Comparative Analysis

Anahit Baghdasaryan

2023-12-04

Comparative Analysis Plan: Good vs. Normal Performing Students

Objective:

The goal of this comparative analysis is to examine the differences between good and normal performing students based on their GPAs. Good performers are defined as students with a GPA equal to or higher than 3, while normal performers have a GPA lower than 3. We will investigate the impact of various factors, including lifestyle choices, future plans and goals, study-related factors, and demographic factors, on academic performance.

Steps:

1. **Data Preparation:**
   - Ensure the dataset is cleaned and missing values are handled appropriately.
   - Create a binary variable for performance: "Good" (GPA >= 3) and "Normal" (GPA < 3).
2. **Descriptive Statistics:**
   - Provide summary statistics for GPA, considering both the overall distribution and the distribution within each performance group.
3. **Data Visualization:**
   - Generate visualizations to compare the distributions of key features between good and normal performers.
   - Utilize histograms, box plots, and bar charts to visually assess differences.
4. **Statistical Tests:**
   - Perform appropriate statistical tests to assess the associations between performance and various factors:
   - Chi-Squared Tests: For categorical variables (e.g., gender, academic level, attendance).
   - ANOVA: For continuous variables with more than two groups (e.g., GPA distribution among performance levels).
5. **Linear Regression:**
   - Conduct linear regression to identify significant predictors of GPA.
   - Include relevant features.
6. **Interpretation:**
   - Interpret the results of statistical tests and regression analyses.
   - Discuss any significant differences or associations found between good and normal performers.
   - Consider the practical implications of the findings and potential areas for further research.
7. **Limitations and Recommendations:**
   - Highlight the limitations of the study, such as the potential impact of a small sample size.
   - Provide recommendations for future research based on the outcomes of this analysis.

Data Preparation

```
data <- read_csv("data_tidy.csv")
```

```
## Rows: 95 Columns: 59
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (20): timestamp, gender, age, academic_level, field_of_study, university...
## dbl (39): id, study_env_Campus Common Spaces, study_env_Classroom, study_env...
##
## 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.
```

```
head(data)
```

```
## # A tibble: 6 x 59
##      id study_~1 study~2 study~3 study~4 study~5 study~6 study~7 backg~8 backg~9
##   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1     1        0       0       0       1       1       0       0       1       0
## 2     2        1       0       0       0       1       0       0       1       0
## 3     3        0       0       0       1       1       0       0       1       0
## 4     4        0       1       1       0       1       0       0       1       1
## 5     5        0       0       0       1       1       1       0       1       1
## 6     6        0       0       0       1       0       0       0       0       0
## # ... with 49 more variables:
## #   `background_noise_Prefer a bustling environment` <dbl>,
## #   `background_noise_Prefer complete silence` <dbl>,
## #   study_time_Afternoon <dbl>, `study_time_Early morning` <dbl>,
## #   study_time_Evening <dbl>, `study_time_Late morning` <dbl>,
## #   study_time_Night <dbl>, `study_time_No specific preference` <dbl>,
## #   `resources_Interactive simulations or applications` <dbl>, ...
```

```r
columns_to_exclude <- c("id", "timestamp", "field_of_study", "university", "gpa")

# List of columns to convert to categorical
columns_to_convert <- setdiff(names(data), columns_to_exclude)

# Convert selected columns to categorical
data[columns_to_convert] <- lapply(data[columns_to_convert], as.factor)
```

```r
# Create a new column 'performance' based on GPA
data$performance <- ifelse(data$gpa < 3.0, "normal", "good")

# Convert the 'performance' column to a factor for better representation
data$performance <- factor(data$performance, levels = c("normal", "good"))

# View specific columns
head(data[c('id', 'gpa', 'performance')])
```

```
## # A tibble: 6 x 3
##      id   gpa performance
##   <dbl> <dbl> <fct>
## 1     1  3.7  good
## 2     2  3.3  good
## 3     3  2.15 normal
## 4     4  2.2  normal
## 5     5  2.41 normal
## 6     6  2.5  normal
```

```r
# Create a new dataset excluding rows with NA values in 'performance'
data_no_na_performance <- data[complete.cases(data$performance), ]
```

Descriptive Statistics

```r
# Summary statistics for the overall distribution of GPA
overall_summary <- summary(data_no_na_performance$gpa)

# Summary statistics for the distribution within each performance group
group_summary <- data_no_na_performance %>%
  group_by(performance) %>%
  summarise(
    Mean_GPA = mean(gpa),
    Median_GPA = median(gpa),
    SD_GPA = sd(gpa),
    Min_GPA = min(gpa),
    Max_GPA = max(gpa)
  )

# Display the results
cat("Overall GPA Summary:\n", overall_summary, "\n\n")
```

```
## Overall GPA Summary:
##  2 2.9 3.4 3.288352 3.8 4
```

```r
cat("GPA Summary within Each Performance Group:\n")
```

```
## GPA Summary within Each Performance Group:
```

```r
print(group_summary)
```

```
## # A tibble: 2 x 6
##   performance Mean_GPA Median_GPA SD_GPA Min_GPA Max_GPA
##   <fct>          <dbl>      <dbl>  <dbl>   <dbl>   <dbl>
## 1 normal          2.54       2.55  0.300       2    2.98
## 2 good            3.60       3.7   0.303       3    4
```

Overall GPA Summary:

Minimum GPA: 2.0

1st Quartile (25th percentile): 2.9

Median (50th percentile): 3.4

Mean: 3.288352

3rd Quartile (75th percentile): 3.8

Maximum GPA: 4.0

GPA Summary within Each Performance Group:

For "Normal" Performance Group:

Mean GPA: 2.535231

Median GPA: 2.55

Standard Deviation: 0.3002016

Minimum GPA: 2.0

Maximum GPA: 2.98

For "Good" Performance Group:

Mean GPA: 3.604177

Median GPA: 3.70

Standard Deviation: 0.3026630

Minimum GPA: 3.0

Maximum GPA: 4.0

Interpretation:

The overall GPA distribution has a mean of approximately 3.29, with a median of 3.4.

The "Normal" performance group has a lower mean GPA of approximately 2.54, indicating lower academic performance, while the "Good" performance group has a higher mean GPA of approximately 3.60, indicating better academic performance.

The standard deviation within each group suggests the degree of variability in GPA scores. "Good" performance has a slightly higher standard deviation than the "Normal" group.

The median is the middle value in the distribution, and it is lower in the "Normal" group (2.55) compared to the "Good" group (3.70).

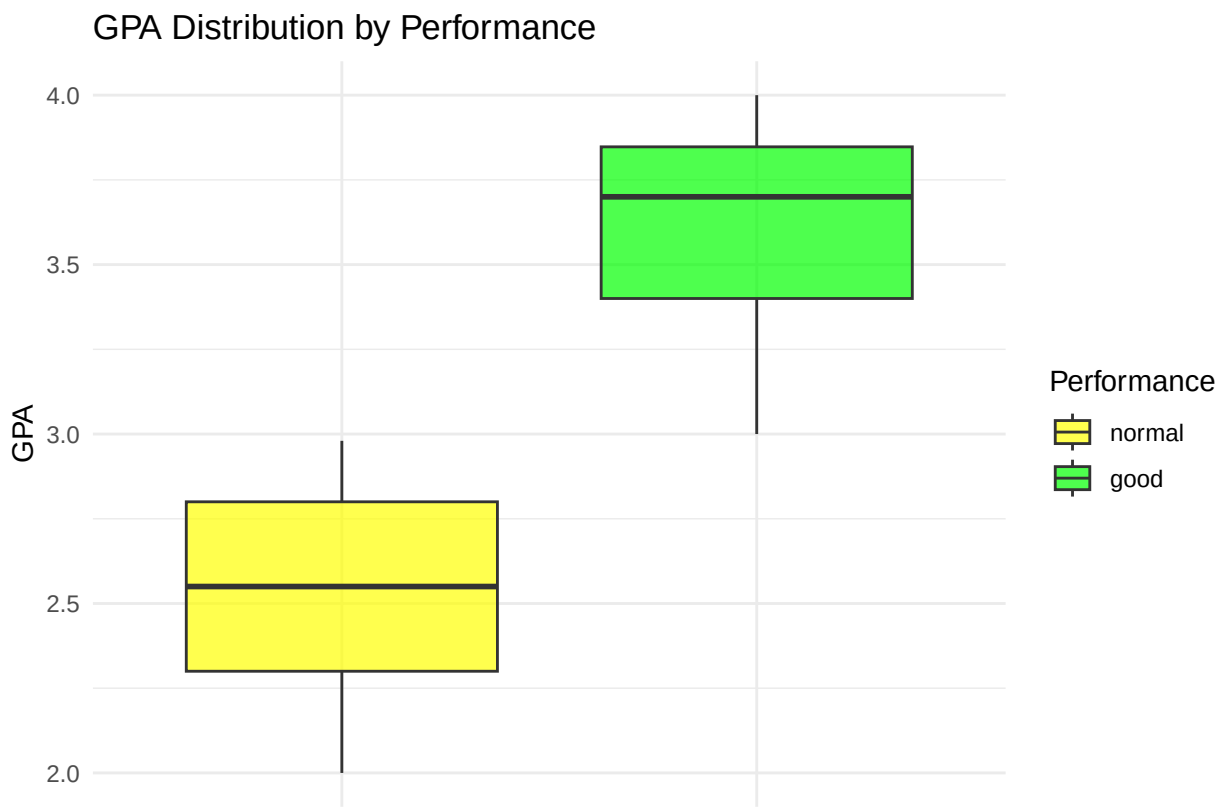
Data Visualization and Statistical Tests (Chi-Squared Tests)

Performance Distribution

```r
# Create a bar plot
ggplot(data_no_na_performance, aes(x = performance, fill = performance)) +
  geom_bar(alpha = 0.7) +
  labs(title = "Performance Comparison",
       x = "Performance",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("yellow", "green"))
```

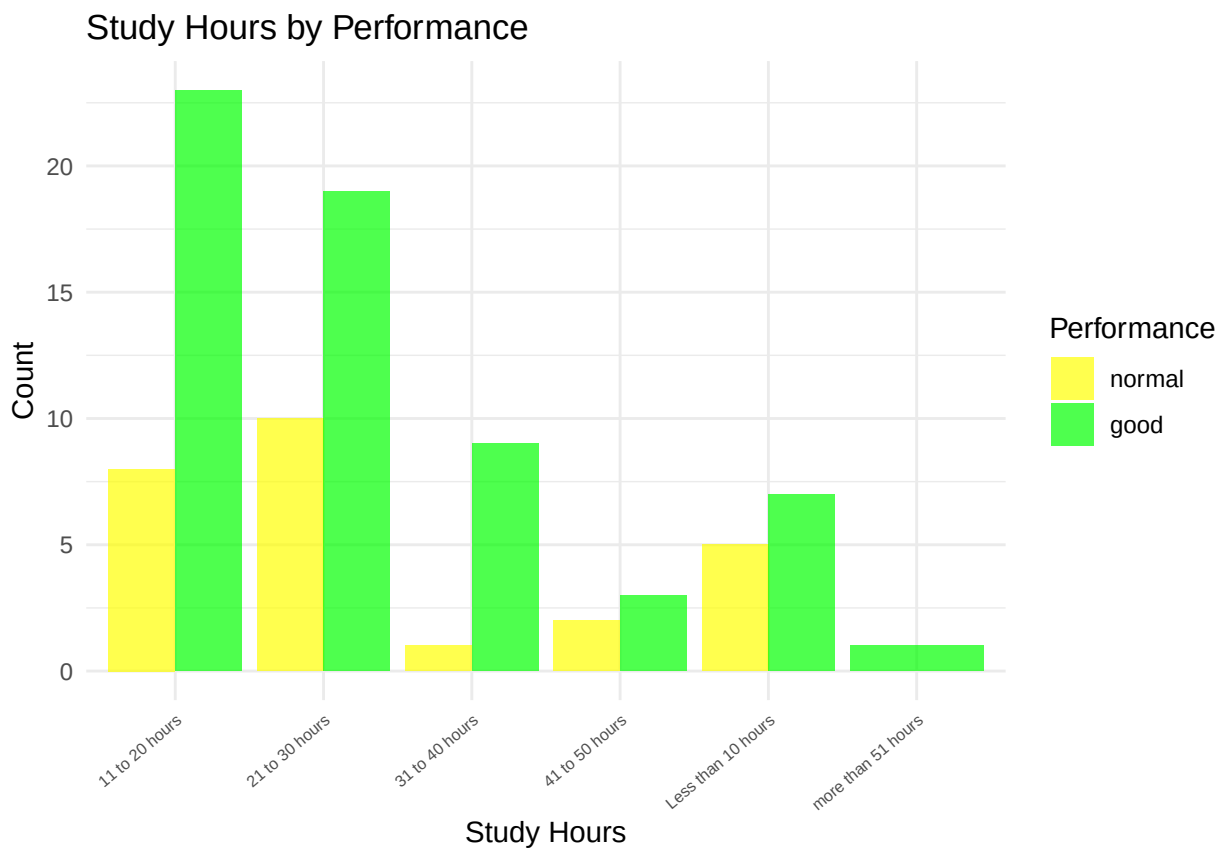
GPA Distribution and Performance

```
# Create a box plot
ggplot(data_no_na_performance, aes(x = performance, y = gpa, fill = performance)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "GPA Distribution by Performance",
       x = "",
       y = "GPA",
       fill = "Performance") +
  theme_minimal() +
  theme(axis.text.x = element_blank())+
  scale_fill_manual(values = c("yellow", "green"))
```



