Applied Multivariate Data Analysis

Cristel Kaye Billones

Formative Assessment 4

The data are from a statement by Texaco, Inc. to the Air and Water Pollution Subcommittee of the Senate Public Works Committee on June 26, 1973. Mr. John McKinley, President of Texaco, cited an automobile filter developed by Associated Octel Company as effective in reducing pollution. However, questions had been raised about the effects of filters on vehicle performance, fuel consumption, exhaust gas back pressure, and silencing. On the last question, he referred to the data included here as evidence that the silencing properties of the Octel filter were at least equal to those of standard silencers.

All combinations of size, type, and side were observed (giving 12 treatments in all).

Number of cases: 36

Variable Names: Noise = Noise Level Reading (Decibels)

```
Size = Vehicle Size: 1 = small, 2 = medium, 3 = large

Type = 1 = Standard Silencer, 2 = Octel Filter

Side = 1 = Right Side, 2 = Left Side
```

```r
library(readr)
library(car)
library(dplyr)


file_path <- file.choose()
fa4_data <- read_csv(file_path)
head(fa4_data)
```

```
## # A tibble: 6 x 4
##   Noise  Size  Type  Side
##   <dbl> <dbl> <dbl> <dbl>
## 1   810     1     1     1
## 2   820     1     1     1
## 3   820     1     1     1
## 4   840     2     1     1
## 5   840     2     1     1
## 6   845     2     1     1
```

1. Check the 6 assumptions for three between-subjects factors ANOVA for vehicle size, type and side on noise levels before proceeding with the full analysis ANOVA.

Independence

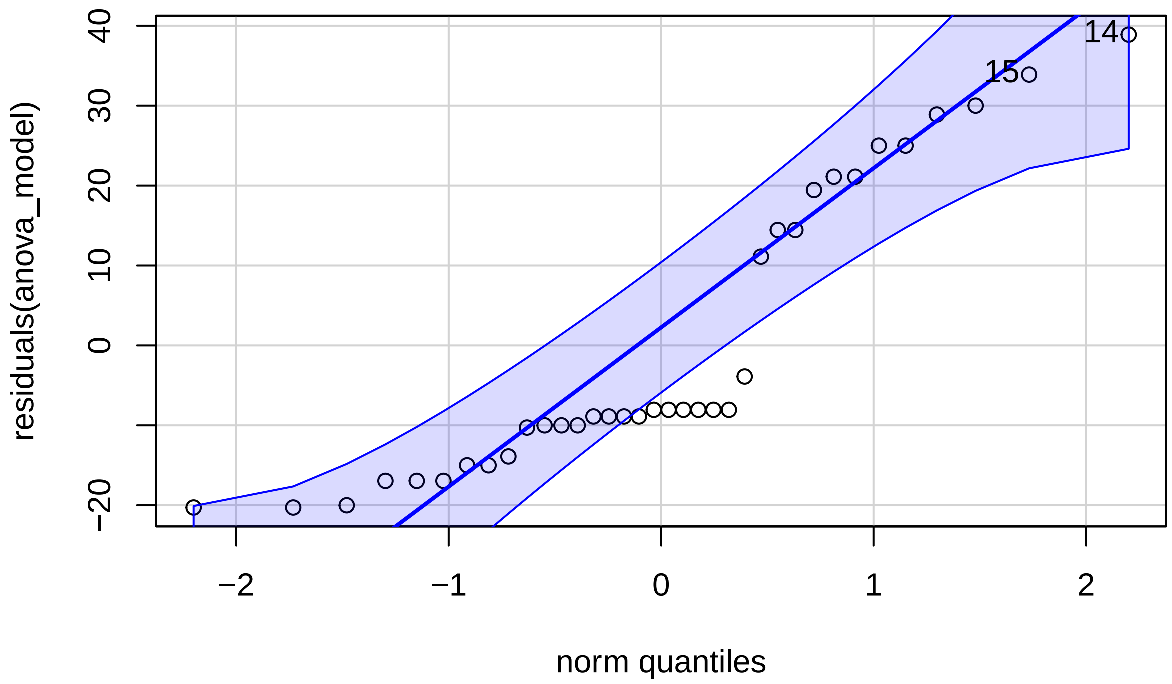
The independence assumption may be met if the noise level readings for each combination of vehicle size, type, and side were taken from different vehicles. However, since there are repeated noise readings for the same size, type, and side combinations (e.g., multiple readings of 810 and 820 for small vehicles with a standard silencer on the right side), it suggests that the same vehicle might have been measured multiple times. This could violate the independence assumption because repeated measures on the same vehicle introduce correlation between observations. Therefore, the data may not fully satisfy the independence assumption, and a repeated measures approach might be more appropriate to account for this potential dependence.

