STAT7123/STAT8123

Statistical Graphics Assignment 3

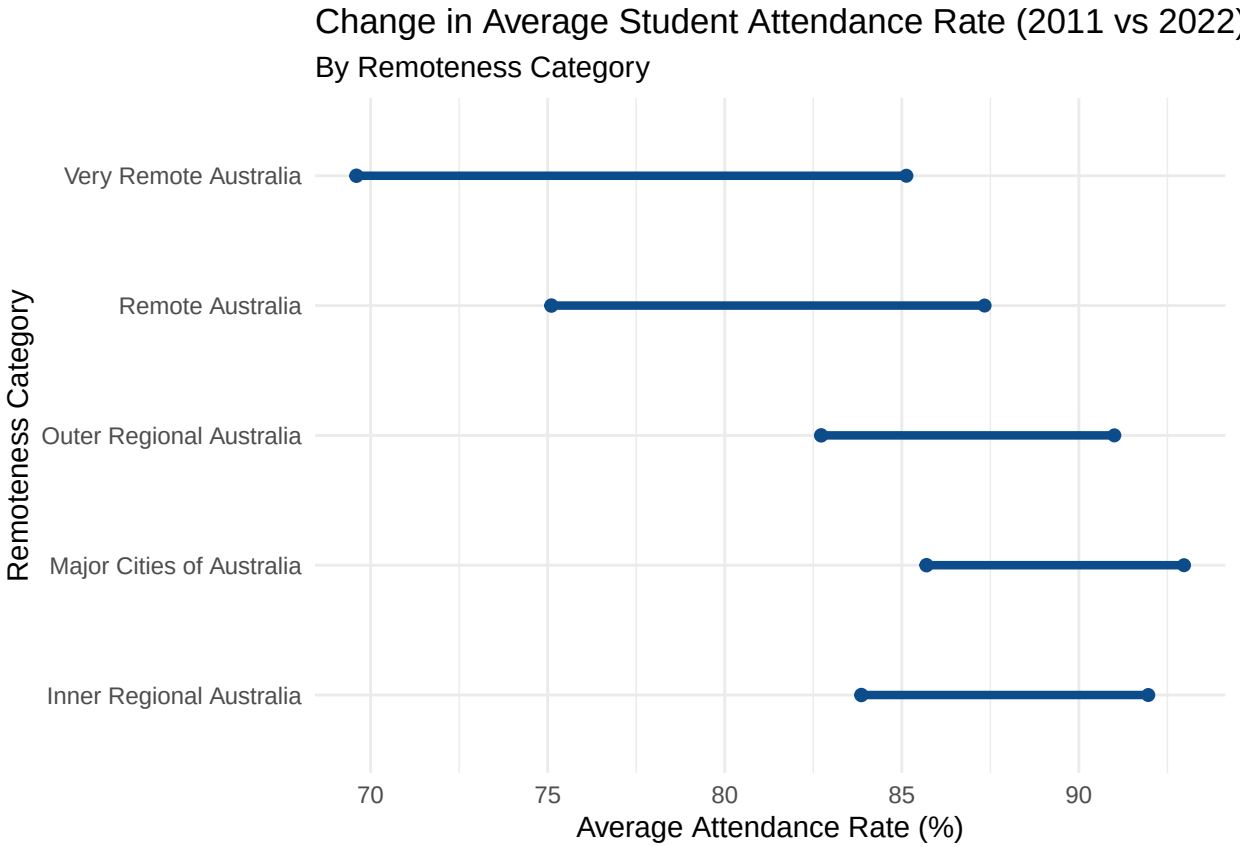
Osbert Bryan T. Villasis

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

Question 1

a) Make a dumbbell plot that shows that change in average attendance between 2011 and 2022 for the 5 different school remoteness categories. In 2-4 sentences, comment on what the dumbbell plot shows in regards to the changes in student attendance rates.

