STAT7123/STAT8123

Statistical Graphics Assignment 3

Umut Demirhan

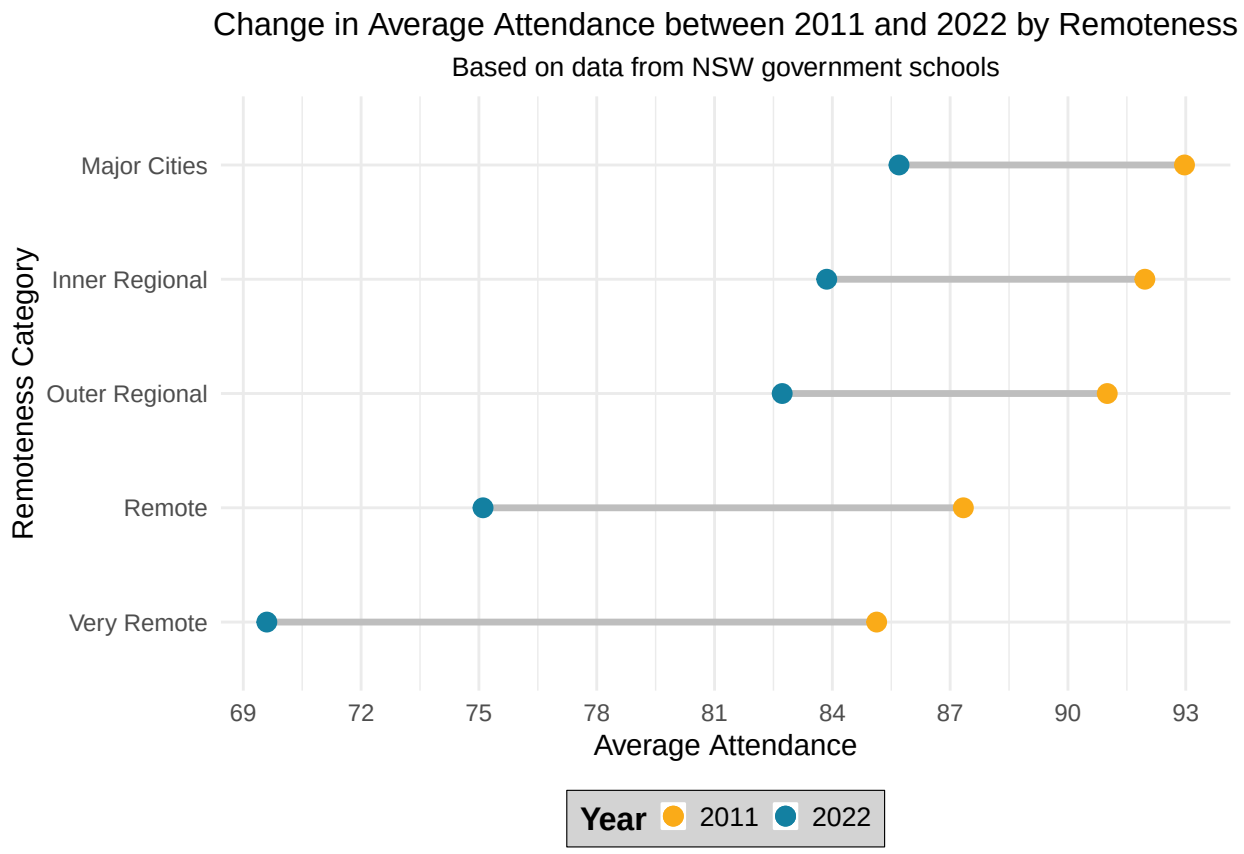
Due 11:55 pm, Friday November 3rd, 2023

Question 1

a) A dumbbell plot to show the change in average attendance between 2011 and 2022 for the 5 different school remoteness categories

```
# Read  the CSV data
school <- read.csv("school.csv")

# Modify the asgs_remoteness column to remove "Australia" and "of"
school$asgs_remoteness <- str_replace_all(school$asgs_remoteness,
                                          c(" Australia" = "", " of" = ""))
# Aggregating the data
agg_data <- school %>%
  group_by(asgs_remoteness) %>%
  summarise(
    avg_2011 = mean(attend_2011, na.rm = TRUE),
    avg_2022 = mean(attend_2022, na.rm = TRUE)
  )
# Reshaping the data
long_data <- agg_data %>%
  gather(key = "year", value = "attendance", avg_2011, avg_2022)

# Plotting the dumbbell
ggplot(agg_data, aes(y = reorder(asgs_remoteness, avg_2011))) +
  geom_segment(aes(x = avg_2011, xend = avg_2022, yend = asgs_remoteness),
               color = "grey", size = 1.2) +
  geom_point(aes(x = avg_2011, color = "2011"), size = 3) +