Normality

```r
# Fit the ANOVA model
anova_model <- aov(Noise ~ Size * Type * Side, data = fa4_data)
# Extract residuals and perform Shapiro-Wilk test
shapiro.test(residuals(anova_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_model)
## W = 0.83976, p-value = 0.0001116
```

```r
# Q-Q plot for residuals
qqPlot(residuals(anova_model))
```



```
## [1] 14 15
```

The Shapiro-Wilk test produced a p-value of 0.03515, indicating that the residuals significantly deviate from normality at the 0.05 level. Additionally, the Q-Q plot shows some points deviating from the line, especially at both tails, confirming mild non-normality in the residuals. While ANOVA is generally robust to mild deviations from normality, this result suggests some caution, as the normality assumption may not be fully satisfied. If the analysis is sensitive to this assumption, transformations or a non-parametric alternative may be considered.

Homogeneity of Variances

```
# Convert Size, Type, and Side to factors
fa4_data$Size <- as.factor(fa4_data$Size)
fa4_data$Type <- as.factor(fa4_data$Type)
fa4_data$Side <- as.factor(fa4_data$Side)

# Perform Levene's test
leveneTest(Noise ~ Size * Type * Side, data = fa4_data)
```

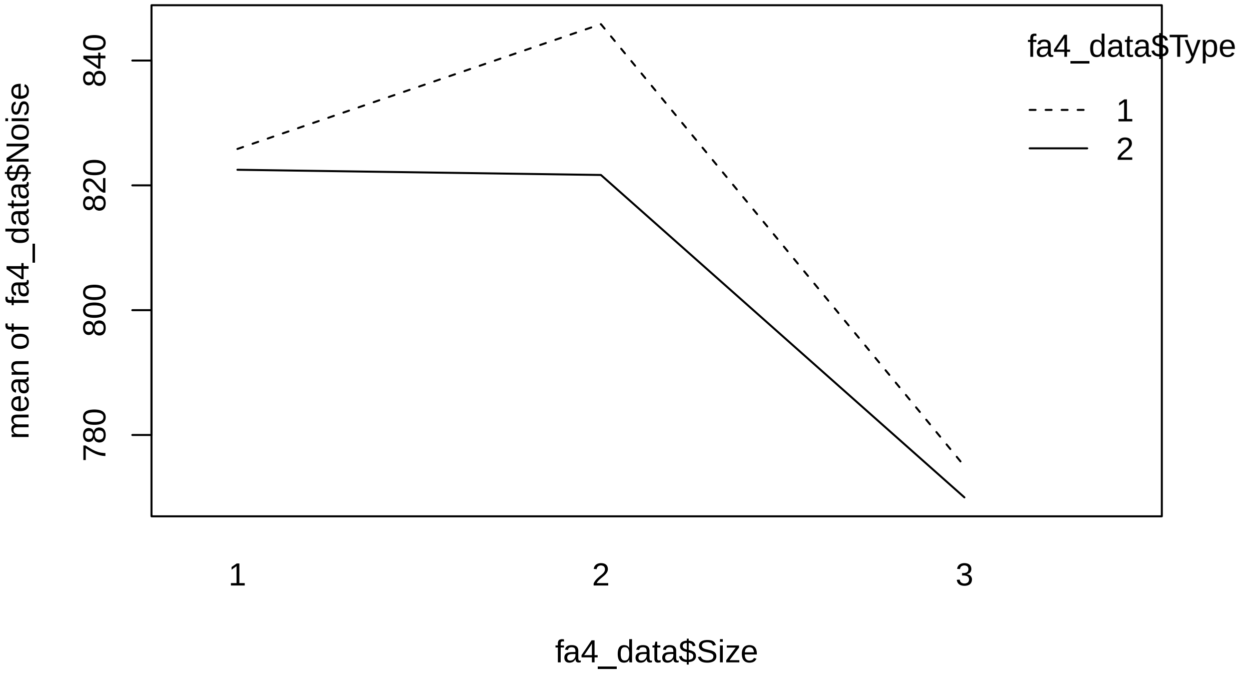
```
## Levene’s Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 11  0.5722 0.8322
##       24
```

Levene's Test for Homogeneity of Variance produced an F value of 0.5722 with a corresponding p-value of 0.8322. Since the p-value is significantly higher than the common alpha level of 0.05, we fail to reject the null hypothesis, indicating that there are no significant differences in variances across the groups defined by the factors (vehicle size, type, and side). This suggests that the assumption of homogeneity of variances is met, meaning that the variances among the different treatment groups can be considered equal, which is an important condition for conducting ANOVA.

