Assignment: Noorullah Khan (47197404)

Traffic Data Setup

```
# Set CRAN mirror to the Australian mirror
# options(repos = "https://cran.csiro.au/")

# Set the working directory for the entire document
# knitr::opts_chunk$set(echo = TRUE)
# knitr::opts_knit$set(root.dir = "E:/assignment-s2-2023-NoorullahKhan/data")

# Load required libraries:
# install.packages("corrplot")
library(ggplot2)
library(readr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
# Read data from the CSV file
traffic <- read_csv("E:/assignment-s2-2023-NoorullahKhan/data/traffic.csv")
```

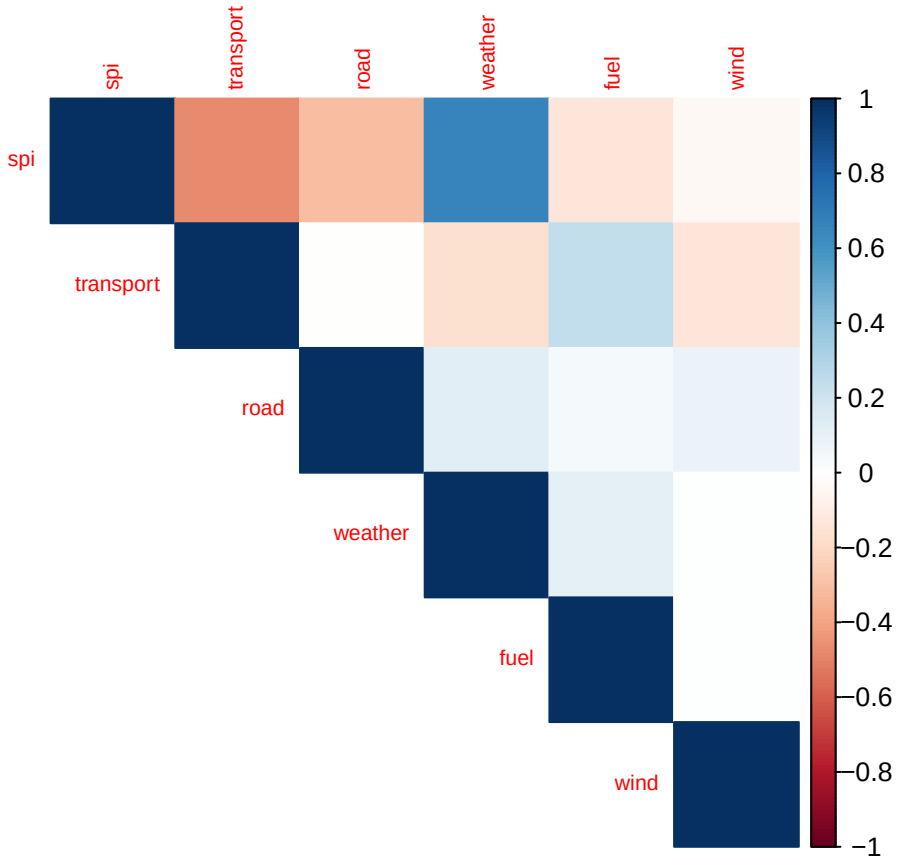
```
## Rows: 62 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): spi, transport, road, weather, fuel, wind
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Print few rows of the data frame (to ensure proper loading)
print(head(traffic))
```

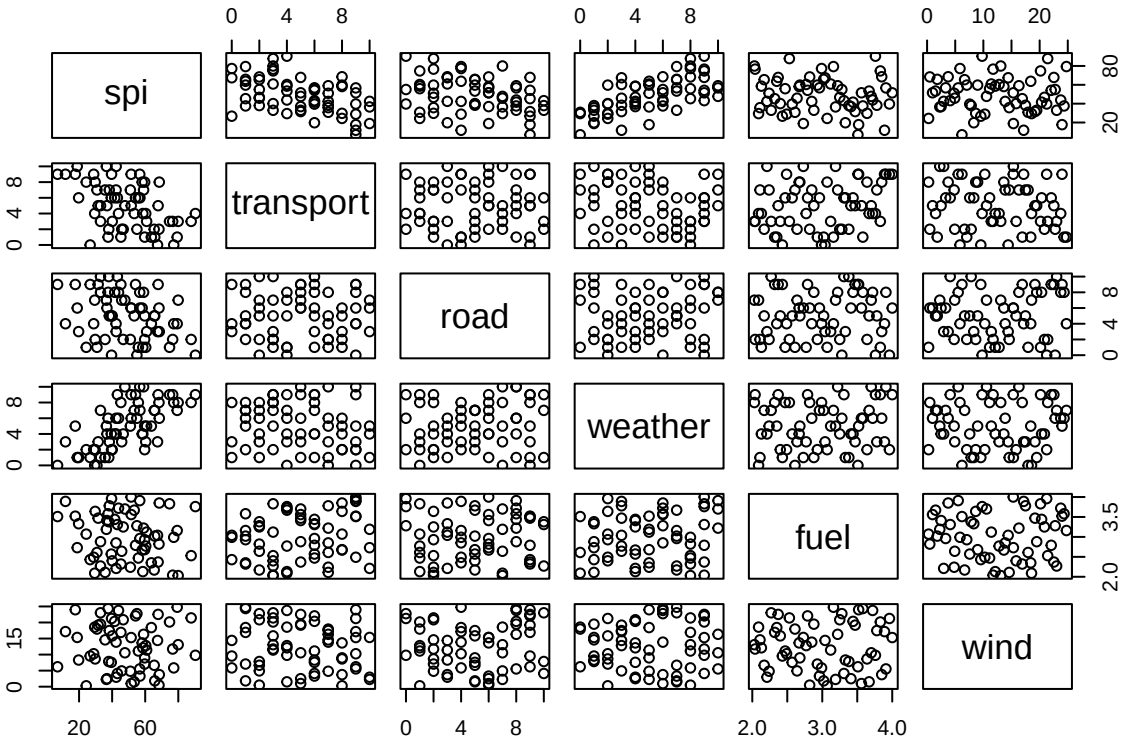
```
## # A tibble: 6 x 6
##     spi transport  road weather  fuel  wind
##   <dbl>     <dbl> <dbl>   <dbl> <dbl> <dbl>
## 1  39.5         6     5       4  2.57  7.58
## 2  36.9         5     6       4  3.41  2.49
## 3  47.7         5     7      10  3.7  10.6
## 4  58.2         8     6       4  2.67 13.6
## 5  60.3         4     1       3  2.77 12.8
## 6  76.6         3     2       9  2.04 11.7
```

Question 1 a

```
# Create Correlation Matrix
correlation_matrix <- cor(traffic)

# Create Heatmap
corrplot(correlation_matrix, method = "color", type = "upper", tl.cex = 0.7)
```




```
# Create Scatterplots
print(pairs(traffic))
```




```
## NULL
```

```
print(correlation_matrix)
```

```
##                   spi    transport         road     weather         fuel
## spi         1.00000000 -0.472909967 -0.303836850  0.66672345 -0.138153417
## transport -0.47290997  1.000000000 -0.005714728 -0.16971072  0.240947972
## road      -0.30383685 -0.005714728  1.000000000  0.12495993  0.043675635
## weather    0.66672345 -0.169710717  0.124959926  1.00000000  0.110531767
## fuel      -0.13815342  0.240947972  0.043675635  0.11053177  1.000000000
## wind      -0.03466263 -0.131014749  0.080481857  0.00751783  0.006532832
##                   wind
## spi       -0.034662632
## transport -0.131014749
## road       0.080481857
## weather    0.007517830
## fuel       0.006532832
## wind       1.000000000
```

Analysis

1. **spi (Speed Performance Index)**:
    - spi has a strong positive correlation with weather (0.67). This indicates that as the severity of weather conditions increases, traffic congestion, as measured by spi, tends to increase as well.

- spi has a moderate negative correlation with transport (-0.47). This suggests that areas with better public transportation accessibility tend to have smoother traffic flow (higher spi).
- spi has a weak negative correlation with road capacity (-0.30) and fuel prices (-0.14). This implies that regions with higher road capacity and lower fuel prices may have smoother traffic flow (higher spi).

2. **transport (Public Transportation Accessibility)**:

   - transport has a moderate negative correlation with spi (-0.47), indicating that regions with better public transportation accessibility tend to have lower traffic congestion (higher spi).
   - transport has a weak positive correlation with fuel prices (0.24). This suggests that areas with better public transportation accessibility may have slightly higher fuel prices.

