The Effects of Alcohol

2022-12-31

Data Access

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Load the data
data <- read_csv("Stats survey.csv")
```

```
## Rows: 406 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (15): Timestamp, Your Sex?, What year were you in last year (2023) ?, Wh...
## dbl  (2): Your Matric (grade 12) Average/ GPA (in %), Your 2023 academic yea...
##
## 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.
```

```
# Inspect and view the data structure
glimpse(data)
```

```
## Rows: 406
## Columns: 17
## $ Timestamp
## $ `Your Sex?`
```

```
## $ `Your Matric (grade 12) Average/ GPA (in %)`
## $ `What year were you in last year (2023) ?`
## $ `What faculty does your degree fall under?`
## $ `Your 2023 academic year average/GPA in % (Ignore if you are 2024 1st year student)`
## $ `Your Accommodation Status Last Year (2023)`
## $ `Monthly Allowance in 2023`
## $ `Were you on scholarship/bursary in 2023?`
## $ `Additional amount of studying (in hrs) per week`
## $ `How often do you go out partying/socialising during the week?`
## $ `On a night out, how many alcoholic drinks do you consume?`
## $ `How many classes do you miss per week due to alcohol reasons, (i.e: being hungover or too tired?)
## $ `How many modules have you failed thus far into your studies?`
## $ `Are you currently in a romantic relationship?`
## $ `Do your parents approve alcohol consumption?`
## $ `How strong is your relationship with your parent/s?`
```

```r
names(data)
```

```
##  [1] "Timestamp"
##  [2] "Your Sex?"
##  [3] "Your Matric (grade 12) Average/ GPA (in %)"
##  [4] "What year were you in last year (2023) ?"
##  [5] "What faculty does your degree fall under?"
##  [6] "Your 2023 academic year average/GPA in % (Ignore if you are 2024 1st year student)"
##  [7] "Your Accommodation Status Last Year (2023)"
##  [8] "Monthly Allowance in 2023"
##  [9] "Were you on scholarship/bursary in 2023?"
## [10] "Additional amount of studying (in hrs) per week"
## [11] "How often do you go out partying/socialising during the week?"
## [12] "On a night out, how many alcoholic drinks do you consume?"
## [13] "How many classes do you miss per week due to alcohol reasons, (i.e: being hungover or too tire
## [14] "How many modules have you failed thus far into your studies?"
## [15] "Are you currently in a romantic relationship?"
## [16] "Do your parents approve alcohol consumption?"
## [17] "How strong is your relationship with your parent/s?"
```

#Data Wrangling

```r
# Renaming and data transformation
data <- rename(data,
               Year_in_2023 = `What year were you in last year (2023) ?`,
               Drinks_Consumed = `On a night out, how many alcoholic drinks do you consume?`,
               Classes_Missed = `How many classes do you miss per week due to alcohol reasons, (i.e: be
               Socialising_Frequency = `How often do you go out partying/socialising during the week?`,
               Avg_GPA = `Your 2023 academic year average/GPA in % (Ignore if you are 2024 1st year stu

# Handling missing data and type transformations
df <- data %>%
  mutate(
    Year_in_2023 = ifelse(is.na(Year_in_2023), "HS", Year_in_2023),
    Drinks_Consumed = as.numeric(gsub("\\D", "", Drinks_Consumed)),
    Socialising_Frequency = as.numeric(case_when(
      Socialising_Frequency == "Only weekends" ~ "3",
```