Study Hours and Performance

```
# Create a box plot of study hours by performance
ggplot(data_no_na_performance, aes(x = hours_spend_studying, fill = performance)) +
  geom_bar(alpha = 0.7, position = "dodge") +
  labs(title = "Study Hours by Performance",
       x = "Study Hours",
       y = "Count",
       fill = "Performance") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 40, size = 6, h = 1))+
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_study_hours <- table(data_no_na_performance$performance,
                                 data_no_na_performance$hours_spend_studying)

# Display the contingency table
print(performance_study_hours)
```

```
##
##          11 to 20 hours 21 to 30 hours 31 to 40 hours 41 to 50 hours
##   normal              8             10              1              2
##   good               23             19              9              3
```

```
##
##          Less than 10 hours more than 51 hours
##   normal                  5                  0
##   good                    7                  1
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_study_hours <- chisq.test(performance_study_hours)
```

```
## Warning in chisq.test(performance_study_hours): Chi-squared approximation may be
## incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_study_hours)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_study_hours
## X-squared = 3.9119, df = 5, p-value = 0.5622
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$hours_spend_studying, correct = TRUE)
```

```
## [1] 0.2917584
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$hours_spend_studying))
```

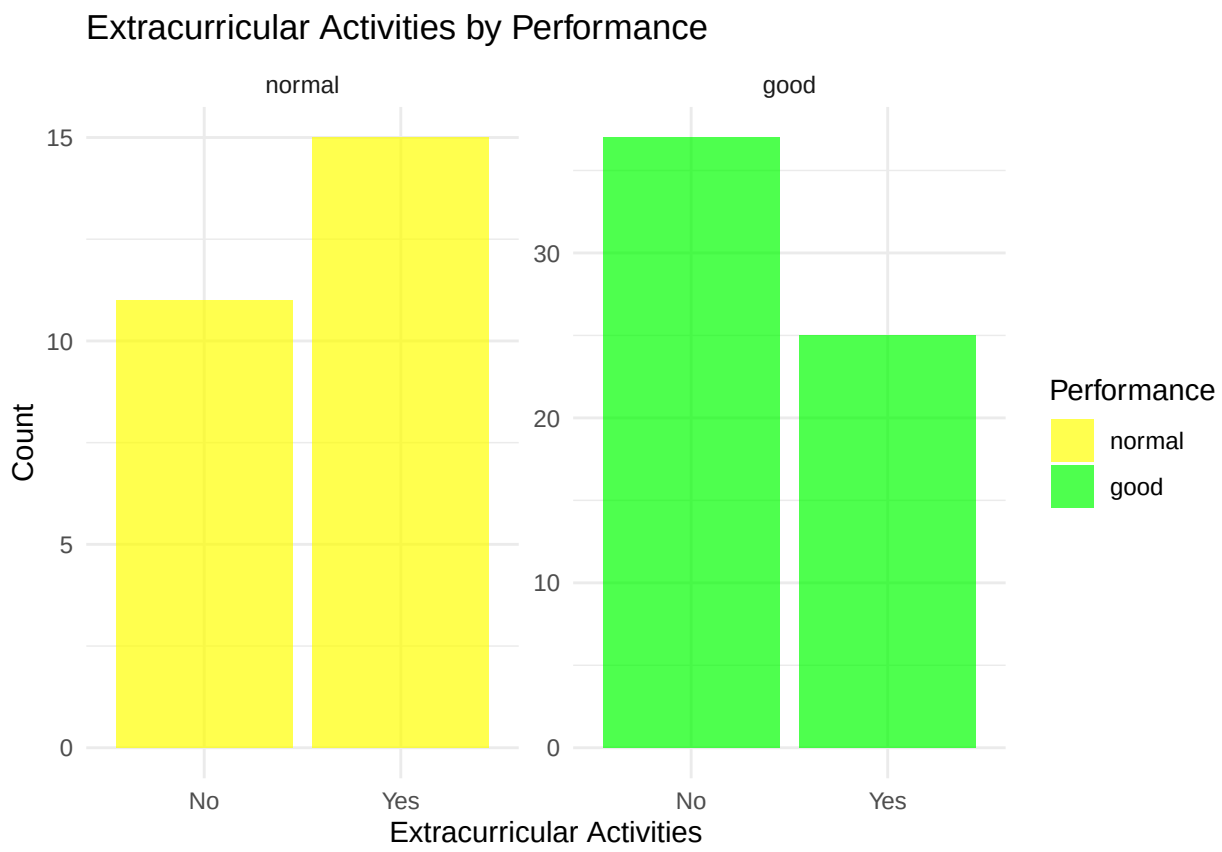
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 4.5270  5  0.47628
## Pearson          3.9119  5  0.56217
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.206
## Cramer's V        : 0.211
```

Overall Interpretation:

The chi-squared test examining the relationship between study hours and academic performance suggests that there is no statistically significant association in the given dataset (p-value = 0.5622). The contingency coefficient and Cramer's V indicate a moderate, yet not particularly strong, association between the two variables.

Extracurricular Activities and Performance

```
# Create a bar plot of extracurricular activities by performance
ggplot(data_no_na_performance, aes(x = extracurricular_activities, fill = performance)) +
  geom_bar(position = "stack", alpha = 0.7) +
  labs(title = "Extracurricular Activities by Performance",
       x = "Extracurricular Activities",
       y = "Count",
       fill = "Performance") +
  facet_wrap(~performance, scales = "free_y")+
  theme_minimal()+
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_extracurricular_activities <- table(data_no_na_performance$performance,
                          data_no_na_performance$extracurricular_activities)

# Display the contingency table
print(performance_extracurricular_activities)
```

```
##
##          No Yes
##   normal 11  15
##   good   37  25
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_extracurricular_activities <-
  chisq.test(performance_extracurricular_activities)

# Display the chi-squared test result
print(chi_squared_performance_extracurricular_activities)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  performance_extracurricular_activities
## X-squared = 1.5836, df = 1, p-value = 0.2082
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$extracurricular_activities, correct = TRUE)
```

```
## [1] 0.2222838
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$extracurricular_activities))
```

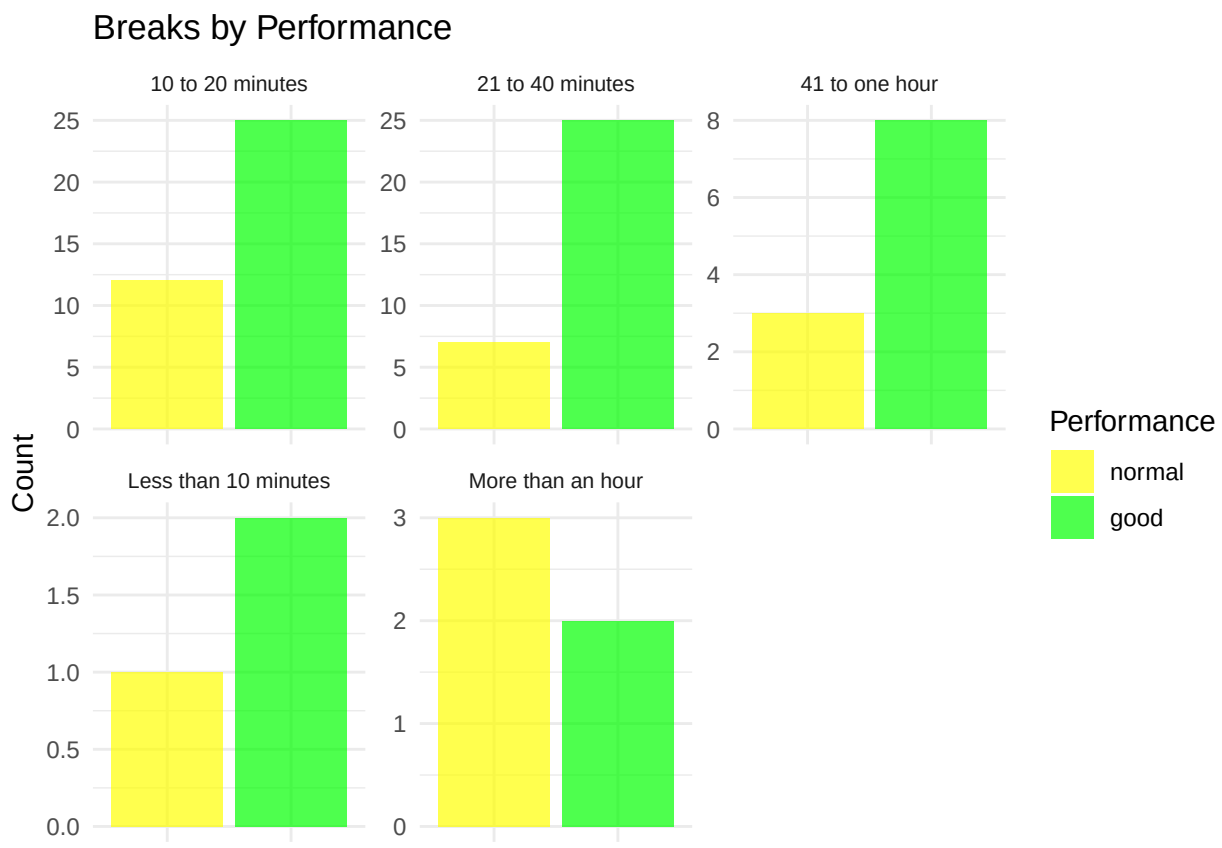
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.2269  1  0.13563
## Pearson          2.2291  1  0.13543
##
## Phi-Coefficient   : 0.159
## Contingency Coeff.: 0.157
## Cramer's V        : 0.159
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and participation in extracurricular activities reveals no statistically significant relationship (p-value = 0.2082). The contingency coefficient and Cramer's V suggest a weak to moderate association, indicating that academic performance may have some dependence on involvement in extracurricular activities, though this association is not particularly strong.

Breaks by Performance

```
# Create a bar plot of breaks by performance
ggplot(data_no_na_performance, aes(x = performance, fill = performance)) +
  geom_bar(position = "stack", alpha = 0.7) +
  labs(title = "Breaks by Performance",
       x = "",
       y = "Count",
       fill = "Performance") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),  # Remove x-axis text
        axis.ticks.x = element_blank(),  # Remove x-axis ticks
        strip.text.x = element_text(size = 8, angle = 0)) +  # Adjust facet text size and angle
  facet_wrap(~breaks, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_breaks <- table(data_no_na_performance$performance,
                            data_no_na_performance$breaks)

# Display the contingency table
print(performance_breaks)
```

```
##
```

```
##          10 to 20 minutes 21 to 40 minutes 41 to one hour Less than 10 minutes
##   normal               12                7              3                     1
##   good                 25               25              8                     2
##
##          More than an hour
##   normal                 3
##   good                   2
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_breaks <-
  chisq.test(performance_breaks)
```

```
## Warning in chisq.test(performance_breaks): Chi-squared approximation may be
## incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_breaks)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_breaks
## X-squared = 3.3284, df = 4, p-value = 0.5044
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$breaks, correct = TRUE)
```

```
## [1] 0.269978
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$breaks))
```

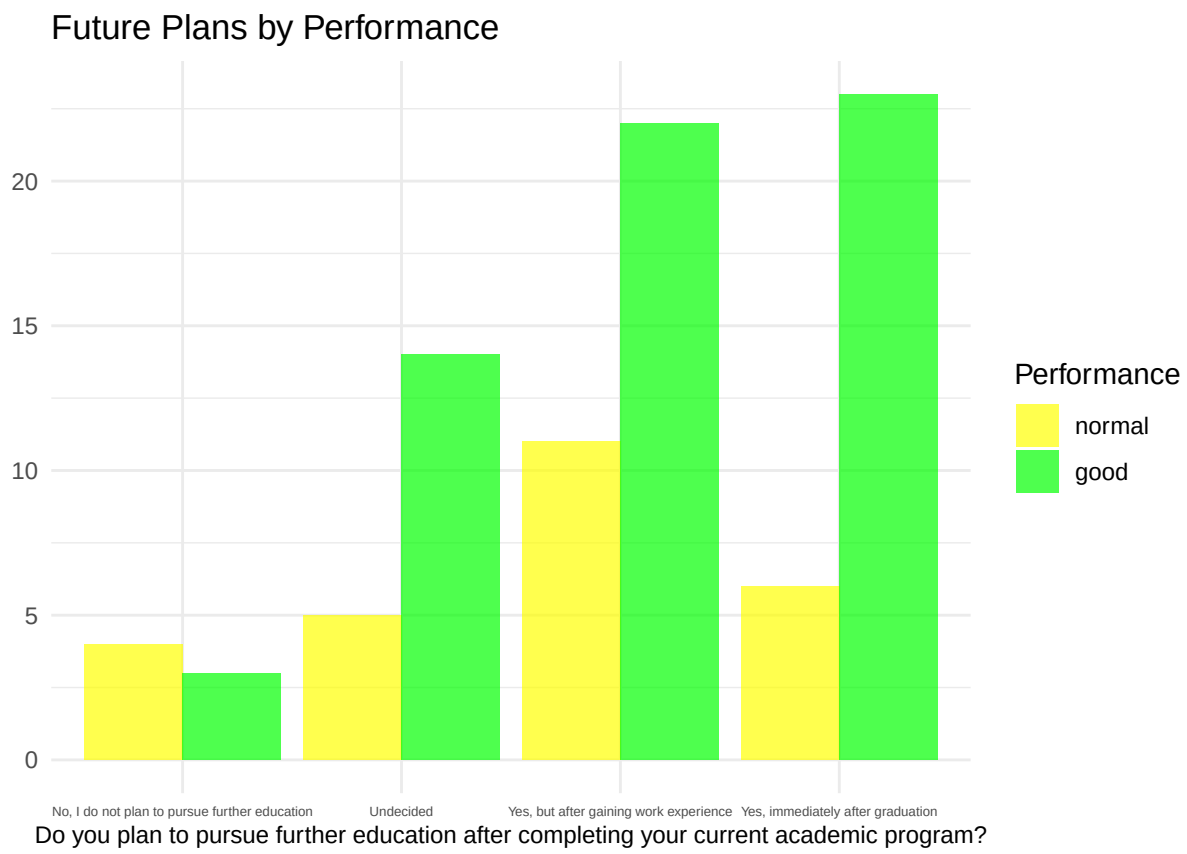
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 3.1385  4  0.53492
## Pearson          3.3284  4  0.50445
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.191
## Cramer's V        : 0.194
```

Overall Interpretation:

The chi-squared test investigating the relationship between academic performance and preferences for break duration did not yield statistically significant results (p-value = 0.5044). The contingency coefficient and Cramer's V indicate a weak association, suggesting that the choice of break duration may have some influence on academic performance, but this relationship is not strong.