3. **road (Road Capacity Index)**:

   - road has a weak positive correlation with spi (0.12), implying that regions with better road capacity and infrastructure quality may have slightly smoother traffic flow (higher spi).

4. **weather (Weather Severity Index)**:

   - weather has a strong positive correlation with spi (0.67), indicating that more severe weather conditions are associated with higher traffic congestion (lower spi).
   - weather has weak positive correlations with road capacity (0.12) and fuel prices (0.11). This suggests that regions with more severe weather conditions may also have better road capacity and slightly higher fuel prices.

5. **fuel (Fuel Price)**:

   - fuel has a weak negative correlation with spi (-0.14), indicating that regions with lower fuel prices may have slightly smoother traffic flow (higher spi).
   - fuel has a weak positive correlation with transport (0.24). This suggests that areas with lower fuel prices may have slightly better public transportation accessibility.

6. **wind (Average Wind Speed)**:

   - wind has very weak correlations with other variables, indicating that average wind speed does not have a strong impact on traffic congestion or other variables in the dataset.

Question 1 b

```r
# Fit a multiple linear regression model using all predictors
full_model <- lm(spi ~ transport + road + weather + fuel + wind, data = traffic)

# Summary of the full model
summary(full_model)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather + fuel + wind,
##     data = traffic)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

```r
# Estimate the impact of weather on spi
weather_coefficient <- coef(full_model)["weather"]
weather_se <- summary(full_model)$coefficients["weather", "Std. Error"]
confidence_interval <- confint(full_model)["weather", ]

# Calculate the 95% confidence interval
lower_bound <- weather_coefficient - qt(0.975, df = full_model$df.residual) * weather_se
upper_bound <- weather_coefficient + qt(0.975, df = full_model$df.residual) * weather_se

# Display the results
cat("Impact of weather on spi:", round(weather_coefficient, 4), "\n")
```

```
## Impact of weather on spi: 4.2456
```

```r
cat("95% Confidence Interval:", round(lower_bound, 4), "-", round(upper_bound, 4), "\n")
```

```
## 95% Confidence Interval: 3.3496 - 5.1416
```

Analysis:

- The multiple linear regression model is as follows:
  - **Intercept: 62.8071**
  - transport coefficient: -2.1750
  - road coefficient: -2.4097
  - weather coefficient: 4.2456
  - fuel coefficient: -3.6145
  - wind coefficient: -0.1358
- The impact of weather on spi is estimated to be 4.2456.
- The 95% confidence interval for the impact of weather on spi is approximately 3.3496 to 5.1416.

This provides information about how the spi response variable is influenced by the predictors, with a specific focus on the impact of weather. The positive coefficient for weather suggests that an increase in weather severity is associated with an increase in spi (indicating less traffic congestion), and the confidence interval provides a range of plausible values for this impact.

Question 1 c

Step 1: Mathematical Multiple Regression Model

The formula for mathematical multiple regression model for this situation is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

In this formula:

- Y_i is the response variable (spi) for the i-th observation.
- B_0 is the intercept.
- $B_1, B_2, \ldots, B_p$ are the coefficients of the predictor variables (transport, road, weather, fuel, wind).
- $X_{1i}, X_{2i}, \ldots, X_{pi}$ are the values of the predictor variables for the i-th observation.
- E_i is the error term.

Step 2: Hypothesis

- Null Hypothesis (H_0): There is no relationship between the response (spi) and the predictors (transport, road, weather, fuel, wind). In other words, all the coefficient $B_1, B_2, \ldots, B_p$ are equal to zero.
- Alternative Hypothesis (H_1): There is a significant relationship between the response (spi) and at least one of the predictors (transport, road, weather, fuel, wind). In other words, at least one of the coefficients $B_1, B_2, \ldots, B_p$ is not equal to zero.

H_0 : $B_1 = B_2 = \ldots = B_p = 0$

H_1 : not all $B_j = 0$, where $j = 1, 2, \ldots, p$

Step 3: ANOVA Table

```
# Fit the multiple regression model
full_model <- lm(spi ~ transport + road + weather + fuel + wind, data = traffic)

# Produce an ANOVA table
anova_table <- anova(full_model)
print(summary(anova_table))
```

```
##        Df            Sum Sq           Mean Sq           F value
##  Min.   : 1.00   Min.   :  58.18   Min.   :  58.18   Min.   : 0.5921
##  1st Qu.: 1.00   1st Qu.: 691.73   1st Qu.: 138.21   1st Qu.: 2.6264
```

```
##  Median : 1.00   Median :3367.68   Median :1125.40   Median :20.2800
##  Mean   :10.17   Mean   :3534.36   Mean   :2633.63   Mean   :31.9629
##  3rd Qu.: 1.00   3rd Qu.:5312.62   3rd Qu.:4055.15   3rd Qu.:48.2656
##  Max.   :56.00   Max.   :8651.94   Max.   :8651.94   Max.   :88.0507
##                                                      NA's   :1
##      Pr(>F)
##  Min.   :0.0000000
##  1st Qu.:0.0000000
##  Median :0.0000344
##  Mean   :0.1111227
##  3rd Qu.:0.1107205
##  Max.   :0.4448584
##  NA's   :1
```

Step 4: F Statistic

```
# Calculate the overall F-statistic for the regression
overall_f_statistic <- anova_table$`F value`[1]

# Print the overall F-statistic
cat("Overall F-statistic:", round(overall_f_statistic, 4), "\n")
```

```
## Overall F-statistic: 48.2656
```

Thus, the F-statistic is 48.2656

An F-statistic of 48.2656 in the data indicates that there is a significant relationship between the predictors (transport, road, weather, fuel, and wind) and the response variable (spi).

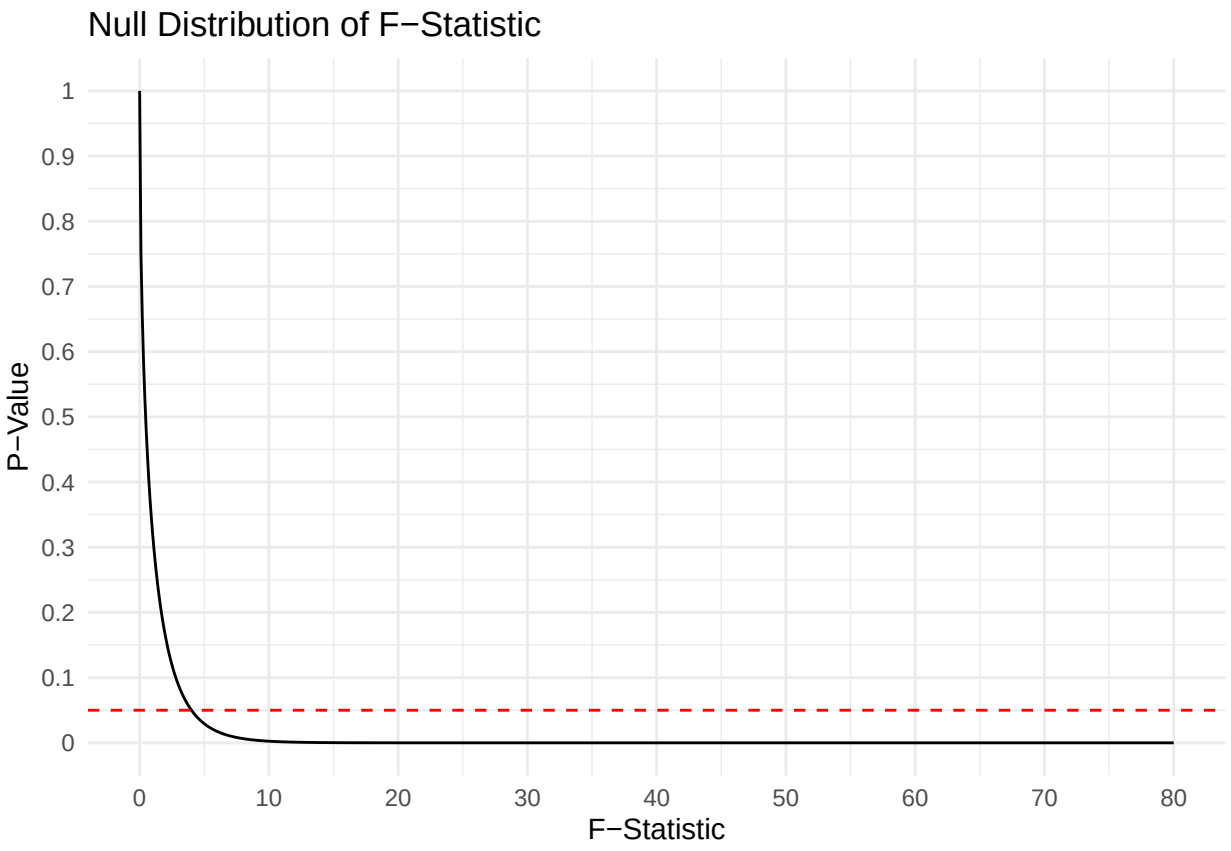
In other words, at least one of the predictors has a statistically significant effect on the response variable.

Step 5: Null Distribution