```
      TRUE ~ as.character(Socialising_Frequency)
    ))
  )
```

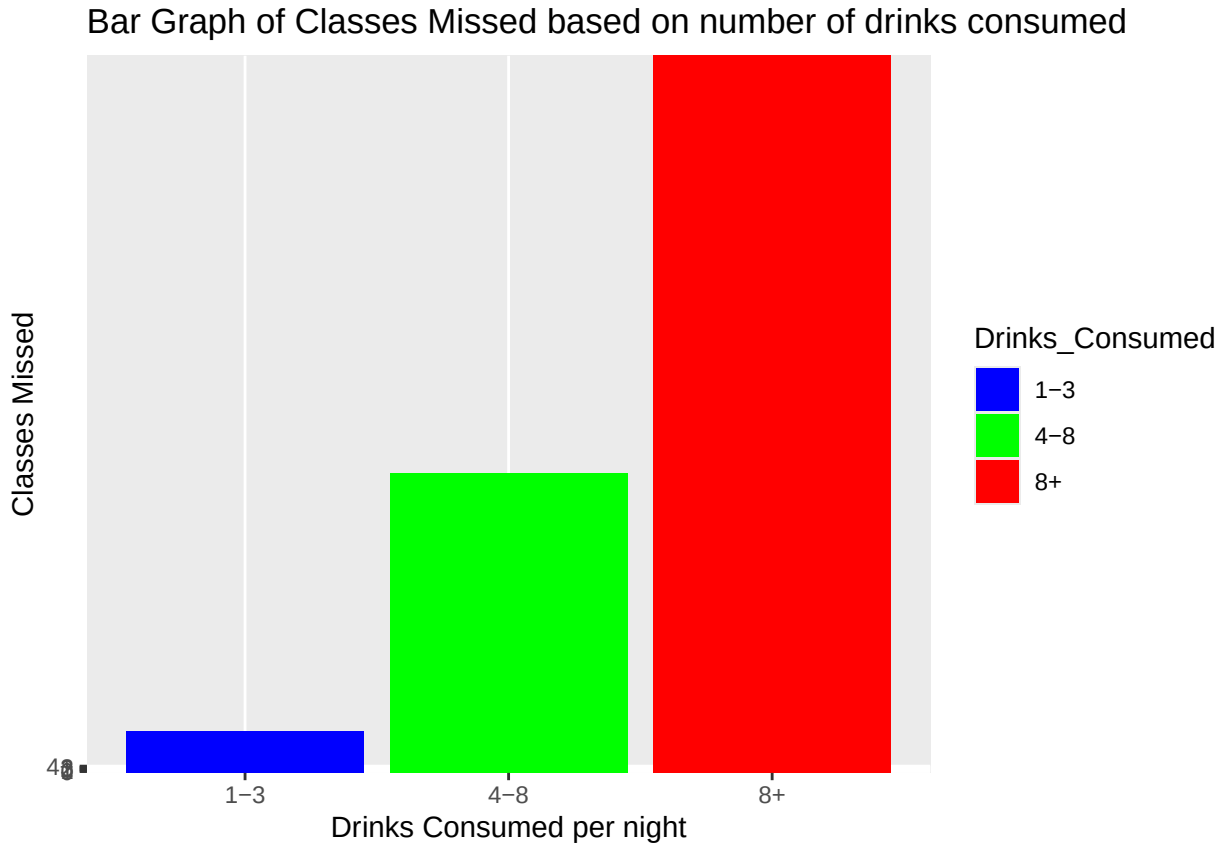
```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Socialising_Frequency = as.numeric(...)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Check the transformed data
glimpse(df)
```

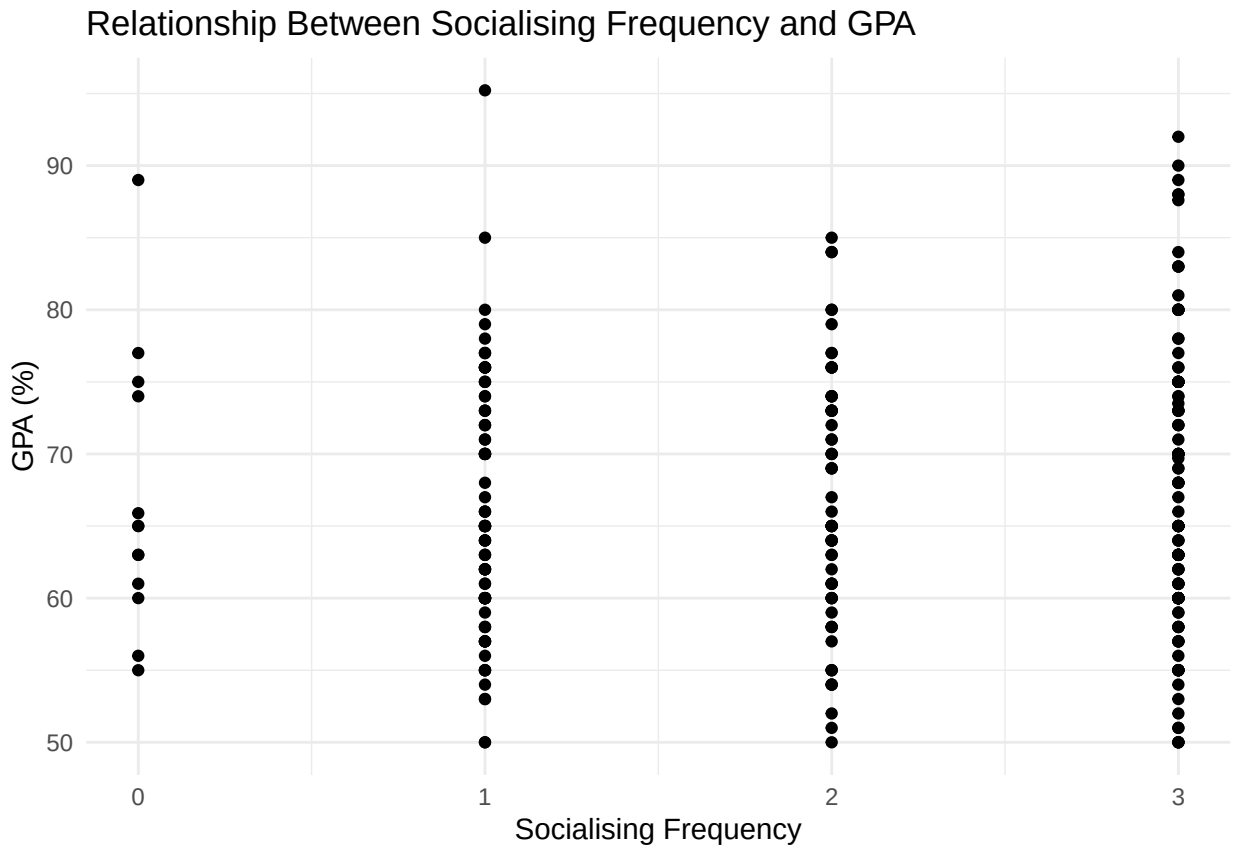
```
## Rows: 406
## Columns: 17
## $ Timestamp                                                    <chr> "2024/0~
## $ `Your Sex?`                                                  <chr> "Female~
## $ `Your Matric (grade 12) Average/ GPA (in %)`                 <dbl> 76, 89,~
## $ Year_in_2023                                                 <chr> "2nd Ye~
## $ `What faculty does your degree fall under?`                  <chr> "Arts &~
## $ Avg_GPA                                                      <dbl> 72, 75,~
## $ `Your Accommodation Status Last Year (2023)`                 <chr> "Privat~
## $ `Monthly Allowance in 2023`                                  <chr> "R 4001~
## $ `Were you on scholarship/bursary in 2023?