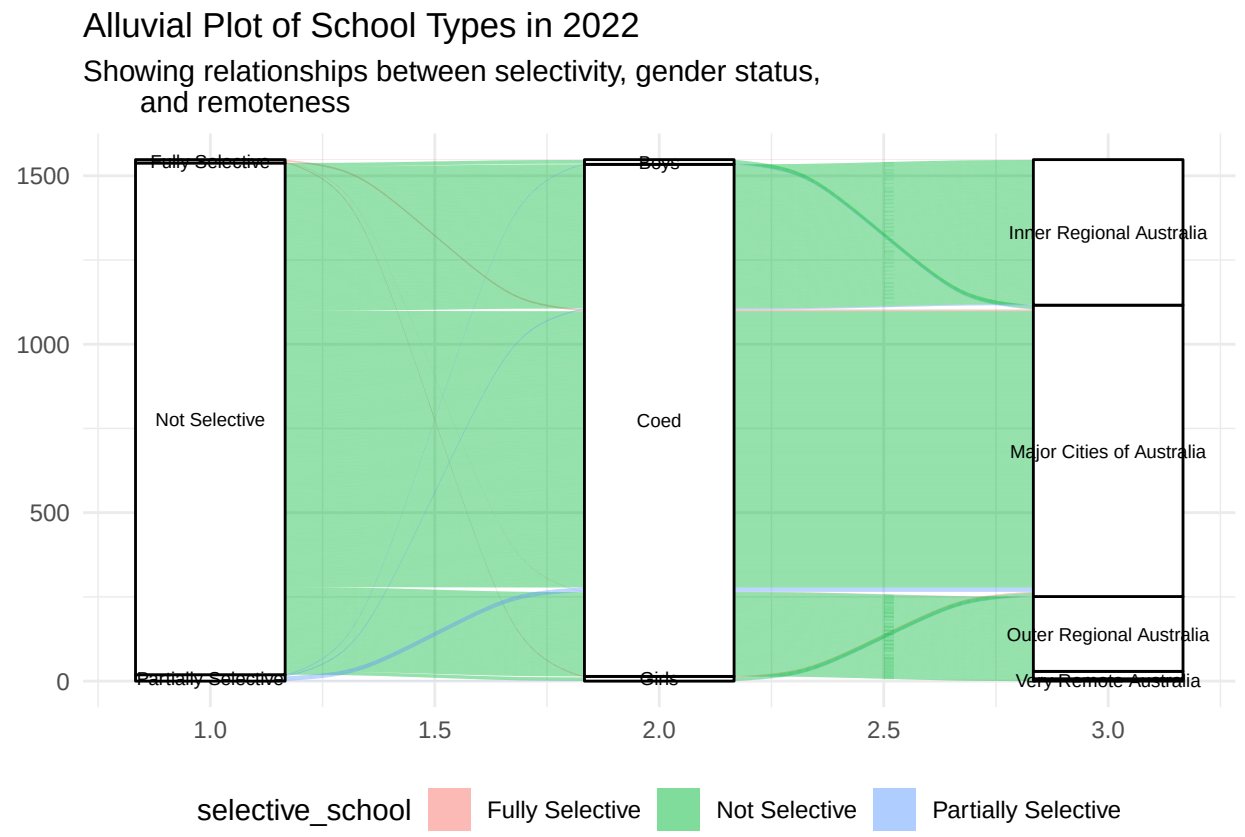
```
# Load the data
schools_data <- read_csv("school.csv")

# 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()
```



The graph visually represents shifts in average student attendance rates across different remoteness categories in Australia, comparing data from 2011 and 2022. Notably, all remoteness categories seem to exhibit a noticeable change in attendance rates over the decade. Inner Regional Australia and Major Cities of Australia have relatively narrow variations between the two years, implying smaller fluctuations in attendance. In contrast, Very Remote Australia and Remote Australia have more pronounced differences, suggesting potentially significant external factors or policies impacting attendance in those areas during this timeframe.

b) Create an alluvial plot with 2022 data, using three categorical variables on the axes: selective_school, school_gender, and asgs_remoteness. In 2-4 sentences, comment on what the alluvial plot shows in regards to school types.

```
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")
```