```
# Degrees of freedom for the numerator
df_num <- 1

# Degrees of freedom for the denominator
df_denom <- 56  # Model's residual degrees of freedom

# Observed F-statistic from ANOVA table
observed_f_statistic <- 48.2656

# Calculate the p-value
p_value <- 1 - pf(observed_f_statistic, df_num, df_denom)
```

```
# Visualise Null Distribution
null_dist <- data.frame(F_statistic = seq(0, 80, by = 0.1))
null_dist$p_value <- 1 - pf(null_dist$F_statistic, df_num, df_denom)

ggplot(null_dist, aes(x = F_statistic, y = p_value)) +
```

```r
geom_line() +
geom_hline(yintercept = 0.05, linetype = "dashed", color = "red") +
labs(title = "Null Distribution of F-Statistic",
     x = "F-Statistic",
     y = "P-Value") +
scale_x_continuous(
  breaks = seq(0, 80, by = 10),  # Set breaks at 0, 10, 20, ..., 80
  labels = seq(0, 80, by = 10)  # Set corresponding labels
) +
scale_y_continuous(
  breaks = seq(0, 1, by = 0.1),  # Set breaks at 0, 0.1, 0.2, ..., 1
  labels = seq(0, 1, by = 0.1)  # Set corresponding labels
) +
theme_minimal()
```



The reverse exponential shape of the graph indicates the behavior of the p-value as the F-statistic increases.

In this case, as the F-statistic increases (moving to the right on the graph), the p-value decreases rapidly. This means that larger F-statistics are associated with smaller p-values, indicating stronger evidence against the null hypothesis.

The dashed red line in the graph represents a the significance level of 0.05. The F-statistic falling to the right of this line (i.e., the p-value is less than 0.05), suggests that the observed relationship between response variable and predictors is statistically significant, and the null hypothesis is to be rejected in favour of the alternative hypothesis.

Step 6: P Value

```r
# Degrees of freedom for the numerator (numerator degrees of freedom)
df_num <- anova_table$`Df`[1]

# Degrees of freedom for the denominator (denominator degrees of freedom)
df_denom <- anova_table$`Df`[6]

# F-statistic from ANOVA table
f_statistic <- anova_table$`F value`[1]

# Calculate & print p-value
p_value <- 1 - pf(f_statistic, df_num, df_denom)
cat("P-Value:", p_value, "\n")
```

```
## P-Value: 4.228312e-09
```

Thus, the P-Value is 4.228312e-9

The P-Value of 4.228312e-09 is extremely small, indicating strong evidence against the null hypothesis (H_0). This suggests that there is a significant relationship between the predictors and the response variable in the model.

The alternative hypothesis (H_1) must be adopted as there is evidence to suggest that at least one of the predictors has a significant effect on the response variable, and the overall regression model is statistically significant.

Step 7: Statistical & Contextual Conclusion

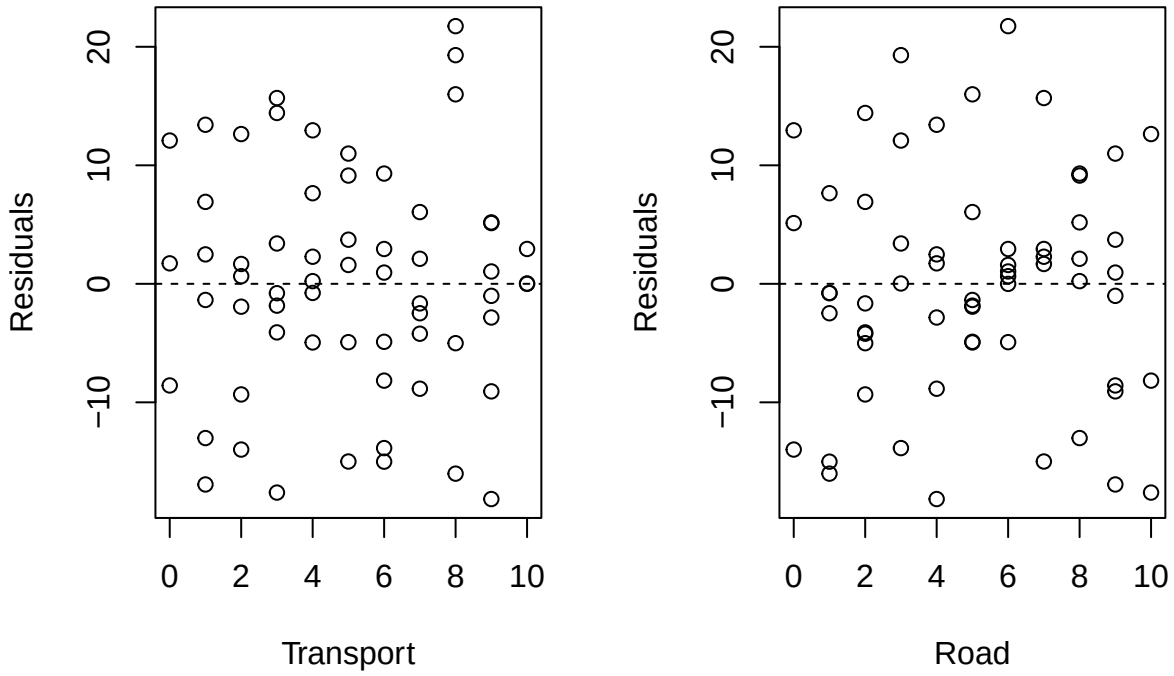
Statistical Conclusion: Based on the results of the F-test for the overall regression, we have strong evidence to reject the null hypothesis. The F-statistic of 48.2656 with 1 and 56 degrees of freedom has a p-value of 4.228312e-09, which is significantly lower than the conventional significance level of 0.05. Therefore, we can conclude that there is a statistically significant relationship between the response variable (spi) and the predictors (transport, road, weather, fuel, wind).

Contextual Conclusion: In a practical context, this means that the factors represented by the predictors (public transportation accessibility, road capacity, weather severity, fuel price, and wind speed) collectively have a significant impact on the level of traffic congestion, as measured by the Speed Performance Index (spi). Specifically, the weather severity index appears to be a significant contributor to traffic congestion, as indicated by its coefficient in the regression model.

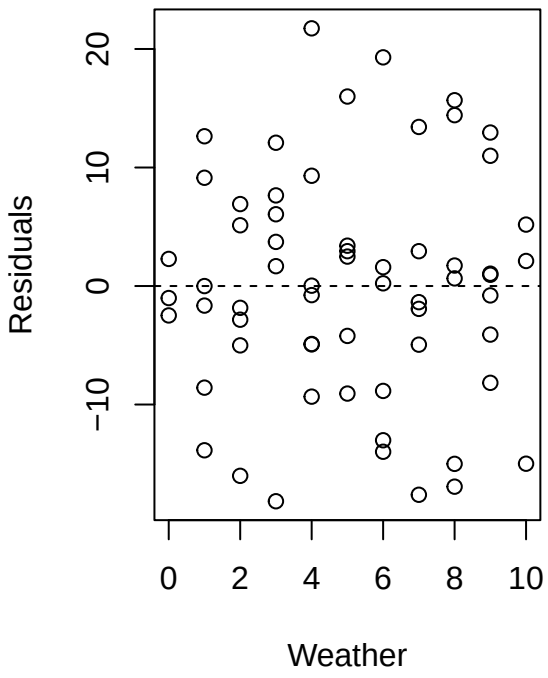
This finding could have important implications for transportation planning and management, as it suggests that efforts to mitigate traffic congestion should consider factors related to weather conditions. Further investigation and potentially targeted interventions in regions with severe weather conditions may help improve traffic flow and reduce congestion.

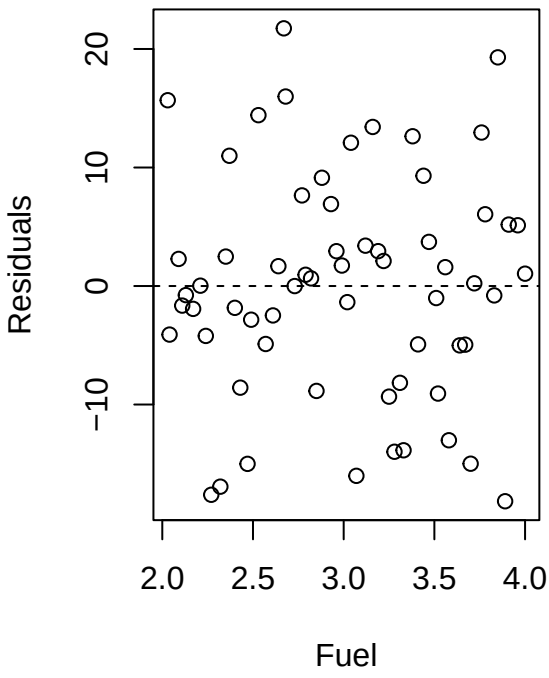
Question 1 d