  geom_point(aes(x = avg_2022, color = "2022"), size = 3) +
  scale_color_manual(values = c("2011" = "#FAAB18", "2022" = "#1380A1"),
                     name = "Year") +
  scale_x_continuous(breaks = seq(69, 96, 3)) +
  theme_minimal() +
  labs(
    title = "Change in Average Attendance between 2011 and 2022 by Remoteness",
    subtitle = "Based on data from NSW government schools",
    x = "Average Attendance",
    y = "Remoteness Category"
  ) +
  theme(
```

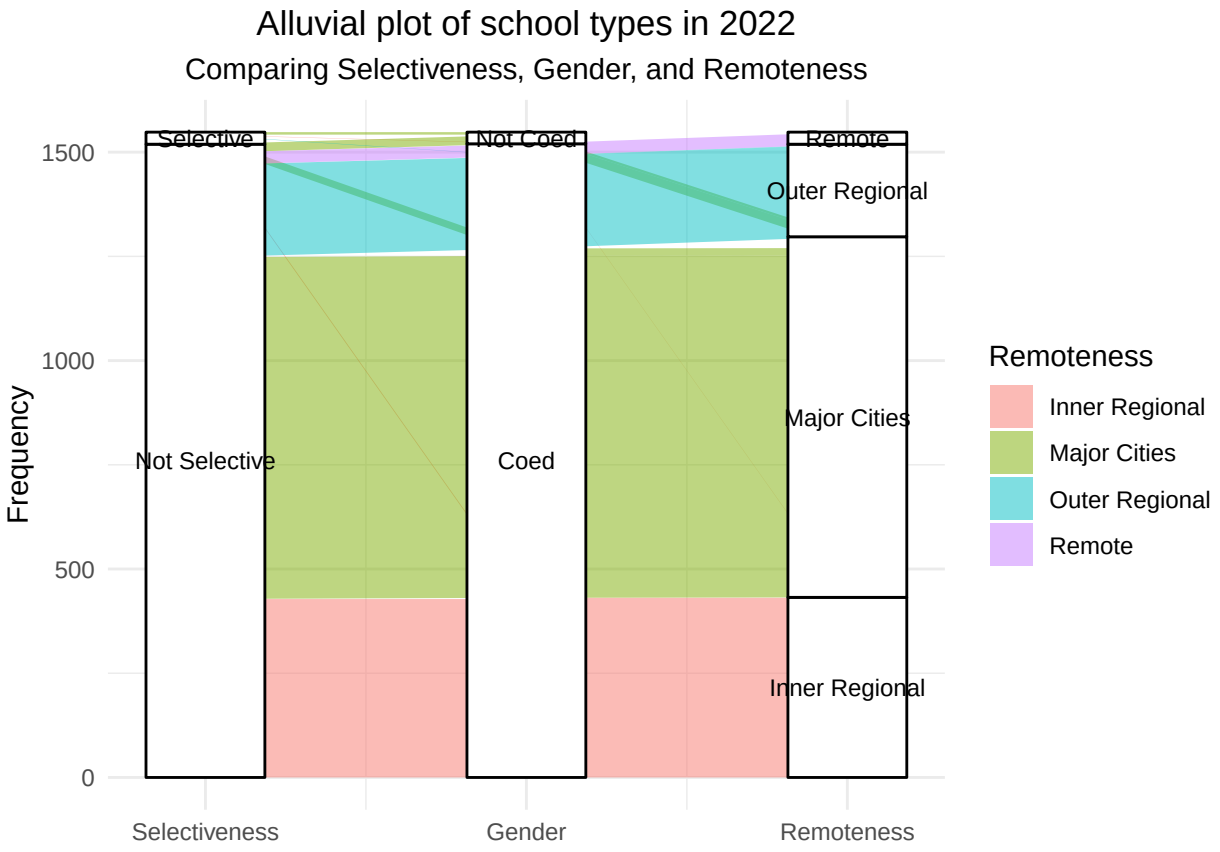
```
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5, size = 12),
    plot.subtitle = element_text(hjust = 0.5, size = 10),
    legend.title = element_text(face = "bold", size = 12),
    legend.text = element_text(size = 10),
    legend.key.size = unit(0.5, "lines"),
    legend.background = element_rect(fill = "lightgray", color = "black",
                                     size = 0.2, linetype = "solid"),
    legend.margin = margin(t = 5, r = 5, b = 5, l = 5),
    legend.spacing.x = unit(0.2, 'cm'),
    legend.key = element_rect(fill = "white", colour = "white")
    )
```