Future Plans and Performance

```
# Bar plot for future plans
ggplot(data_no_na_performance, aes(x = future_plans, fill = performance)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Future Plans by Performance",
       x = "Do you plan to pursue further education after completing your current academic program?",
       y = "",
       fill = "Performance") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, size = 5),
        axis.title.x=element_text(size=9))+
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_future_plans <- table(data_no_na_performance$performance,
                                  data_no_na_performance$future_plans)

# Display the contingency table
print(performance_future_plans)
```

```
##
##          No, I do not plan to pursue further education Undecided
##   normal                                             4         5
```

```
##   good                                              3        14
##
##          Yes, but after gaining work experience
##   normal                                        11
##   good                                          22
##
##          Yes, immediately after graduation
##   normal                                     6
##   good                                      23
```

```
# Perform a chi-squared test of independence
chi_squared_performance_future_plans <-
  chisq.test(performance_future_plans)
```

```
## Warning in chisq.test(performance_future_plans): Chi-squared approximation may
## be incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_future_plans)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_future_plans
## X-squared = 3.9764, df = 3, p-value = 0.264
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$future_plans, correct = TRUE)
```

```
## [1] 0.2940504
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$future_plans))
```

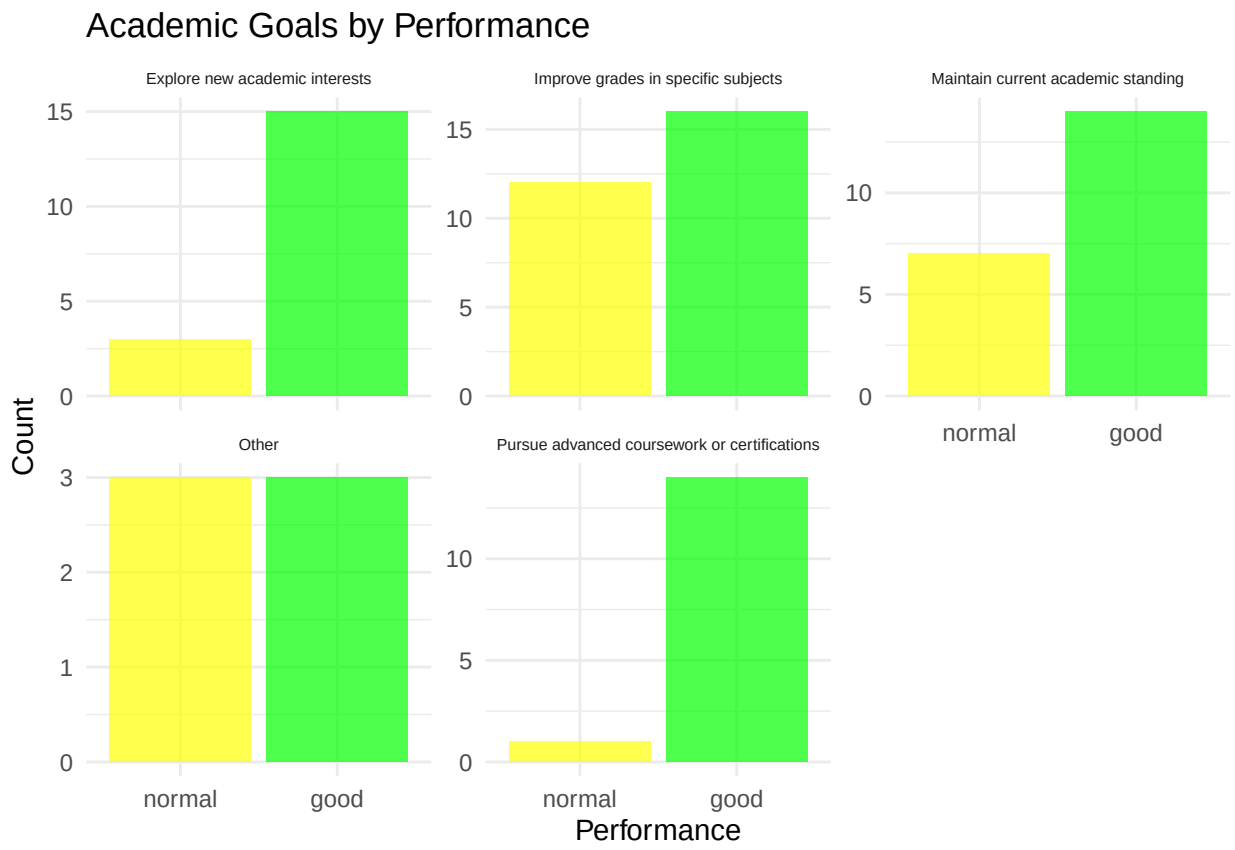
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 3.7849  3  0.28564
## Pearson          3.9764  3  0.26402
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.208
## Cramer's V        : 0.213
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and future plans did not reveal statistically significant results (p-value = 0.264). The contingency coefficient and Cramer's V suggest a weak association, implying that the future plans of students may have some influence on their academic performance, but this relationship is not substantial.

Academic Goals and Performance

```
# Stacked bar plot for academic goals
ggplot(data_no_na_performance, aes(x = performance, fill = performance)) +
  geom_bar(alpha = 0.7) +
  labs(title = "Academic Goals by Performance",
       x = "Performance",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "none",
        strip.text = element_text(size = 6)) +  # Adjust facet text size and angle
  facet_wrap(~academic_goals, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_academic_goals <- table(data_no_na_performance$performance,
                                    data_no_na_performance$academic_goals)

# Display the contingency table
print(performance_academic_goals)
```

```
##
##          Explore new academic interests Improve grades in specific subjects
##   normal                              3                                   12
```

```
##   good                                15                                    16
##
##          Maintain current academic standing Other
##   normal                                   7     3
##   good                                    14     3
##
##          Pursue advanced coursework or certifications
##   normal                                            1
##   good                                             14
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_academic_goals <-
  chisq.test(performance_academic_goals)
```

```
## Warning in chisq.test(performance_academic_goals): Chi-squared approximation may
## be incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_academic_goals)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_academic_goals
## X-squared = 8.9404, df = 4, p-value = 0.06261
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$academic_goals, correct = TRUE)
```

```
## [1] 0.4294776
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$academic_goals))
```

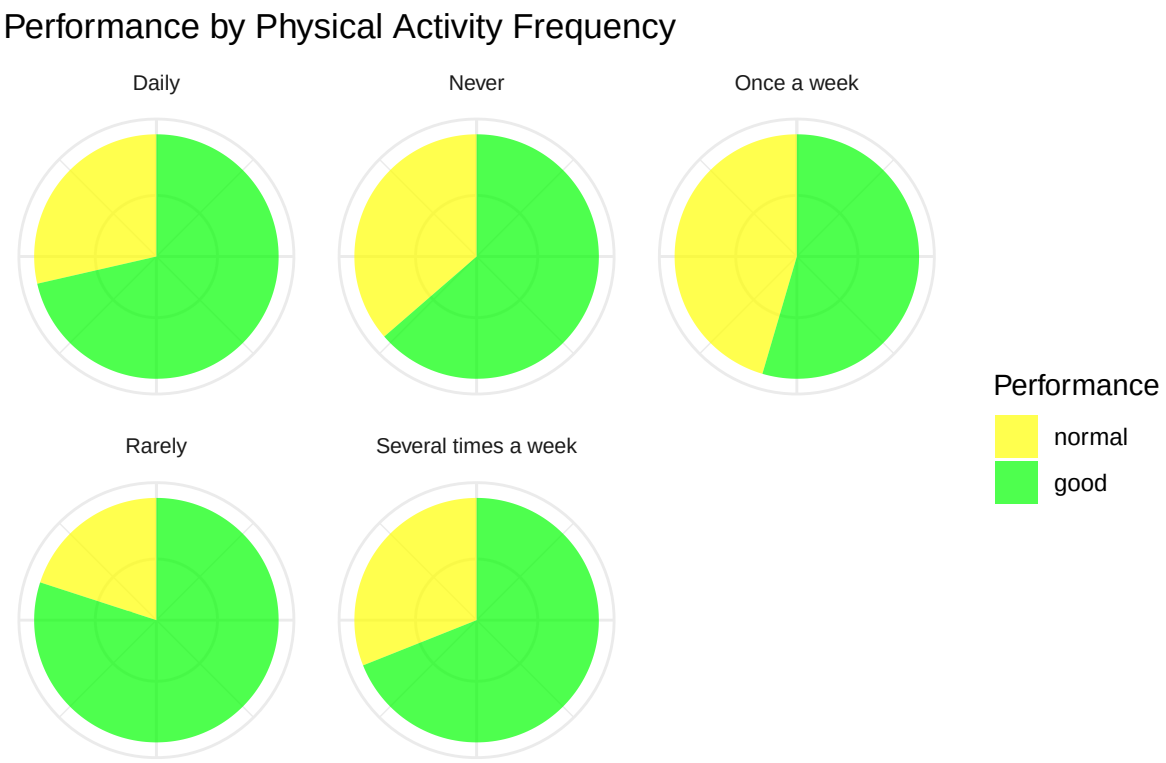
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 9.9633  4 0.041051
## Pearson          8.9404  4 0.062607
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.304
## Cramer's V        : 0.319
```

Overall Interpretation:

The chi-squared test investigating the association between academic performance and academic goals yielded interesting but marginally significant results (p-value = 0.06261). The contingency coefficient and Cramer's V indicate a moderate association. The data suggests that students with different academic goals may exhibit variations in their academic performance. While the findings are not strongly significant, they hint at a potential relationship that warrants further exploration.

Performance and Physical Activity Frequency

```
# Pie chart for physical activity frequency
ggplot(data_no_na_performance, aes(x = "", fill = performance)) +
  geom_bar(width = 1, position = "fill", alpha = 0.7) +
  coord_polar("y") +
  labs(title = "Performance by Physical Activity Frequency",
       fill = "Performance",
       x = "",
       y = "") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),  # Remove x-axis text
        axis.ticks.x = element_blank(),  # Remove x-axis ticks
        strip.text.x = element_text(size = 8, angle = 0)) +  # Adjust facet text size and angle
  facet_wrap(~physical_activity_freq) +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_physical_activity_freq <- table(data_no_na_performance$performance,
                                            data_no_na_performance$physical_activity_freq)

# Display the contingency table
print(performance_physical_activity_freq)
```

```
##
##          Daily Never Once a week Rarely Several times a week
##   normal     2     4           5      6                    9
##   good       5     7           6     24                   20
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_physical_activity_freq <-
  chisq.test(performance_physical_activity_freq)
```

```
## Warning in chisq.test(performance_physical_activity_freq): Chi-squared
## approximation may be incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_physical_activity_freq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_physical_activity_freq
## X-squared = 2.9304, df = 4, p-value = 0.5695
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$physical_activity_freq, correct = TRUE)
```

```
## [1] 0.2538755
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$physical_activity_freq))
```

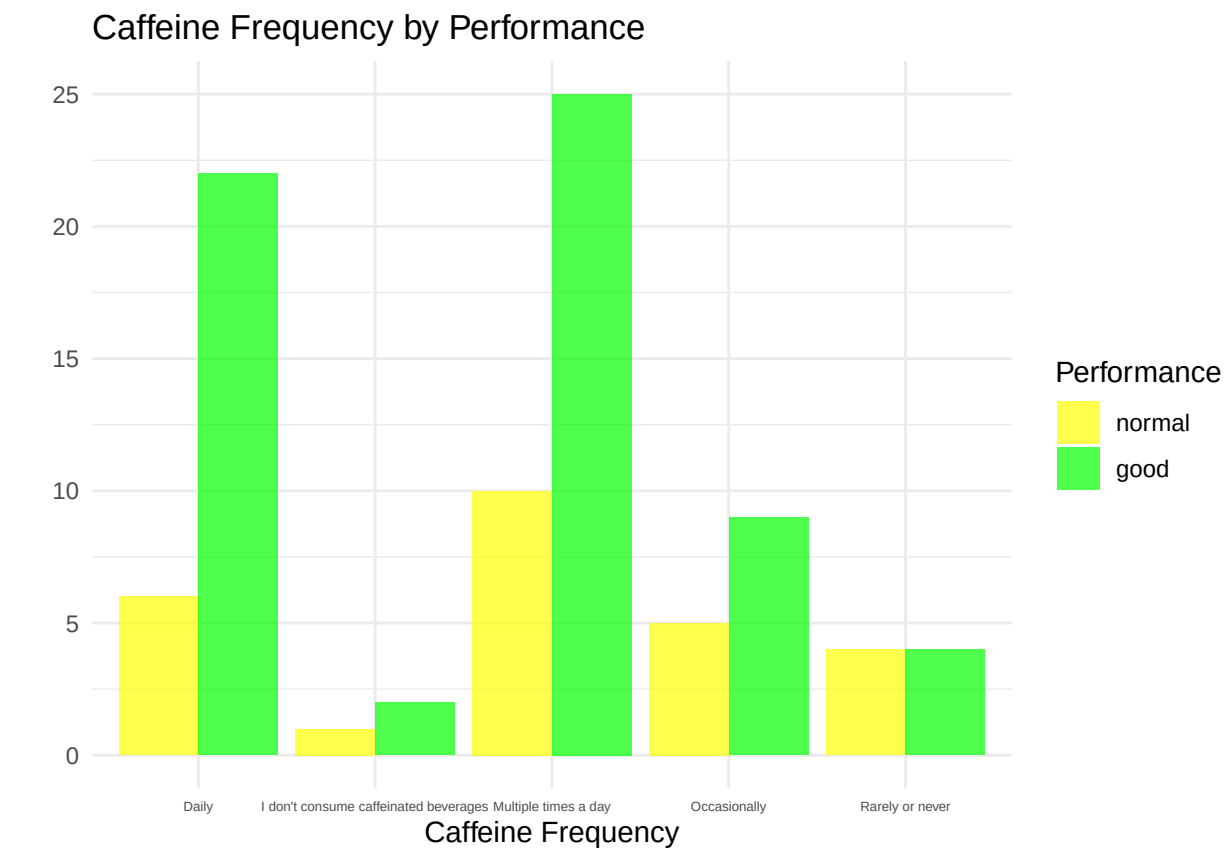
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.9230  4  0.57078
## Pearson          2.9304  4  0.56955
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.18
## Cramer's V        : 0.182
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and physical activity frequency did not reveal a significant relationship (p-value = 0.5695). The contingency coefficient and Cramer's V values indicate a weak association, suggesting that there is no strong evidence to support a link between physical activity frequency and academic performance in the provided dataset.

Caffeine Frequency and Performance