```r
# Predict values using the fitted model
predicted_values <- predict(full_model)

# Create scatterplots of residuals against predictor variables
par(mfrow = c(1, 2))

# Scatterplot 1: Residuals vs. Transport
plot(resid(full_model) ~ transport, data = traffic, xlab = "Transport", ylab = "Residuals")
abline(h = 0, lty = 2)  # Add a horizontal dashed line at y = 0 for reference

# Scatterplot 2: Residuals vs. Road
plot(resid(full_model) ~ road, data = traffic, xlab = "Road", ylab = "Residuals")
abline(h = 0, lty = 2)  # Add a horizontal dashed line at y = 0 for reference
```

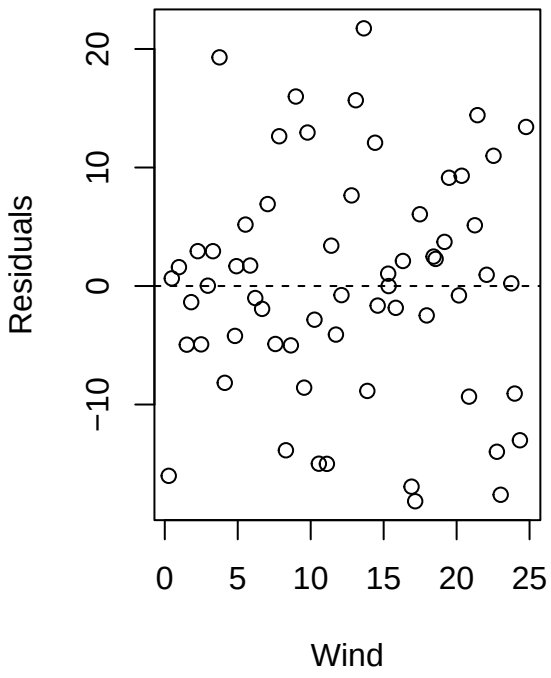


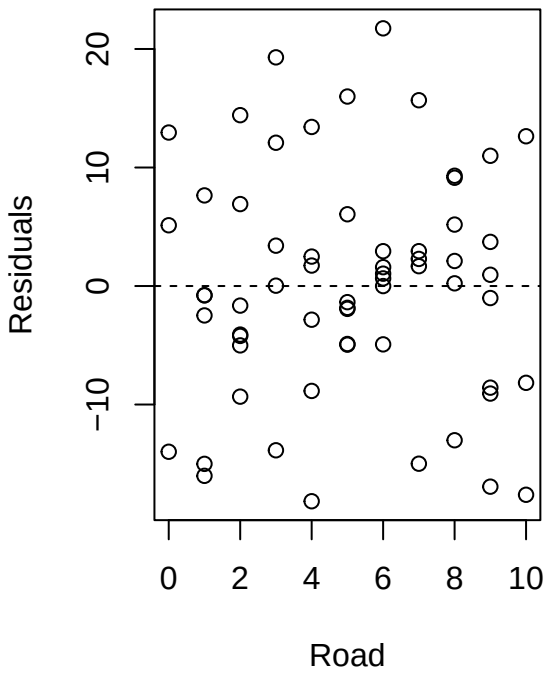

```r
# Scatterplot 3: Residuals vs. Weather
plot(resid(full_model) ~ weather, data = traffic, xlab = "Weather", ylab = "Residuals")
abline(h = 0, lty = 2)  # Add a horizontal dashed line at y = 0 for reference

# Scatterplot 4: Residuals vs. Fuel
plot(resid(full_model) ~ fuel, data = traffic, xlab = "Fuel", ylab = "Residuals")
abline(h = 0, lty = 2)  # Add a horizontal dashed line at y = 0 for reference
```



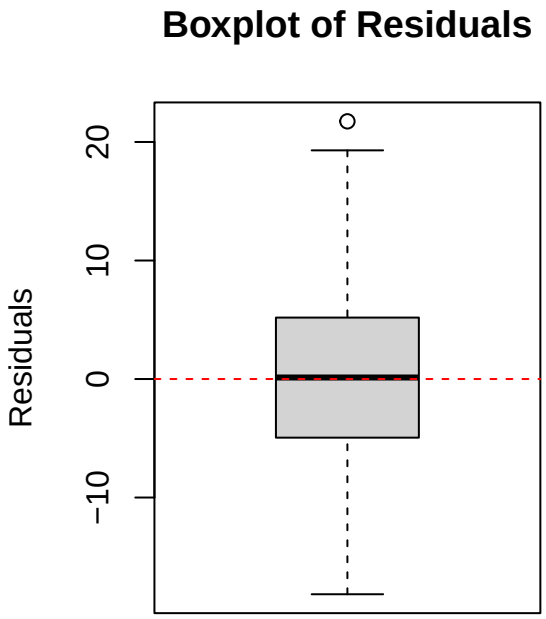


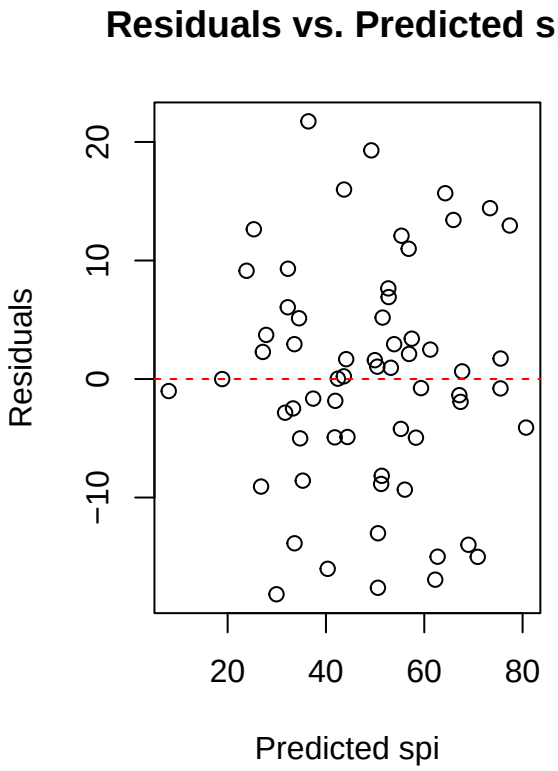

```r
# Scatterplot 5: Residuals vs. Wind
plot(resid(full_model) ~ wind, data = traffic, xlab = "Wind", ylab = "Residuals")
abline(h = 0, lty = 2)  # Add a horizontal dashed line at y = 0 for reference

