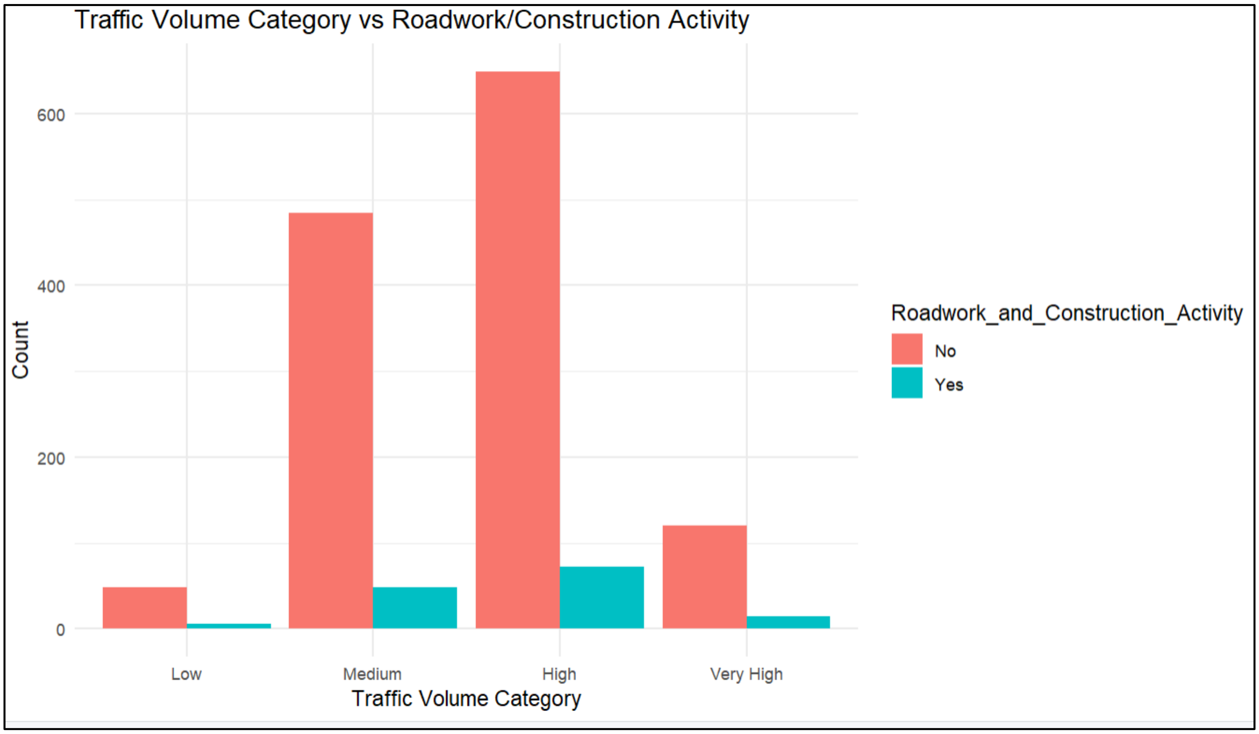
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 0.68579, df = 3, p-value = 0.8765
```

```r
# Results Interpretation
if (chi_sq_test$p.value < 0.05) {
  cat("There is a significant relationship between Roadwork/Constructi
on Activity and Traffic Volume Category.\n")
} else {
  cat("No significant relationship between Roadwork/Construction Activ
ity and Traffic Volume Category.\n")
}
```

```
## No significant relationship between Roadwork/Construction Activity
and Traffic Volume Category.
```

```r
# Visualization: Bar plot for contingency table
library(ggplot2)
ggplot(data_clean, aes(x = Traffic_Volume_Category, fill = Roadwork_an
d_Construction_Activity)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Traffic Volume Category vs Roadwork/Construction Activity
",
```

```
    x = "Traffic Volume Category",
    y = "Count"
  ) +
  theme_minimal()
```



Interpretation:

The **p-value** from the Chi-squared test is **0.8765**, which is much larger than the commonly used significance level of **0.05**. This indicates that there is **no significant association** between **Roadwork and Construction Activity** and **Traffic Volume Category**. The high p-value suggests that the variables are likely **independent** of each other, meaning the occurrence of roadwork and construction activity does not appear to have a significant effect on the distribution of traffic volume categories.

Conclusion:

Since the **p-value (0.8765)** is greater than **0.05**, we **fail to reject the null hypothesis**. Therefore, we conclude that there is **no significant relationship** between **Roadwork and Construction Activity** and **Traffic Volume Category**. The data suggests that roadwork and construction activity does not significantly affect the categorization of traffic volume in this dataset.

Graphical Interpretation:

The bar chart shows the distribution of traffic volume categories (Low, Medium, High, Very High) for cases with and without roadwork and construction activity. Both groups dis play similar patterns, with "High" and "Medium" categories being the most frequent, regardless of roadwork activity. The Chi-squared test (p-value = 0.8765) indicates no significant association between roadwork activity and traffic volume category, suggesting that these variables are independent. Thus, roadwork and construction activity do not significantly influence the categorization of traffic volumes in this dataset.

Objective 3: Two-Way ANOVA to Compare Traffic Volume across Area Names when Roadwork and Construction Activity is Present vs Absent

Answer

Let us set up the null hypothesis
H_0: There is no significant interaction between **Area Name** and **Roadwork and Construction Activity** on **Traffic Volume**.

Against the alternative hypothesis
H_1: There is a significant interaction between **Area Name** and **Roadwork and Construction Activity** on **Traffic Volume**.
Under H_0 the F-statistic is given by:

For the interaction effect in the Two-Way ANOVA, the test statistic (F-value) is calculated as:

$$F = \frac{\text{Mean Square for Interaction}}{\text{Mean Square for Error}}$$

Where the **Mean Square for Interaction** is the variance due to the interaction between **Area Name** and **Roadwork and Construction Activity**, and the **Mean Square for Error** is the variance not explained by the model.

```
# Assumptions for Two-Way ANOVA:

# 1. Normality of residuals for each group (Shapiro-Wilk test)
cat("\nChecking Assumptions for Two-Way ANOVA:\n")

##
## Checking Assumptions for Two-Way ANOVA:

cat("1. Normality of residuals (Shapiro-Wilk Test):\n")

## 1. Normality of residuals (Shapiro-Wilk Test):

shapiro_test_anova <- shapiro.test(residuals(lm(Traffic_Volume ~ Area_Name * Roadwork_and_Construction_Activity, data = data)))
print(shapiro_test_anova)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  residuals(lm(Traffic_Volume ~ Area_Name * Roadwork_and_Const
ruction_Activity, data = data))
## W = 0.99863, p-value = 0.3027

if (shapiro_test_anova$p.value > 0.05) {
  cat("Residuals are normally distributed (p-value =", shapiro_test_an
ova$p.value, ").\n")
} else {
  cat("Residuals are NOT normally distributed (p-value =", shapiro_tes
t_anova$p.value, ").\n")
}

## Residuals are normally distributed (p-value = 0.3027306 ).

# 2. Homogeneity of variances (Bartlett test)
cat("\n2. Homogeneity of variances (Bartlett Test):\n")

##
## 2. Homogeneity of variances (Bartlett Test):

bartlett_test_anova <- bartlett.test(Traffic_Volume ~ interaction(Area
_Name, Roadwork_and_Construction_Activity), data = data)
print(bartlett_test_anova)

##
##  Bartlett test of homogeneity of variances
##
## data:  Traffic_Volume by interaction(Area_Name, Roadwork_and_Constr
uction_Activity)
## Bartlett's K-squared = 4.809, df = 9, p-value = 0.8506

if (bartlett_test_anova$p.value > 0.05) {
  cat("Variances are homogeneous (p-value =", bartlett_test_anova$p.va
lue, ").\n")
} else {
  cat("Variances are NOT homogeneous (p-value =", bartlett_test_anova$
p.value, ").\n")
}

## Variances are homogeneous (p-value = 0.8506295 ).

# Main Test: Two-Way ANOVA
cat("\nConducting Two-Way ANOVA:\n")

##
## Conducting Two-Way ANOVA:

anova_result <- aov(Traffic_Volume ~ Area_Name * Roadwork_and_Construc
tion_Activity, data = data)
```

```
anova_summary <- summary(anova_result)
print(anova_summary)

##                                                 Df    Sum Sq   Mean 
Sq F value
## Area_Name                                        4 1.159e+09 2897079
87   1.723
## Roadwork_and_Construction_Activity               1 6.988e+07  698812
20   0.416
## Area_Name:Roadwork_and_Construction_Activity     4 2.644e+09 6610985
18   3.931
## Residuals                                     1464 2.462e+11 1681811
59
##                                               Pr(>F)   
## Area_Name                                    0.14243   
## Roadwork_and_Construction_Activity           0.51929   
## Area_Name:Roadwork_and_Construction_Activity 0.00353 **
## Residuals                                              
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Extract the F-value and tabulated F-value
f_value <- anova_summary[[1]]$`F value`[3]  # Interaction term F value
f_tabulated <- qf(0.95, df1 = anova_summary[[1]]$Df[3], df2 = anova_su
mmary[[1]]$Df[4])

cat("\nF-value (Interaction Term):", f_value, "\nF-tabulated (Critical
):", f_tabulated, "\n")

## 
## F-value (Interaction Term): 3.930871 
## F-tabulated (Critical): 2.378007

# Check if Tukey's test is needed
if (f_value >= f_tabulated) {
  cat("F value is greater than or equal to the tabulated value. Procee
ding with Tukey's test.\n")
  
  # Tukey's Test
  tukey_result <- TukeyHSD(anova_result)
  print(tukey_result)
} else {
  cat("F value is less than the tabulated value. No Tukey test require
d.\n")
}
```