The dumbbell plot illustrates a decline in student attendance rates across all remoteness categories from 2011 to 2022. Major Cities had the highest attendance rates in both years, closely followed by Inner and Outer Regional areas. Remote and Very Remote areas consistently showed lower attendance rates with the most significant drop observed in the Very Remote category. The disparity in attendance rates between urban and remote regions has widened over the years.

b) An alluvial plot with 2022 data, using three categorical variables on the axes: **selective_school**, **school_gender**, and **asgs_remoteness**

```
# Only 14 schools for each "Boys"and "Girls"
# table(school$school_gender)
# Coed values overwhelming the data, I combined the rest together as "Not-Coed"
# Adjust the school_gender column
school$school_gender <- ifelse(school$school_gender %in%
                          c("Boys", "Girls"), "Not Coed", school$school_gender)

# Same goes 'Remote' and 'Very Remote'
# table(school$asgs_remoteness)
# So Combining them as 'Remote'
school$asgs_remoteness <- ifelse(school$asgs_remoteness %in%
       c("Remote","Very Remote"), "Remote", school$asgs_remoteness)

# Same goes 'Fully Selective' and 'Partially selective'
#table(school$selective_school)
# So Combining them as 'Selective'
school$selective_school <- ifelse(school$selective_school %in%
 c("Fully Selective","Partially Selective"), "Selective", school$selective_school)

# Create a dataset suitable for alluvial plotting
alluvial_data <- school %>%
  group_by(selective_school, school_gender, asgs_remoteness) %>%
  summarise(Frequency = n())

# Create alluvial plot
ggplot(data = alluvial_data,
       aes(axis1 = selective_school, axis2 = school_gender,
           axis3 = asgs_remoteness, y= Frequency)) +
  geom_alluvium(aes(fill =asgs_remoteness),
                width = 0, knot.pos = 0, reverse = FALSE) +
  geom_stratum(width = 1/2.7, reverse = FALSE) +
  geom_text(stat = "stratum", aes(label = after_stat(stratum)), size = 3,
            reverse = FALSE) +
  scale_x_continuous(breaks = 1:3, labels = c("Selectiveness", "Gender", "Remoteness")) +
  theme_minimal() + theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)) + # Centers the title
  labs(
    title = "Alluvial plot of school types in 2022",
    subtitle = "Comparing Selectiveness, Gender, and Remoteness",
    fill = "Remoteness")
```



Comment on the alluvial plot

As all columns are 2022 data except attendance columns from 2011 to 2021, I chose to fill the plot by Remoteness. The alluvial plot illustrates how selectiveness, gender, and remoteness categories of schools interrelate. While each vertical axis represents a category, the horizontal flows depict the proportions of schools in each category combination. Notably, the dominant representation of "Not Selective", "Coed", and "Major Cities" underscores an imbalance in the data, making distinctions for less frequent categories more elusive. While the alluvial plot serves to showcase the relationships between categorical variables, its efficacy is diminished in this context due to the data imbalance. The predominance of certain categories creates large, overwhelming flows that can obscure the nuances of less represented categories.

c) **Random Selection:**

- 2 schools from "Major Cities".
- 1 school from "Inner Regional".
- 1 school from "Inner Regional".
- 1 school from "Remote".