`                   <chr> "No", "~
## $ `Additional amount of studying (in hrs) per week`            <chr> "8+", "~
## $ Socialising_Frequency                                        <dbl> 3, 3, 2~
## $ Drinks_Consumed                                              <dbl> 8, 35, ~
## $ Classes_Missed                                               <chr> "3", "4~
## $ `How many modules have you failed thus far into your studies?` <chr> "0", "0~
## $ `Are you currently in a romantic relationship?`              <chr> "Yes", ~
## $ `Do your parents approve alcohol consumption?`               <chr> "Yes", ~
## $ `How strong is your relationship with your parent/s?`        <chr> "Very c~
```

#Data Visulization

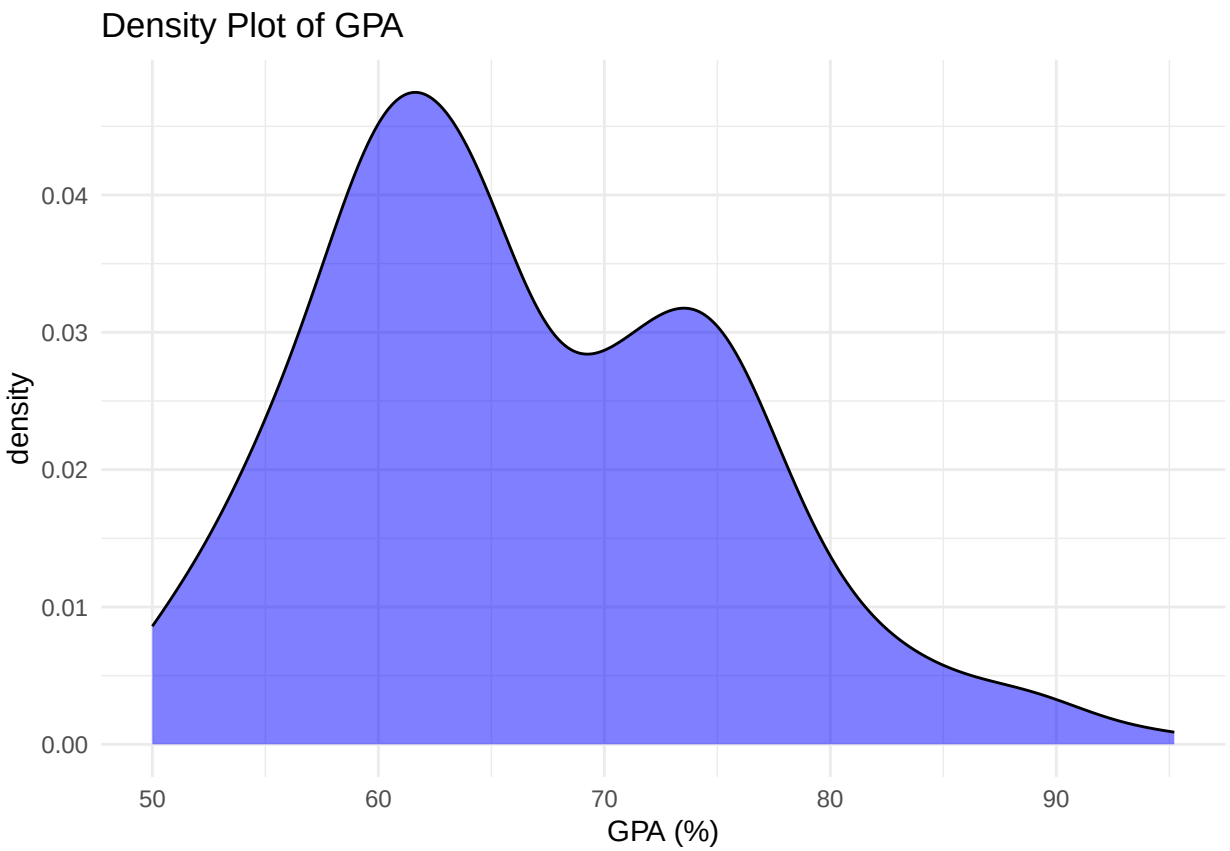
```
# Filter data and create visualizations
df <- df %>%
  filter(!is.na(Drinks_Consumed), !is.na(Classes_Missed))
df$Drinks_Consumed <- cut(df$Drinks_Consumed, breaks = c(-Inf, 3, 8, Inf), labels = c("1-3", "4-8", "8+