```
## F value is greater than or equal to the tabulated value. Proceeding
with Tukey's test.
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Traffic_Volume ~ Area_Name * Roadwork_and_Constr
uction_Activity, data = data)
##
## $Area_Name
##                                    diff        lwr       upr     p a
dj
## Indiranagar-Electronic City  -160.2775 -3607.0043 3286.4493 0.99994
14
## Jayanagar-Electronic City    -970.1668 -4641.9489 2701.6153 0.95155
26
## Koramangala-Electronic City  1289.0375 -2270.3757 4848.4507 0.86032
31
## M.G. Road-Electronic City   -1110.7518 -4611.5785 2390.0749 0.90909
01
## Jayanagar-Indiranagar        -809.8893 -3627.3383 2007.5596 0.93497
78
## Koramangala-Indiranagar      1449.3150 -1220.0402 4118.6702 0.57381
09
## M.G. Road-Indiranagar        -950.4743 -3541.1928 1640.2442 0.85451
59
## Koramangala-Jayanagar        2259.2043  -695.0325 5213.4412 0.22546
78
## M.G. Road-Jayanagar          -140.5850 -3023.9651 2742.7951 0.99992
93
## M.G. Road-Koramangala       -2399.7893 -5138.6429  339.0642 0.11771
18
##
## $Roadwork_and_Construction_Activity
##            diff       lwr      upr     p adj
## Yes-No 731.2553 -1500.454 2962.965 0.5204902
##
## $`Area_Name:Roadwork_and_Construction_Activity`
##                                                diff          lwr
upr
## Indiranagar:No-Electronic City:No        -469.8757  -4629.74490  36
89.9935
## Jayanagar:No-Electronic City:No         -1759.5530  -6174.64746  26
55.5415
## Koramangala:No-Electronic City:No        1104.4806  -3224.63063  54
33.5919
## M.G. Road:No-Electronic City:No          -697.1161  -4931.22557  35
```

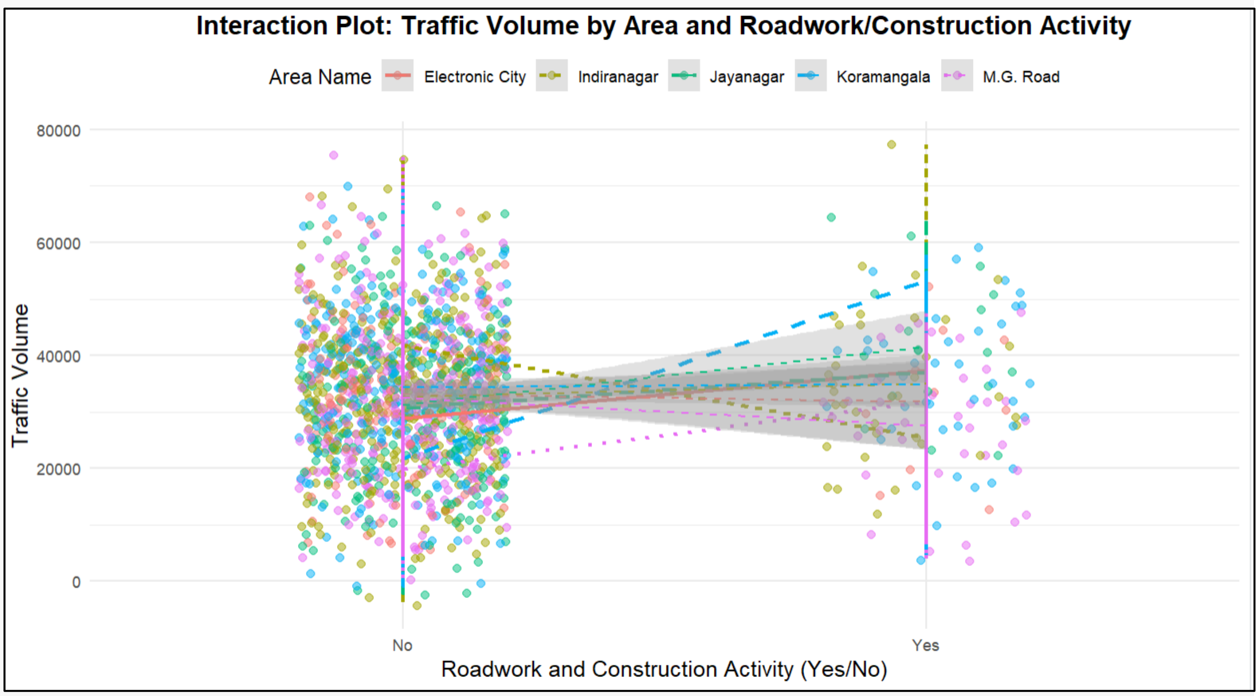
```
36.9933
## Electronic City:Yes-Electronic City:No  -1398.4501 -14868.98017 120
72.0801
## Indiranagar:Yes-Electronic City:No       1804.2733  -5827.06373  94
35.6103
## Jayanagar:Yes-Electronic City:No         8355.3034  -1960.16565 186
70.7725
## Koramangala:Yes-Electronic City:No       1764.4618  -5502.20304  90
31.1266
## M.G. Road:Yes-Electronic City:No        -5583.7541 -13215.09108  20
47.5830
## Jayanagar:No-Indiranagar:No             -1289.6773  -4694.88839  21
15.5339
## Koramangala:No-Indiranagar:No            1574.3563  -1718.60704  48
67.3197
## M.G. Road:No-Indiranagar:No              -227.2404  -3394.27157  29
39.7907
## Electronic City:Yes-Indiranagar:No       -928.5744 -14102.69995 122
45.5512
## Indiranagar:Yes-Indiranagar:No           2274.1490  -4820.89916  93
69.1971
## Jayanagar:Yes-Indiranagar:No             8825.1791  -1100.10706 187
50.4653
## Koramangala:Yes-Indiranagar:No           2234.3375  -4466.91771  89
35.5927
## M.G. Road:Yes-Indiranagar:No            -5113.8784 -12208.92651  19
81.1698
## Koramangala:No-Jayanagar:No              2864.0336   -745.97407  64
74.0412
## M.G. Road:No-Jayanagar:No                1062.4368  -2433.07941  45
57.9531
## Electronic City:Yes-Jayanagar:No          361.1029 -12895.82482 136
18.0306
## Indiranagar:Yes-Jayanagar:No             3563.8263  -3683.81169 108
11.4642
## Jayanagar:Yes-Jayanagar:No              10114.8564     79.92473 201
49.7881
## Koramangala:Yes-Jayanagar:No             3524.0148  -3338.59181 103
86.6213
## M.G. Road:Yes-Jayanagar:No              -3824.2011 -11071.83904  34
23.4368
## M.G. Road:No-Koramangala:No             -1801.5967  -5187.85998  15
84.6665
## Electronic City:Yes-Koramangala:No      -2502.9307 -15731.47095 107
25.6096
## Indiranagar:Yes-Koramangala:No            699.7927  -6495.78935  78
```

```
95.3747
## Jayanagar:Yes-Koramangala:No              7250.8228  -2746.57677 172
48.2224
## Koramangala:Yes-Koramangala:No             659.9812  -6147.62588  74
67.5882
## M.G. Road:Yes-Koramangala:No             -6688.2347 -13883.81671   5
07.3473
## Electronic City:Yes-M.G. Road:No          -701.3339 -13899.08965 124
96.4218
## Indiranagar:Yes-M.G. Road:No              2501.3894  -4637.43956  96
40.2184
## Jayanagar:Yes-M.G. Road:No                9052.4196   -904.21024 190
09.0494
## Koramangala:Yes-M.G. Road:No              2461.5779  -4286.01367  92
09.1695
## M.G. Road:Yes-M.G. Road:No               -4886.6379 -12025.46691  22
52.1910
## Indiranagar:Yes-Electronic City:Yes      3202.7234 -11442.76665 178
48.2134
## Jayanagar:Yes-Electronic City:Yes        9753.7535  -6453.09615 259
60.6032
## Koramangala:Yes-Electronic City:Yes      3162.9119 -11295.90866 176
21.7324
## M.G. Road:Yes-Electronic City:Yes       -4185.3040 -18830.79401 104
60.1860
## Jayanagar:Yes-Indiranagar:Yes            6551.0301  -5257.61495 183
59.6752
## Koramangala:Yes-Indiranagar:Yes           -39.8115  -9304.77854  92
25.1555
## M.G. Road:Yes-Indiranagar:Yes           -7388.0274 -16941.69163  21
65.6369
## Koramangala:Yes-Jayanagar:Yes           -6590.8416 -18167.16283  49
85.4795
## M.G. Road:Yes-Jayanagar:Yes            -13939.0575 -25747.70259 -21
30.4124
## M.G. Road:Yes-Koramangala:Yes           -7348.2159 -16613.18289  19
16.7512
##                                            p adj
## Indiranagar:No-Electronic City:No      0.9999984
## Jayanagar:No-Electronic City:No        0.9615518
## Koramangala:No-Electronic City:No      0.9984702
## M.G. Road:No-Electronic City:No        0.9999585
## Electronic City:Yes-Electronic City:No 0.9999992
## Indiranagar:Yes-Electronic City:No     0.9991645
## Jayanagar:Yes-Electronic City:No       0.2346323
## Koramangala:Yes-Electronic City:No     0.9989660
```