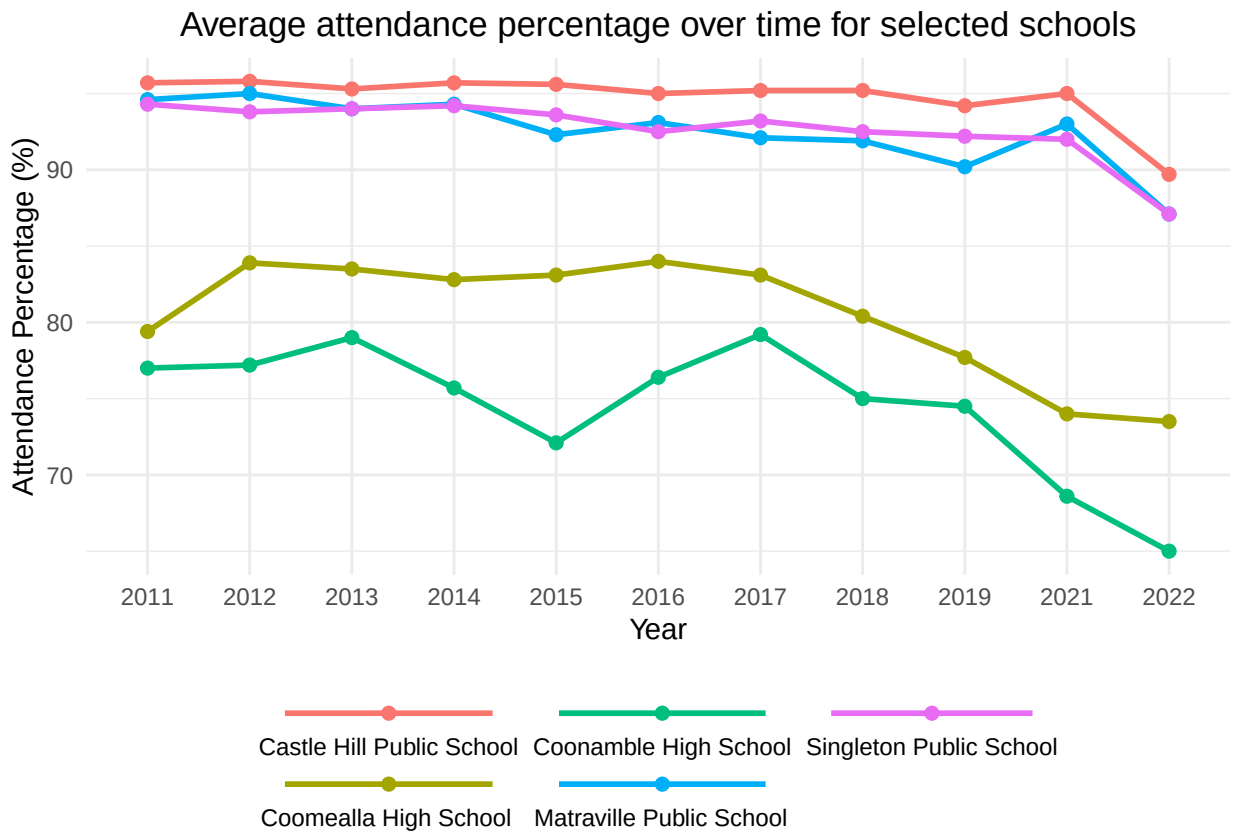
```
# Set a seed for reproducibility
set.seed(12345)

# Sampling for each category
schools_major_cities <- school %>%
  filter(asgs_remoteness == "Major Cities") %>%
  sample_n(2) %>%
  pull(school_name)
schools_inner_regional <- school %>%
  filter(asgs_remoteness == "Inner Regional") %>%
  sample_n(1) %>%
  pull(school_name)
schools_outer_regional <- school %>%
  filter(asgs_remoteness == "Outer Regional") %>%
  sample_n(1) %>%
  pull(school_name)
schools_remote <- school %>%
  filter(asgs_remoteness == "Remote") %>%
  sample_n(1) %>%
  pull(school_name)

# Combining the sampled schools
selected_schools <- c(schools_major_cities, schools_inner_regional,