# Bar graph for classes missed
ggplot(df, aes(x = Drinks_Consumed, y = Classes_Missed, fill = Drinks_Consumed)) +
  geom_bar(stat = "identity") +
  labs(x = "Drinks Consumed per night", y = "Classes Missed", title = "Bar Graph of Classes Missed base
  scale_x_discrete(limits = c("1-3", "4-8", "8+")) +
  scale_fill_manual(values = c("1-3" = "blue", "4-8" = "green", "8+" = "red"))
```



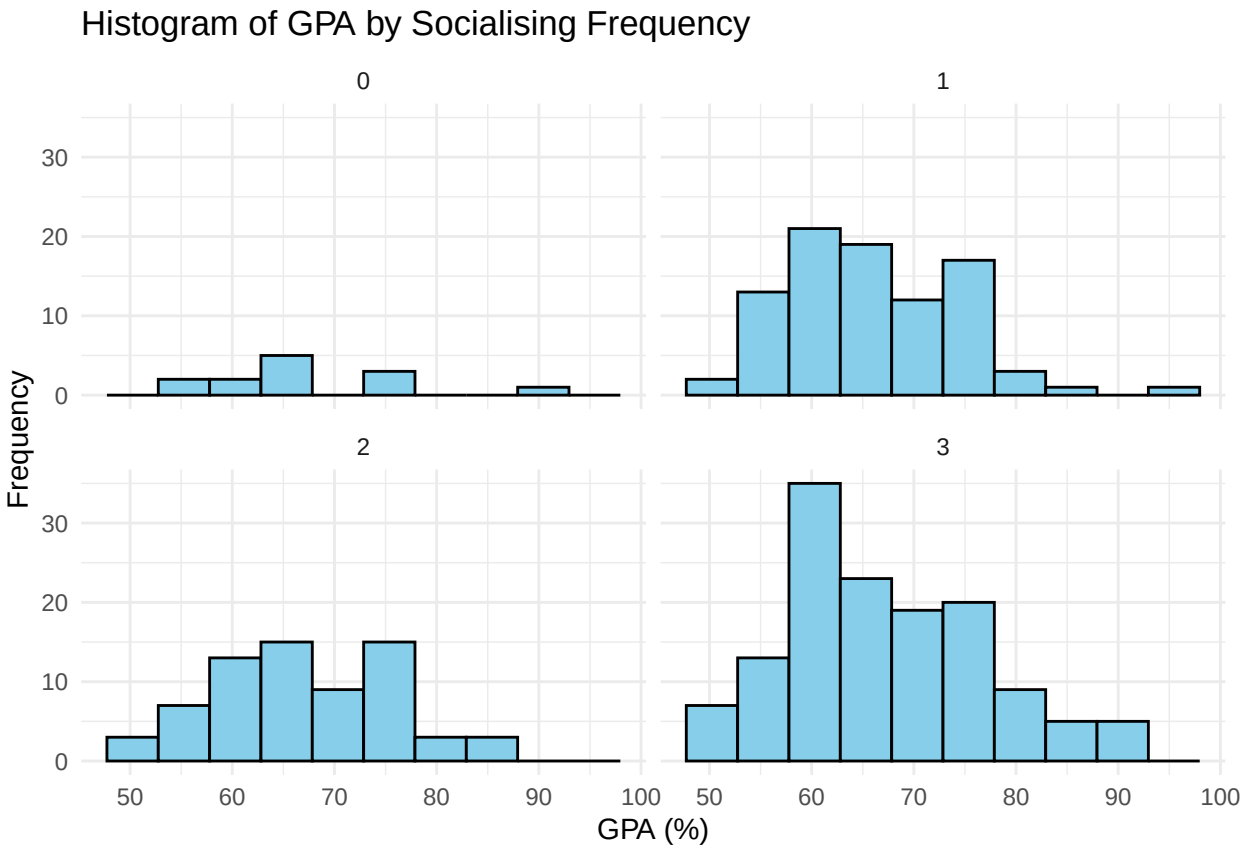