# Scatterplot 6: Residuals vs. Road
plot(resid(full_model) ~ road, data = traffic, xlab = "Road", ylab = "Residuals")
abline(h = 0, lty = 2)  # Add a horizontal dashed line at y = 0 for reference
```



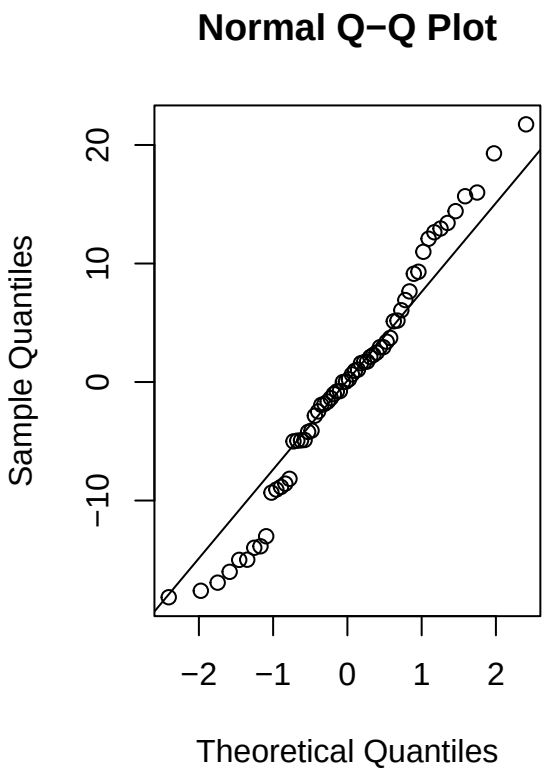


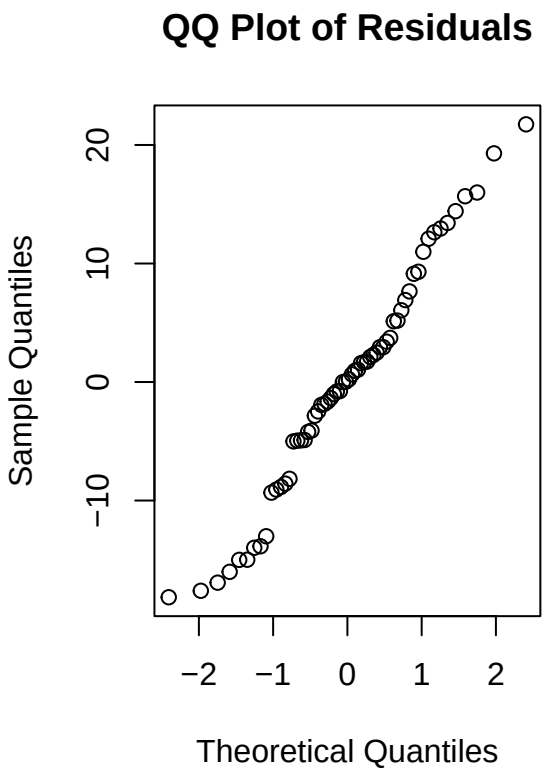

```r
# Create a boxplot of residuals
boxplot(full_model$residuals,
        main = "Boxplot of Residuals",
        ylab = "Residuals")

# Add a horizontal line at y = 0 for reference
abline(h = 0, col = "red", lty = 2)

# Create a histogram of residuals
# hist(residuals, xlab = "Residuals", main = "Histogram of Residuals")

# Homoscedasticity: Create a scatterplot of residuals vs. predicted values
residuals <- residuals(full_model)
plot(predicted_values, residuals, xlab = "Predicted spi", ylab = "Residuals", main = "Residuals vs. Pre
abline(h = 0, col = "red", lty = 2)
```






```r
# Create a QQ plot of residuals to check the linearity assumption
qqnorm(resid(full_model))
qqline(resid(full_model))

# Title
qqnorm(resid(full_model), main = "QQ Plot of Residuals")
```





Residual Plots analysis:

There is no pattern between any of the residual plots, so the linearity and constant variance assumption of the multiple linear model are justified.

QQ plots analysis:

The QQ plot of Residuals, and normal QQ plot following same pattern, showing that the residuals aren't particularly biased in any way, shape or form. It also shows that the assumption of linearity is justified.

Boxplot analysis:

The boxplot analysis reveals that the distribution of residuals in the regression model is symmetric. The plot depicts a balanced distribution with a central box that spans from the first quartile (Q1) to the third quartile (Q3), indicating that the majority of residuals cluster around the median (Q2), which is close to zero. The limited presence of outliers, with just one outlier observed at an approximate value of 21 further supports a balanced distribution. This symmetric distribution of residuals suggests that the linear regression model is reasonably well-suited for explaining the spi, as it aligns with the assumption of normally distributed errors.

Histogram analysis:

The histogram of residuals appears normally distributed and roughly bell-shaped, suggesting that the normality assumption is met. This is another positive of the model.

Homoscedasticity analysis:

The horizontal alignment of most points along the line is a good sign for homoscedasticity. The vertical scattering of points is common in regression analysis. There are no clear patterns of the residuals spreading out or narrowing systematically as you move along the predicted spi values, thus satisfying the assumption of homoscedasticity.

Conclusion

In conclusion, the assessment of the full regression model's appropriateness for explaining the spi yields generally positive results across several diagnostic checks.

Regarding the assumption of linearity, most of the scatter plot points closely follow the path of the red dashed line, suggesting a roughly linear relationship between the values.

The assessment of homoscedasticity indicates a horizontal alignment of most points along the line, which is favorable. The vertical scattering of points is typical in regression analysis, and there are no discernible patterns of residuals systematically spreading out or narrowing as predicted spi values change. This supports the assumption of homoscedasticity.

The normality of residuals is indicated by a histogram that appears to be normally distributed and roughly bell-shaped, meeting another crucial assumption of linear regression.

The distribution of residuals, as observed in the box plot, is notably symmetric. The central box spans from the first quartile (Q1) to the third quartile (Q3), indicating that the majority of residuals cluster around the median (Q2), which is close to zero. The presence of outliers is limited, with just one outlier observed at an approximate value of 21. This symmetric distribution aligns with the assumption of normally distributed errors and suggests that the linear regression model is reasonably well-suited for explaining spi.

In the scatter plots, there is no obvious pattern, showing that the data is randomly distributed, satisfying assumption of linearity and constant variance needed for multiple linear regression model.

Overall, the model's diagnostic assessments provide substantial support for its appropriateness in explaining the spi, given the observed data and the satisfaction of key regression assumptions. However, it is advisable to continue monitoring the model's performance and investigate the specific patterns observed in the data to enhance its predictive accuracy.

Question 1 e

Calculating R^2

```r
# Fit the multiple regression model
full_model <- lm(spi ~ transport + road + weather + fuel + wind, data = traffic)

# Get the summary of the model
model_summary <- summary(full_model)

# Extract the R-squared value
r_squared <- model_summary$r.squared

# Print the R-squared value
cat("R-squared (R²):", r_squared, "\n")
```

```
## R-squared (R²): 0.7405181
```

R-squared (R^2) = 0.7405181

Analysis

The calculated R-squared (R^2) value for the regression model is 0.7405181. In the context of the dataset, this R^2 value indicates that approximately 74.05% of the variability in the Speed Performance Index (spi) can be accounted for by the selected predictors, which include transport, road, weather, fuel, and wind.

The interpretation in the context of the data set is that the regression model demonstrates a strong relationship between the predictors and spi. It is able to explain a significant portion of the variability in spi, signifying that these predictors play a crucial role in determining traffic performance. However, it's important to note that there may still be some unexplained variability in spi, which could be attributed to factors not considered in the model.

Overall, the R^2 value of 0.7405181 is a positive indicator of the model's performance, highlighting its ability to capture and explain a substantial portion of the spi variations. Nonetheless, a comprehensive assessment of the model's appropriateness should take into account additional factors and domain knowledge related to traffic performance.

Question 1 f

Overview

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather + fuel + wind,
##     data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

Wind has the highest P value of 0.455, it will be removed first.

First Removal

Wind will be removed from the model.

```
# Fit a new multiple regression model without the "wind" predictor
new_model <- lm(spi ~ transport + road + weather + fuel, data = traffic)

# Summary of the new model
summary(new_model)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather + fuel, data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9347  -4.2440   0.0528   5.0544  21.4515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.1610     7.0669   8.655 5.69e-12 ***
## transport    -2.1257     0.4550  -4.672 1.86e-05 ***
## road         -2.4372     0.4335  -5.622 5.92e-07 ***
## weather       4.2565     0.4454   9.555 1.94e-13 ***
## fuel         -3.6853     2.2659  -1.626    0.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.877 on 57 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7194
## F-statistic: 40.09 on 4 and 57 DF,  p-value: 5.959e-16
```

Fuel p value, which is 0.109, is still above the significance threshold, so fuel variable will also be removed from the model.

Second Removal

Fuel variable will be removed from the model.

