STAT7123/STAT8123

Statistical Graphics Assignment 3

Change this to your name

Due 11:55 pm, Friday November 3rd, 2023

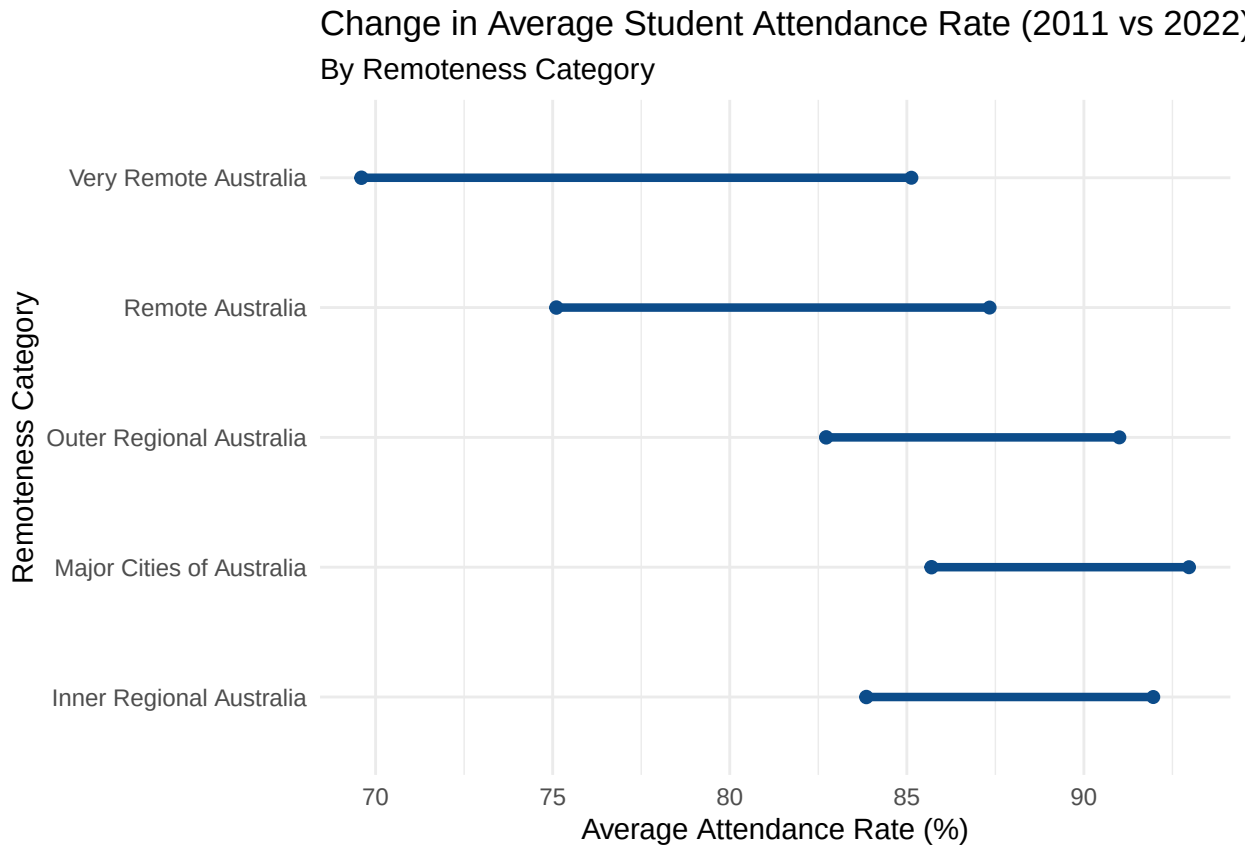
Question 1

a)

```
# Load the data
schools_data <- read_csv("school.csv")

# Calculating the averages for 2011 and 2022
attendance_avg <- schools_data %>%
  group_by(asgs_remoteness) %>%
  summarize(average_2011 = mean(attend_2011, na.rm = TRUE),
            average_2022 = mean(attend_2022, na.rm = TRUE)) %>%
  ungroup()

# Creating a dumbbell plot
ggplot(attendance_avg, aes(y = asgs_remoteness)) +
  geom_dumbbell(aes(x = average_2011, xend = average_2022),
                size = 1.5, color = "#0c4c8a") +
  labs(x = "Average Attendance Rate (%)", y = "Remoteness Category",
       title = "Change in Average Student Attendance Rate (2011 vs 2022)",
       subtitle = "By Remoteness Category") +
  theme_minimal()
```