```
## M.G. Road:Yes-Electronic City:No       0.3774794
## Jayanagar:No-Indiranagar:No            0.9724233
## Koramangala:No-Indiranagar:No          0.8864199
## M.G. Road:No-Indiranagar:No            1.0000000
## Electronic City:Yes-Indiranagar:No     1.0000000
## Indiranagar:Yes-Indiranagar:No         0.9913680
## Jayanagar:Yes-Indiranagar:No           0.1314440
## Koramangala:Yes-Indiranagar:No         0.9885368
## M.G. Road:Yes-Indiranagar:No           0.4001639
## Koramangala:No-Jayanagar:No            0.2617068
## M.G. Road:No-Jayanagar:No              0.9941555
## Electronic City:Yes-Jayanagar:No       1.0000000
## Indiranagar:Yes-Jayanagar:No           0.8679283
## Jayanagar:Yes-Jayanagar:No             0.0463377
## Koramangala:Yes-Jayanagar:No           0.8349078
## M.G. Road:Yes-Jayanagar:No             0.8112973
## M.G. Road:No-Koramangala:No            0.8036460
## Electronic City:Yes-Koramangala:No     0.9998657
## Indiranagar:Yes-Koramangala:No         0.9999996
## Jayanagar:Yes-Koramangala:No           0.3907414
## Koramangala:Yes-Koramangala:No         0.9999996
## M.G. Road:Yes-Koramangala:No           0.0943669
## Electronic City:Yes-M.G. Road:No       1.0000000
## Indiranagar:Yes-M.G. Road:No           0.9837432
## Jayanagar:Yes-M.G. Road:No             0.1118351
## Koramangala:Yes-M.G. Road:No           0.9785559
## M.G. Road:Yes-M.G. Road:No             0.4788485
## Indiranagar:Yes-Electronic City:Yes    0.9995556
## Jayanagar:Yes-Electronic City:Yes      0.6643792
## Koramangala:Yes-Electronic City:Yes    0.9995544
## M.G. Road:Yes-Electronic City:Yes      0.9963211
## Jayanagar:Yes-Indiranagar:Yes          0.7616182
## Koramangala:Yes-Indiranagar:Yes        1.0000000
## M.G. Road:Yes-Indiranagar:Yes          0.2967934
## Koramangala:Yes-Jayanagar:Yes          0.7328070
## M.G. Road:Yes-Jayanagar:Yes            0.0072997
## M.G. Road:Yes-Koramangala:Yes          0.2621132
```

```r
# Enhanced Interaction Plot


ggplot(data, aes(x = Roadwork_and_Construction_Activity,
                 y = Traffic_Volume,
                 color = Area_Name,
                 group = Area_Name)) +
  geom_point(position = position_jitter(width = 0.2), alpha = 0.5, siz
```

```
e = 2) +
  geom_line(aes(linetype = Area_Name), size = 1) +
  geom_smooth(method = "lm", se = TRUE, linetype = "dashed", size = 0.
5, alpha = 0.3) +
  labs(title = "Interaction Plot: Traffic Volume by Area and Roadwork/
Construction Activity",
       x = "Roadwork and Construction Activity (Yes/No)",
       y = "Traffic Volume",
       color = "Area Name",
       linetype = "Area Name") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold
"),
        axis.title = element_text(size = 12))

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2
3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warni
ng was
## generated.
```




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

Interpretation:

The **Two-Way ANOVA** reveals a **significant interaction** between **Area Name** and **Roadwork and Construction Activity**, suggesting that roadwork's impact on traffic volume varies across different areas. However, the main effects of **Area Name** and **Roadwork and Construction Activity** individually are not significant (p-values 0.14243 and 0.51929, respectively).

Conclusion:

Since the interaction effect is significant, we conclude that the impact of roadwork on traffic volume differs across areas. Specific comparisons, such as between **M.G. Road** and **Koramangala** during roadwork conditions, show notable differences in traffic volume. Thus, traffic volume should be analyzed with consideration of both area and construction activity together, rather than independently.

Graphical Interpretation:
The interaction plot shows traffic volume by roadwork/construction activity (Yes/No) across different areas (Electronic City, Indiranagar, Jayanagar, Koramangala, and M.G. Road). The overlapping data points and trends indicate no significant main effects of either roadwork or area individually, as their p-values are above the 0.05 threshold. However, the significant interaction effect from the Two-Way ANOVA suggests that the impact of roadwork on traffic volume varies by area. For instance, some areas like M.G. Road show more pronounced changes in traffic volume during roadwork compared to others like Indiranagar. This emphasizes the importance of jointly analyzing area and roadwork activity when assessing traffic volume.

 # Question 4: Exploring the Effect of Weather Conditions on Public Transport Usage

 # Objective 1: One-Way ANOVA for Public Transport Usage by Weather Conditions.

Answer:
Let us set up the null hypothesis
H_0: There is no difference in the mean public transport usage across the different weather conditions

i.e. $\mu_{Clear} = \mu_{Rainy}=\mu_{Foggy} = \mu_{Snowy}=\mu_{Cloudy}$

Against the alternative hypothesis

H_1: At least one of the means is different from the others. This means that public transport usage differs depending on the weather condition.
Under H_0 the test statistic is

$$F = \frac{\text{Between-group variability (Mean Square Between)}}{\text{Within-group variability (Mean Square Error)}}$$

Where:

- $\text{Mean Square Between} = \frac{\text{Sum of Squares Between Groups (SSB)}}{\text{Degrees of Freedom Between Groups } (\text{df}_B)}$
- $\text{Mean Square Error} = \frac{\text{Sum of Squares Within Groups (SSW)}}{\text{Degrees of Freedom Within Groups } (\text{df}_W)}$

```
# Assumptions for One-Way ANOVA:
# 1. Normality of residuals for each group (Shapiro-Wilk test).
cat("\nPerforming Shapiro-Wilk Test for Normality of Residuals:\n")

##
## Performing Shapiro-Wilk Test for Normality of Residuals:

shapiro_tests <- by(data$Public_Transport_Usage, data$Weather_Conditio
ns, function(x) shapiro.test(x)$p.value)
cat("\nShapiro-Wilk Test Results (p-values):\n")

##
## Shapiro-Wilk Test Results (p-values):

print(shapiro_tests)

## data$Weather_Conditions: Clear
## [1] 0.5334393
## ------------------------------------------------------------
## data$Weather_Conditions: Fog
## [1] 0.045456
## ------------------------------------------------------------
## data$Weather_Conditions: Overcast
## [1] 0.3388047
## ------------------------------------------------------------
## data$Weather_Conditions: Rain
## [1] 0.1484458
## ------------------------------------------------------------
## data$Weather_Conditions: Windy
## [1] 0.3600069
```

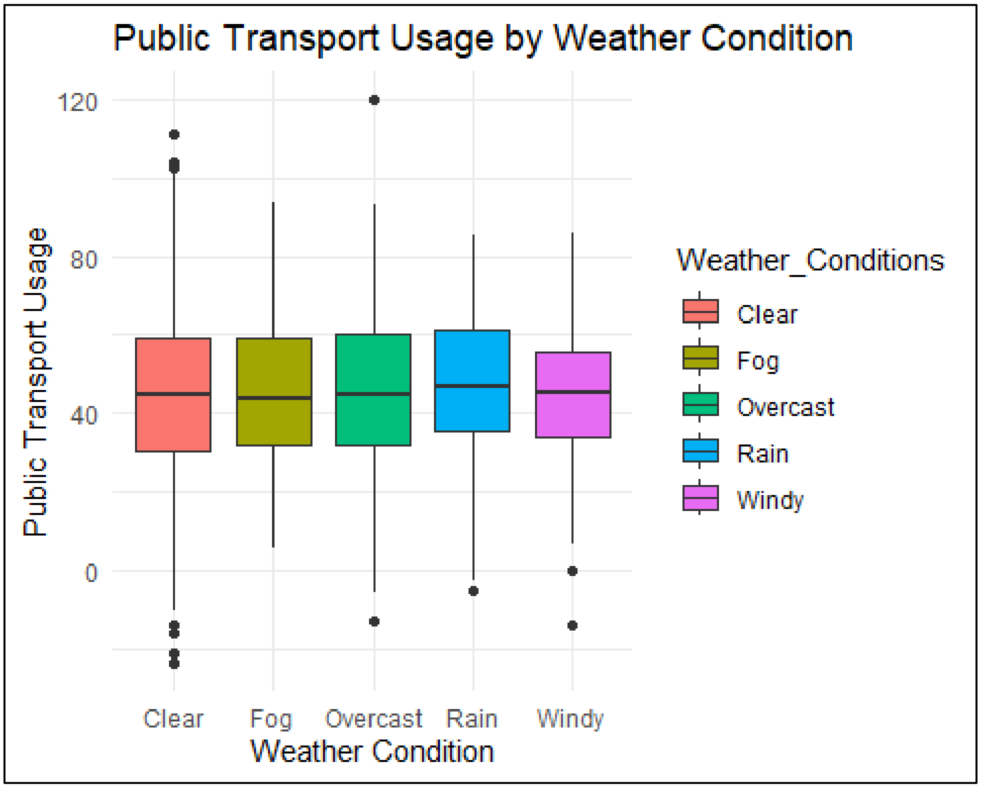
```r
# 2. Homogeneity of variances (Bartlett test).
cat("\nPerforming Bartlett Test for Homogeneity of Variances:\n")

##
## Performing Bartlett Test for Homogeneity of Variances:

bartlett_test <- bartlett.test(Public_Transport_Usage ~ Weather_Condit
ions, data = data)
cat("\nBartlett Test Results:\n")

##
## Bartlett Test Results:

print(bartlett_test)

##
##  Bartlett test of homogeneity of variances
##
## data:  Public_Transport_Usage by Weather_Conditions
## Bartlett's K-squared = 1.937, df = 4, p-value = 0.7473