                      schools_outer_regional, schools_remote)

# Display the selected school names using kable
selected_schools_df <- data.frame(School_Name = selected_schools)
kable(selected_schools_df, caption = "Selected Schools Based on Remoteness",
      row.names = TRUE, col.names = rep("School Names", ncol(selected_schools_df)),
      format = "latex", booktabs = TRUE, escape = FALSE) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 1: Selected Schools Based on Remoteness

	School Names
1	Matraville Public School
2	Castle Hill Public School
3	Singleton Public School
4	Coomealla High School
5	Coonamble High School

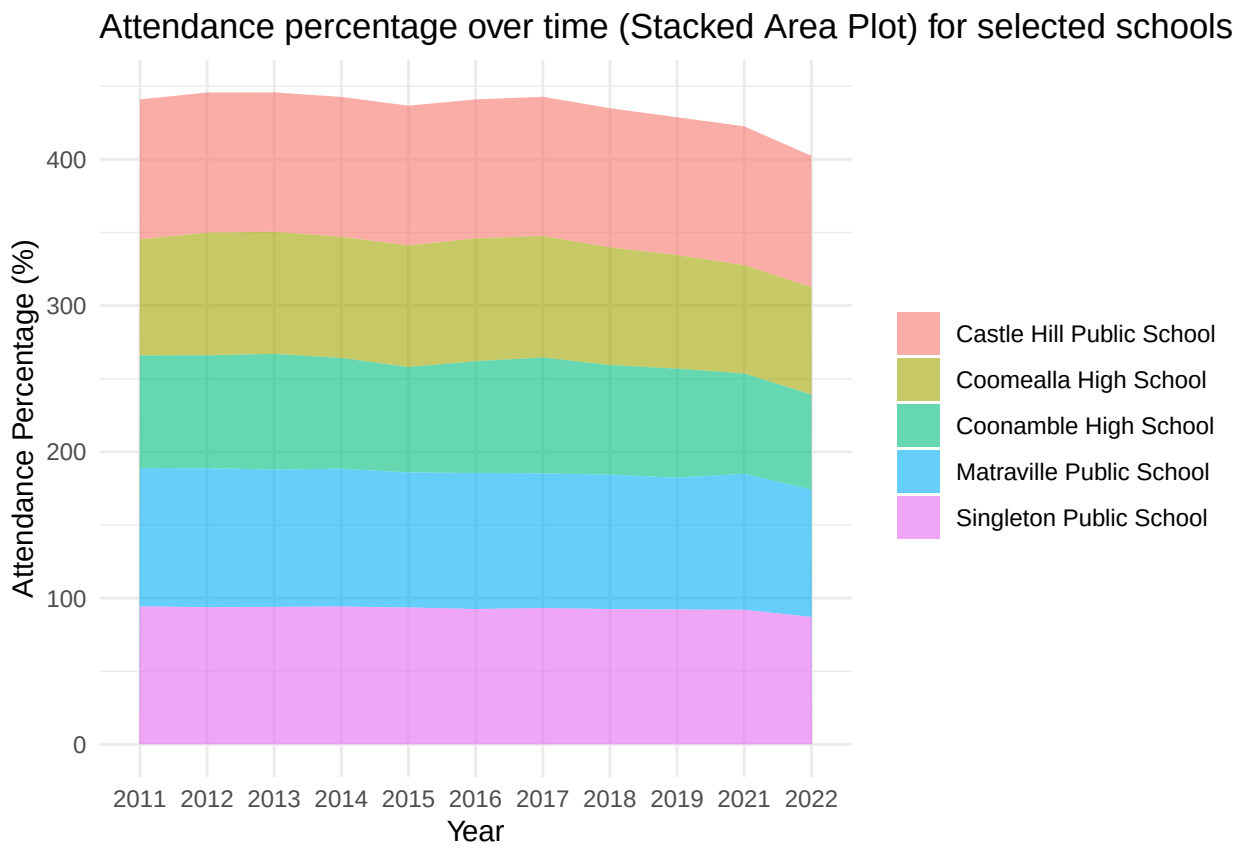
(i) a line plot