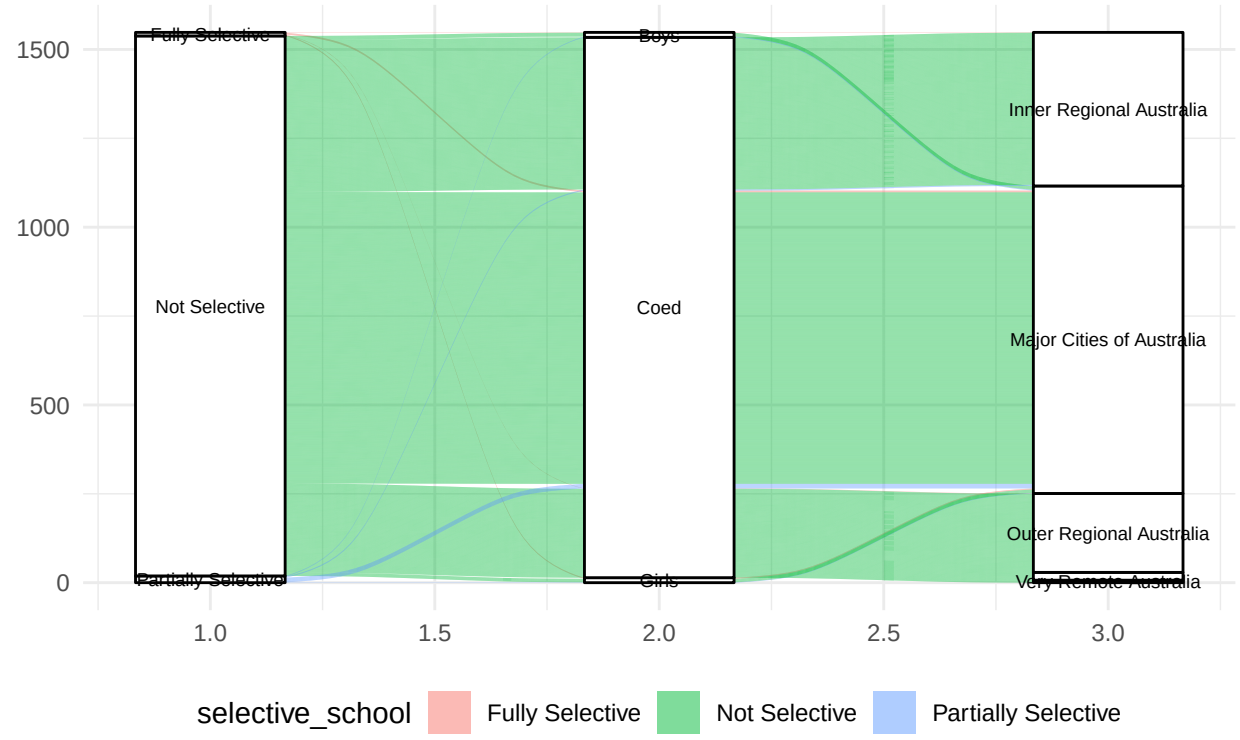


add your answers here

b)

```
schools_2022 <- schools_data %>%
  select(selective_school, school_gender, asgs_remoteness)

# Alluvial Plot
ggplot(schools_2022, aes(axis1 = selective_school, axis2 = school_gender, axis3 = asgs_remoteness)) +
  geom_alluvium(aes(fill = selective_school)) +
  geom_stratum() +
  geom_text(stat = "stratum", aes(label = after_stat(stratum)),
            size = 2.3, check_overlap = TRUE) + # further reduce size
  theme_minimal() +
  theme(legend.position = "bottom") + # move legend to the bottom
  labs(title = "Alluvial Plot of School Types in 2022",
       subtitle = "Showing relationships between selectivity, gender status, and remoteness")
```