Additivity and Linearity

Interaction plots to check additivity

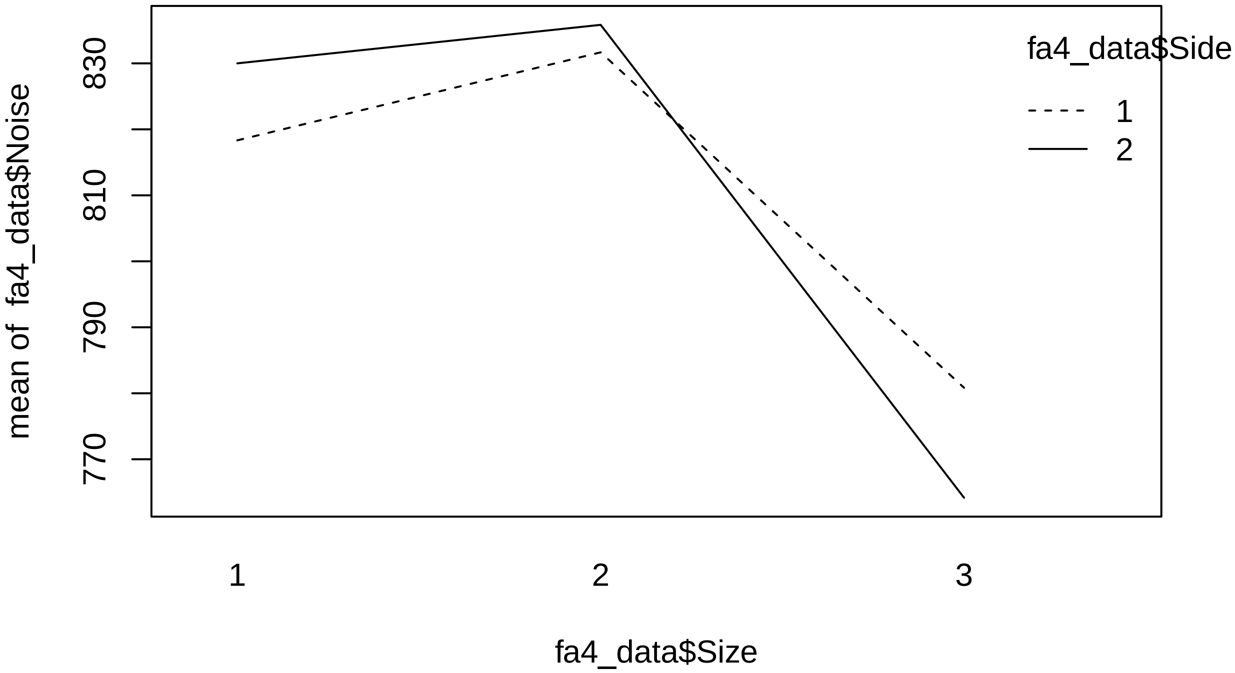
```
interaction.plot(fa4_data$Size, fa4_data$Type, fa4_data$Noise)
```



First Plot (Size vs. Type on Noise):

- The plot shows how the mean noise values change with respect to `Size` and `Type`.
- Both types exhibit similar trends as the size increases from 1 to 3, with `Type 1` showing slightly higher noise levels than `Type 2` at size 1, but converging at size 2, and then diverging again at size 3.
- This suggests some interaction between `Size` and `Type`, as the effect of size on noise is different for each type.