```r
# Reshape the data
long_data <- school %>%
  filter(school_name %in% selected_schools) %>%
  gather(key = "year", value = "attendance_pct", attend_2011:attend_2022)

# Removing Unnecessary Prefixes
long_data$year <- gsub("attend_", "", long_data$year)

# Line plot
line_plot <- ggplot(long_data, aes(x = year, y = attendance_pct,
                                   color = school_name, group = school_name)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "Average attendance percentage over time for selected schools",
       x = "Year", y = "Attendance Percentage (%)") +
  theme_minimal() +
  theme(legend.title = element_blank(),
        legend.position = "bottom", # Moves legend to the bottom
        legend.key.size = unit(0.2, "cm"), # Adjusts size of legend keys
        plot.title = element_text(hjust = 0.5)) + # Centers the title
  guides(color = guide_legend(nrow = 2, title.position = "top",
                              label.position = "bottom",
                             title.hjust = 0.0, label.hjust = 0.2,
                             keywidth = 0.5, keyheight = 0.5))
line_plot
```



(ii) a stacked area plot

```
# Stacked area plot
area_plot <- ggplot(long_data, aes(x = year, y = attendance_pct,
                                   fill = school_name, group = school_name)) +
  geom_area(position = "stack", alpha = 0.6) +
  labs(title = "Attendance percentage over time (Stacked Area Plot) for selected schools",
       x = "Year", y = "Attendance Percentage (%)") +
  theme_minimal() +
  theme(legend.title = element_blank())
area_plot
```



The line Plot:

- **In 2018**, **Cast Hill Public School** had the highest attendance percentage in these 5 randomly selected schools.
- Provides a clear visualization of trends over time for each school, making it easy to spot fluctuations and compare attendance rates across different schools.
- With many schools or overlapping data points, the plot can become cluttered, making it difficult to differentiate between individual lines.

Stacked Area Plot:

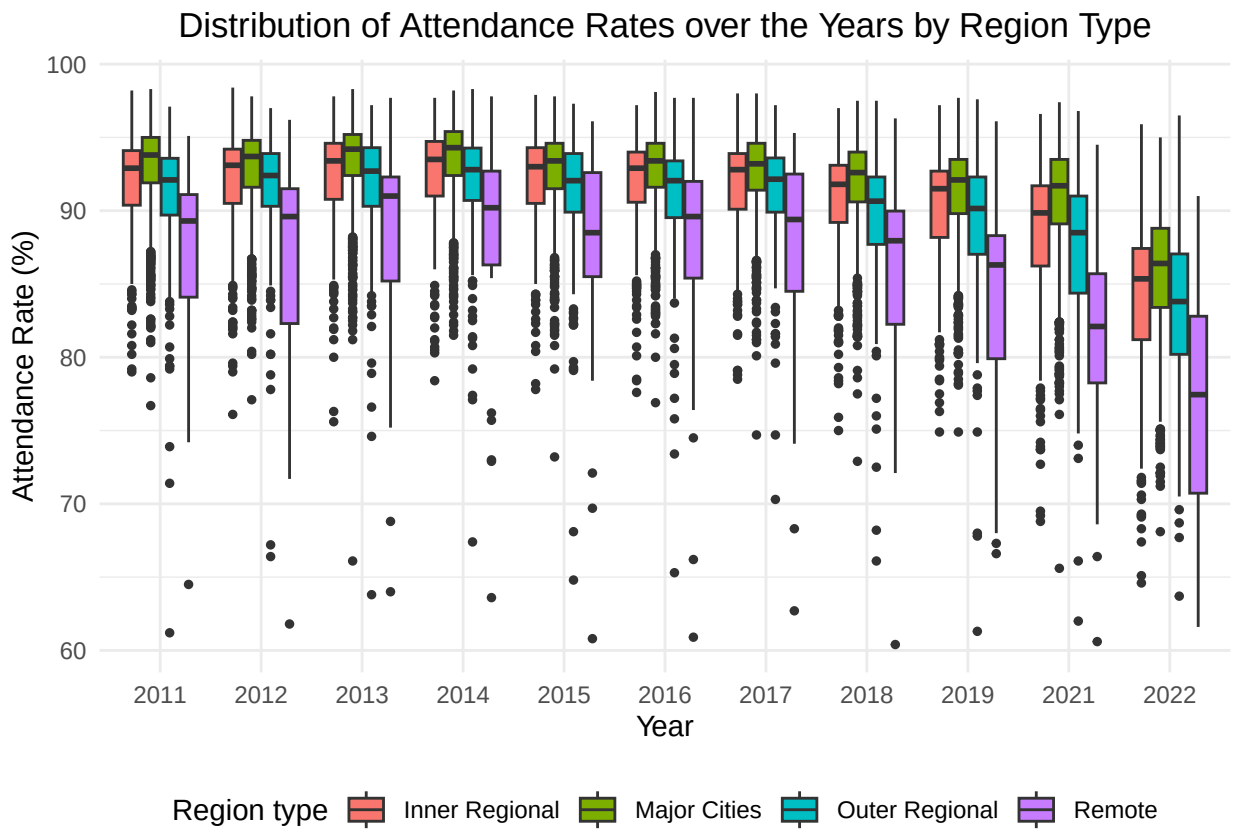
- It obscures individual trends, making it challenging to discern exact values for individual schools.

- A stacked area plot clearly is not a good choice for attendance percentages for different schools. The sum of attendance percentages for the five schools at any given year doesn't represent a meaningful total.

d)

Plot 1:Box plot of Distribution of Attendance Rates over the Years by Region Type