# Main Test: One-Way ANOVA
cat("\nPerforming One-Way ANOVA:\n")

##
## Performing One-Way ANOVA:

anova_result <- aov(Public_Transport_Usage ~ Weather_Conditions, data
= data)
anova_summary <- summary(anova_result)
cat("\nOne-Way ANOVA Summary:\n")

##
## One-Way ANOVA Summary:

print(anova_summary)

##                      Df Sum Sq Mean Sq F value Pr(>F)
## Weather_Conditions    4   1249   312.3   0.787  0.534
## Residuals          1469 583018   396.9

# Check F-value and determine if Tukey's post hoc test is needed
f_value <- anova_summary[[1]]$`F value`[1]  # Extract the first F-valu
e
f_tabulated <- qf(0.95, df1 = anova_summary[[1]]$Df[1], df2 = anova_su
mmary[[1]]$Df[2])

cat("\nF-value from ANOVA:", f_value, "\nF-tabulated:", f_tabulated, "
\n")
```

```
##
## F-value from ANOVA: 0.7867829
## F-tabulated: 2.377986
```

```r
if (f_value >= f_tabulated) {
  cat("\nF value is greater than or equal to the tabulated value. Proc
eeding with Tukey's test.\n")
  tukey_result <- TukeyHSD(anova_result)
  cat("\nTukey's Post Hoc Test Results:\n")
  print(tukey_result)
} else {
  cat("\nF value is less than the tabulated value. No Tukey test requi
red.\n")
}
```

```
##
## F value is less than the tabulated value. No Tukey test required.
```

```r
# Visualization: Boxplot for Public Transport Usage across Weather Con
ditions

ggplot(data, aes(x = Weather_Conditions, y = Public_Transport_Usage, f
ill = Weather_Conditions)) +
  geom_boxplot() +
  labs(title = "Public Transport Usage by Weather Condition", x = "Wea
ther Condition", y = "Public Transport Usage") +
  theme_minimal()
```



Interpretation:

The One-Way ANOVA test was conducted to examine whether there are significant differences in public transport usage across different weather conditions. The F-value from the ANOVA (0.787) was found to be smaller than the F-tabulated value (2.378), indicating that the variation between the groups is not greater than the variation within the groups. Additionally, the p-value for the ANOVA was 0.534, which is much higher than the typical significance level of 0.05. These results suggest that there is no significant difference in public transport usage between the different weather conditions.

Conclusion:

Based on the results of the One-Way ANOVA, we fail to reject the null hypothesis, meaning that weather conditions do not have a statistically significant effect on public transport usage in this dataset. The data indicates that differences in public transport usage across weather conditions are likely due to random variation rather than any meaningful or systematic differences. Therefore, there is no strong evidence to suggest that weather conditions influence how people use public transportation.

Graphical Interpretation:

The graph shows that public transport usage remains fairly consistent across different weather conditions, with similar medians and overlapping distributions. This suggests that weather conditions, such as Clear, Fog, Overcast, Rain, and Windy, do not significantly im pact public transport usage, as any differences appear minor and within the range of random variation.

Objective 2: Spearman Rank Correlation between Environmental Impact and Public Transport Usage

Answer

Let us set up the null hypothesis

H_0: There is no monotonic relationship between **Environmental Impact** and **Public Transport Usage**, i.e . $\rho = 0$

Against the alternative hypothesis

H_1: There is a monotonic relationship between **Environmental Impact** and **Public Transport Usage** ($\rho \neq 0$).

Under H_0 the test statistic is given by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where:

- d_i is the difference in ranks between the paired observations.
- n is the number of paired observations.

```r
# Assumption: Monotonic relationship between variables.
cat("\nAssumption Check: Monotonic relationship between Environmental Impact and Public Transport Usage.\n")
```

```
##
## Assumption Check: Monotonic relationship between Environmental Impact and Public Transport Usage.
```

```r
cat("Proceeding with Spearman Rank Correlation calculation.\n")
```

```
## Proceeding with Spearman Rank Correlation calculation.
```

```r
# Calculate Spearman rank correlation
spearman_corr <- cor.test(data$Environmental_Impact, data$Public_Transport_Usage, method = "spearman")

# Print the Spearman correlation result
cat("\nSpearman Rank Correlation Test Results:\n")
```

```
##
## Spearman Rank Correlation Test Results:
```

```r
print(spearman_corr)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Environmental_Impact and data$Public_Transport_Usage
## S = 539285728, p-value = 0.6909
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.01036415
```

```r
# Check the p-value and correlation coefficient
cat("\nSpearman's rho (Correlation Coefficient):", spearman_corr$estim
ate, "\n")
```

```
##
## Spearman's rho (Correlation Coefficient): -0.01036415
```

```r
cat("p-value:", spearman_corr$p.value, "\n")
```

```
## p-value: 0.6909392
```

```r
cat("Null hypothesis: rho = 0 (no monotonic relationship)\n")
```

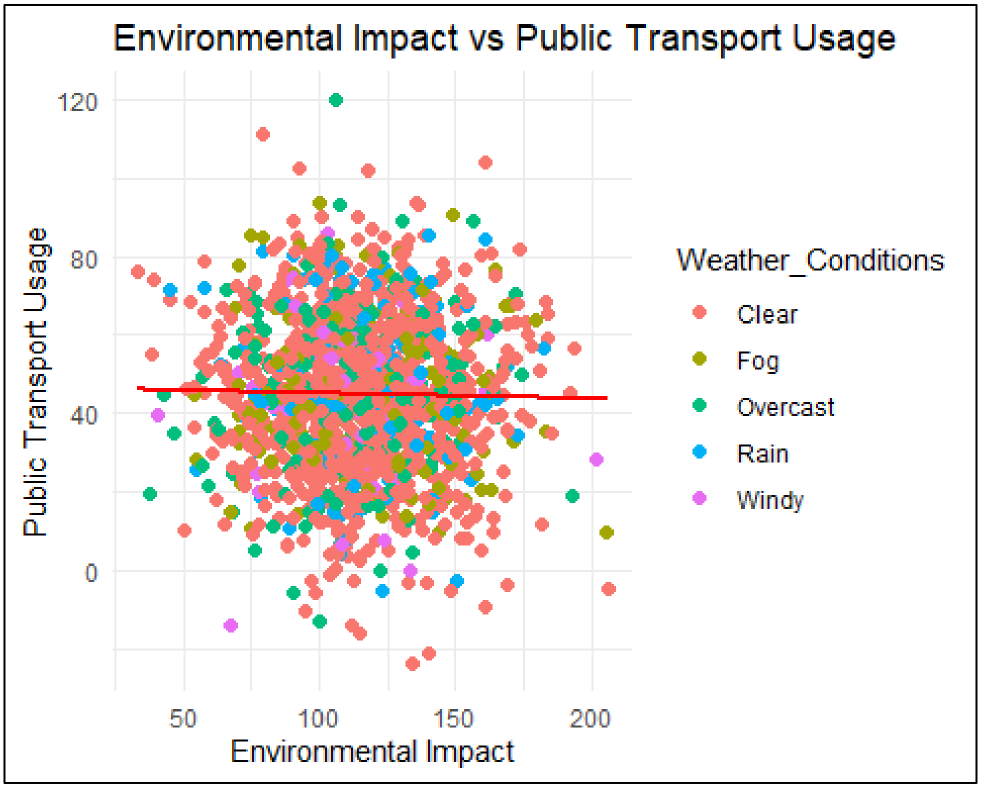
```
## Null hypothesis: rho = 0 (no monotonic relationship)
```

```r
if (spearman_corr$p.value < 0.05) {
  cat("The correlation is statistically significant. We reject the nul
l hypothesis.\n")
} else {
  cat("The correlation is not statistically significant. We fail to re
ject the null hypothesis.\n")
}
```

```
## The correlation is not statistically significant. We fail to reject
the null hypothesis.
```

```r
# Visualization: Scatter plot for Environmental Impact vs Public Trans
port Usage