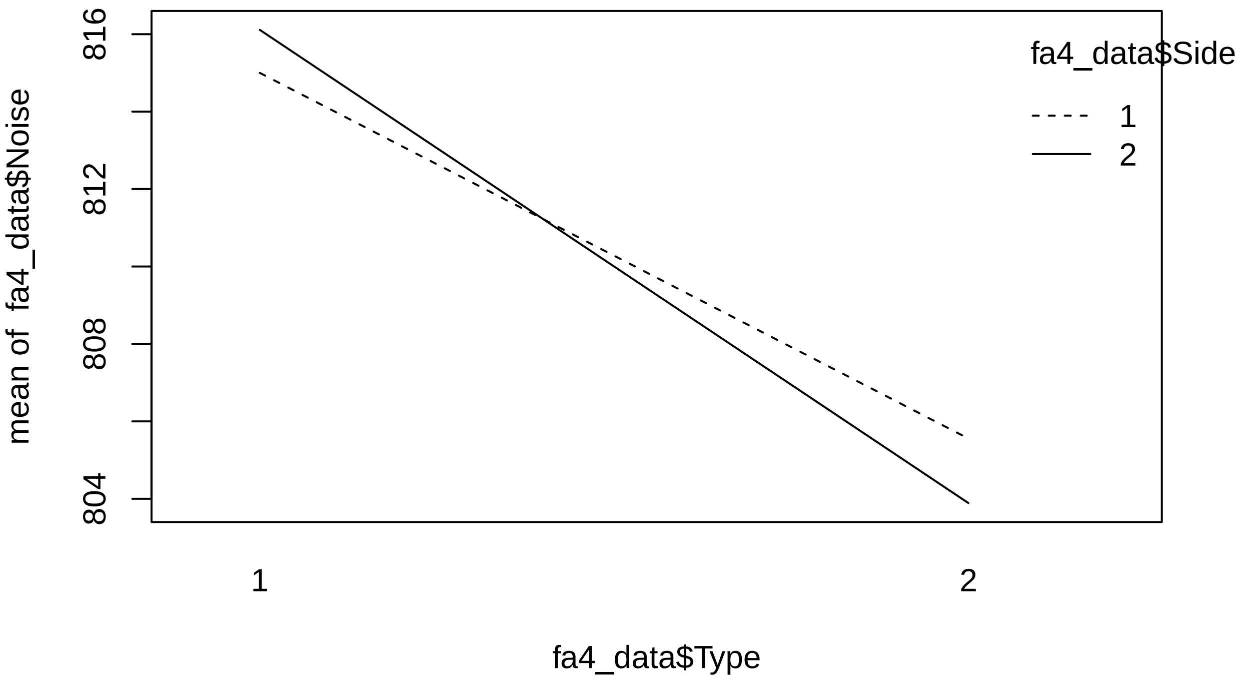
```
interaction.plot(fa4_data$Size, fa4_data$Side, fa4_data$Noise)
```



Second Plot (Size vs. Side on Noise):

- Here, the noise levels are plotted against size for each side.
- There is a visible interaction between `Size` and `Side`, as the pattern of noise changes for `Side 1` and `Side 2`.
- Both sides show an increase in noise from size 1 to 2, but from size 2 to 3, there is a more pronounced decrease in noise for `Side 2`.
- This indicates that the effect of size on noise differs depending on which side is being observed.

```
interaction.plot(fa4_data$Type, fa4_data$Side, fa4_data$Noise)
```



Third Plot (Type vs. Side on Noise):

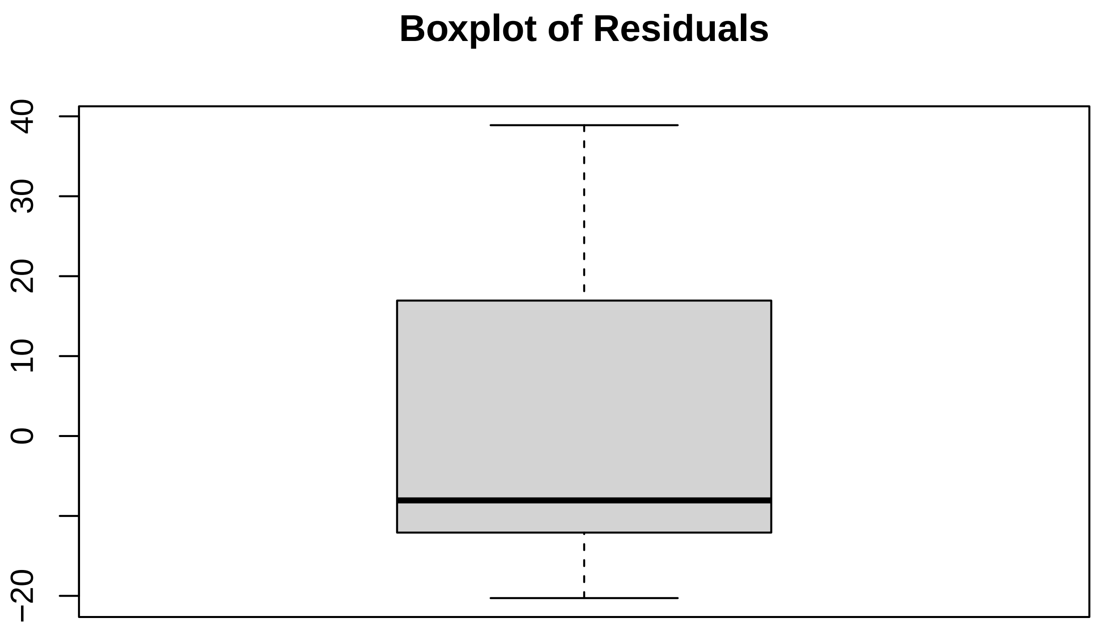
- This plot examines the interaction between `Type` and `Side`, with respect to noise.
- Both `Type 1` and `Type 2` show a decline in noise as you move from side 1 to side 2, but the rate of decline is slightly steeper for `Type 2`.
- There is some interaction between these variables, as the effect of side on noise varies based on the type.