```
# Bar plot for caffeine frequency
ggplot(data_no_na_performance, aes(x = caffeine_freq, fill = performance)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Caffeine Frequency by Performance",
       x = "Caffeine Frequency",
       y = "",
       fill = "Performance") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, size = 5))+
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_caffeine_freq <- table(data_no_na_performance$performance,
                                   data_no_na_performance$caffeine_freq)

# Display the contingency table
print(performance_caffeine_freq)
```

```
##
##          Daily I don't consume caffeinated beverages Multiple times a day
##   normal     6                                     1                    10
##   good      22                                     2                    25
```

```
##
##          Occasionally Rarely or never
##   normal            5               4
##   good              9               4
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_caffeine_freq <-
  chisq.test(performance_caffeine_freq)
```

```
## Warning in chisq.test(performance_caffeine_freq): Chi-squared approximation may
## be incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_caffeine_freq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_caffeine_freq
## X-squared = 2.7867, df = 4, p-value = 0.5941
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$caffeine_freq, correct = TRUE)
```

```
## [1] 0.247771
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$caffeine_freq))
```

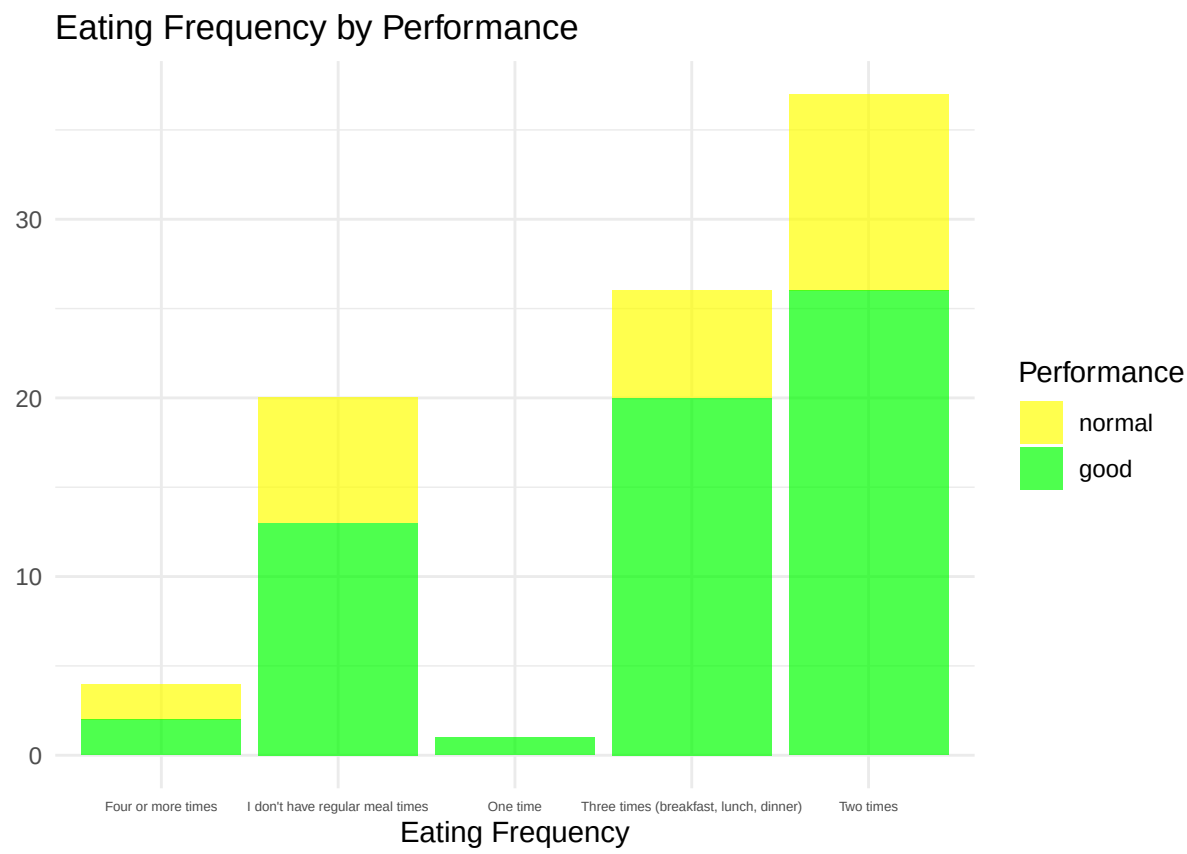
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.6916  4  0.61068
## Pearson          2.7867  4  0.59413
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.175
## Cramer's V        : 0.178
```

Overall Interpretation:

The chi-squared test assessing the association between academic performance and caffeine consumption frequency did not yield a statistically significant relationship (p-value = 0.5941). The contingency coefficient and Cramer's V values, both indicating a weak association, suggest that there is no strong evidence supporting a link between caffeine consumption frequency and academic performance in the given dataset.

Eating Frequency and Performance

```
# Stacked bar plot for eating frequency
ggplot(data_no_na_performance, aes(x = eat_freq, fill = performance)) +
  geom_bar(position = "stack", alpha = 0.7) +
  labs(title = "Eating Frequency by Performance",
       x = "Eating Frequency",
       y = "",
       fill = "Performance") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, size = 5))+
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_eat_freq <- table(data_no_na_performance$performance,
                              data_no_na_performance$eat_freq)

# Display the contingency table
print(performance_eat_freq)
```

```
##
##          Four or more times I don't have regular meal times One time
##   normal                  2                                7        0
##   good                    2                               13        1
```

```
##
##          Three times (breakfast, lunch, dinner) Two times
##   normal                                      6         11
##   good                                       20         26
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_eat_freq<-
  chisq.test(performance_eat_freq)
```

```
## Warning in chisq.test(performance_eat_freq): Chi-squared approximation may be
## incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_eat_freq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_eat_freq
## X-squared = 2.0324, df = 4, p-value = 0.7298
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$eat_freq, correct = TRUE)
```

```
## [1] 0.2124812
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$eat_freq))
```

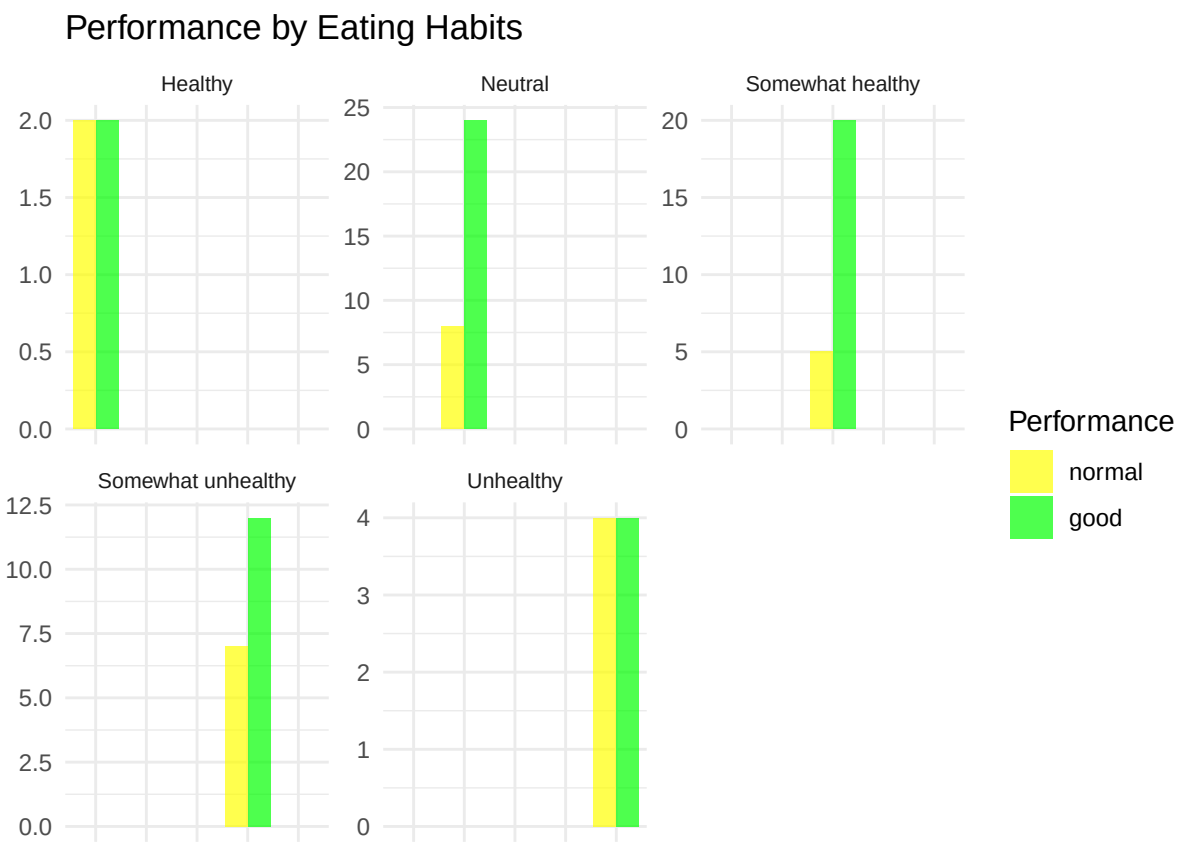
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.2587  4  0.68829
## Pearson          2.0324  4  0.72980
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.15
## Cramer's V        : 0.152
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and eating frequency did not reveal statistically significant results (p-value = 0.7298). The contingency coefficient and Cramer's V values, both indicating a weak association, suggest that there is no strong evidence supporting a link between eating frequency and academic performance in the given dataset.

Performance and Eating Habits

```
# Stacked bar chart for eating habits
ggplot(data_no_na_performance, aes(x = eating_habits, fill = performance)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Performance by Eating Habits",
       fill = "Performance",
       x = "",
       y = "") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),  # Remove x-axis text
        axis.ticks.x = element_blank(),  # Remove x-axis ticks
        strip.text.x = element_text(size = 8, angle = 0)) +
  facet_wrap(~eating_habits, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_eating_habits <- table(data_no_na_performance$performance,
                                   data_no_na_performance$eating_habits)

# Display the contingency table
print(performance_eating_habits)
```

```
##
```

```
##          Healthy Neutral Somewhat healthy Somewhat unhealthy Unhealthy
##   normal       2       8                5                  7         4
##   good         2      24               20                 12         4
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_eating_habits<-
  chisq.test(performance_eating_habits)
```

```
## Warning in chisq.test(performance_eating_habits): Chi-squared approximation may
## be incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_eating_habits)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_eating_habits
## X-squared = 4.3098, df = 4, p-value = 0.3657
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$eating_habits, correct = TRUE)
```

```
## [1] 0.3055757
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$eating_habits))
```

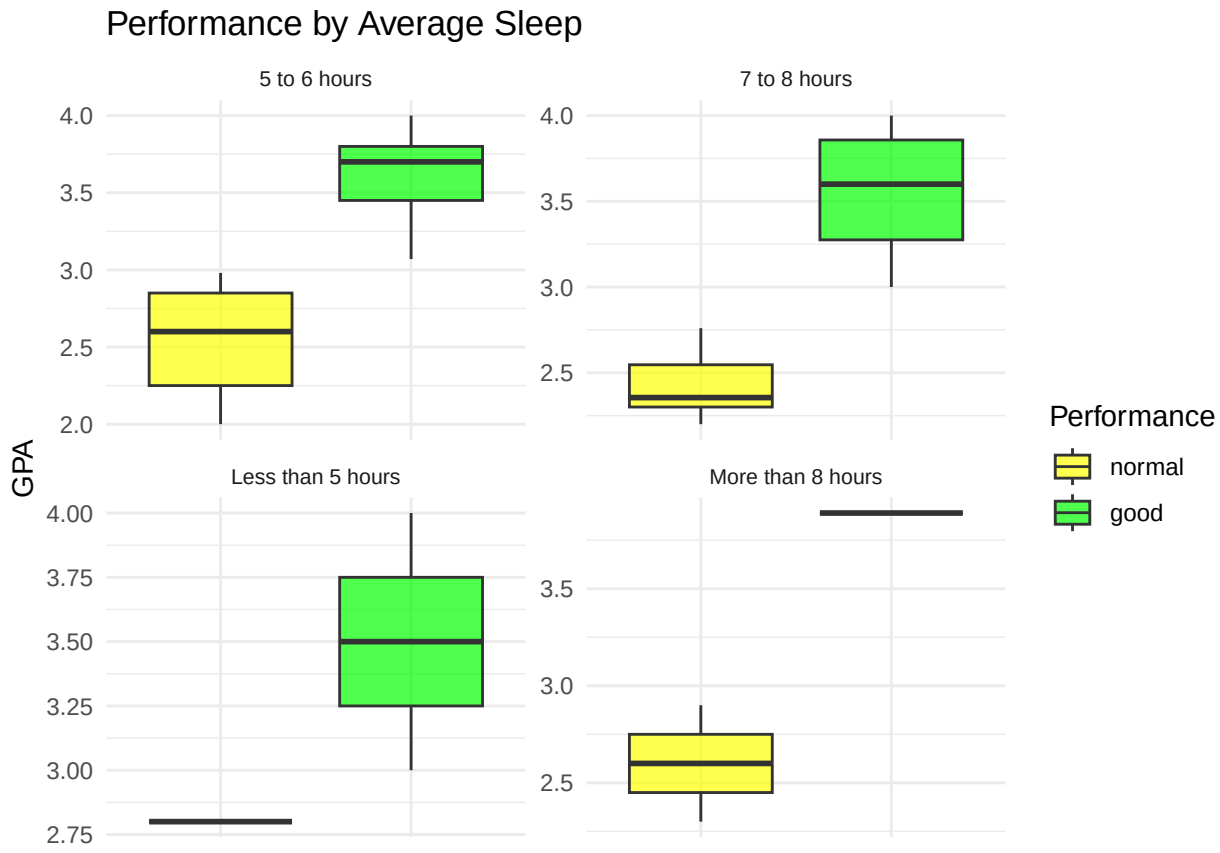
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 4.1723  4  0.38319
## Pearson          4.3098  4  0.36570
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.216
## Cramer's V        : 0.221
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and eating habits did not reveal statistically significant results (p-value = 0.3657). The contingency coefficient and Cramer's V values, both indicating a weak association, suggest that there is no strong evidence supporting a link between eating habits and academic performance in the given dataset.