```
# Gather the attendance data for all years into long format
long_school_data <- school %>%
  gather(key = "year", value = "attendance_rate", attend_2011:attend_2022)

# Removing Unnecessary Prefixes
long_school_data$year <- gsub("attend_", "", long_school_data$year)

# Filter out extreme values for better visualization
long_school_data <- long_school_data %>%
  filter(attendance_rate >= 60)

ggplot(long_school_data, aes(x = year, y = attendance_rate, fill = asgs_remoteness)) +
  geom_boxplot(outlier.size = 1) +
  labs(title = "Distribution of Attendance Rates over the Years by Region Type",
       y = "Attendance Rate (%)",
       x = "Year",
       fill = "Region type") +
  theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust = 0.5))
```

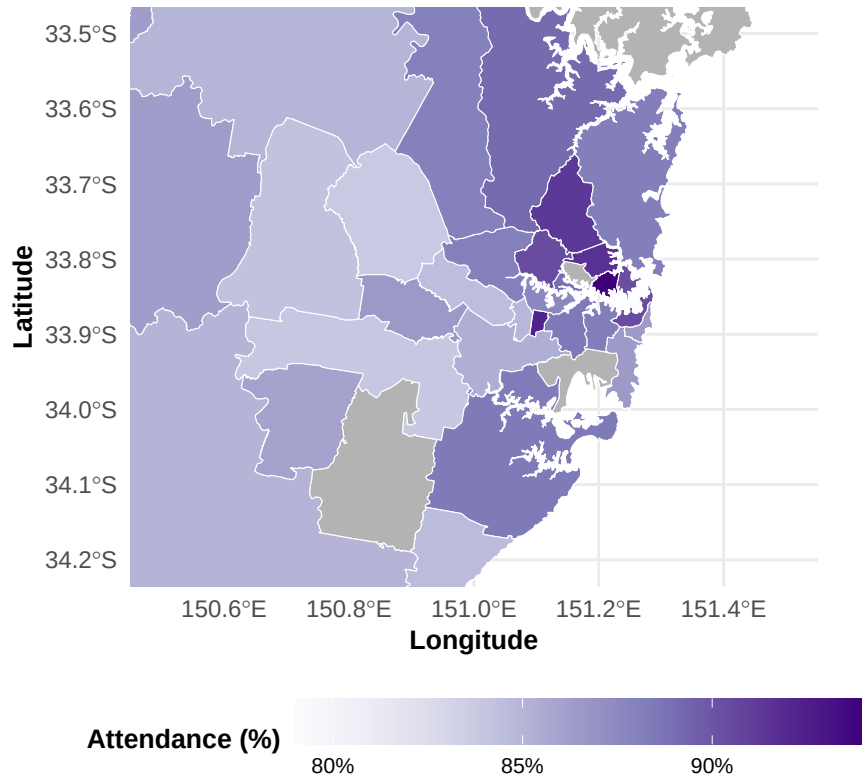


Plot 2: The choropleth map for Average Attendance by LGA in Sydney Region (2022)

```
# Read the shapefile- this shapefile has been also used in Assignment 2
shapefile_data <- st_read("LGA_2021_AUST_GDA2020_SHP/LGA_2021_AUST_GDA2020.shp")
# Filter for New South Wales
shapefile_data_nsw <- shapefile_data %>% filter(STE_NAME21 == "New South Wales")
# The results are hidded for aesthetic reason
```