ggplot(data, aes(x = Environmental_Impact, y = Public_Transport_Usage)
) +
  geom_point(aes(color = Weather_Conditions), size = 2) +
  labs(title = "Environmental Impact vs Public Transport Usage", x = "
Environmental Impact", y = "Public Transport Usage") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE, color = "red")  # Add trend l
ine
```

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



Interpretation

The Spearman rank correlation test returned a test statistic (ρ) of **-0.0104**, suggesting an e xtremely weak negative monotonic relationship between **Environmental Impact** and **Pu blic Transport Usage**. With a **p-value** of **0.6909**, which is much greater than the signific ance level of 0.05, we conclude that there is no significant correlation between the two va riables. This indicates that changes in **Environmental Impact** do not consistently relate t o changes in **Public Transport Usage** in a monotonic fashion in this dataset.

Conclusion:

Given the **Spearman's rho** of **-0.0104** and a **p-value** of **0.6909**, we fail to reject the null hypothesis that there is no monotonic relationship between **Environmental Impact** and **Public Transport Usage**. The correlation is not statistically significant, meaning there is no evidence to support a consistent or meaningful monotonic relationship between the tw o variables in this dataset.

Graphical Interpretation:

The scatterplot shows no clear trend or pattern between Environmental Impact and Public Transport Usage, with points widely scattered across the plot and no discernible monoton ic relationship. The nearly flat red regression line further indicates an extremely weak or negligible association between the two variables, consistent with the statistical finding of a non-significant correlation. This visual representation supports the conclusion that chan ges in Environmental Impact do not meaningfully relate to changes in Public Transport Usage.

 # Objective 3: Poisson Regression to predict Public Transport Usage by Weather Conditions

Answer:

Let us set up the null hypothesis

H_0: There is no relationship between Weather Conditions and Public Transport Usage. This means that the coefficients for the weather conditions (Fog, Overcast, Rain, Windy) are all zero.

i.e. $\beta_{\text{Weather Conditions}} = 0$ (no effect)

Against the alternative hypothesis

H_1: There is a relationship between Weather Conditions and Public Transport Usage. This implies that at least one of the coefficients for the weather conditions is non-zero.

i.e. $\beta_{\text{Weather Conditions}} \neq 0$ (effect exists)

Under H_0 the test statistic is given by:

In Poisson regression, the test statistic for each coefficient is based on the **Wald z-test**. The z-value for each coefficient is computed as:

$$z = \frac{\hat{\beta}}{\mathrm{SE}(\hat{\beta})}$$

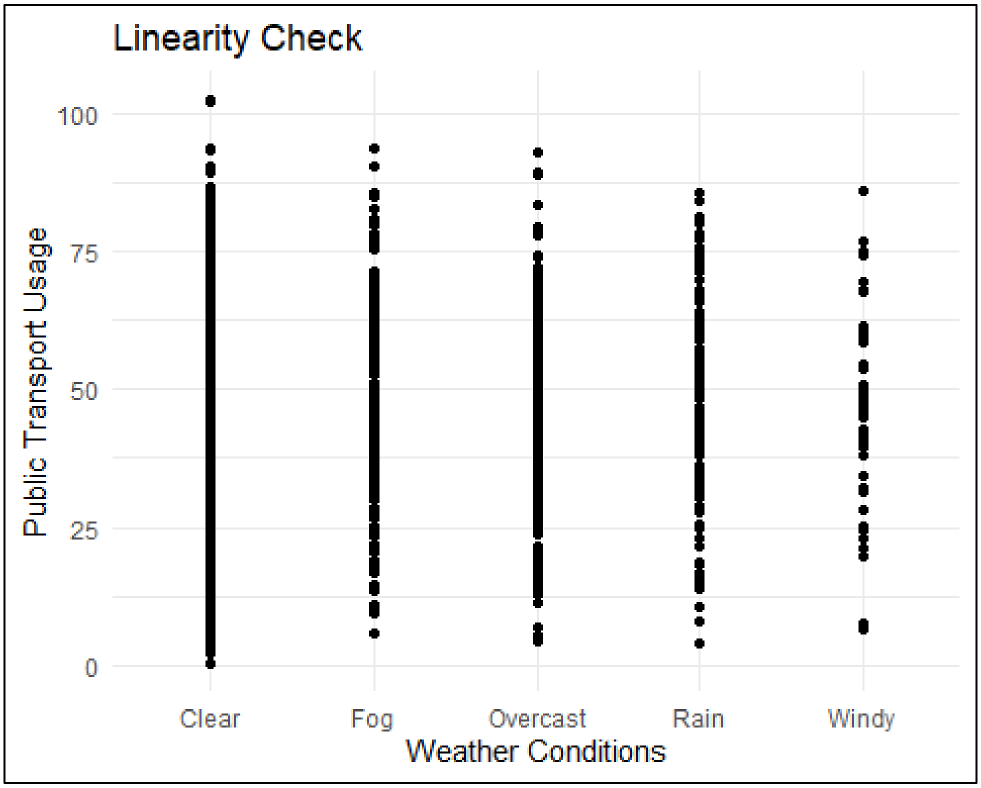
Where:

- $\hat{\beta}$ is the estimated coefficient for the predictor (e.g., Weather Conditions).
- $\mathrm{SE}(\hat{\beta})$ is the standard error of the coefficient estimate.

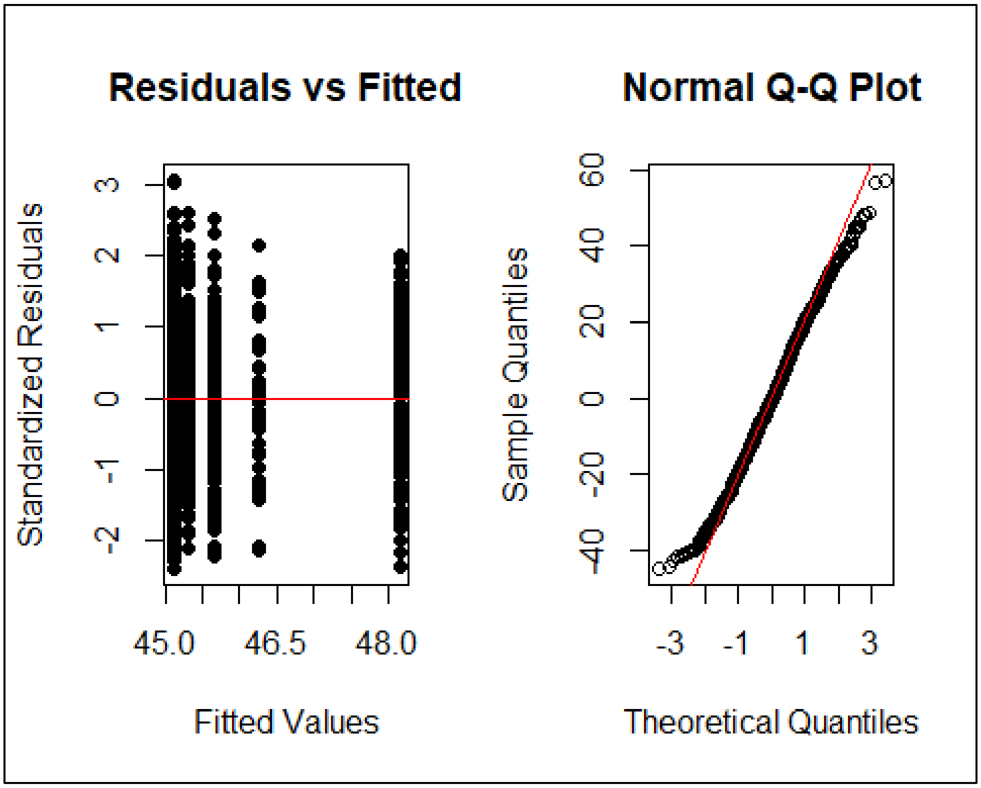
```
# Step 1: Remove rows with negative or extreme Public_Transport_Usage values
cat("\nStep 1: Removing rows with negative or extreme Public_Transport_Usage values...\n")
```

```
##
## Step 1: Removing rows with negative or extreme Public_Transport_Usage values...
```

```
data <- data[data$Public_Transport_Usage >= 0, ]
cat("Removed rows with negative Public Transport Usage values.\n")

## Removed rows with negative Public Transport Usage values.

# Remove extreme outliers (values > 3 SD from the mean)
outlier_threshold <- mean(data$Public_Transport_Usage) + 3 * sd(data$P
ublic_Transport_Usage)
cat("\nOutlier threshold for Public_Transport_Usage (mean + 3 SD):", o
utlier_threshold,"\n")

##
## Outlier threshold for Public_Transport_Usage (mean + 3 SD): 102.566
5

data <- data[data$Public_Transport_Usage <= outlier_threshold, ]
cat("Removed rows with extreme outliers in Public Transport Usage.\n")

## Removed rows with extreme outliers in Public Transport Usage.

# Step 2: Visualizations and Assumption Checks

# 2.1: Check linearity assumption using a scatter plot with smooth lin
e
cat("\nStep 2.1: Checking linearity assumption...\n")

##
## Step 2.1: Checking linearity assumption...

ggplot(data, aes(x = Weather_Conditions, y = Public_Transport_Usage))
+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Linearity Check", x = "Weather Conditions", y = "Publi
c Transport Usage") +
  theme_minimal()

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




```
# 2.2: Homoscedasticity check using residuals vs fitted values
cat("\nStep 2.2: Checking homoscedasticity...\n")

##
## Step 2.2: Checking homoscedasticity...

linear_model <- lm(Public_Transport_Usage ~ Weather_Conditions, data =
data
par(mfrow = c(1, 2))  # Set up 1 row, 2 columns for plots
plot(linear_model$fitted.values, rstandard(linear_model),
     xlab = "Fitted Values", ylab = "Standardized Residuals",
     main = "Residuals vs Fitted", pch = 19)
abline(h = 0, col = "red")

# 2.3: Normality check for residuals using QQ plot and Shapiro-Wilk te
st
cat("\nStep 2.3: Checking normality of residuals...\n")

##
## Step 2.3: Checking normality of residuals...

qqnorm(residuals(linear_model))
qqline(residuals(linear_model), col = "red")
```




```
cat("\nShapiro-Wilk Test for Normality of Residuals:\n")

##
## Shapiro-Wilk Test for Normality of Residuals:

shapiro_test <- shapiro.test(residuals(linear_model))
print(shapiro_test)

##
##  Shapiro-Wilk normality test
##
## data:  residuals(linear_model)
## W = 0.99495, p-value = 8.451e-05

# Step 3: Fit Poisson regression model
cat("\nStep 3: Fitting Poisson regression model...\n")

##
## Step 3: Fitting Poisson regression model...

poisson_model <- glm(Public_Transport_Usage ~ Weather_Conditions, fami
ly = poisson(), data = data)

cat("\nPoisson Model Summary:\n")
```

```
##
## Poisson Model Summary:
```

```r
summary(poisson_model)
```

```
##
## Call:
## glm(formula = Public_Transport_Usage ~ Weather_Conditions, family =
poisson(),
##     data = data)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.809159   0.005019 758.964  < 2e-16 ***
## Weather_ConditionsFog        0.004649   0.012156   0.382    0.702
## Weather_ConditionsOvercast   0.012253   0.011512   1.064    0.287
## Weather_ConditionsRain       0.065160   0.013214   4.931 8.17e-07 ***
## Weather_ConditionsWindy      0.025212   0.022250   1.133    0.257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12104  on 1448  degrees of freedom
## Residual deviance: 12079  on 1444  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```r
# Check for overdispersion (Residual deviance / df should be close to
1)
dispersion_ratio <- summary(poisson_model)$deviance / summary(poisson_
model)$df.residual
cat("\nDispersion Ratio for Poisson Model:", dispersion_ratio, "\n")
```

```
##
## Dispersion Ratio for Poisson Model: 8.364846
```

```r
if (dispersion_ratio > 1.5) {
  cat("\nWarning: Potential overdispersion. Consider using Negative Bi
nomial regression.\n")
}
```

```
##
## Warning: Potential overdispersion. Consider using Negative Binomial
regression.
```

```r
# Step 4: Fit Negative Binomial regression model to address overdisper
sion
cat("\nStep 4: Fitting Negative Binomial regression model...\n")
```

```
##
## Step 4: Fitting Negative Binomial regression model...
```