Performance and Average Sleep

```
# Box plot for average sleep
ggplot(data_no_na_performance, aes(x = performance, y = gpa, fill = performance)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Performance by Average Sleep",
       y = "GPA",
       fill = "Performance") +
  theme_minimal() +
   theme(axis.title.x = element_blank(),
         axis.text.x = element_blank(),  # Remove x-axis text
         axis.ticks.x = element_blank(),  # Remove x-axis ticks
         strip.text.x = element_text(size = 8, angle = 0)) +
  facet_wrap(~avg_sleep, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_avg_sleep <- table(data_no_na_performance$performance,
                               data_no_na_performance$avg_sleep)

# Display the contingency table
print(performance_avg_sleep)
```

```
##
```

```
##          5 to 6 hours 7 to 8 hours Less than 5 hours More than 8 hours
##   normal           15           8                 1                 2
##   good             35          24                 2                 1
```

```
# Perform a chi-squared test of independence
chi_squared_performance_avg_sleep<-
  chisq.test(performance_avg_sleep)
```

```
## Warning in chisq.test(performance_avg_sleep): Chi-squared approximation may be
## incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_avg_sleep)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_avg_sleep
## X-squared = 2.3292, df = 3, p-value = 0.507
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$avg_sleep, correct = TRUE)
```

```
## [1] 0.2270932
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$avg_sleep))
```

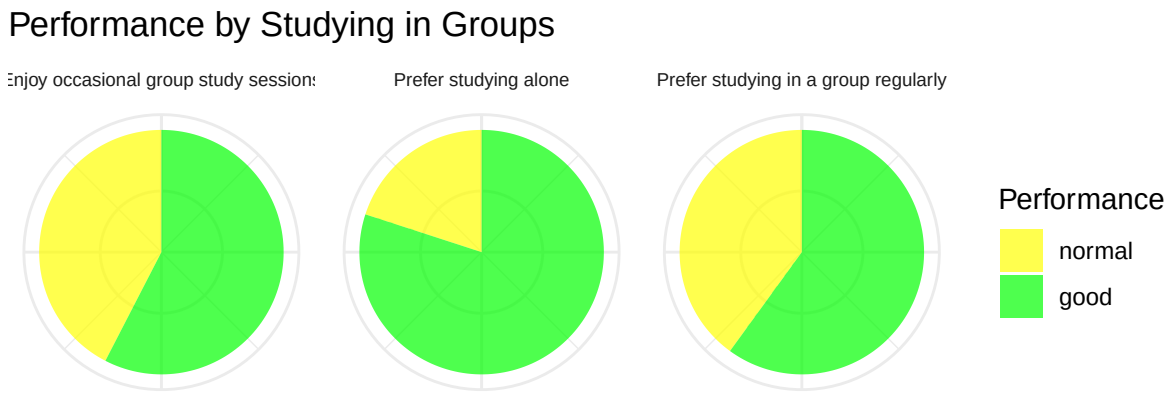
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.1115  3  0.54958
## Pearson          2.3292  3  0.50695
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.161
## Cramer's V        : 0.163
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and average sleep duration did not reveal statistically significant results (p-value = 0.507). The contingency coefficient and Cramer's V values, both indicating a weak association, suggest that there is no strong evidence supporting a significant link between average sleep duration and academic performance in the given dataset.

Performance and Studying in Groups

```
# Pie chart for studying in groups
ggplot(data_no_na_performance, aes(x = "", fill = performance)) +
  geom_bar(width = 1, position = "fill", alpha = 0.7) +
  coord_polar("y") +
  labs(title = "Performance by Studying in Groups",
       fill = "Performance",
       x = "") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),  # Remove x-axis text
        axis.ticks.x = element_blank(),  # Remove x-axis ticks
        strip.text.x = element_text(size = 7, angle = 0)) +
  facet_wrap(~studying_in_groups) +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_studying_in_groups <- table(data_no_na_performance$performance,
                                        data_no_na_performance$studying_in_groups)

# Display the contingency table
print(performance_studying_in_groups)
```

```
##
```

```
##          Enjoy occasional group study sessions Prefer studying alone
##   normal                                    14                    10
##   good                                      19                    40
##
##          Prefer studying in a group regularly
##   normal                                    2
##   good                                      3
```

```r
# Perform a chi-squared test of independence
chi_squared_performance_studying_in_groups<-
  chisq.test(performance_studying_in_groups)
```

```
## Warning in chisq.test(performance_studying_in_groups): Chi-squared approximation
## may be incorrect
```

```r
# Display the chi-squared test result
print(chi_squared_performance_studying_in_groups)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_studying_in_groups
## X-squared = 5.0806, df = 2, p-value = 0.07884
```

```r
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$studying_in_groups, correct = TRUE)
```

```
## [1] 0.330401
```

```r
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$studying_in_groups))
```

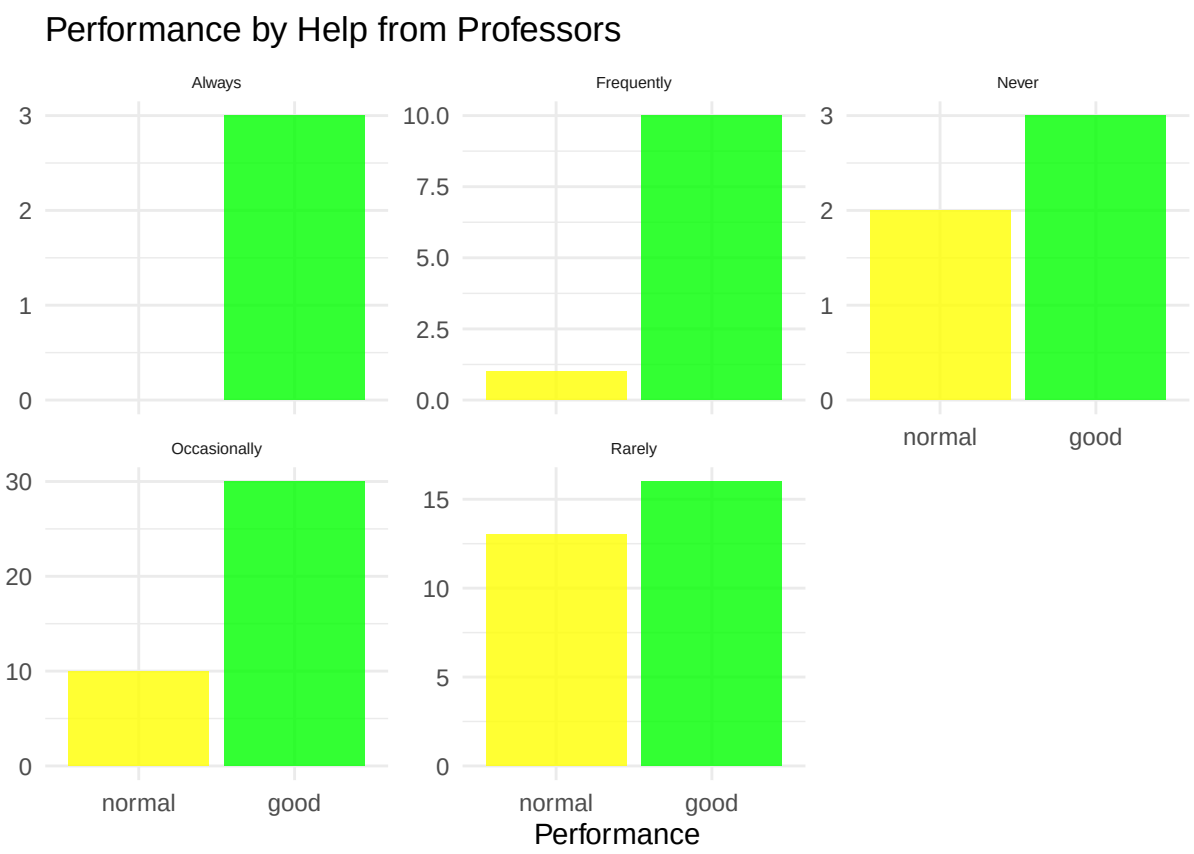
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 5.0680  2 0.079340
## Pearson          5.0806  2 0.078844
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.234
## Cramer's V        : 0.24
```

Overall Interpretation:

The chi-squared test investigating the association between academic performance and studying preferences (studying alone, enjoying occasional group study sessions, or regularly studying in a group) resulted in a p-value of 0.07884, suggesting that the relationship is not statistically significant. While the p-value is close to the conventional significance threshold of 0.05, caution is warranted due to the warning about the chi-squared approximation being incorrect, likely due to a smaller sample size. The contingency coefficient and Cramer's V values imply a moderate association, emphasizing the importance of further exploration with a larger and more diverse sample to draw more reliable conclusions about the impact of studying preferences on academic performance.

Performance and Help from Professors

```
# Stacked bar chart for help from professors
ggplot(data_no_na_performance, aes(x = performance, fill=performance))+
  geom_bar(alpha = 0.8)+
  labs(title = "Performance by Help from Professors",
       x = "Performance",
       y = "") +
  theme_minimal() +
    theme(legend.position = "none",
        strip.text = element_text(size = 6)) +  # Adjust facet text size and angle
  facet_wrap(~help_from_professors, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_help_from_professors <- table(data_no_na_performance$performance,
                          data_no_na_performance$help_from_professors)

# Display the contingency table
print(performance_help_from_professors)
```

```
##
##          Always Frequently Never Occasionally Rarely
##   normal      0          1     2           10     13
```

```
##   good        3         10     3           30     16
```

```
# Perform a chi-squared test of independence
chi_squared_performance_help_from_professors<-
  chisq.test(performance_help_from_professors)
```

```
## Warning in chisq.test(performance_help_from_professors): Chi-squared
## approximation may be incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_help_from_professors)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_help_from_professors
## X-squared = 7.3822, df = 4, p-value = 0.117
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$help_from_professors, correct = TRUE)
```

```
## [1] 0.3934349
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$help_from_professors))
```

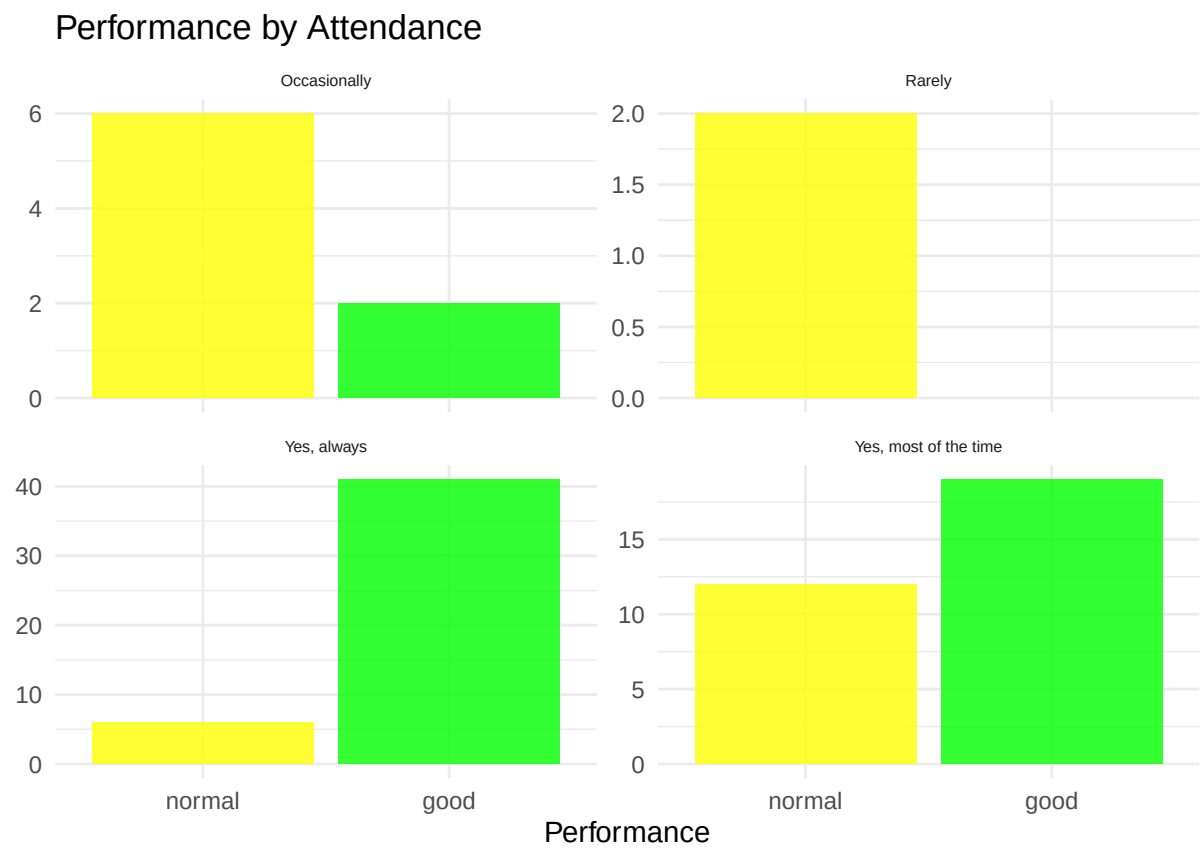
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 8.5150  4 0.074433
## Pearson          7.3822  4 0.117020
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.278
## Cramer's V        : 0.29
```

Overall Interpretation:

The chi-squared test examining the relationship between academic performance and the frequency of seeking help from professors yielded a p-value of 0.117, indicating that there is no statistically significant association between these two variables. The contingency coefficient and Cramer's V values suggest a moderate association, but the cautionary note about the chi-squared approximation being incorrect should be taken into consideration, potentially due to a smaller sample size.