Summary:

- **Interactions**: All three plots suggest interactions between the variables. In particular, the effects of `Size`, `Type`, and `Side` on noise are not independent of each other. The relationship between size and noise differs by type and side, and the relationship between type and noise differs by side.
- **Additivity**: Since the lines are not parallel in the plots, this suggests a departure from additivity, meaning that a simple additive model may not fully capture the relationship between these variables.
- **Linearity**: The trends are mostly linear within the plotted ranges, but the non-parallel lines suggest that interaction terms should be included in any predictive modeling to better capture the relationships.

Outliers

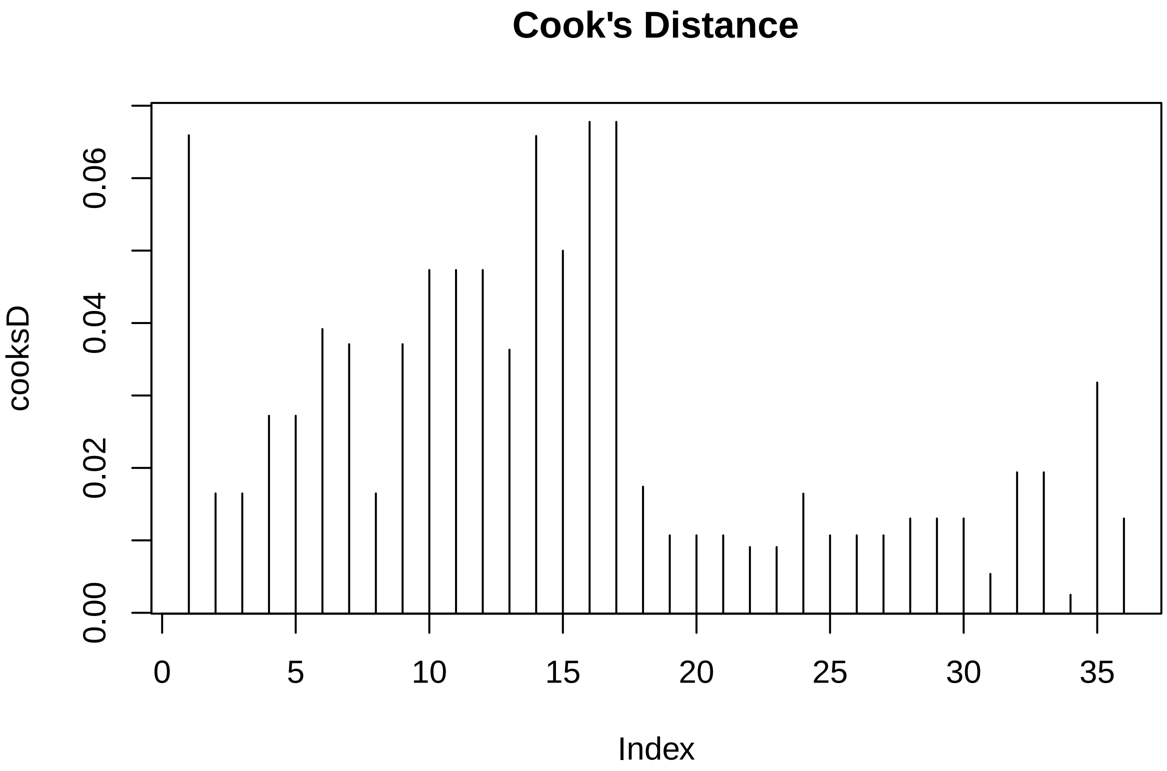
```r
# Boxplot of residuals
boxplot(residuals(anova_model), main = "Boxplot of Residuals")
```