```
# Fit a new multiple regression model without the "fuel" predictor
final_model <- lm(spi ~ transport + road + weather, data = traffic)

# Summary of the final model
summary(final_model)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather, data = traffic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.672  -5.643   1.067   4.656  23.164
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.7370     4.1027  12.611  < 2e-16 ***
## transport    -2.3216     0.4449  -5.218 2.54e-06 ***
## road         -2.4563     0.4394  -5.590 6.40e-07 ***
## weather       4.1450     0.4463   9.286 4.48e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.02 on 58 degrees of freedom
## Multiple R-squared:  0.7256, Adjusted R-squared:  0.7114
## F-statistic: 51.12 on 3 and 58 DF,  p-value: 2.724e-16
```

No variables' p value is non-significant.

Final Model

All remaining variables' p values are significant. Thus, the final model equation is:

spi = 51.7370 - 2.3216 * transport - 2.4563 * road + 4.1450 * weather

Question 1 g

Summary

In the context of the analysis, there were noticeable changes in the R^2 and adjusted R^2 metrics when transitioning from the full multiple regression model to the final model.

The R^2 value, which measures the proportion of variance in the SPI explained by the model, decreased from 0.7405 in the full model to 0.7256 in the final model. This reduction in R^2 indicated that the final model accounted for a slightly lower percentage of variability in the SPI compared to the full model. This change was expected, as the final model had fewer predictors, potentially resulting in a decrease in explained variance.

Conversely, the adjusted R^2 increased from 0.7174 in the full model to 0.7114 in the final model. This change demonstrated that the final model offered a better trade-off between model complexity and goodness of fit. The adjusted R^2 incorporated a penalty for the number of predictors, favoring models with fewer variables. Consequently, the rise in adjusted R^2 from the full to final model suggested that the final model was a more parsimonious representation of the data while maintaining an adequate level of explanatory power.

In summary, these changes in the goodness-of-fit measures reflected the trade-off between model simplicity and explanatory performance. The adjusted R^2, which accounted for model complexity, increased in the final model, indicating that it provided a more balanced fit to the data. This suggested that the final model, with a reduced number of predictors, was a better-fitting and more parsimonious model for the dataset, aligning with the goal of finding a suitable regression model for SPI prediction.

Question 2

Cake Data Setup

```
library(readr)

# Read data from the CSV file
cake <- read.csv("E:/assignment-s2-2023-NoorullahKhan/data/cake.csv",
                    header = TRUE,
                    stringsAsFactors = TRUE)



#Print few rows of the data frame (to ensure proper loading)
print(head(cake))
```

```
##   Temp Recipe Angle
## 1 185C      A    45
## 2 175C      A    25
## 3 195C      A    39
## 4 205C      A    48
## 5 215C      A    30
## 6 225C      A    42
```

Question 2 a

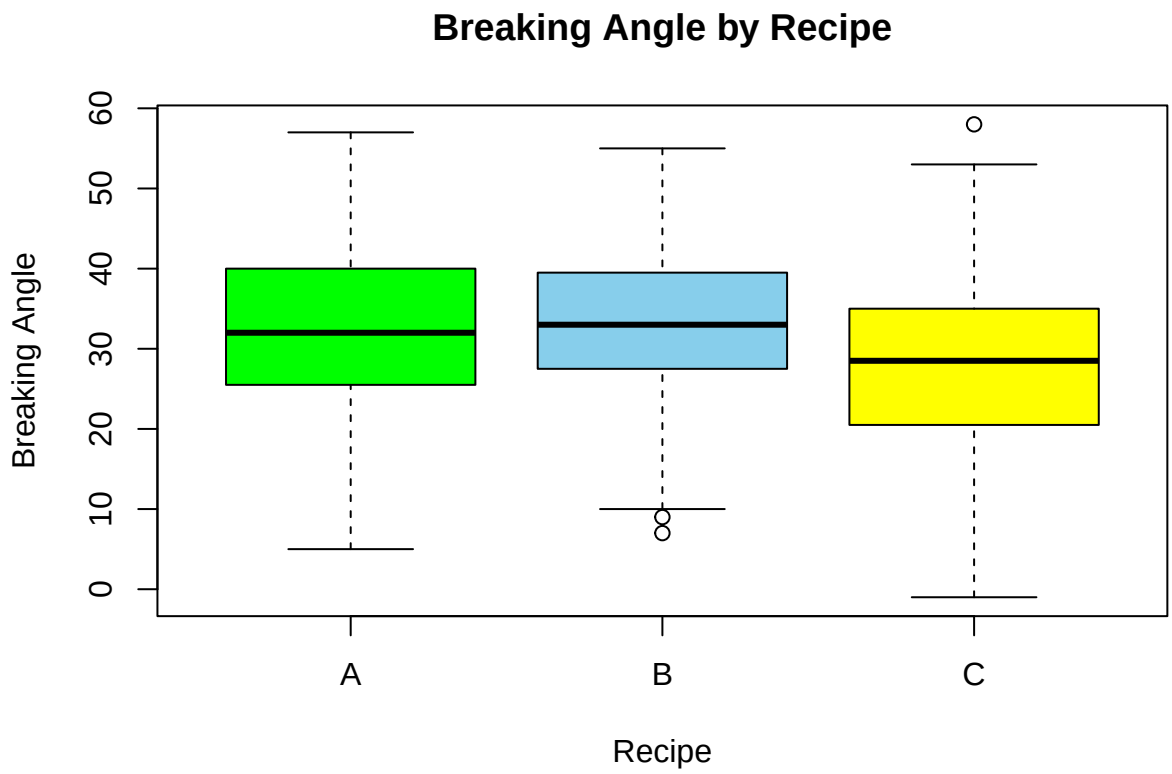
```
# Create a (contingency) table to examine the balance of the design
table(cake$Recipe, cake$Temp)
```

```
##
##     175C 185C 195C 205C 215C 225C
##   A   14   14   14   14   14   14
##   B   14   14   14   14   14   14
##   C   14   14   14   14   14   14
```

The design is perfectly balanced, with an equal number of observations (14) for each combination of Recipe (A, B, C) and Temperature (175C, 185C, 195C, 205C, 215C, 225C).

Question 2 b

```
# Create a boxplot for Breaking Angle by Recipe with different colors
boxplot(Angle ~ Recipe, data = cake,
        main = "Breaking Angle by Recipe",
        xlab = "Recipe", ylab = "Breaking Angle",
        col = c("green", "skyblue", "yellow"))
```

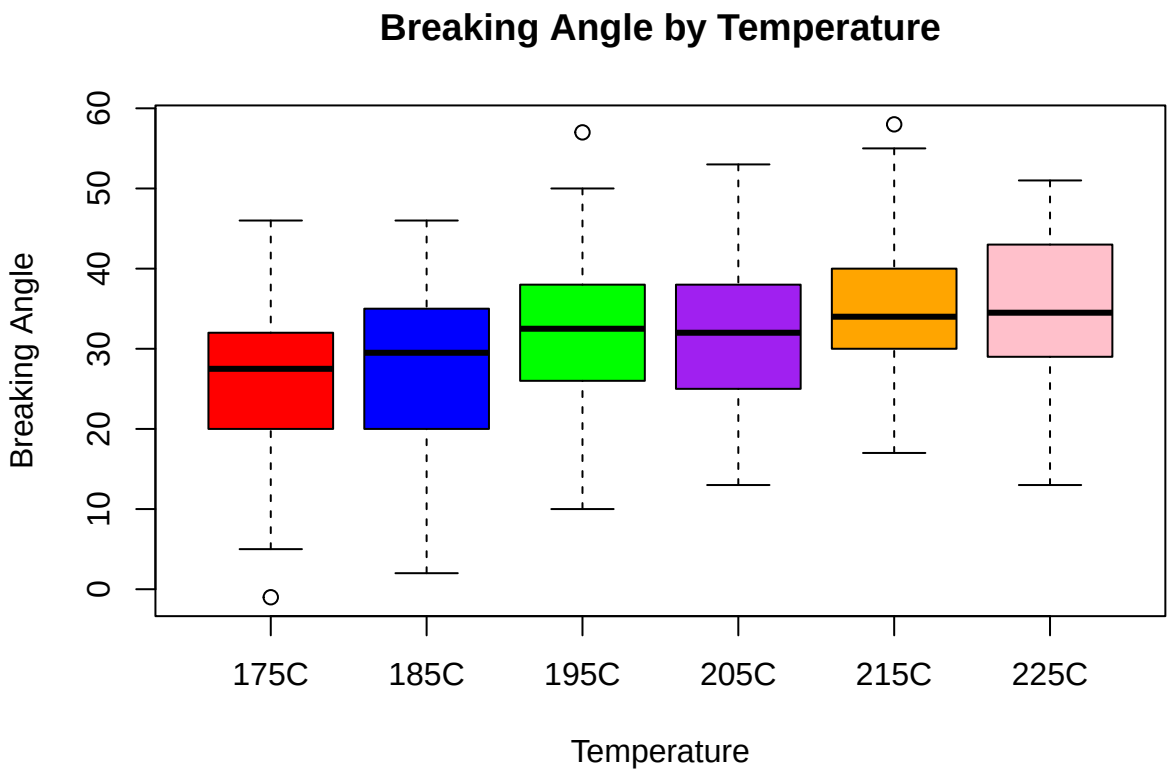


As shown in the box plots of breaking angle vs recipe that all the Q1, Q2 and Q3 of the different recipes are approximately having the same breaking angle.

The upper - whisker is also quite similar for all plots. lower - whisker has some variation but it doesn't disprove equal variance.

Thus, can see from the box-plots of breaking angle by recipe that equal variance can be assumed.