Performance and Attendance

```
# Bar chart for attendance
ggplot(data_no_na_performance, aes(x = performance, fill=performance))+
  geom_bar(alpha = 0.8)+
  labs(title = "Performance by Attendance",
       x = "Performance",
       y = "") +
  theme_minimal() +
    theme(legend.position = "none",
        strip.text = element_text(size = 6)) +  # Adjust facet text size and angle
  facet_wrap(~attendance, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_attendance <- table(data_no_na_performance$performance,
                                data_no_na_performance$attendance)

# Display the contingency table
print(performance_attendance)
```

```
##
##          Occasionally Rarely Yes, always Yes, most of the time
##   normal            6      2          6                    12
```

```
##   good              2      0          41                    19
```

```
# Perform a chi-squared test of independence
chi_squared_performance_attendance<-
  chisq.test(performance_attendance)
```

```
## Warning in chisq.test(performance_attendance): Chi-squared approximation may be
## incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_attendance)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_attendance
## X-squared = 20.317, df = 3, p-value = 0.0001459
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$attendance, correct = TRUE)
```

```
## [1] 0.6124918
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$attendance))
```

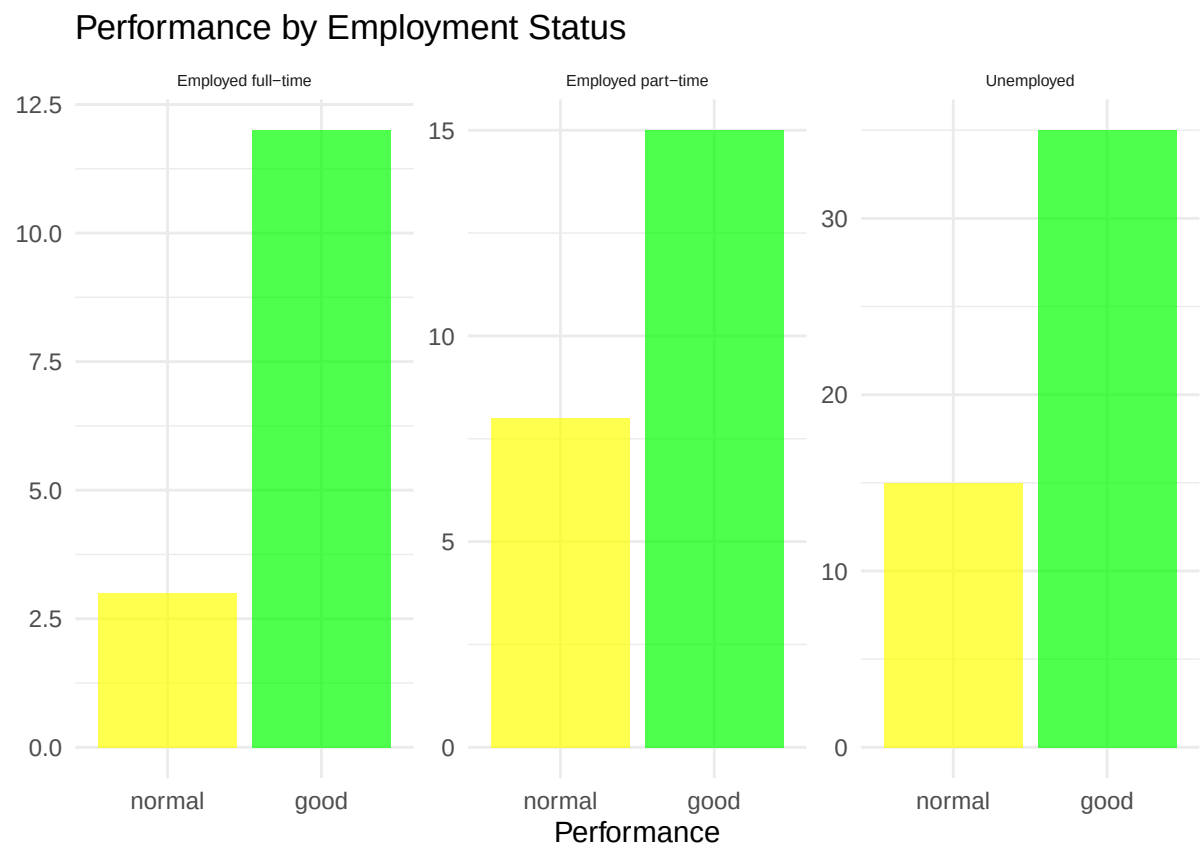
```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 20.548  3 0.00013069
## Pearson          20.317  3 0.00014588
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.433
## Cramer's V        : 0.48
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and attendance revealed a highly significant p-value of 0.0001459, indicating a strong relationship between these variables. Both the contingency coefficient (0.433) and Cramer's V (0.48) underscore this significant association, suggesting that attendance plays a substantial role in predicting academic performance. The warning about the chi-squared approximation being incorrect might be due to the sample size, emphasizing the need for cautious interpretation. Nonetheless, based on this analysis, maintaining regular attendance seems to be a crucial factor positively impacting academic performance.

Performance and Employment Status

```
# Stacked bar chart for employment status
ggplot(data_no_na_performance, aes(x = performance, fill = performance)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Performance by Employment Status",
       x = "Performance",
       y = "") +
  theme_minimal() +
  theme(legend.position = "none",
        strip.text = element_text(size = 6)) +  # Adjust facet text size and angle
  facet_wrap(~employment_status, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_employment_status <- table(data_no_na_performance$performance,
                          data_no_na_performance$employment_status)

# Display the contingency table
print(performance_employment_status)
```

```
##
##          Employed full-time Employed part-time Unemployed
##   normal                  3                  8         15
```

```
##   good                  12                 15         35
```

```
# Perform a chi-squared test of independence
chi_squared_performance_employment_status<-
  chisq.test(performance_employment_status)
```

```
## Warning in chisq.test(performance_employment_status): Chi-squared approximation
## may be incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_employment_status)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_employment_status
## X-squared = 0.96459, df = 2, p-value = 0.6174
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$employment_status, correct = TRUE)
```

```
## [1] 0.1472577
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$employment_status))
```

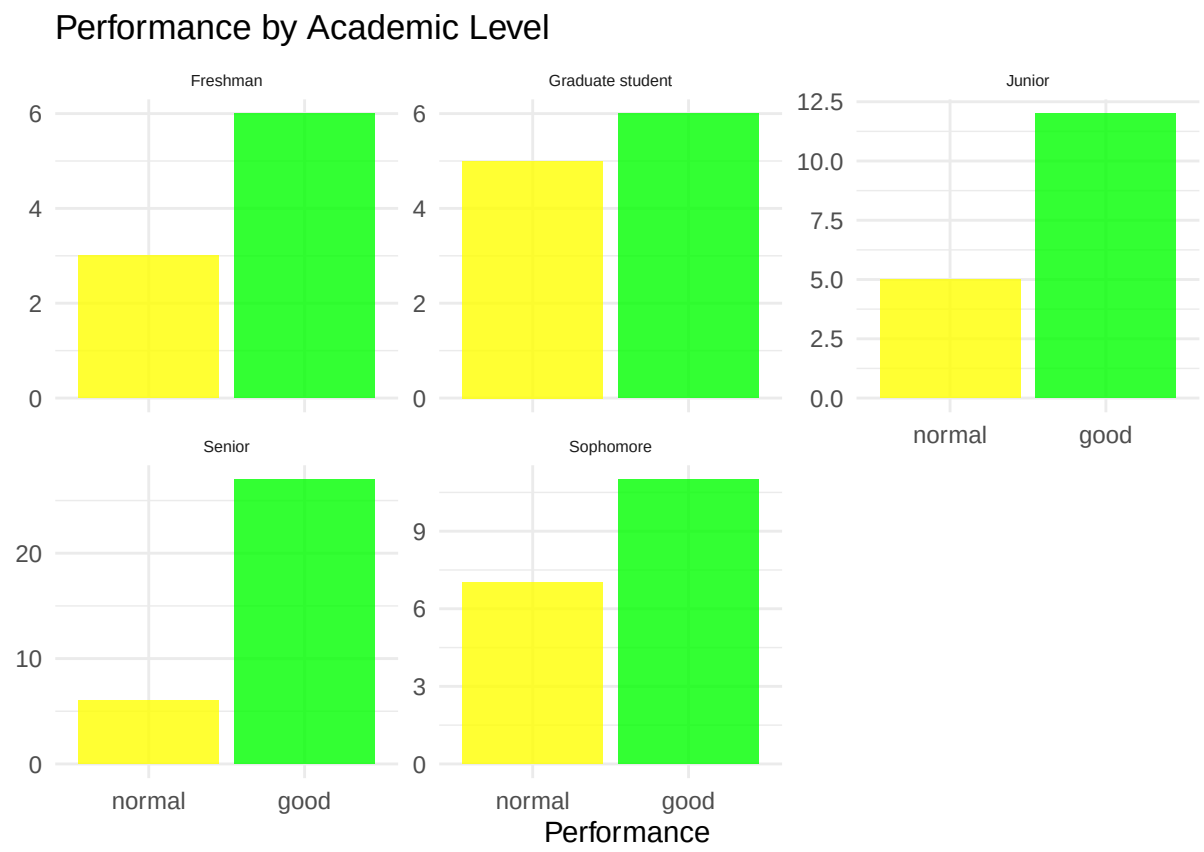
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 1.00693  2  0.60443
## Pearson          0.96459  2  0.61736
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.104
## Cramer's V        : 0.105
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and employment status resulted in a p-value of 0.6174, indicating no significant relationship between these variables. The contingency coefficient (0.104) and Cramer's V (0.105) values further support the notion of a weak association.

Performance and Academic Level

```
# Stacked bar chart for academic level
ggplot(data_no_na_performance, aes(x = performance, fill=performance))+
  geom_bar(alpha = 0.8)+
  labs(title = "Performance by Academic Level",
       x = "Performance",
       y = "") +
  theme_minimal() +
    theme(legend.position = "none",
        strip.text = element_text(size = 6)) +  # Adjust facet text size and angle
  facet_wrap(~academic_level, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_academic_level <- table(data_no_na_performance$performance,
                                    data_no_na_performance$academic_level)

# Display the contingency table
print(performance_academic_level)
```

```
##
##          Freshman Graduate student Junior Senior Sophomore
##   normal        3               5      5      6         7
```

```
##   good          6                6     12     27        11
```

```
# Perform a chi-squared test of independence
chi_squared_performance_academic_level<-
  chisq.test(performance_academic_level)
```

```
## Warning in chisq.test(performance_academic_level): Chi-squared approximation may
## be incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_academic_level)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_academic_level
## X-squared = 4.2017, df = 4, p-value = 0.3794
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$academic_level, correct = TRUE)
```