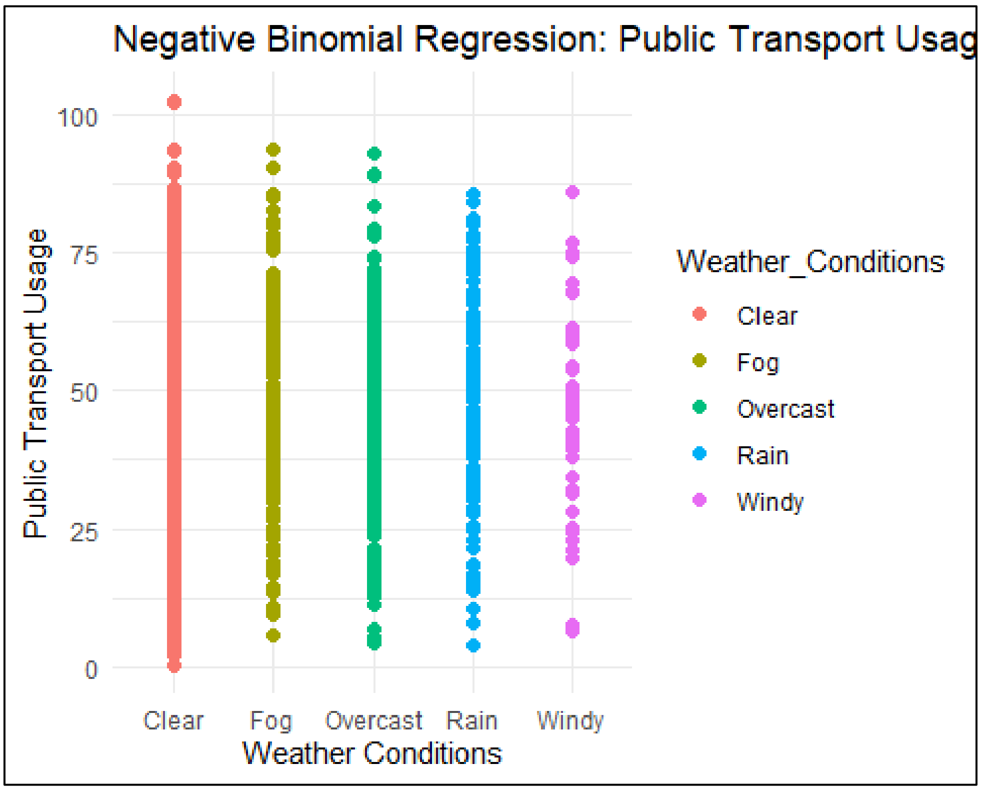
```r
negbinom_model <- glm.nb(Public_Transport_Usage ~ Weather_Conditions,
data = data)

cat("\nNegative Binomial Model Summary:\n")
```

```
##
## Negative Binomial Model Summary:
```

```r
summary(negbinom_model)
```

```
##
## Call:
## glm.nb(formula = Public_Transport_Usage ~ Weather_Conditions,
##     data = data, init.theta = 5.443490081, link = log)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.809159   0.015295 249.041   <2e-16 ***
## Weather_ConditionsFog        0.004649   0.037109   0.125    0.900
## Weather_ConditionsOvercast 0.012253   0.035239   0.348    0.728
## Weather_ConditionsRain       0.065160   0.041291   1.578    0.115
## Weather_ConditionsWindy      0.025212   0.068538   0.368    0.713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(5.4435) family taken to
be 1)
##
##     Null deviance: 1531.3  on 1448  degrees of freedom
## Residual deviance: 1528.7  on 1444  degrees of freedom
## AIC: 12737
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  5.443
##           Std. Err.:  0.228
##
##  2 x log-likelihood:  -12725.254
```

```
# Step 5: Visualize the Negative Binomial regression results

ggplot(data, aes(x = Weather_Conditions, y = Public_Transport_Usage))
+
  geom_point(aes(color = Weather_Conditions), size = 2) +
  geom_smooth(method = "glm", method.args = list(family = "poisson"),
se = TRUE, color = "blue") +
  labs(title = "Negative Binomial Regression: Public Transport Usage v
s. Weather Conditions",
       x = "Weather Conditions", y = "Public Transport Usage") +
  theme_minimal()

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



Interpretation:

The **Poisson regression model** results show that most weather conditions do not have a s ignificant effect on **Public Transport Usage** except for **Rain**. The coefficient for **Rain** is statistically significant with a p-value of **< 0.0001**, suggesting that rainy weather increase s **Public Transport Usage**. The coefficients for **Fog**, **Overcast**, and **Windy** weather con ditions have high p-values (greater than 0.05), meaning these weather conditions do not h ave a statistically significant effect on **Public Transport Usage** in this model. The disper sion ratio of **8.36** suggests overdispersion, indicating that the Poisson model may not be t he best fit, and a Negative Binomial regression model is more appropriate.

Conclusion:

Based on the **Poisson regression** results, we fail to reject the null hypothesis for most we ather conditions except for **Rain**, which is significantly associated with **Public Transpor t Usage**. The significant increase in usage during rainy weather suggests that weather con ditions do influence public transport patterns. However, due to overdispersion (dispersion ratio > 1.5), we recommend using a **Negative Binomial regression** to better model the da ta, as it accounts for the overdispersion observed in the Poisson model.

Graphical Representation:

1. **Linearity Check :**

   - The scatterplot shows that public transport usage varies widely across weather conditions (Clear, Fog, Overcast, Rain, and Windy), with no apparent linear trend.
   - This suggests that the effect of weather conditions on transport usage is likely non-linear or complex, warranting a more flexible model to capture the patterns.

2. **Model Diagnostics:**

   - **Residuals vs. Fitted Plot**: The residuals are not evenly distributed and sho w signs of clustering, indicating that the Poisson regression model fails to fully explain the variability in the data.
   - **Normal Q-Q Plot:** Residuals deviate from the normal distribution, especially in the tails, further supporting the idea that the Poisson model may not be appropriate due to overdispersion (variance much greater than the mean).

3. **Negative Binomial Regression :**

   - The Negative Binomial regression model accounts for overdispersion in the data (evident in Model Diagnostics).
   - This plot reiterates that rainy weather significantly increases public transpor t usage, while other weather conditions (Fog, Overcast, and Windy) show le ss or no statistically significant effect.
   - The clustering of points suggests that public transport usage during rainy w eather is higher and more consistent, while the variation is more widespread under other weather conditions.

Key Insights:

- **Effect of Rain:** Rainy weather significantly increases public transport usage, likely because people opt for public transport over walking or other means during rain.
- **Other Conditions:** Fog, Overcast, and Windy conditions show no significant effect on transport usage.
- **Model Suitability:** The Poisson regression model is inadequate due to overdispersion, as indicated by the high dispersion ratio (8.36) and residual diagnostics. The Negative Binomial regression is a better fit, as it accounts for the overdispersion and better reflects the variability in the data.

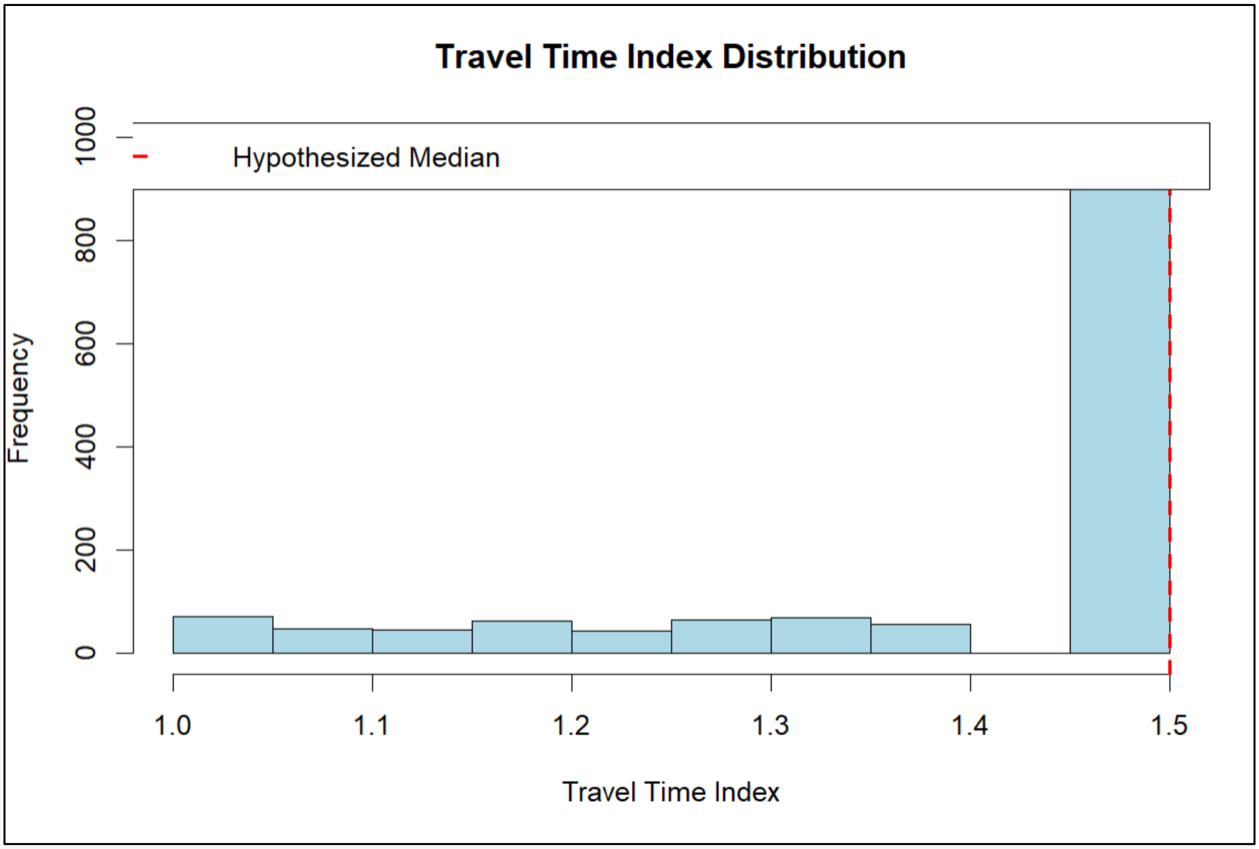
 # Question 5: Analyzing the Relationship Between Congestion Level and Travel Time Index
 # Objective 1: Determine if the median Travel Time Index differs significantly from a hypothesized value using a Wilcoxon Signed-Rank Test.

Answer
Let us set up the null hypothesis
H_0: The median Travel Time Index is equal to the hypothesized value ($\mu=1.5$).

Against the alternative hypothesis
H_1: The median Travel Time Index is not equal to the hypothesized value ($\mu \neq 1.5$)