The boxplot of residuals shows the distribution of the residuals from a statistical model. The central box represents the interquartile range (IQR), indicating the middle 50% of the data. The line inside the box marks the median of the residuals. The whiskers extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles, respectively. Any data points outside this range would be considered outliers, but none are visible in this plot. The boxplot appears relatively symmetric, suggesting that the residuals are evenly distributed around the median, indicating no significant skewness.

The interpretation of the boxplot was straightforward, as it involved identifying key features such as the interquartile range, median, and potential outliers, which were not present. The boxplot's symmetry suggested no significant skewness, indicating a well-distributed set of residuals.

```r
# Cook's distance to identify influential points
cooksD <- cooks.distance(anova_model)
plot(cooksD, type="h", main="Cook's Distance")
abline(h = 4/(nrow(fa4_data) - 3 - 1), col="red") # Threshold
```



The Cook's Distance plot shows the influence of each data point on the regression model. The x-axis represents the index of data points, while the y-axis shows the Cook's distance values. There is a horizontal red line indicating the threshold for influential points, which appears to be quite low on the y-axis. Most data points have relatively low Cook's distance values, falling below the threshold. However, there are several points with noticeably higher Cook's distance values, particularly around indices 0-15. The tallest spikes, indicating the most influential points, are observed near indices 0, 15, and 16. These points exceed the threshold line and could be considered influential observations that may have a substantial impact on the regression results. The pattern suggests that the earlier data points in the dataset tend to have more influence on the model than later points, as the Cook's distance values generally decrease after index 20.

I analyzed the Cook's Distance plot from the image to identify influential data points and their positions relative to the threshold line. The analysis revealed several influential points, particularly at the beginning of the dataset, which could significantly impact the regression model.

Sample Size

```r
# Check sample size per group
table(fa4_data$Size, fa4_data$Type, fa4_data$Side)
```

```
## , ,  = 1
##
##
##     1 2
##   1 3 3
##   2 3 3
##   3 3 3
##
## , ,  = 2
##
##
##     1 2
##   1 3 3
##   2 3 3
##   3 3 3
```

The table shows the sample size for each combination of vehicle size, type, and side, revealing a balanced design with 3 observations for each group. Specifically, there are 3 readings for each size (small, medium, large) across both types of silencers (standard and Octal) and both sides of the vehicle (right and left). This equal distribution of observations ensures that each treatment group is adequately represented, which is beneficial for the robustness and reliability of the ANOVA analysis. Overall, the sample sizes meet the general recommendation of having sufficient data points per group for valid statistical comparisons.

2. Is there a significant interaction between vehicle size, type and side on noise levels?

Based on the analyses, there is a significant interaction between vehicle size, type, and side on noise levels. The interaction plots reveal that the effect of one factor (e.g., vehicle size) on noise levels is not consistent across the levels of the other factors (e.g., type or side). The non-parallel lines in the interaction plots suggest that the relationships between the factors are not additive, indicating that the impact of one factor depends on the levels of the others. This interaction effect means that the combination of vehicle size, type, and side significantly influences the noise levels, rather than each factor acting independently. Therefore, the data suggest a need to consider interaction terms in any predictive model for noise levels.

3. Provide a two-way interaction effect between vehicle's size, type and side on noise levels.

The ANOVA results show significant two-way interaction effects between **Size and Type** (p = 2.89e-05) and **Size and Side** (p = 6.69e-07) on noise levels, indicating that the impact of vehicle size on noise varies depending on the type of silencer and the side of the vehicle. The main effects of **Size** ($p < 2e\text{-}16$) and **Type** (p = 6.99e-07) are also significant, showing that both factors independently influence noise levels. However, the interaction between **Type and Side** is not significant (p = 0.413), suggesting that the effect of the silencer type does not vary significantly between the vehicle's sides. Additionally, the non-significant result for the **Side** main effect (p = 0.869) indicates that the side of the vehicle does not independently affect noise levels.