```
## [1] 0.3018961
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$academic_level))
```

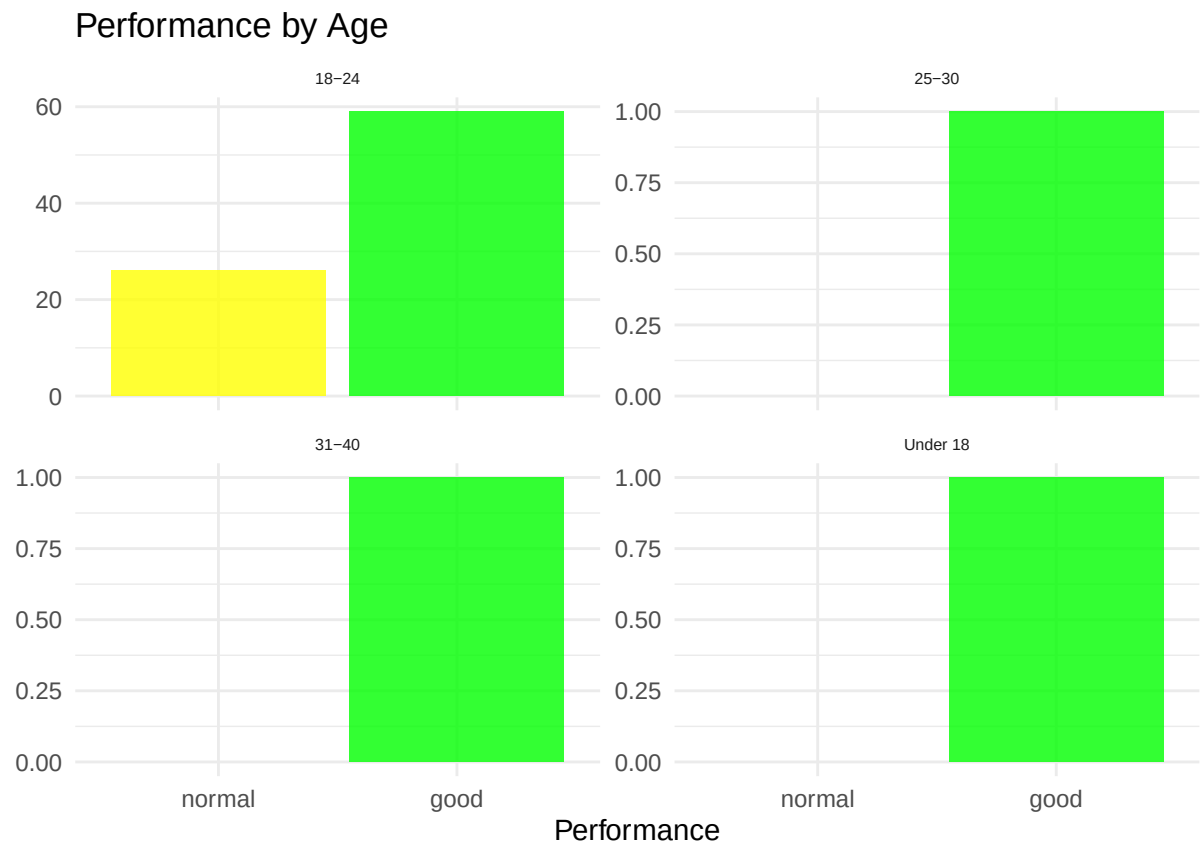
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 4.2629  4   0.3716
## Pearson          4.2017  4   0.3794
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.213
## Cramer's V        : 0.219
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and academic level resulted in a p-value of 0.3794, indicating no significant relationship between these variables. The contingency coefficient (0.213) and Cramer's V (0.219) values further support the notion of a weak association.

Performance and Age

```
# Stacked bar chart for age
ggplot(data_no_na_performance, aes(x = performance, fill=performance))+
  geom_bar(alpha = 0.8)+
  labs(title = "Performance by Age",
       x = "Performance",
       y = "") +
  theme_minimal() +
    theme(legend.position = "none",
        strip.text = element_text(size = 6)) +  # Adjust facet text size and angle
  facet_wrap(~age, scales = "free_y") +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_age <- table(data_no_na_performance$performance,
                         data_no_na_performance$age)

# Display the contingency table
print(performance_age)
```

```
##
##          18-24 25-30 31-40 Under 18
##   normal    26     0     0        0
```

```
##   good      59     1     1        1
```

```
# Perform a chi-squared test of independence
chi_squared_performance_age<-
  chisq.test(performance_age)
```

```
## Warning in chisq.test(performance_age): Chi-squared approximation may be
## incorrect
```

```
# Display the chi-squared test result
print(chi_squared_performance_age)
```

```
##
##  Pearson's Chi-squared test
##
## data:  performance_age
## X-squared = 1.3025, df = 3, p-value = 0.7285
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$age, correct = TRUE)
```

```
## [1] 0.1707916
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$age))
```

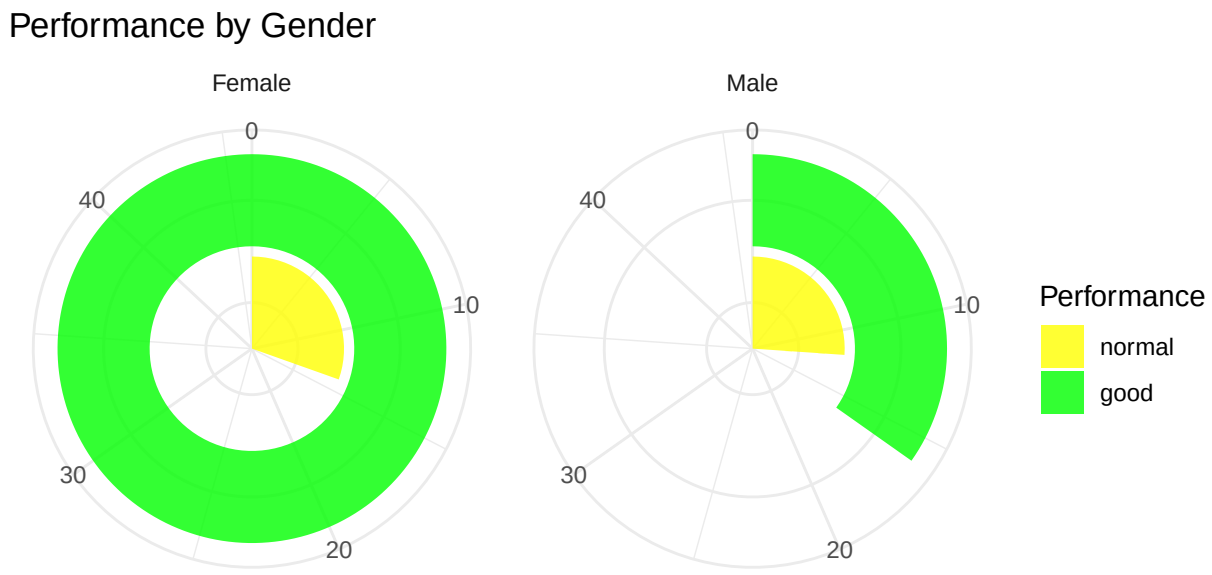
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.1453  3  0.54280
## Pearson          1.3025  3  0.72855
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.121
## Cramer's V        : 0.122
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and age resulted in a p-value of 0.7285, indicating no significant relationship between these variables. The contingency coefficient (0.121) and Cramer's V (0.122) values suggest a weak association.

Performance and Gender

```
# Stacked bar chart for gender
ggplot(data_no_na_performance, aes(x = performance, fill=performance))+
  geom_bar(alpha = 0.8)+
  coord_polar("y") +
  labs(title = "Performance by Gender",
       x = "",
       y = "",
       fill = "Performance") +
  theme_minimal() +
    theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        strip.text.y = element_text(size = 7, angle = 0)) +
  facet_wrap(~gender) +
  scale_fill_manual(values = c("yellow", "green"))
```




```
# Create a contingency table
performance_gender <- table(data_no_na_performance$performance,
                            data_no_na_performance$gender)

# Display the contingency table
print(performance_gender)
```

```
##
##          Female Male
##   normal     14   12
##   good       46   16
```

```
# Perform a chi-squared test of independence
chi_squared_performance_gender<-
  chisq.test(performance_gender)

# Display the chi-squared test result
print(chi_squared_performance_gender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  performance_gender
## X-squared = 2.6209, df = 1, p-value = 0.1055
```

```
# Contingency coefficient C
ContCoef(data_no_na_performance$performance,
         data_no_na_performance$gender, correct = TRUE)
```

```
## [1] 0.2764345
```

```
#Cramer's V and more
assocstats(xtabs(~data_no_na_performance$performance +
                   data_no_na_performance$gender))
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 3.3900  1 0.065593
## Pearson          3.4959  1 0.061522
##
## Phi-Coefficient   : 0.199
## Contingency Coeff.: 0.195
## Cramer's V        : 0.199
```

Overall Interpretation:

The chi-squared test examining the association between academic performance and gender yielded a p-value of 0.1055, suggesting no statistically significant association between these variables. The contingency coefficient (0.195) and Cramer's V (0.199) values indicate a weak association.

Conclusion

Overall Conclusion on Associations with Performance: In the analyses exploring the association between academic performance and various factors, including study-related habits, lifestyle choices, and demographic variables, the chi-squared tests consistently yielded p-values suggesting no statistically significant association. The contingency coefficients and Cramer's V values were generally low, indicating weak associations.

Study-Related Factors:

Study Hours: No significant association (p = 0.5622).

Extracurricular Activities: No significant association (p = 0.2082).

Breaks: No significant association (p = 0.5044).

Studying in Groups: Marginally significant association (p = 0.07884).

Help from Professors: No significant association (p = 0.117).

Attendance: Significant association (p = 0.0001459).

Lifestyle Choices:

Physical Activity Frequency: No significant association (p = 0.5695).

Caffeine Consumption: No significant association (p = 0.5941).

Eating Frequency: No significant association (p = 0.7298).

Eating Habits: No significant association (p = 0.3657).

Average Sleep: No significant association (p = 0.507).

Future Plans and Goals:

Future Education Plans: No significant association (p = 0.264).

Academic Goals: Marginally significant association (p = 0.06261).

Other Demographic Factors:

Age: No significant association (p = 0.7285).

Gender: No significant association (p = 0.1055).

Academic Level: No significant association (p = 0.3794).

Employment Status: No significant association (p = 0.6174).

In summary, based on the analyses, there is a lack of compelling evidence to suggest a strong association between academic performance and the examined variables. It's important to note the potential limitations of these findings, such as the small sample size and the assumption of independence between observations. Future investigations with larger and more diverse samples may provide more robust insights into the complex interplay of factors influencing academic performance.

Linear Regression

```r
# Create a linear regression model
#Through iterative training process this is our best model

linear_model <-
  lm(gpa ~ .,data =data_no_na_performance[, c("gpa", "attendance",
                                              "academic_goals",
                                              "studying_in_groups",
                                              "gender",
                                              "study_time_Night",
                                              "resources_Interactive simulations or applications",
                                              "consumed_beverages_Coffee",
                                              "manage_stress_Meditation",
                                              "manage_stress_Other")])

# Display the summary of the linear regression model
summary(linear_model)
```

```
##
## Call:
## lm(formula = gpa ~ ., data = data_no_na_performance[, c("gpa",
##     "attendance", "academic_goals", "studying_in_groups", "gender",
##     "study_time_Night", "resources_Interactive simulations or applications",
##     "consumed_beverages_Coffee", "manage_stress_Meditation",
##     "manage_stress_Other")])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27136 -0.21436  0.01966  0.22973  0.80403
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                               2.32465    0.24095
## attendanceRarely                                         -0.20938    0.35193
## attendanceYes, always                                     0.77279    0.17887
## attendanceYes, most of the time                           0.47738    0.18611
## academic_goalsImprove grades in specific subjects        -0.30964    0.13704
## academic_goalsMaintain current academic standing         -0.03141    0.14556
## academic_goalsOther                                      -0.09110    0.22685
## academic_goalsPursue advanced coursework or certifications  0.19828    0.15393
## studying_in_groupsPrefer studying alone                   0.27055    0.10067
## studying_in_groupsPrefer studying in a group regularly    0.26994    0.22684
## genderMale                                               -0.10224    0.10768
## study_time_Night1                                         0.31634    0.09920
## `resources_Interactive simulations or applications`1     -0.22185    0.12374
## consumed_beverages_Coffee1                                0.21302    0.10202
## manage_stress_Meditation1                                 0.33948    0.13980
## manage_stress_Other1                                      0.41063    0.16121
##                                                          t value Pr(>|t|)
## (Intercept)                                                9.648 1.31e-14 ***
## attendanceRarely                                          -0.595  0.55375
```