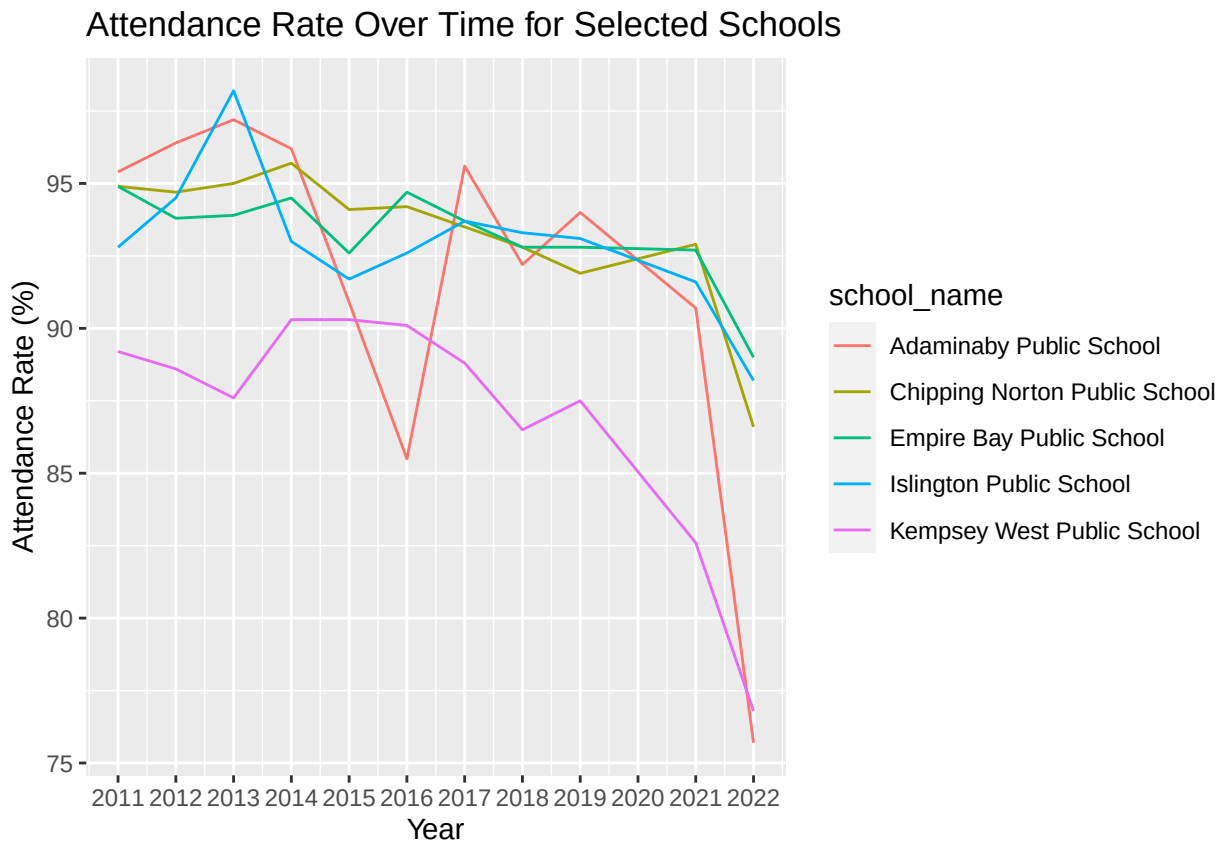


The alluvial plot offers a nuanced visualization of school types in 2022, interlinking selectivity, gender status, and geographical remoteness. A noteworthy observation is the significant number of non-selective schools, with most of them being coeducational. These coed, non-selective institutions predominantly lie within 'Major Cities of Australia' and 'Inner Regional Australia', highlighting urban and semi-urban areas' inclination towards more inclusive educational environments. Conversely, fully selective schools exhibit a more even gender distribution but are still primarily concentrated within major cities.

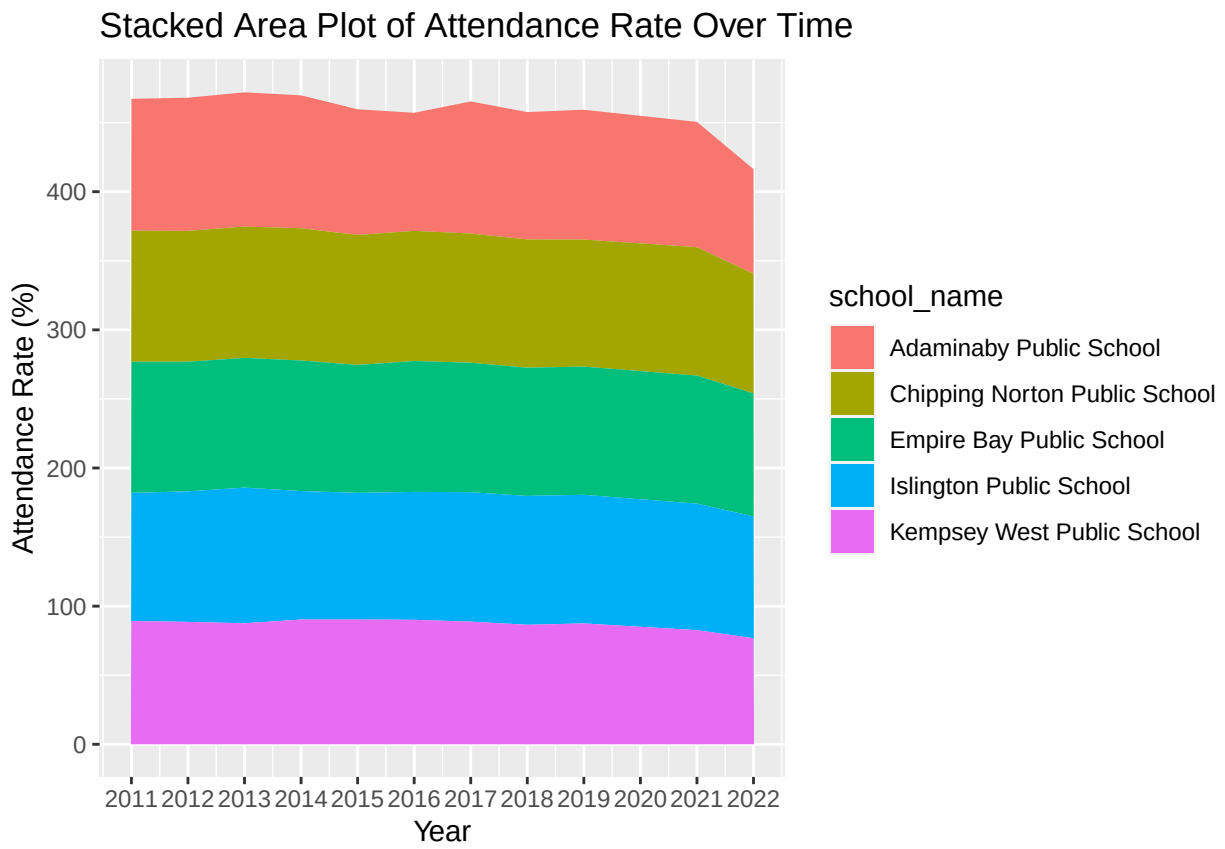
c) For this question, select 5 different schools (You are required to show which schools you picked). Create two plots; (i) a line plot, and (ii) a stacked area plot for the 5 schools showing the changes in percentage attendance over time. Which school had the highest attendance percentage in 2018? Provide 2-4 sentences which comment on the benefits and drawbacks of each plot.

```
# Selecting five schools
selected_schools <- c("Adaminaby Public School", "Chipping Norton Public School",
                      "Empire Bay Public School", "Islington Public School",
                      "Kempsey West Public School")
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))
```