```
# Step 1: Assumption - Normality Check
cat("\nChecking the assumption of normality for Travel Time Index...\n
")

##
## Checking the assumption of normality for Travel Time Index...

shapiro_test <- shapiro.test(data$Travel_Time_Index)
print(shapiro_test)

##
##  Shapiro-Wilk normality test
##
## data:  data$Travel_Time_Index
## W = 0.65383, p-value < 2.2e-16

if (shapiro_test$p.value < 0.05) {
  cat("Result: The data significantly deviates from normality (p < 0.0
5).\n")
  cat("Proceeding with Wilcoxon Signed-Rank Test (non-parametric).\n")
} else {
```

```
  cat("Result: The data does not significantly deviate from normality
(p ≥ 0.05).\n")
  cat("Consider using a parametric test if assumptions are met.\n")
}

## Result: The data significantly deviates from normality (p < 0.05).
## Proceeding with Wilcoxon Signed-Rank Test (non-parametric).

# Step 2: Main Test - Wilcoxon Signed-Rank Test for a Single Median
cat("\nConducting Wilcoxon Signed-Rank Test...\n")

##
## Conducting Wilcoxon Signed-Rank Test...

mu <- 1.5  # Hypothesized median

# Perform Wilcoxon Signed-Rank Test
wilcox_test <- wilcox.test(data$Travel_Time_Index, mu = mu, alternativ
e = "two.sided")
print(wilcox_test)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data$Travel_Time_Index
## V = 0, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 1.5

# Step 3: Visualization - Histogram with Hypothesized Median

hist(data$Travel_Time_Index,
     main = "Travel Time Index Distribution",
     xlab = "Travel Time Index",
     col = "lightblue",
     border = "black")
abline(v = mu, col = "red", lwd = 2, lty = 2)  # Add line for hypothes
ized median
legend("topright", legend = c("Hypothesized Median"), col = c("red"),
lty = 2, lwd = 2)
```



Interpretation:
The Shapiro-Wilk test shows that the data significantly deviates from normality ($p<0.05$), confirming the appropriateness of using the non-parametric Wilcoxon Signed-Rank Test. The test statistic $V=0$ and an extremely small p-value (<0.05) indicate strong evidence against the null hypothesis. This suggests that the true median Travel Time Index significantly differs from the hypothesized value of 1.5.

Conclusion:

Since the p-value is much smaller than the typical significance level of 0.05, we **reject the null hypothesis**. This implies that the median Travel Time Index is statistically different from 1.5. The histogram further shows the distribution of the data, and the red dashed line at the hypothesized median highlights this deviation visually.

Graphical Interpretation:

The histogram shows the distribution of the Travel Time Index, with most of the data concentrated well below the hypothesized median value of 1.5 (indicated by the red dashed line). This suggests a significant deviation from the hypothesized value. The distribution is highly skewed, with a sharp increase in frequency near the hypothesized value, which aligns with the statistical results from the Wilcoxon Signed-Rank Test and Shapiro-Wilk tes

t. Together, these findings strongly indicate that the actual median Travel Time Index is statistically different from 1.5.

 # Objective 2: Investigate the monotonic relationship between Congestion Level and Travel Time Index using Spearman's Rank Correlation.

 # Main Test: Spearman's Rank Correlation

Answer

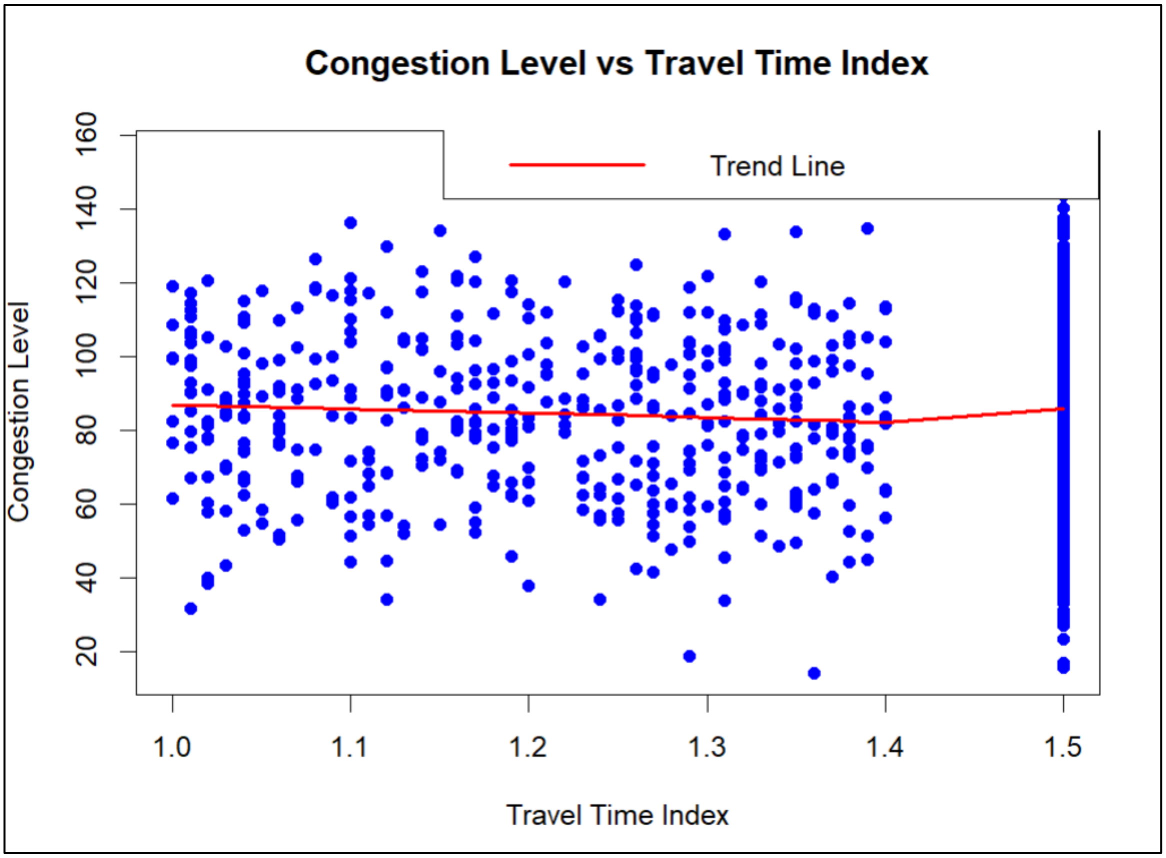
Let us set up the null hypothesis

H_0: There is no monotonic relationship between Congestion Level and Travel Time Index

Against the alternative hypothesis

H_1: There is a monotonic relationship between Congestion Level and Travel Time Index.