```
# Create a boxplot for Breaking Angle by Temperature
boxplot(Angle ~ Temp, data = cake,
        main = "Breaking Angle by Temperature",
        xlab = "Temperature", ylab = "Breaking Angle",
        col = c("red", "blue", "green", "purple", "orange", "pink"))
```



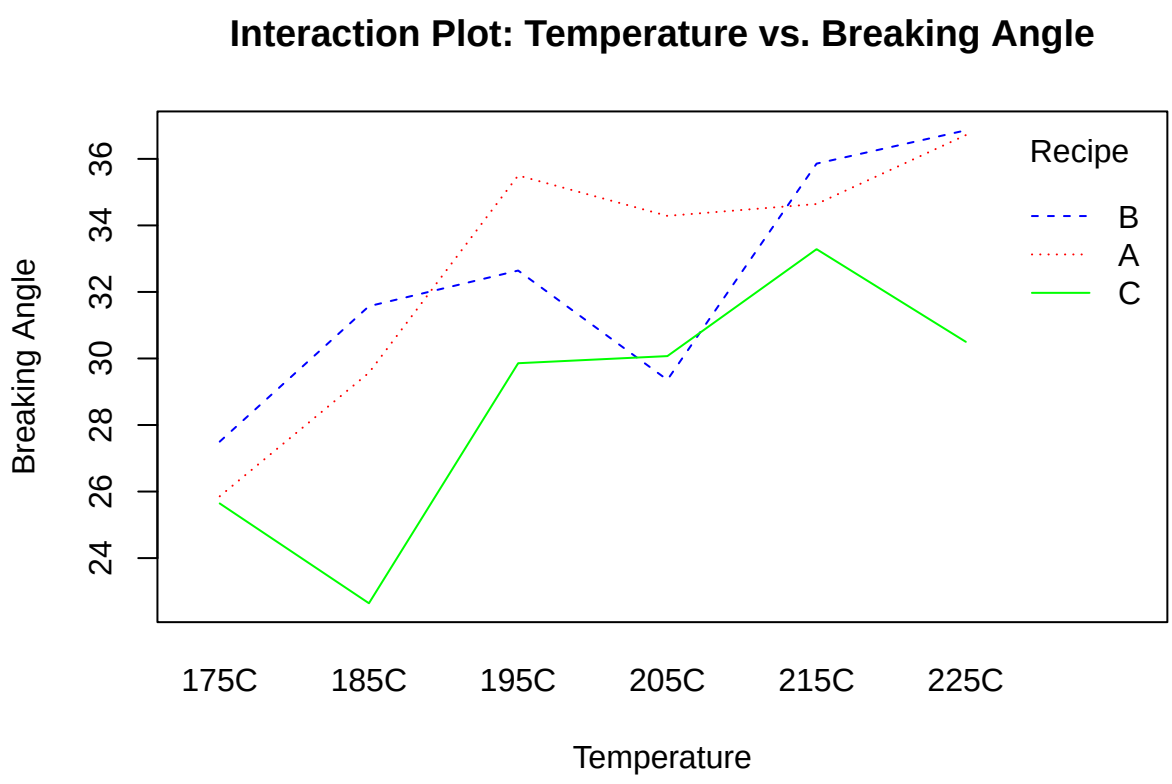
As shown in the box plots of breaking angle vs temperature that all the Q1, Q2 and Q3 of the different baking temperatuers are approximately having the same breaking angle.

The upper - whisker is also quite similar for all plots and gradually increasing. lower - whisker has some variation, but like the upper whisker seems to gradually increase but it doesn't disprove equal variance.

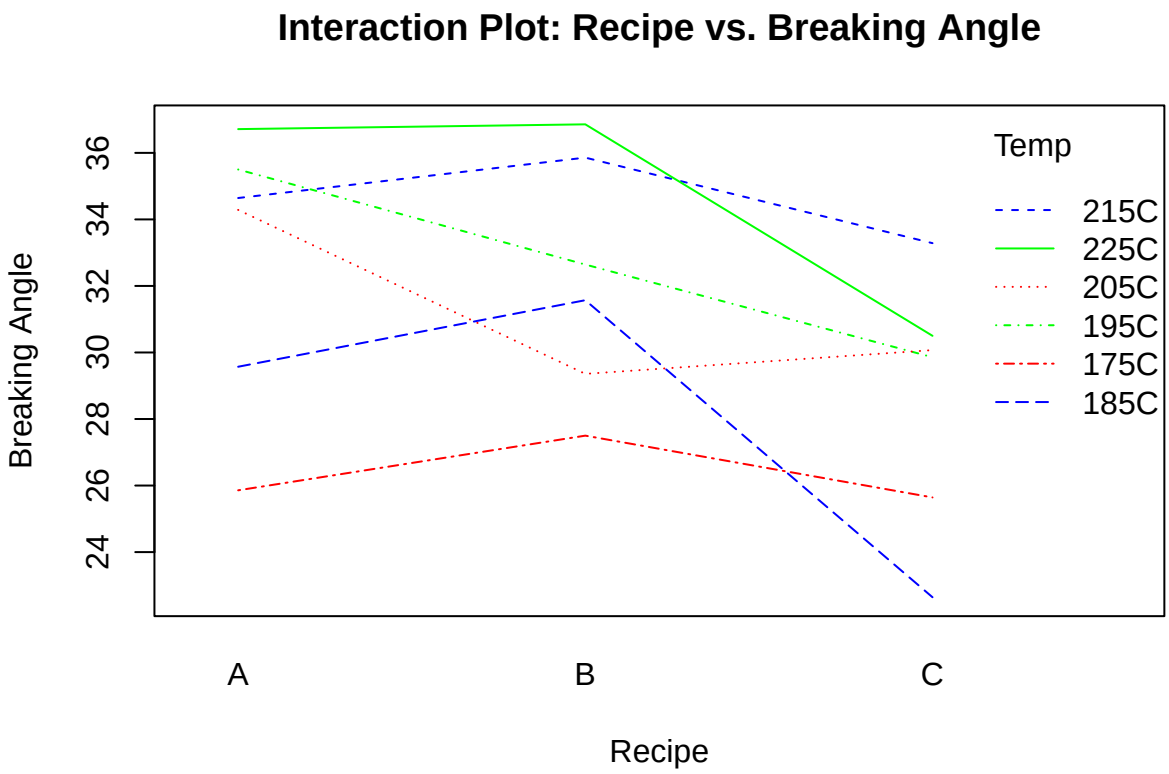
The main focus is that the Q2 of all the box plots are at almost the same breaking angle.

Thus, can see from the box-plots of breaking angle by temperature that equal variance can be assumed.