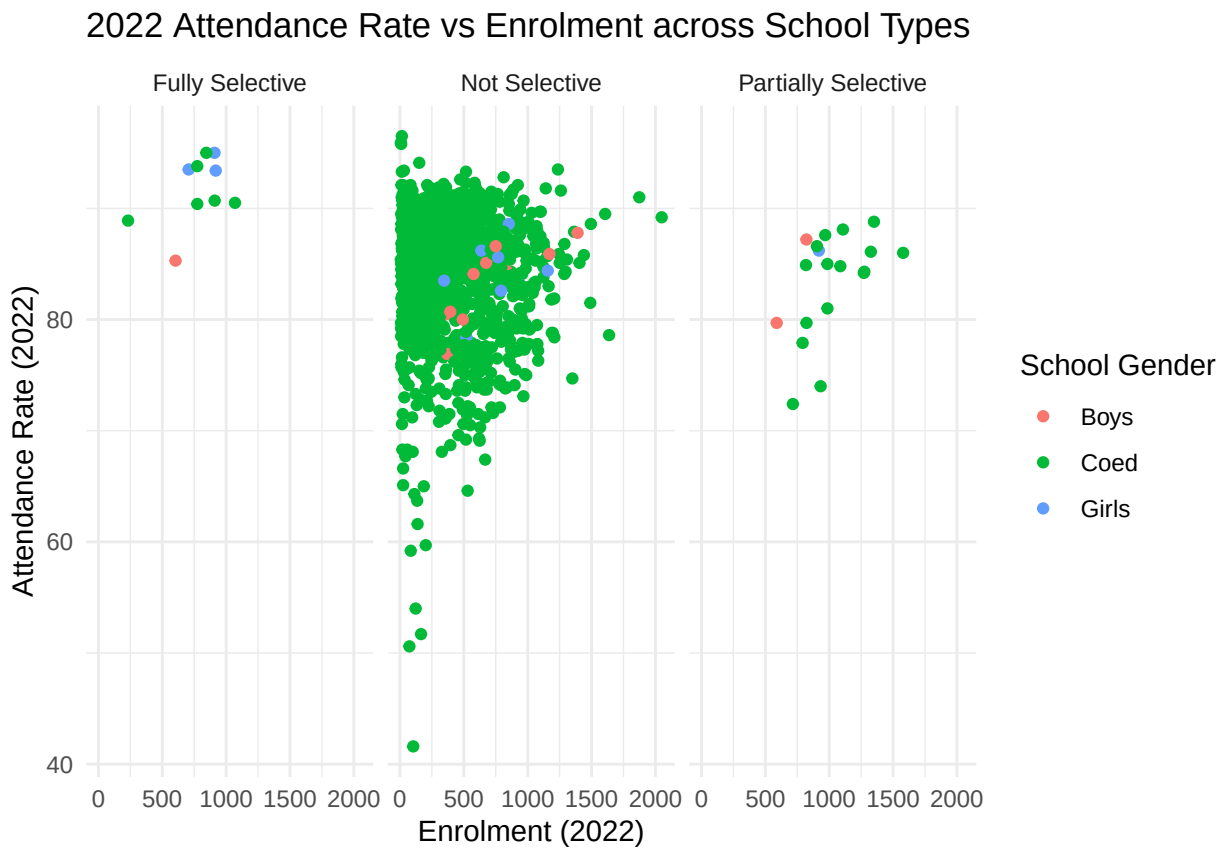

```
# Stacked area plot
ggplot(schools_subset, aes(x = year, y = attendance_rate, fill = school_name)) +
  geom_area(position = 'stack') +
  labs(title = "Stacked Area Plot of Attendance Rate Over Time",
       x = "Year", y = "Attendance Rate (%)") +
  scale_x_continuous(breaks = seq(min(schools_subset$year),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



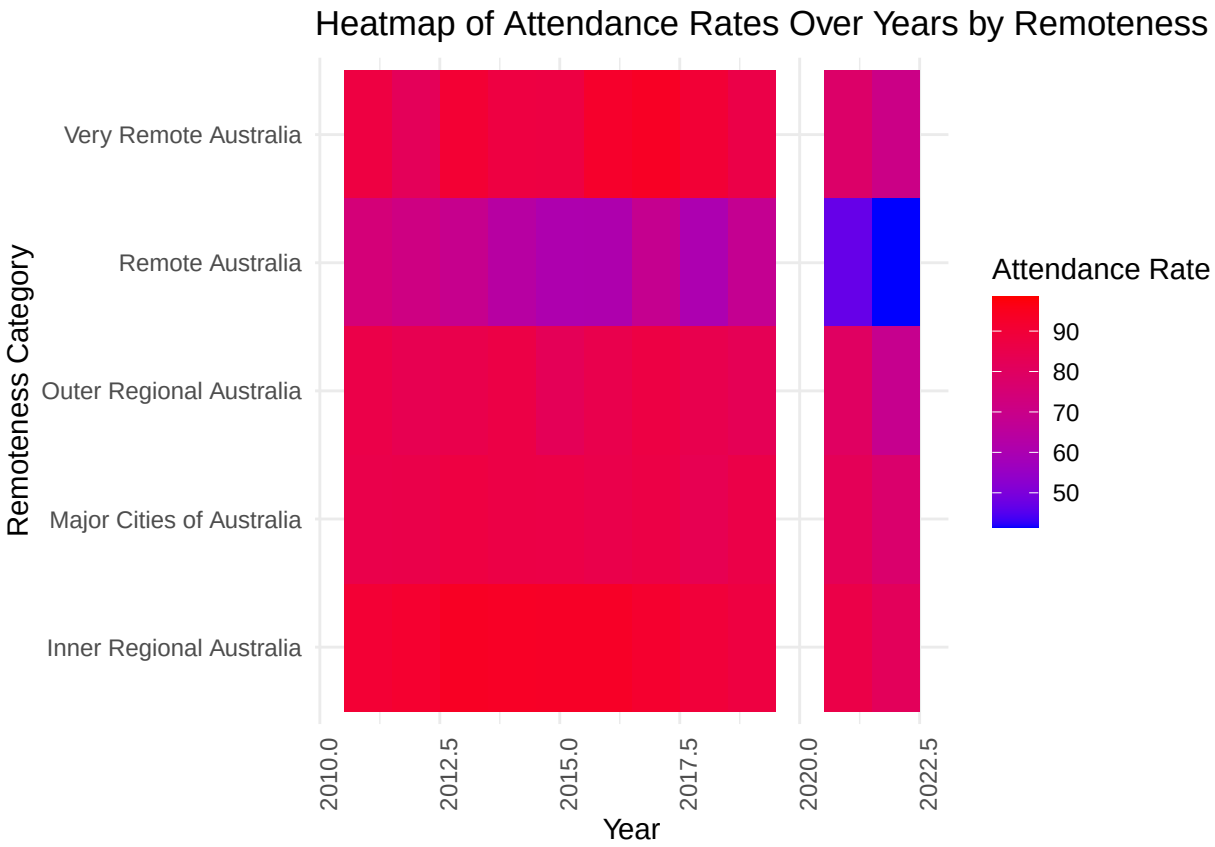
The two plots present an analysis of attendance rates over time for five selected Australian schools. In 2018, "Adaminaby Public School" had the highest attendance percentage among the selected schools, as depicted by the line plot. The line plot offers clear visual representation of trends over time for each individual school, making it easier to compare attendance rates across schools for any specific year. However, it might get cluttered when more schools are added, making the lines harder to distinguish. The stacked area plot, on the other hand, provides a collective visualization of attendance percentages, emphasizing the overall trend rather than individual school trends. This makes it challenging to determine specific values for each school, especially when the areas overlap. However, it offers a concise representation of how attendance percentages for all schools cumulatively change over time.

d) Create two different plots of your choice to explore the school attendance data further.