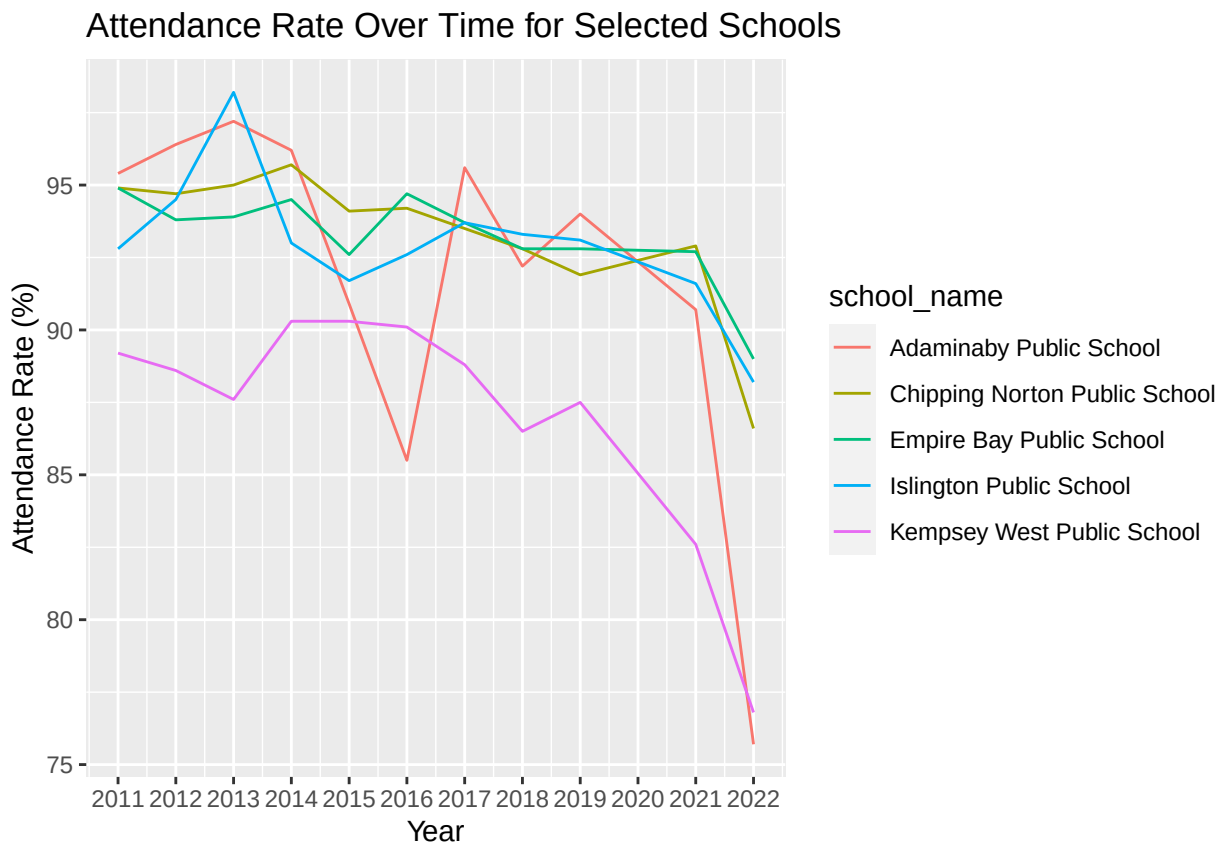


add your answers here

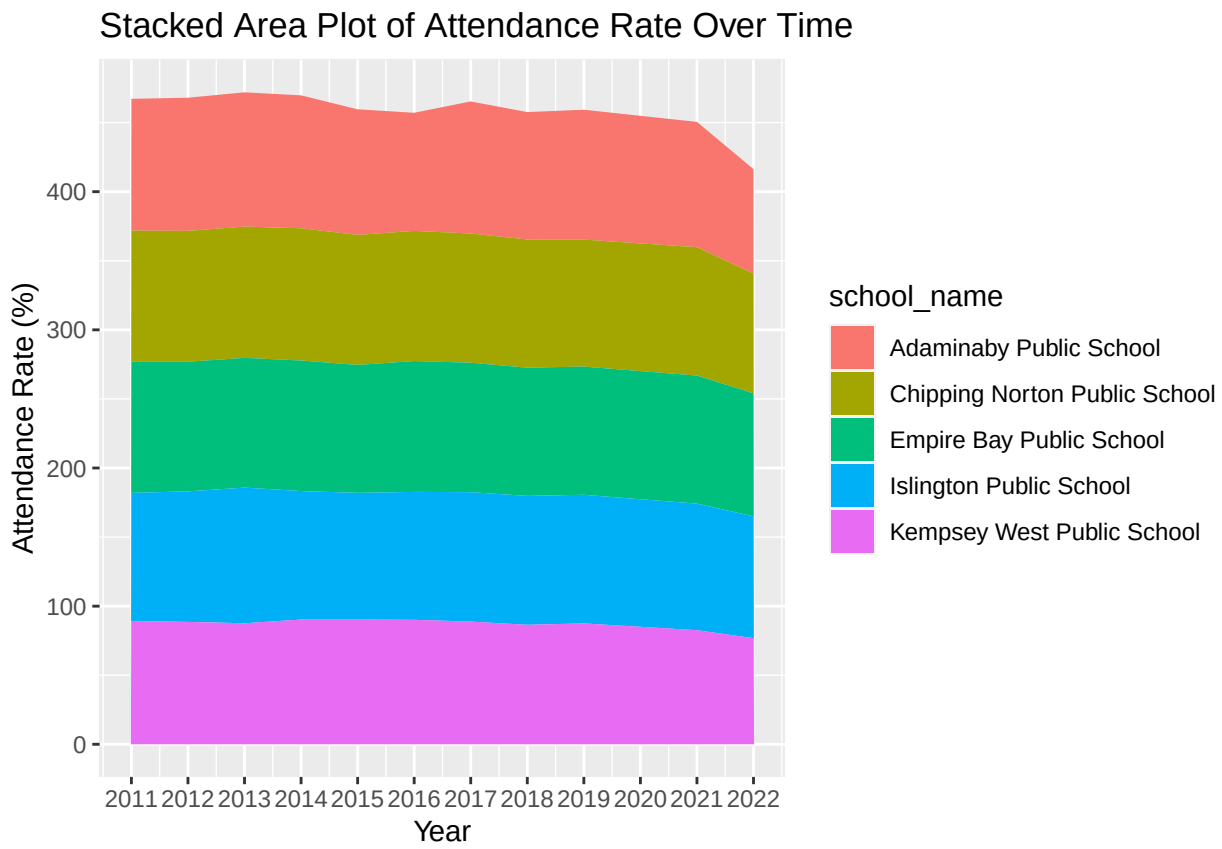
c)

```
# Selecting five schools
selected_schools <- c("Adaminaby Public School", "Chipping Norton Public School", "Empire Bay Public Sc
schools_subset <- schools_data %>%
  filter(school_name %in% selected_schools) %>%
  pivot_longer(cols = starts_with("attend_"), names_to = "year", values_to = "attendance_rate") %>%
  # Extract the numeric part of the 'year' column and convert it to numeric type
  mutate(year = as.numeric(str_extract(year, "\\d+")))

# Line plot
ggplot(schools_subset, aes(x = year, y = attendance_rate, color = school_name)) +
  geom_line() +
  labs(title = "Attendance Rate Over Time for Selected Schools",
       x = "Year", y = "Attendance Rate (%)") +
  scale_x_continuous(breaks = seq(min(schools_subset$year), max(schools_subset$year), by = 1))
```