```
# Point plot for GPA and socializing frequency
cleaned_df <- df %>%
  filter(!is.na(Socialising_Frequency) & !is.na(Avg_GPA))

ggplot(cleaned_df, aes(x = Socialising_Frequency, y = Avg_GPA)) +
  geom_point() +
  labs(title = "Relationship Between Socialising Frequency and GPA", x = "Socialising Frequency", y = "
  theme_minimal()
```




```
# Density and histogram plots for GPA
ggplot(cleaned_df, aes(x = `Avg_GPA`)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Density Plot of GPA", x = "GPA (%)") +
  theme_minimal()
```




```r
ggplot(cleaned_df, aes(x = `Avg_GPA`)) +
  geom_histogram(bins = 10, fill = "skyblue", color = "black") +
  facet_wrap(~ Socialising_Frequency) +
  labs(title = "Histogram of GPA by Socialising Frequency", x = "GPA (%)", y = "Frequency") +
  theme_minimal()
```




```
# Group and summarize GPA by socializing frequency
gpa_summary <- cleaned_df %>%
  group_by(Socialising_Frequency) %>%
  summarize(Mean_GPA = mean(Avg_GPA, na.rm = TRUE), Median_GPA = median(Avg_GPA, na.rm = TRUE), Min_GPA
print(gpa_summary)
```

```
## # A tibble: 4 x 6
##   Socialising_Frequency Mean_GPA Median_GPA Min_GPA Max_GPA SD_GPA
##                   <dbl>    <dbl>      <dbl>   <dbl>   <dbl>  <dbl>
## 1                     0     66.8         65      55    89     9.52
## 2                     1     65.8         65      50    95.2   8.41
## 3                     2     66.6         65      50    85     8.60
## 4                     3     66.6         65      50    92     9.57
```

Statistical analysis

```
library(dplyr)
library(stats)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.3.3