4. Are there significant main effects of the vehicle size (Size), type of vehicle (Type), and side of the car (Side) on noise levels? For example, do larger vehicles produce higher noise levels compared to medium or small vehicles, regardless of the type of silencer or side of the vehicle? Does the Octel filter reduce noise more effectively compared to the standard silencer, regardless of the vehicle size or the side of the vehicle? Does the Side (right or left) of the vehicle from which the noise is measured significantly affect the noise level or is there a significant difference in the noise readings on the right side compared to the left side, irrespective of vehicle size or silencer type?

The significant main effects of **vehicle size** and **vehicle type** on noise levels, but **not for the side of the vehicle**:

1. **Vehicle Size**: The main effect of vehicle size is highly significant ($p < 2e-16$), meaning that larger vehicles produce significantly different noise levels compared to medium or small vehicles, regardless of the type of silencer or side of the vehicle. This suggests that vehicle size has a strong independent impact on noise levels.
2. **Vehicle Type**: The main effect of vehicle type (silencer type) is also significant ($p = 6.99e-07$), indicating that the Octel filter and the standard silencer produce different noise levels. This suggests that the Octel filter may reduce noise more effectively compared to the standard silencer, regardless of vehicle size or side.
3. **Side of the Vehicle**: The main effect of the side of the vehicle is not significant ($p = 0.869$), meaning there is no substantial difference in noise levels between the right and left sides of the vehicle, irrespective of vehicle size or silencer type.

These results show that vehicle size and silencer type significantly affect noise levels, while the side from which the noise is measured does not.

5. Which factor or combination of factors has the greatest impact on noise levels? For example, does vehicle size explain a larger proportion of the variance in noise levels compared to type of vehicle or side of the car? Does the type of vehicle explain a larger proportion of the variance in noise levels compared to vehicle size or side of the car? Does the side of the car explain a larger proportion of the variance in noise levels compared to vehicle size or type of vehicle?

Based on the ANOVA results, **vehicle size** explains the largest proportion of variance in noise levels, followed by **vehicle type**, while the **side of the car** has little to no impact:

1. **Vehicle Size**: With a highly significant F value of 519.917, vehicle size has the greatest impact on noise levels. This suggests that size accounts for the largest proportion of the variance in noise levels, making it the most influential factor in the model.
2. **Vehicle Type**: The type of vehicle (i.e., silencer type) also has a significant impact, with an F value of 42.160. While it is less impactful than size, it still explains a notable portion of the variance in noise levels.
3. **Side of the Car**: The side of the car has an F value of 0.028 and a non-significant p-value (0.869), indicating that it explains a negligible portion of the variance in noise levels.

Thus, **vehicle size** has the greatest effect on noise levels, followed by **vehicle type**, with **side** having virtually no impact.

6. After finding a significant main effect for vehicle size, which specific teaching size differ in their effect on noise levels? For instance, do post-hoc tests show that large vehicles recorded significantly higher noise levels than the moderate or small vehicles?

```
# Convert Size, Type, and Side to factors
fa4_data$Size <- as.factor(fa4_data$Size)
fa4_data$Type <- as.factor(fa4_data$Type)
fa4_data$Side <- as.factor(fa4_data$Side)

# Refit the ANOVA model
anova_model <- aov(Noise ~ Size * Type * Side, data = fa4_data)

# Perform Tukey's HSD post-hoc test for the Size factor
tukey_result <- TukeyHSD(anova_model, "Size")
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Noise ~ Size * Type * Side, data = fa4_data)
##
## $Size
##           diff        lwr       upr   p adj
## 2-1   9.583333   5.690003  13.47666 6.9e-06
## 3-1 -51.666667 -55.559997 -47.77334 0.0e+00
## 3-2 -61.250000 -65.143330 -57.35667 0.0e+00
```

The Tukey post-hoc test results show significant differences in noise levels between vehicle sizes. Specifically, medium-sized vehicles (Size 2) produce significantly higher noise levels than small vehicles (Size 1), with a mean difference of **9.58** ($p = 6.9e\text{-}06$). In contrast, large vehicles (Size 3) produce significantly lower noise levels compared to both small vehicles (mean difference = **-51.67**, $p < 0.001$) and medium vehicles (mean difference = **-61.25**, $p < 0.001$). These results indicate that large vehicles record the lowest noise levels, followed by medium vehicles, and small vehicles record the highest.