```
# Step 1: Perform Spearman's correlation test
correlation_result <- cor.test(data$Congestion_Level, data$Travel_Time_Index,
                               method = "spearman")

## Warning in cor.test.default(data$Congestion_Level, data$Travel_Time_Index, :
## Cannot compute exact p-value with ties

# Step 2: Display the results
cat("Spearman's Rank Correlation Coefficient:", correlation_result$estimate, "\n")

## Spearman's Rank Correlation Coefficient: 0.01747416

cat("P-value:", correlation_result$p.value, "\n")

## P-value: 0.5062786

# Step 3: Plot - Scatterplot with Lowess Trend Line
plot(data$Travel_Time_Index, data$Congestion_Level,
     main = "Congestion Level vs Travel Time Index",
     xlab = "Travel Time Index", ylab = "Congestion Level",
     pch = 19, col = "blue")
lines(lowess(data$Travel_Time_Index, data$Congestion_Level), col = "red", lwd = 2)
legend("topright", legend = c("Trend Line"), col = c("red"), lwd = 2)
```




```
# Handle warning: If ties are present, it's normal, no need to worry
if (correlation_result$p.value >= 0.05) {
  cat("Result: No significant monotonic relationship (p-value >= 0.05)
.\n")
} else {
  cat("Result: Significant monotonic relationship found (p-value < 0.0
5).\n")
}

## Result: No significant monotonic relationship (p-value >= 0.05).
```

Interpretation:

The Spearman's Rank Correlation test shows no significant evidence of a monotonic relationship between Congestion Level and Travel Time Index (p-value = 0.5063). The low Spearman's correlation coefficient (ρ = 0.0175) indicates a weak or non-existent monotonic association. The warning about ties is expected and does not affect the interpretation of the results.

Conclusion:

We fail to reject the null hypothesis (H_0) at the 0.05 significance level. This suggests that there is no significant monotonic relationship between Congestion Level and Travel Time Index in the data.

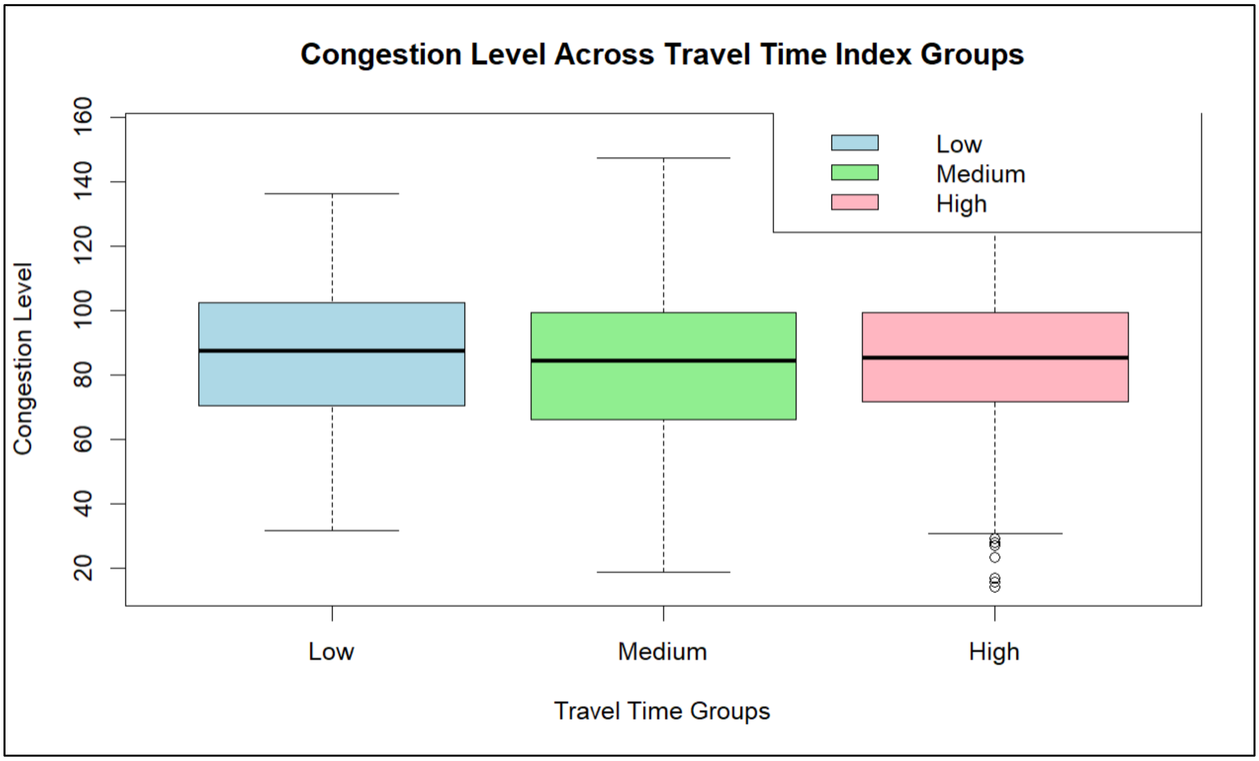
Graphical Interpretation:
The scatterplot illustrates the relationship between Congestion Level (y-axis) and Travel Time Index (x-axis). The data points are widely dispersed, showing no clear pattern or trend between the two variables. The red Lowess trend line is relatively flat, indicating a lack of meaningful monotonic relationship between Congestion Level and Travel Time Index. This visual aligns with the statistical results, where Spearman's correlation coefficient ($\rho = 0.0175$) and a high p-value (0.5063) confirm that the relationship is negligible and not statistically significant.

 # Objective 3: Check for equality of medians of Congestion Level across different ranges of Travel Time Index using the Kruskal-Wallis Test.

Answer
Let us set up the null hypothesis
H_0: The medians of Congestion Level are equal across the three Travel Time Index groups (Low, Medium, High).
Against the alternative hypothesis
H_1: At least one of the medians of Congestion Level is different among the three Travel Time Index groups.

```
# Step 1: Create groups based on Travel Time Index (3 groups: Low, Medium, High)
data$Travel_Time_Groups <- cut(data$Travel_Time_Index, breaks = 3,
                               labels = c("Low", "Medium", "High"))

# Step 2: Perform the Kruskal-Wallis test
kruskal_result <- kruskal.test(Congestion_Level ~ Travel_Time_Groups, data = data)

# Step 3: Display the results of the Kruskal-Wallis test
cat("Kruskal-Wallis Test Statistic:", kruskal_result$statistic, "\n")

## Kruskal-Wallis Test Statistic: 1.465099

cat("P-value:", kruskal_result$p.value, "\n")

## P-value: 0.480682

# Step 4: Interpretation based on p-value
if (kruskal_result$p.value < 0.05) {
  cat("Result: There is a significant difference in the medians of Congestion Level across Travel Time Index groups.\n")
```

```r
} else {
  cat("Result: No significant difference in the medians of Congestion Level a
cross Travel Time Index groups.\n")
}

## Result: No significant difference in the medians of Congestion Level acros
s Travel Time Index groups.

# Step 5: Plot - Boxplot to visualize Congestion Level across groups
boxplot(data$Congestion_Level ~ data$Travel_Time_Groups,
        main = "Congestion Level Across Travel Time Index Groups",
        xlab = "Travel Time Groups", ylab = "Congestion Level",
        col = c("lightblue", "lightgreen", "lightpink"), border = "black")

# Add a legend to the boxplot
legend("topright", legend = c("Low", "Medium", "High"),
       fill = c("lightblue", "lightgreen", "lightpink"), border = "black")
```



Interpretation:

The Kruskal-Wallis test results suggest that there is no statistically significant difference in the medians of Congestion Level among the three Travel Time Index groups (Low, Medium, High). The high p-value (0.4807) implies that the observed differences in the data are likely due to random variation rather than a meaningful relationship.

Conclusion:

We fail to reject the null hypothesis (H_0) at the 0.05 significance level. This indicates that the medians of Congestion Level do not differ significantly across the three Travel Time Index groups. The boxplot visualization supports this result, showing overlapping distributions of Congestion Level across the Low, Medium, and High groups.

Graphical Interpretation:

The boxplot shows the distribution of Congestion Level across the three Travel Time Index groups (Low, Medium, High) and reveals no significant differences in their medians, as the black median lines are close across all groups. The interquartile ranges (IQRs) and overall variability appear similar, with substantial overlap between the groups. A few outliers are present in the "High" group, but they do not indicate a meaningful trend. Overall, the lack of clear separation or trend among the groups visually supports the Kruskal-Wallis test results, confirming no significant differences in the medians of Congestion Level across the three Travel Time Index groups.

Final Conclusion:

The study analyzed the impact of environmental and situational factors on urban traffic in Bengaluru. Employing rigorous statistical methods, the findings revealed:

1. **Congestion Dynamics**: No significant difference in congestion levels was observed between major areas or weather conditions, suggesting uniform traffic behavior across scenarios.
2. **Weather and Traffic Volume**: Weather conditions had minimal impact on traffic volume and road capacity utilization, indicating their limited influence on traffic flow.
3. **Roadwork Effects**: Roadwork and construction activities did not significantly affect traffic volumes, highlighting potential inefficiencies in mitigation strategies.
4. **Public Transport Usage**: Weather conditions did not significantly alter public transport usage, suggesting consistent usage patterns irrespective of environmental changes.
5. **Interdependencies**: Limited correlations between variables like congestion and compliance levels, or environmental impacts and transport usage, underscore the need for holistic policy interventions.

These findings emphasize the need for data-driven urban planning strategies that address systemic inefficiencies while enhancing mobility and sustainability. Future studies should integrate broader datasets and explore advanced modeling techniques to refine these insights.

Future Work:

- **Expansion to Other Cities:** Apply the analysis framework to other metropolitan areas to compare traffic dynamics and develop generalized congestion management strategies.
- **Integration of Real-Time Data:** Incorporate real-time traffic and weather data to create predictive models for dynamic traffic management and forecasting.
- **Impact of Emerging Technologies:** Explore the influence of smart traffic systems, electric vehicles, and ride-sharing services on urban congestion and mobility.
- **Behavioural Analysis:** Study commuter behaviour under varying traffic conditions to design user-focused interventions and policies.

Literary References:

[1] Fundamentals of Mathematical Statistics Book By S.C. Gupta, VK Kapoor

[2] Fundamentals of Applied Statistics Book By S.C. Gupta, VK Kapoor

Dataset References:

[1] Kaggle Datasets

(Link: https://www.kaggle.com/datasets/preethamgouda/banglore-city-traffic-dataset)

[2] R Documentation

(Link: https://www.rdocumentation.org/)

[3] Open Government Data Portal Karnataka

(Link: https://karnataka.data.gov.in/)

[4] Open Government Data (OGD) Platform India

(Link: https://www.data.gov.in/)