```
#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
```



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.

The scatter plot showcases the 2022 attendance rates in relation to school enrolment across different selectivity types. It's evident that non-selective schools, with the most data points, have a wide range of enrolments and attendance rates. Contrastingly, fully selective schools seem to have a narrower band of attendance, albeit with fewer students. The heatmap, on the other hand, reveals a consistent trend where major cities consistently exhibit higher attendance rates over the years compared to remote areas; however, in recent years, even urban centers seem to have dipped in attendance, aligning closer to remote regions.

e) Understanding, analysing, and communicating data is essential for any analytics role. Job descriptions for such positions often emphasise the importance of statistics skills to interpret and communicate data effectively with a broad audience. Within 150 words, provide an informative, coherent and precise summary of your finding from the graphical analysis of the given data set. On top of the 150 words, provide 2-4 insights from the data (these can be done as bullet points).

The graphical analysis of the dataset presents a comprehensive overview of student attendance trends in various Australian schools. Over the past decade, attendance rates have shown noticeable fluctuations. A clear distinction is evident based on school location, with urban schools, specifically those in major cities, consistently boasting higher attendance rates compared to their remote counterparts. However, between 2011 and 2022, all categories of remoteness witnessed a decline in attendance, with "Very Remote Australia" schools experiencing the most pronounced drop. This could signify potential challenges in access to education or other socioeconomic factors at play. Additionally, among the five selected schools, "Adaminaby Public School" showcased the highest attendance percentage in 2018. However, there is a concerning downtrend in the recent years across all these schools.

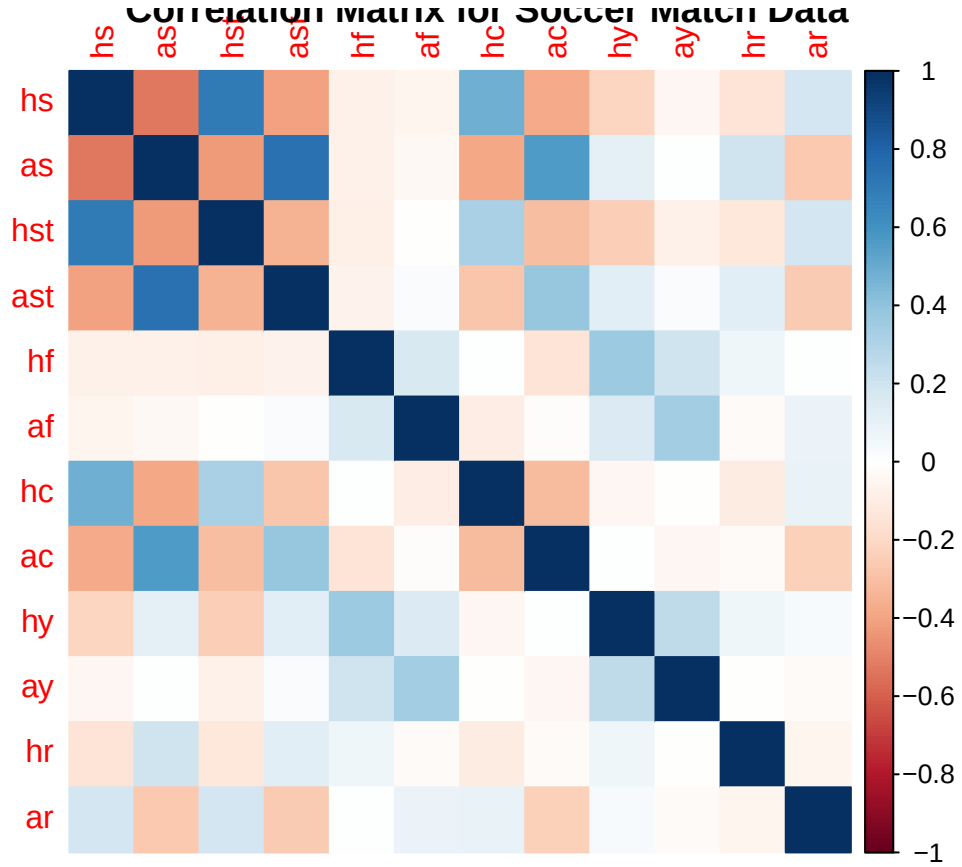
Insights:

- Major city schools have consistently higher attendance compared to remote schools.
- "Very Remote Australia" schools have faced a significant decline in student attendance over the years.
- "Adaminaby Public School" emerged as a frontrunner in 2018, but all selected schools reveal a declining trend in attendance after that year.
- The uniform decline across schools suggests a broader issue affecting attendance, rather than school-specific challenges.

Question 2

a) Plot the correlation matrix for all the appropriate variabes from the soccer dataset. In 2-4 sentences, comment on the different correlations between variables.