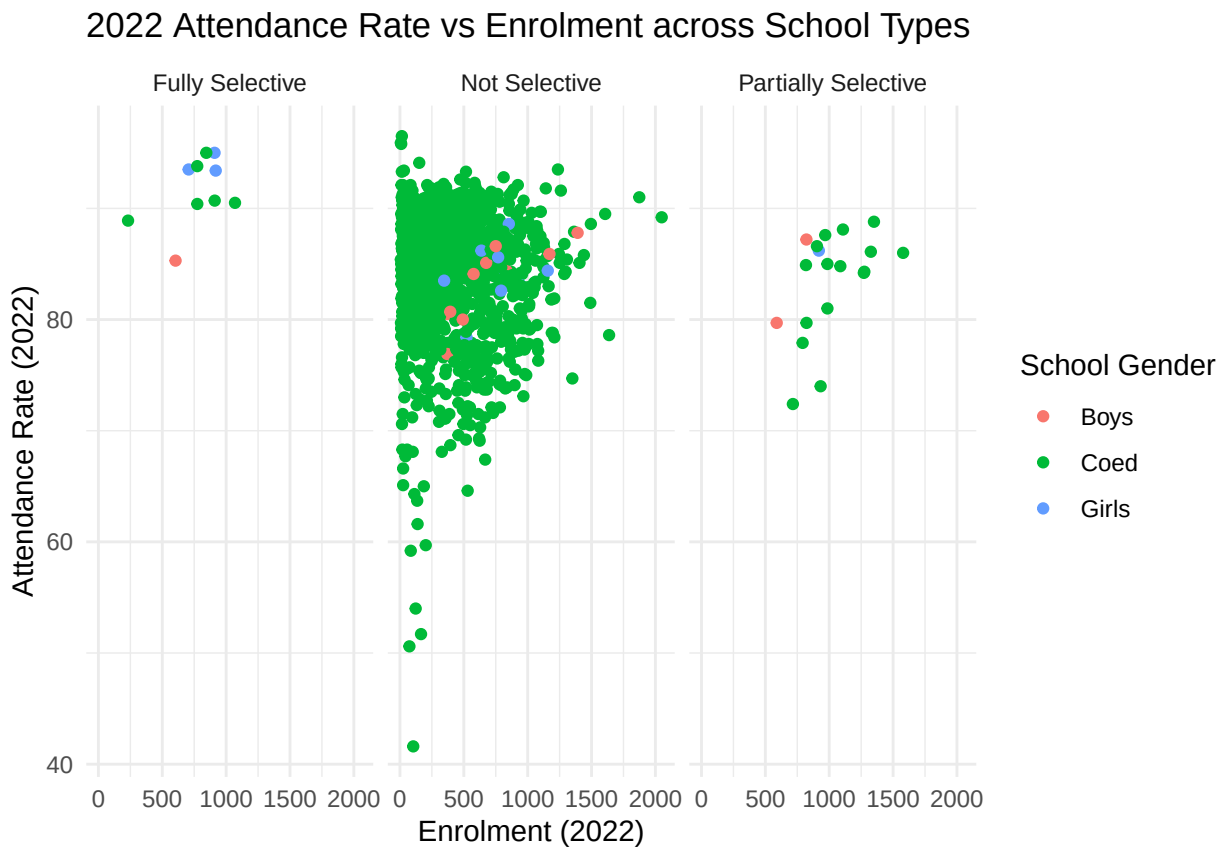

```
# Stacked area plot
ggplot(schools_subset, aes(x = year, y = attendance_rate, fill = school_name)) +
  geom_area(position = 'stack') +
  labs(title = "Stacked Area Plot of Attendance Rate Over Time",
       x = "Year", y = "Attendance Rate (%)") +
  scale_x_continuous(breaks = seq(min(schools_subset$year), max(schools_subset$year), by = 1))
```