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.3.3

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.3.3

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

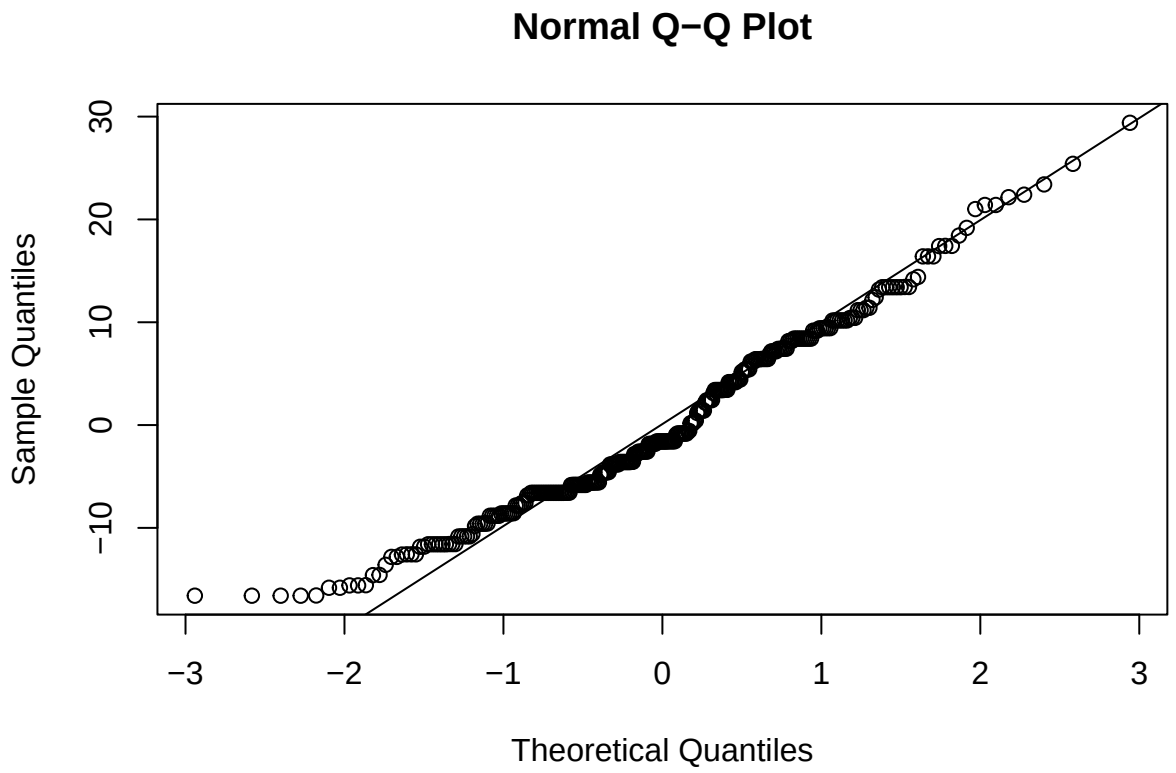
```
library(ggplot2)

# Assuming we have continuous data for GPA and want to perform ANOVA between Socialising Frequency and
prepared_data <- df %>%
  filter(!is.na(Avg_GPA) & !is.na(Socialising_Frequency)) %>%
  mutate(
    GPA = as.numeric(Avg_GPA),
    Socialising_Frequency = as.factor(Socialising_Frequency)
  )
# ANOVA analysis
anova_result <- aov(GPA ~ Socialising_Frequency, data = prepared_data)
summary(anova_result)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## Socialising_Frequency    3     39   13.05    0.16  0.923
## Residuals              302  24619   81.52
```

```
leveneTest(GPA ~ Socialising_Frequency, data = prepared_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.6603  0.577
##       302
```

```
qqnorm(residuals(anova_result))
qqline(residuals(anova_result))
```




```
# Tukey HSD test
tukey_result <- TukeyHSD(anova_result)
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = GPA ~ Socialising_Frequency, data = prepared_data)
##
## $Socialising_Frequency
```

```
##            diff       lwr      upr     p adj
## 1-0 -1.01499568 -7.940569 5.910577 0.9814842
## 2-0 -0.26416290 -7.324715 6.796389 0.9996760
## 3-0 -0.24357466 -7.014910 6.527760 0.9997120
## 2-1  0.75083278 -3.006007 4.507673 0.9551502
## 3-1  0.77142102 -2.408740 3.951582 0.9234480
## 3-2  0.02058824 -3.443696 3.484872 0.9999987
```

```r
# Boxplot for GPA distribution
ggplot(prepared_data, aes(x = Socialising_Frequency, y = GPA)) +
  geom_boxplot() +
  labs(title = "GPA Distribution Across Socialising Frequencies", x = "Socialising Frequency", y = "GPA
  theme_minimal()
```