```
par(mar = c(5, 5, 5, 5))  # Default is c(5, 4, 4, 2) + 0.1
# 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")
```



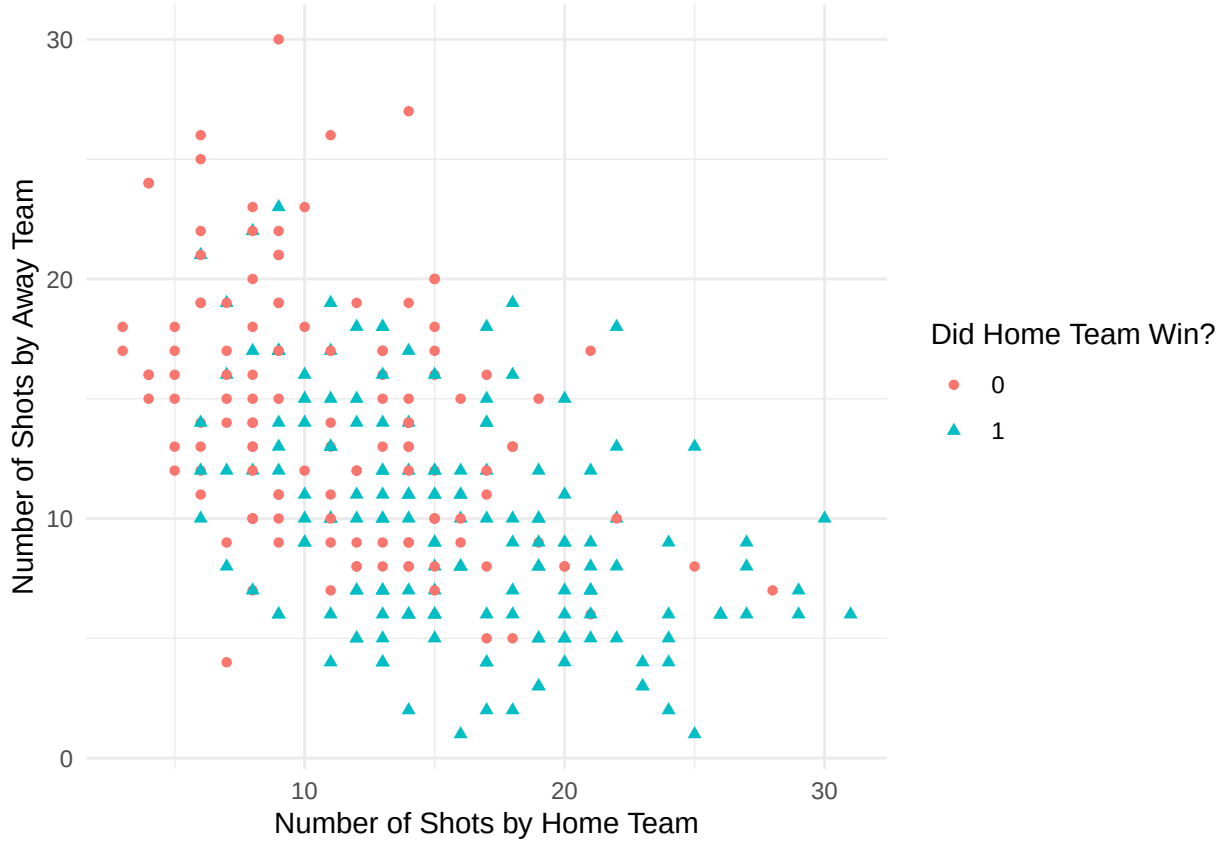
Correlation matrix for soccer match data. Key to abbreviations:

- 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

The provided correlation matrix visually represents the relationships between various soccer match variables. Dark blue signifies a strong positive correlation, while dark red indicates a strong negative correlation. The variable "hs" (probably referring to 'home score' or a similar metric) has a significant positive correlation with "as" (possibly 'away score'). In contrast, "hf" (potentially 'home fouls') shows a notable negative relationship with "ay" (perhaps 'away yellows'). Generally, some variables display strong correlations, while others exhibit weaker or no significant relationships, emphasizing the need to consider multicollinearity and other factors when modeling using these variables.

b) Graphically explore the relationship between hs, as, and home_win.