add your answers here

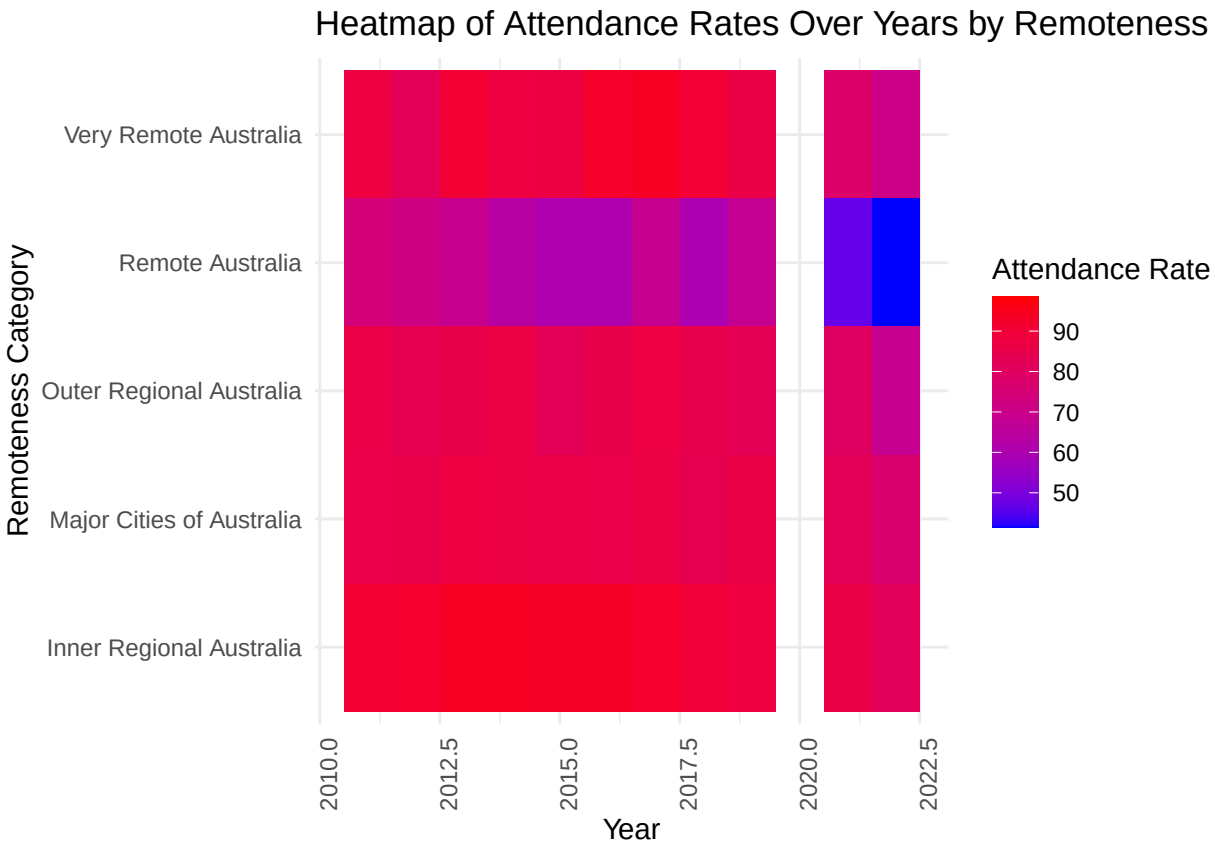
d)

```
#Faceted Scatter Plot comparing Attendance Rate and Enrolment in 2022 across Different School Types
ggplot(schools_data, aes(x = enrolment_2022, y = attend_2022, color = school_gender)) +
  geom_point() +
  facet_wrap(~selective_school) +
  labs(title = "2022 Attendance Rate vs Enrolment across School Types",
       x = "Enrolment (2022)",
       y = "Attendance Rate (2022)",
       color = "School Gender") +
  theme_minimal()
```




```r
#Heatmap showing Attendance Rates over the Years for Different Remoteness Categories
# Prepare data for the heatmap
attendance_long <- schools_data %>%
  pivot_longer(cols = starts_with("attend_"), names_to = "year", values_to = "attendance_rate") %>%
  mutate(year = as.numeric(str_replace(year, "attend_", "")))  # Converting year to numeric

ggplot(attendance_long, aes(x = year, y = asgs_remoteness, fill = attendance_rate)) +
  geom_tile() +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Heatmap of Attendance Rates Over Years by Remoteness",
       x = "Year",
       y = "Remoteness Category",
       fill = "Attendance Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels for readability
```