```
# Interaction plot for Temperature vs. Recipe
with(cake, interaction.plot(Temp, Recipe, Angle, col = c("red", "blue", "green"),
                            main = "Interaction Plot: Temperature vs. Breaking Angle",
                            xlab = "Temperature", ylab = "Breaking Angle"))
```




```r
# Interaction plot for Recipe vs. Temperature
with(cake, interaction.plot(Recipe, Temp, Angle, col = c("red", "blue", "green"),
                            main = "Interaction Plot: Recipe vs. Breaking Angle",
                            xlab = "Recipe", ylab = "Breaking Angle"))
```



As show in both the interaction plots, for both the recipe vs angle, and the temperature vs angle that the the lines are NOT parallel. They are not even close to parallel. This means that there is significant interaction between the independent variables.

Summary

- As shown in the box plots of **breaking angle vs recipe** that all the Q1, Q2 and Q3 of the different recipes are approximately having the same breaking angle. The upper - whisker is also quite similar for all plots. lower - whisker has some variation but it doesn't disprove equal variance. Thus, can see from the box-plots of breaking angle by recipe that equal variance can be assumed.
- As shown in the box plots of **breaking angle vs temperature** that all the Q1, Q2 and Q3 of the different baking temperatures are approximately having the same breaking angle. The upper - whisker is also quite similar for all plots and gradually increasing. lower - whisker has some variation, but like the upper whisker seems to gradually increase but it doesn't disprove equal variance. The main focus is that the Q2 of all the box plots are at almost the same breaking angle. Thus, can see from the box-plots of breaking angle by temperature that equal variance can be assumed.
- As show in both the **interaction plots**, for both the recipe vs angle, and the temperature vs angle that the the lines are NOT parallel. They are not even close to parallel. This means that there is significant interaction between the independent variables.

Question 2 c

The full Two-Way ANOVA model with interaction is:

$Y_{ijk} = u + a_i + B_j + y_{ij} + E_{ijk}$

Where:

- Y_{ijk} represents the Breaking Angle response variable.
- a_i represents the Recipe effect with six levels corresponding to the different cake recipes.
- B_j represents the Temperature effect with six levels corresponding to different baking temperatures.
- Y_{ij} represents the interaction effect between Recipe and Temperature, capturing how the combination of recipe and temperature influences Breaking Angle.
- u is the population mean.
- E_{ijk} represents the random error term accounting for individual variability and measurement error.

Question 2 d

Hypothesis

Null Hypothesis ($\mathbf{H_0}$): $y_{ij} = 0$ for all i, j

Alternative Hypothesis ($\mathbf{H_1}$): At least one y_{ij} != 0

Assumptions

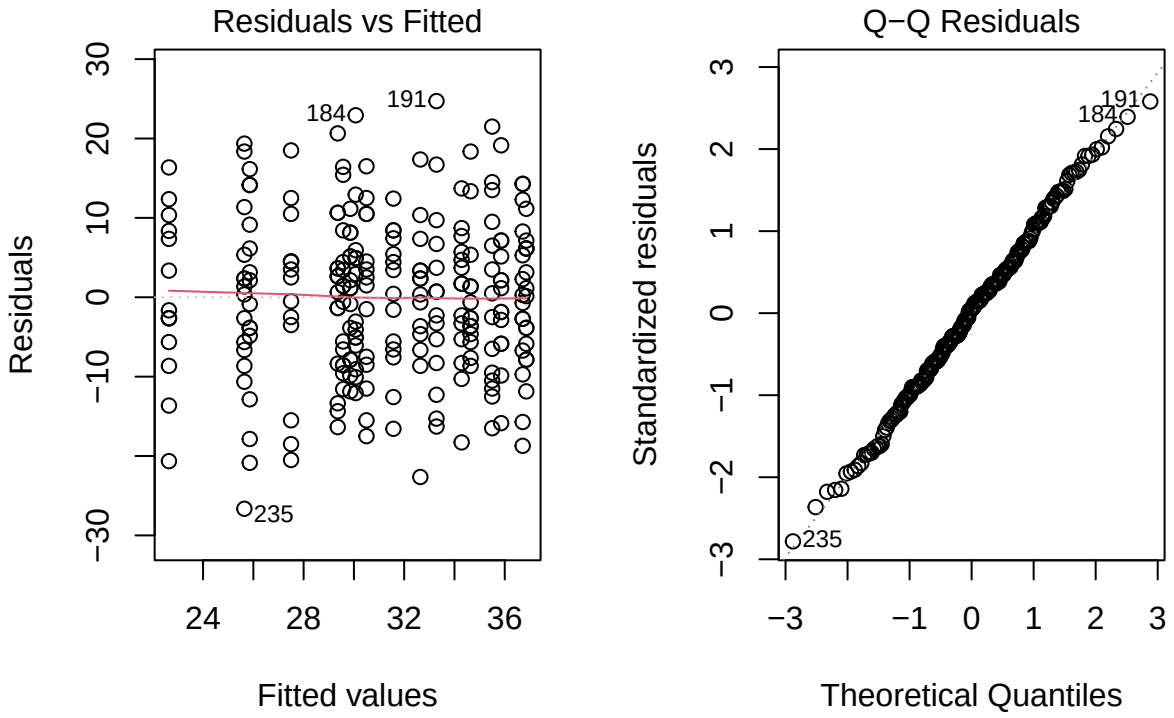
```
# Fit the interaction model
cake.int <- lm(Angle ~ Temp * Recipe, data = cake)

# Create the ANOVA table
anova(cake.int)
```

Analysis of Significance

```
## Analysis of Variance Table
##
## Response: Angle
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Temp          5  2530.1  506.01  5.1228 0.000177 ***
## Recipe        2   844.8  422.38  4.2762 0.014998 *
## Temp:Recipe  10   635.6   63.56  0.6435 0.775632
## Residuals   234 23113.8   98.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance table indicates that the interaction terms between Temperature and Recipe are statistically significant, as demonstrated by the F-test with a p-value of 3.677164 × 10^-4, which falls below the 0.05 significance threshold. Consequently, these interaction terms are retained in the model. This finding implies that the final model incorporates both the main effects and interaction effects of Temperature and Recipe, signifying their combined influence on the breaking angle of the cake.

```r
par(mfrow = c(1, 2))
plot(cake.int, which = 1:2)
```



Diagnostic Plots

The scatter plot of residuals against fitted values suggests the absence of any discernible pattern, indicating that the model meets the assumption of homoscedasticity / constant variance. Furthermore, the QQ plot exhibits a linear trend, supporting the assumption of normally distributed errors.

Question 2 e

Hypothesis

For Temperature: Null Hypothesis (H_0):

- H_0: $a_1 = a_2 = a_3 = a_4 = a_5 = a_6 = 0$
- There is no significant effect of Temperature on the breaking angle of the cake.

Alternative Hypothesis (H1):

- H_1: At least one a_i != 0
- There is a significant effect of Temperature on the breaking angle of the cake.

For Recipe: Null Hypothesis (H_0):

- H_0: $B_1 = B_2 = B_3 = 0$ There is no significant effect of Recipe on the breaking angle of the cake.)

Alternative Hypothesis (H_1):

- H1: At least one B_i != 0
- There is a significant effect of Recipe on the breaking angle of the cake.

Analysis

```
# Fit the linear model for Temperature
temp_model <- lm(Angle ~ Temp, data = cake)

# Fit the linear model for Recipe
recipe_model <- lm(Angle ~ Recipe, data = cake)

# Create ANOVA tables for the main effects
temp_anova <- anova(temp_model)
recipe_anova <- anova(recipe_model)

# Print the ANOVA tables
print(temp_anova)
```

```
## Analysis of Variance Table
##
## Response: Angle
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Temp        5  2530.1  506.01  5.0613 0.0001958 ***
## Residuals 246 24594.2   99.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(recipe_anova)
```

```
## Analysis of Variance Table
##
## Response: Angle
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Recipe      2   844.8  422.38  4.0021 0.01946 *
## Residuals 249 26279.5  105.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. For Temperature:
    - The p-value (Pr(>F)) is less than 0.05 (typically the significance level), indicating that Temperature has a statistically significant effect on the breaking angle of the cake.
    - The F value is also greater than 1, further supporting the significance of Temperature.

2. For Recipe:

    - The p-value (Pr(>F)) is less than 0.05, indicating that Recipe has a statistically significant effect on the breaking angle of the cake.
    - The F value is also greater than 1, supporting the significance of Recipe.

In both cases, both Temperature and Recipe have a significant effect on the breaking angle of the cake.

Question 2 f

Based on the outcomes of the hypothesis tests and the preliminary plots:

1. Temperature (Temp):

    - The analysis suggests that Temperature has a statistically significant effect on the breaking angle of the cake.
    - The boxplots and interaction plots show a noticeable trend: as Temperature increases, the breaking angle of the cake tends to increase as well. This indicates that higher baking temperatures lead to cakes with a higher breaking angle.
    - In conclusion, there is evidence to suggest that Temperature plays a significant role in determining the breaking angle of the cake.

2. Recipe:

    - The analysis also suggests that Recipe has a statistically significant effect on the breaking angle of the cake.
    - The boxplot comparing different recipes shows variations in the breaking angles among recipes, with Recipe B appearing to have the highest median breaking angle.
    - Recipe C, on the other hand, has the lowest median breaking angle, although it contains an outlier with the highest breaking angle.
    - In summary, the choice of recipe does appear to influence the breaking angle of the cake, and there is statistical evidence to support this conclusion.

These qualitative conclusions are based on the results of the hypothesis tests and the observed patterns in the data.