```
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()
```



The scatter plot visualizes the relationship between the number of shots taken by the home and away teams, differentiated by the outcome of whether the home team won. The red circles represent matches where the home team did not win, while the teal triangles depict games where they did. There isn't a distinct linear relationship between the number of shots and the game's outcome, suggesting that simply taking more shots doesn't guarantee a win. Both winning and non-winning matches for the home team are spread throughout the plot, indicating that other factors besides the number of shots likely influence the game's result.

c) Fit a model for home side wins, home_wins, to understand what factors are possibly influencing home side wins. This model should be a fit a glm model with a binomial family (family = "binomial"). Then, use the broom package to tidy up your model output and provide an interpretation of the model. Provide 2-2 sentences on if and how any of the model diagnostic plots look different to those seen in the lecture content and SGTA material.

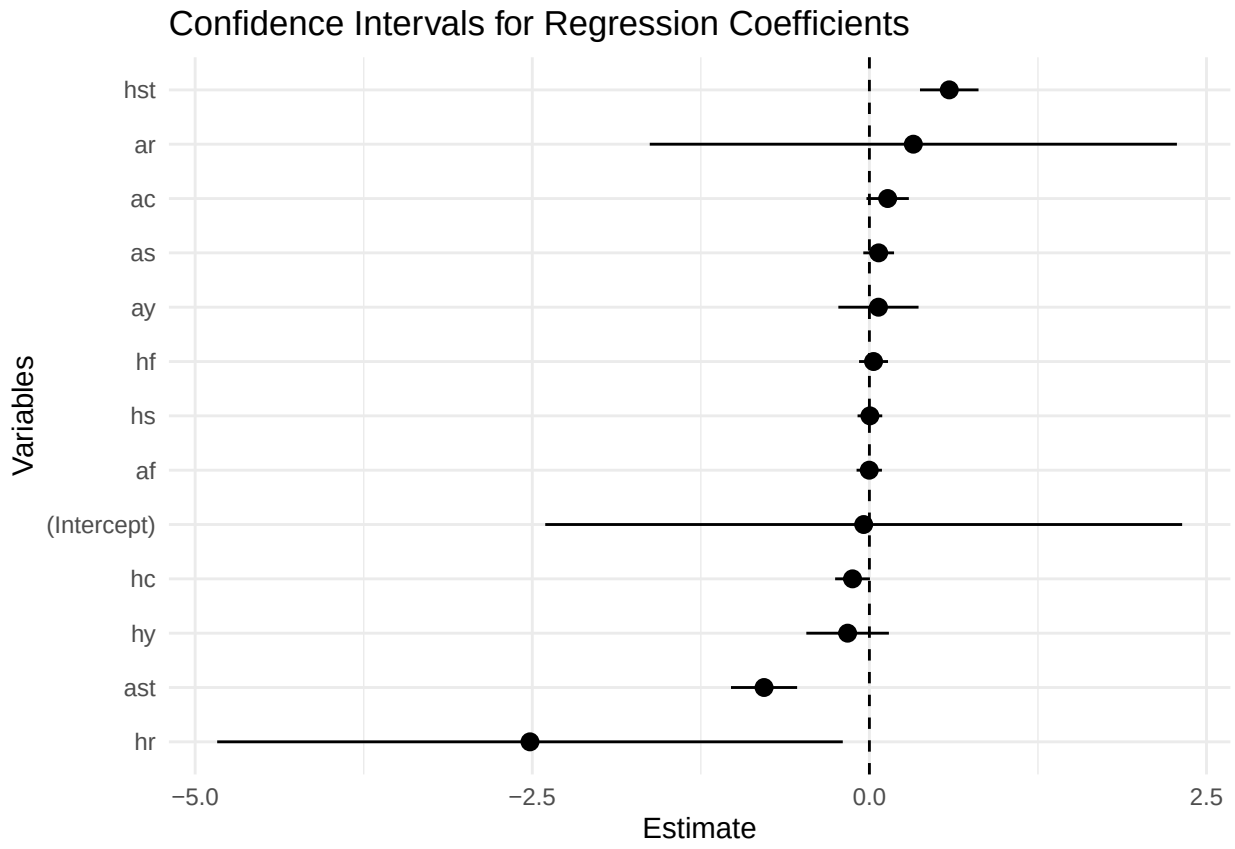
```
# Fit the glm model
model <- glm(home_win ~ hs + as + hst + ast + hf + af + hc + ac + hy + ay + hr + ar,
             data = soccer_data, family = "binomial")

# 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
```

d) Plot the confidence interval(s) for the regression coefficient(s) from the model in part (b). Write 2-3 sentences that discuss what is shown in the plot.

```
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()
```



The plot displays the confidence intervals for the regression coefficients of various predictors in the model. When a confidence interval crosses the dashed line at zero, it indicates that the predictor might not have a statistically significant effect on the outcome. In this visualization, 'hst' stands out with a notably positive estimate, while 'ast' and 'hr' have negative estimates that don't intersect the zero line, suggesting their significance. It's important to integrate this visual information with the actual p-values and other model diagnostics when determining the influence of each predictor.