add your answers here

These two plots offer complexity and variety in graphical representation. The scatter plot with faceting allows examining the relationship between enrolment size and attendance rates, differentiating by school types and genders. The heatmap gives an immediate visual representation of how attendance rates have changed over years across different geographical locations.

e)

Write a concise summary of your findings from the graphical analysis. Remember, the summary should be in your words as per the note.

Question 2

a)

```
# Load the data
soccer_data <- read.csv("soccer.csv")

# Calculate the correlation matrix
cor_matrix <- cor(soccer_data[,1:12]) # Selecting only numeric variables

# Plotting the correlation matrix
corrplot(cor_matrix, method = "color", title = "Correlation Matrix for Soccer Match Data")
```

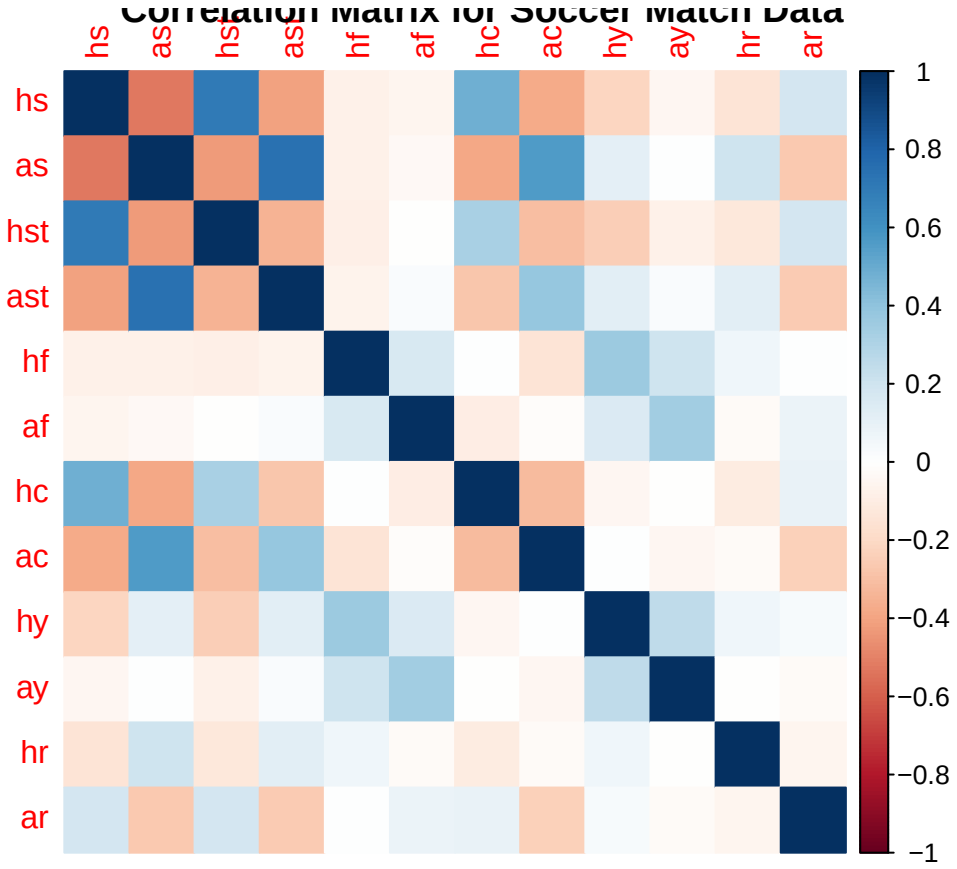
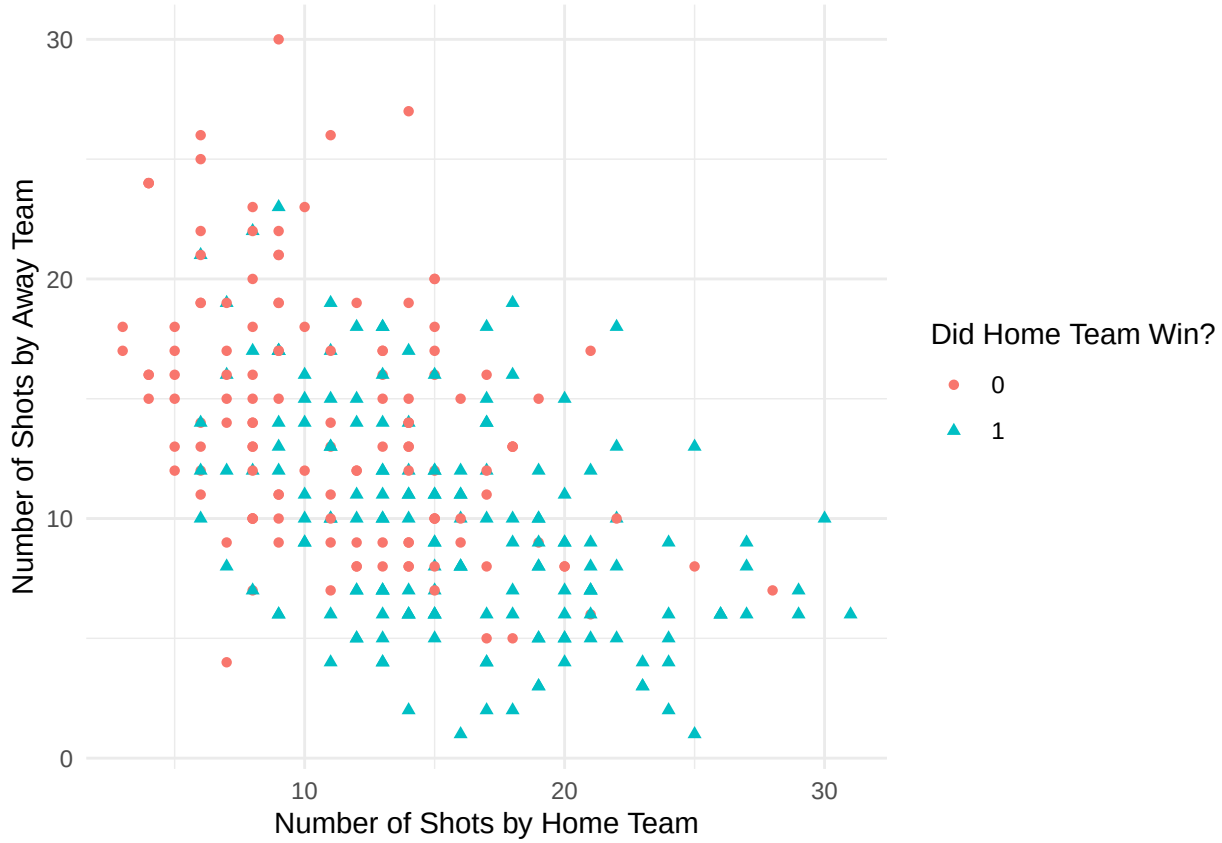


Figure 1: Correlation matrix for soccer match data. Key to abbreviations: - hs: Number of shots taken by the home team
- as: Number of shots taken by the away team
- hst: Number of shots on target by the home team
- ast: Number of shots on target by the away team
- hf: Number of fouls by the home team
- af: Number of fouls by the away team
- hc: Number of corners taken by the home team
- ac: Number of corners taken by the away team
- hy: Number of yellow cards received by the home team
- ay: Number of yellow cards received by the away team
- hr: Number of red cards received by the home team
- ar: Number of red cards received by the away team

b)

```
ggplot(soccer_data, aes(x = hs, y = as, color = factor(home_win))) +
  geom_point(aes(shape = factor(home_win))) +
  labs(x = "Number of Shots by Home Team", y = "Number of Shots by Away Team",
       color = "Did Home Team Win?", shape = "Did Home Team Win?") +
  theme_minimal()
```



add your answers here

c)

```
# Fit the glm model
model <- glm(home_win ~ hs + as + hst + ast + hf + af + hc + ac + hy + ay + hr + ar,
             data = soccer_data, family = "binomial")

# Tidying the model output
tidy_model <- tidy(model)

print(tidy_model)
```