```
## attendanceYes, always                                          4.320 4.91e-05 ***
## attendanceYes, most of the time                                2.565  0.01240 *
## academic_goalsImprove grades in specific subjects             -2.259  0.02688 *
## academic_goalsMaintain current academic standing              -0.216  0.82975
## academic_goalsOther                                           -0.402  0.68919
## academic_goalsPursue advanced coursework or certifications     1.288  0.20184
## studying_in_groupsPrefer studying alone                        2.687  0.00894 **
## studying_in_groupsPrefer studying in a group regularly         1.190  0.23796
## genderMale                                                    -0.949  0.34556
## study_time_Night1                                              3.189  0.00211 **
## `resources_Interactive simulations or applications`1          -1.793  0.07720 .
## consumed_beverages_Coffee1                                     2.088  0.04035 *
## manage_stress_Meditation1                                      2.428  0.01767 *
## manage_stress_Other1                                           2.547  0.01300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4208 on 72 degrees of freedom
## Multiple R-squared:  0.5568, Adjusted R-squared:  0.4645
## F-statistic:  6.03 on 15 and 72 DF,  p-value: 6.436e-08
```

Interpretation:

Students who always or most of the time attend to classes, prefer studying alone and at night time, manage stress through meditation or in other alternative ways, and consume coffee tend to have higher GPAs, according to this model.

Coefficients:

Intercept (2.32465):

- The intercept represents the estimated GPA when all predictor variables are zero.
- In this context, it is the estimated GPA for a hypothetical student who occasionally attends classes, has no specific academic goals, prefers occasionally studying in groups, is female, doesn't study at night, and doesn't use interactive simulations or applications, coffee, meditation, or other stress management methods.

Attendance:

- "attendanceYes, always" (0.77279): Students who always attend have, on average, a GPA higher by 0.77 compared to those who rarely attend.
- "attendanceYes, most of the time" (0.47738): Students who attend most of the time have, on average, a GPA higher by 0.48 compared to those who attend occasionally.

Academic Goals:

- "academic_goalsImprove grades in specific subjects" (-0.30964): Students with the goal of improving grades in specific subjects, on average, have a lower GPA by 0.31 compared to those with other goals.

Studying in Groups:

- "studying_in_groupsPrefer studying alone" (0.27055): Students who prefer studying alone have, on average, a higher GPA by 0.27 compared to those who prefer studying in a group regularly.

Study Time at Night:

- "study_time_Night1" (0.31634): Students who study at night have, on average, a higher GPA by 0.32 compared to those who don't study at night.

Resources - Interactive Simulations or Applications:

- "resources_Interactive simulations or applications1" (-0.22185): Students who use interactive simulations or applications have, on average, a lower GPA by 0.22 compared to those who don't use these resources.

Consumed Beverages - Coffee:

- "consumed_beverages_Coffee1" (0.21302): Students who consume coffee have, on average, a higher GPA by 0.21 compared to those who don't consume coffee.

Manage Stress:

- "manage_stress_Meditation1" (0.33948): Students who manage stress through meditation have, on average, a higher GPA by 0.34 compared to those who don't use meditation for stress management.
- "manage_stress_Other1" (0.41063): Students who manage stress through other methods (not specified) have, on average, a higher GPA by 0.41 compared to those who don't use other stress management methods.

Overall, the model appears to have some degree of explanatory power, as indicated by the significant F-statistic and a moderate multiple R-squared. However, the adjusted R-squared suggests that not all included predictors contribute meaningfully to the model. It's important to consider the context of the study, the assumptions of the linear regression model, and potential limitations when interpreting and applying these results. Further exploration, validation, and consideration of alternative models may be warranted.

ANOVA

An ANOVA (Analysis of Variance) test is typically used to compare the means of three or more groups to determine if there are statistically significant differences among them.

Set Up Hypotheses:

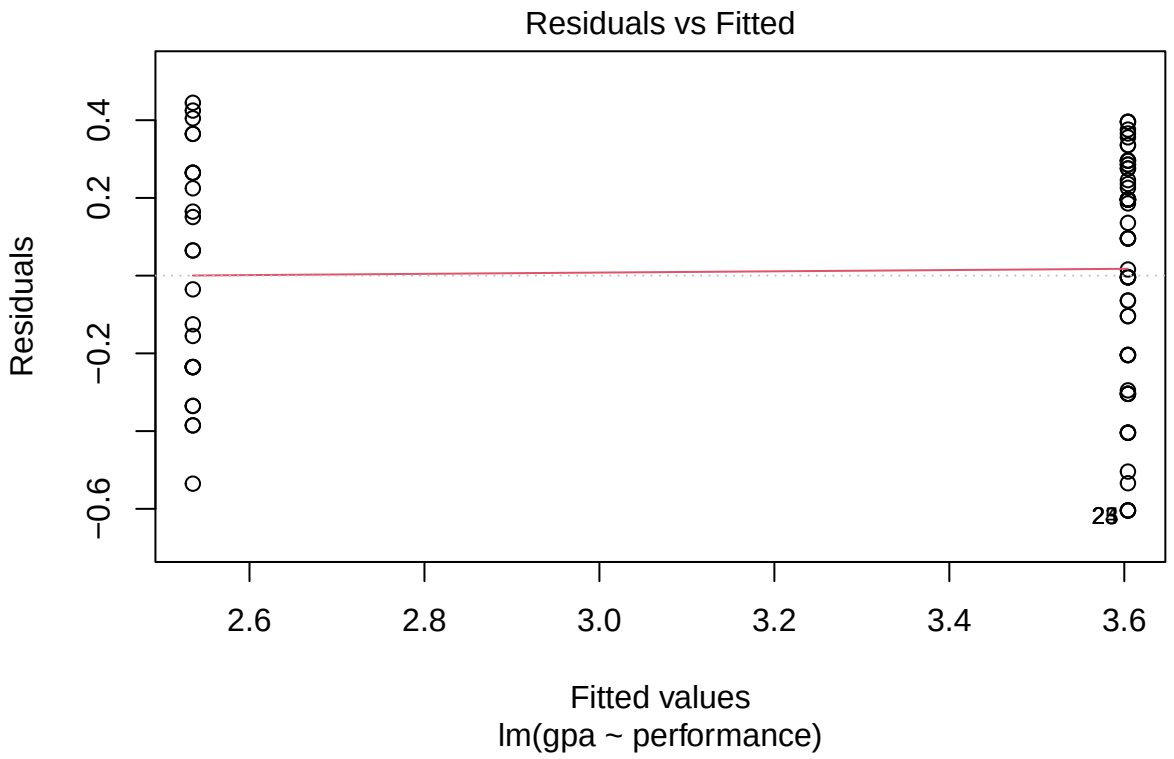
- Null Hypothesis (H0): There is no significant difference in GPA distributions among the "normal" and "good" performance groups.
- Alternative Hypothesis (Ha): There is a significant difference in GPA distributions among the "normal" and "good" performance groups.

Check Assumptions:

ANOVA assumes that the populations being compared have normal distributions and equal variances. You may want to visually inspect the distributions and consider conducting a normality test and a test for homogeneity of variances.

```
model <- lm(gpa ~ performance, data = data_no_na_performance)
```

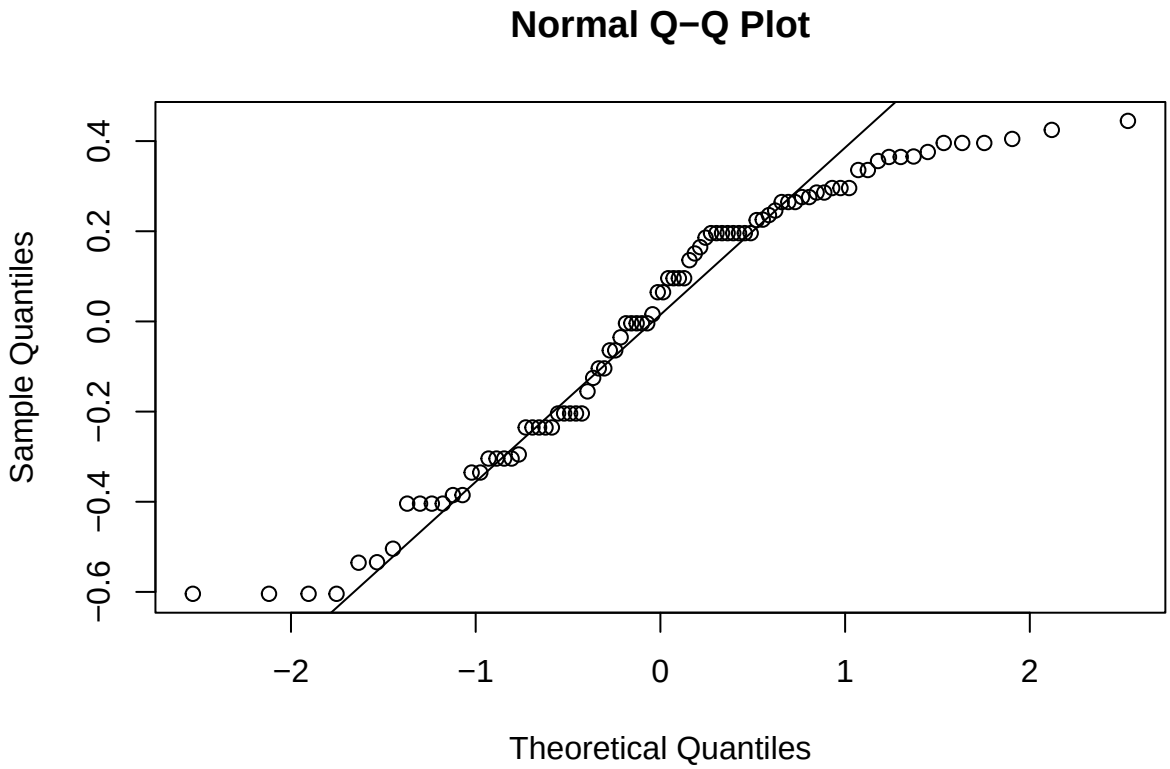
```
# 1. Check Assumption of Homogeneity of Variances
plot(model, which = 1)
```



```r
# 2. Check Assumption of Normality of Residuals
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.93559, p-value = 0.0002833
```

```r
qqnorm(residuals(model))
qqline(residuals(model)) # Check for Linearity
```



Interpretation

Shapiro-Wilk test

The Shapiro-Wilk test is a statistical test used to assess whether a sample of data comes from a normal distribution. In your output, the test statistic (W) is 0.93559, and the p-value is 0.0002833.

Here's what output suggests:

- Null Hypothesis (H0): The data follows a normal distribution.
- Alternative Hypothesis (H1): The data does not follow a normal distribution.

Since the p-value is less than the typical significance level of 0.05, you would reject the null hypothesis. In other words, there is evidence to suggest that the residuals from your model are not normally distributed.

Assumption of Homogeneity of Variances

The assumption of homogeneity of variances (HOV) states that the variance of the dependent variable is the same across all levels of the independent variable. This assumption is important for many statistical tests, such as the independent samples t-test and ANOVA.

One way to check for HOV is to look at a residuals versus fitted values plot. If the points are evenly distributed around the zero line, then HOV is likely met. However, if there is a fan-shaped pattern, then HOV is likely violated.

The residuals versus fitted values plot shows a fan-shaped pattern, indicating that HOV is likely violated. This means that the variance of the dependent variable (GPA performance) is not the same across all levels of the independent variable (fitted values).

We plan to use a statistical test that assumes HOV, we should take steps to address the violation.

QQ-plot

The normal Q-Q plot does not show a straight line, which indicates that the linearity assumption is likely violated. This means that the relationship between the independent and dependent variables is not linear.

Box-Cox Transformation

The Box-Cox transformation is a family of power transformations that are used to stabilize the variance and make a dataset more closely approximate a normal distribution. It is particularly useful when dealing with data that violates the assumption of constant variance (heteroscedasticity) or non-normality.

The Box-Cox transformation is defined as:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

Here, y is the original response variable, and λ is the transformation parameter. The optimal value of λ is chosen to maximize the log-likelihood function, resulting in the most normal and homoscedastic residuals.

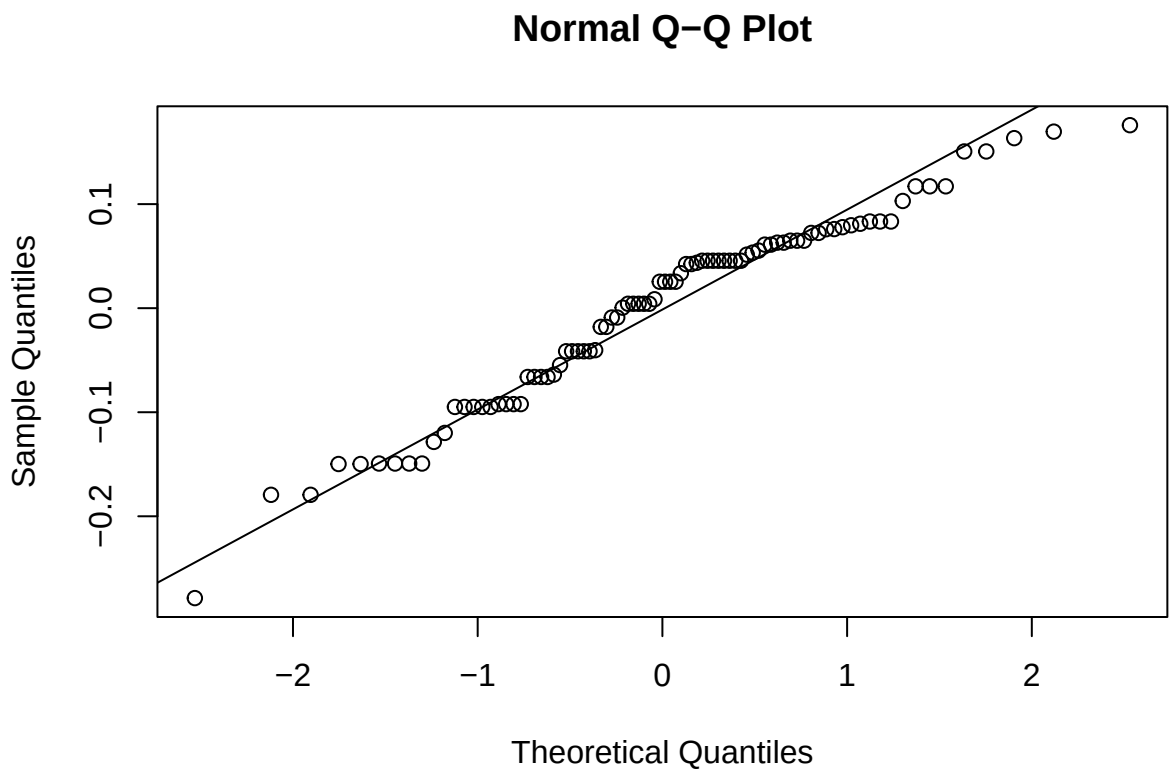
```r
# Perform double normalization transformation
model <- lm(BoxCox(BoxCox(gpa, lambda = 0), lambda = 0) ~ performance, data = data_no_na_performance)

# Shapiro-Wilk normality test on residuals
shapiro_test <- shapiro.test(residuals(model))
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.9656, p-value = 0.01945
```

Shapiro-Wilk normality test

W statistic is 0.9656 and the p-value is 0.01945. This means that there is some evidence to suggest that the data is not normally distributed. However, the p-value is not quite as low as 0.05, which is the conventional level of significance. This means that you cannot reject the null hypothesis that the data is normally distributed with absolute certainty.

```r
qqnorm(residuals(model))
qqline(residuals(model)) # Check for Linearity
```



Perform ANOVA

```
anova_result <- anova(model)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Response: BoxCox(BoxCox(gpa, lambda = 0), lambda = 0)
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## performance  1 2.01004 2.01004     237 < 2.2e-16 ***
## Residuals   86 0.72938 0.00848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

The analysis suggests that there is a significant effect of the 'performance' variable on the Box-Cox transformed GPA. The low p-value (< 0.001) indicates strong evidence against the null hypothesis, supporting the conclusion that there is a meaningful relationship. There is a significant difference in GPA distributions among the "normal" and "good" performance groups. The F-statistic of 237 further reinforces the strength of this relationship.