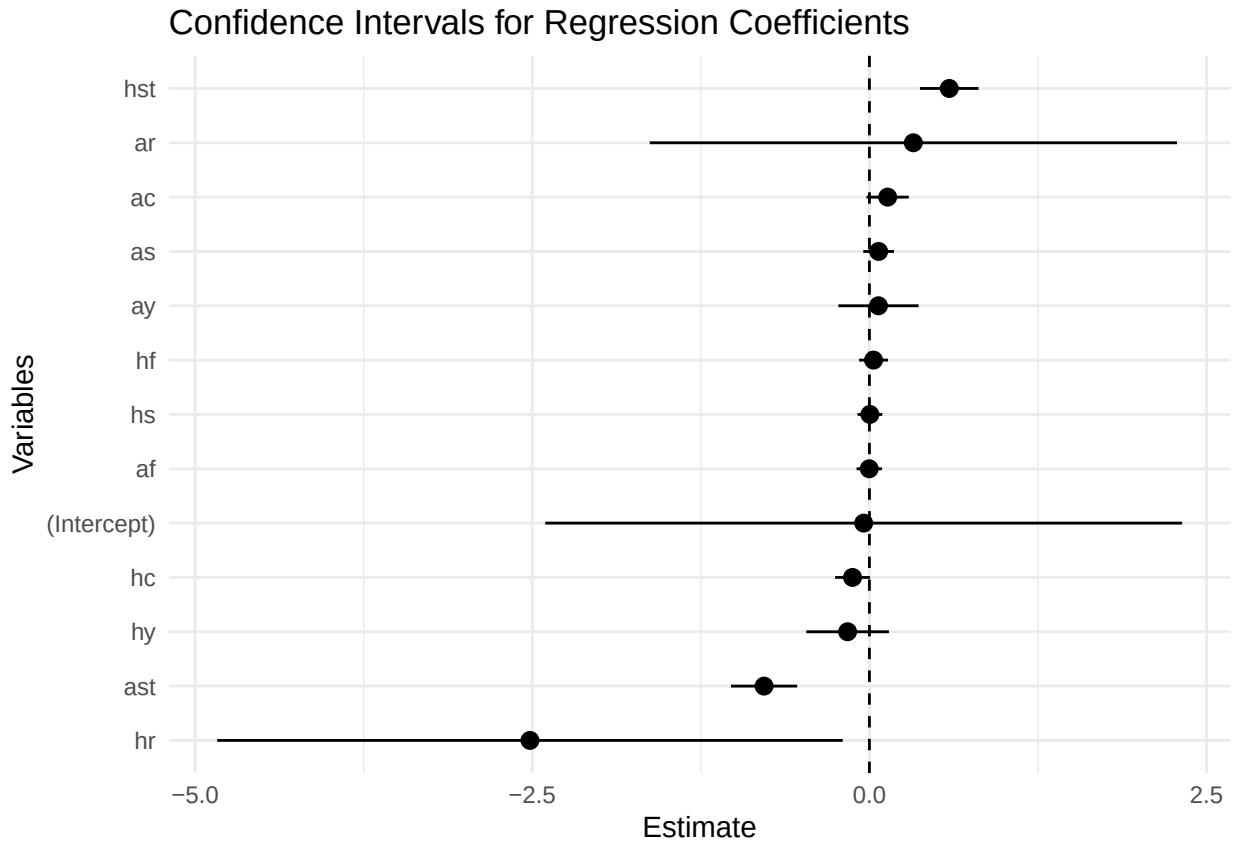
```
## # A tibble: 13 x 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept) -0.0423     1.20     -0.0351 9.72e- 1
##  2 hs           0.00383    0.0467    0.0820 9.35e- 1
##  3 as           0.0688     0.0581    1.18   2.37e- 1
##  4 hst          0.592      0.111     5.35   8.84e- 8
##  5 ast         -0.781      0.125    -6.24   4.28e-10
##  6 hf           0.0308     0.0547    0.563  5.73e- 1
##  7 af          -0.00109    0.0480   -0.0228 9.82e- 1
##  8 hc          -0.125      0.0658   -1.90   5.75e- 2
##  9 ac           0.135      0.0802    1.69   9.10e- 2
```

```
## 10 hy          -0.162      0.156    -1.03   3.01e- 1
## 11 ay           0.0672     0.152     0.443  6.58e- 1
## 12 hr          -2.52       1.18     -2.13   3.34e- 2
## 13 ar           0.326      0.997     0.327  7.44e- 1
```

add your answers here

d)

```
ggplot(tidy_model, aes(x = reorder(term, estimate), y = estimate)) +
  geom_pointrange(aes(ymin = estimate - std.error*1.96,
                      ymax = estimate + std.error*1.96)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  coord_flip() +
  labs(x = "Variables", y = "Estimate",
       title = "Confidence Intervals for Regression Coefficients") +
  theme_minimal()
```



add your answers here