```
# Removing suffixes like (A), (C) etc.
school$lga <- gsub("\\s*\\(.*?\\)\\s*", "", school$lga)
# Making LGA names compatible for merging
shapefile_data$LGA_NAME21 <- tolower(trimws(shapefile_data$LGA_NAME21))
school$lga <- tolower(trimws(school$lga))

# Making LGA names compatible for merging
shapefile_data_nsw$LGA_NAME21 <- tolower(trimws(shapefile_data_nsw$LGA_NAME21))
school$lga <- tolower(trimws(school$lga))

# Removing "(nsw)" from the end of the LGA names in shapefile_data_nsw
shapefile_data_nsw$LGA_NAME21 <- sub("\\(nsw\\)$", "",
                                     shapefile_data_nsw$LGA_NAME21)

# Trim extra whitespace for merging
school$lga <- trimws(school$lga)
shapefile_data_nsw$LGA_NAME21 <- trimws(shapefile_data_nsw$LGA_NAME21)
```

```r
# Manually adjusting the special cases in shapefile_data_nsw
shapefile_data_nsw$LGA_NAME21[shapefile_data_nsw$LGA_NAME21 ==
                      "queanbeyan-palerang regional"] <-"queanbeyan-palerang"
shapefile_data_nsw$LGA_NAME21[shapefile_data_nsw$LGA_NAME21 ==
                                "sutherland shire"] <- "sutherland"
shapefile_data_nsw$LGA_NAME21[shapefile_data_nsw$LGA_NAME21 ==
                                "tamworth regional"] <- "tamworth"
shapefile_data_nsw$LGA_NAME21[shapefile_data_nsw$LGA_NAME21 ==
                                "the hills shire"] <- "the hills"

# Average attendance by LGA for 2022
attendance_by_lga <- school %>%
  group_by(lga) %>%
  summarise(attend_2022 = mean(attend_2022, na.rm = TRUE)) %>%
  rename(lga = lga)

# Merge the spatial data with attendance data
merged_data_attendance <- left_join(shapefile_data, attendance_by_lga,
                          by = c("LGA_NAME21" = "lga"))

# Creating the choropleth map for attendance
ggplot(data = merged_data_attendance, mapping = aes(fill = attend_2022)) +
  geom_sf(color = "white", size = 0.1) +
  scale_fill_gradientn(
    colors = brewer.pal(9, "Purples"),
    na.value = "grey70",
    name = "Attendance (%)",
    limits = c(79, 93.90),  # This will restrict the scale
    breaks = pretty_breaks(n = 5),
    labels = scales::percent_format(scale = 1)
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.title.x = element_text(size = 10, face = "bold"),
    axis.title.y = element_text(size = 10, face = "bold"),
    legend.position = "bottom",
    legend.title = element_text(size = 10, face = "bold"),
    legend.text = element_text(size = 8),
    legend.key.width = unit(1.5, "cm")
  ) +
  labs(
    title = "Average Attendance by LGA in Sydney Region (2022)",
    x = "Longitude",
    y = "Latitude"
  ) +
  coord_sf(xlim = c(150.5, 151.5), ylim = c(-34.2, -33.5)) # Focusing on Sydney
```


Average Attendance by LGA in Sydney Region (2022)
33.5°S
33.6°S
33.7°S
33.8°S
33.9°S
34.0°S
34.1°S
34.2°S
Latitude
150.6°E
150.8°E
151.0°E
151.2°E
151.4°E
Longitude
Attendance (%)
80%
85%
90%

e)

Summary:

The graphical analysis of school attendance data across various regions and years offers a panoramic view of educational trends. Over time, a general decline in attendance rates is observed, with a significant decrease noted between 2021 and 2022. Specifically, in remote areas, attendance rates first saw an uptick from 2011 to 2014, before witnessing a decline. A geographic gradient in attendance emerges when urban centers and their peripheries are inspected; major cities and inner regional schools consistently report higher median attendance compared to their counterparts in more remote locations. The choropleth map further delineates this pattern, with northern suburbs such as North Sydney and Chatswood exhibiting high attendance rates. In contrast, southwestern regions, which are comparatively less urbanized, report decreased attendance in 2022. Lastly, the dataset reveals that the majority of schools are coeducational and non-selective, with a higher concentration in more urbanized regions.

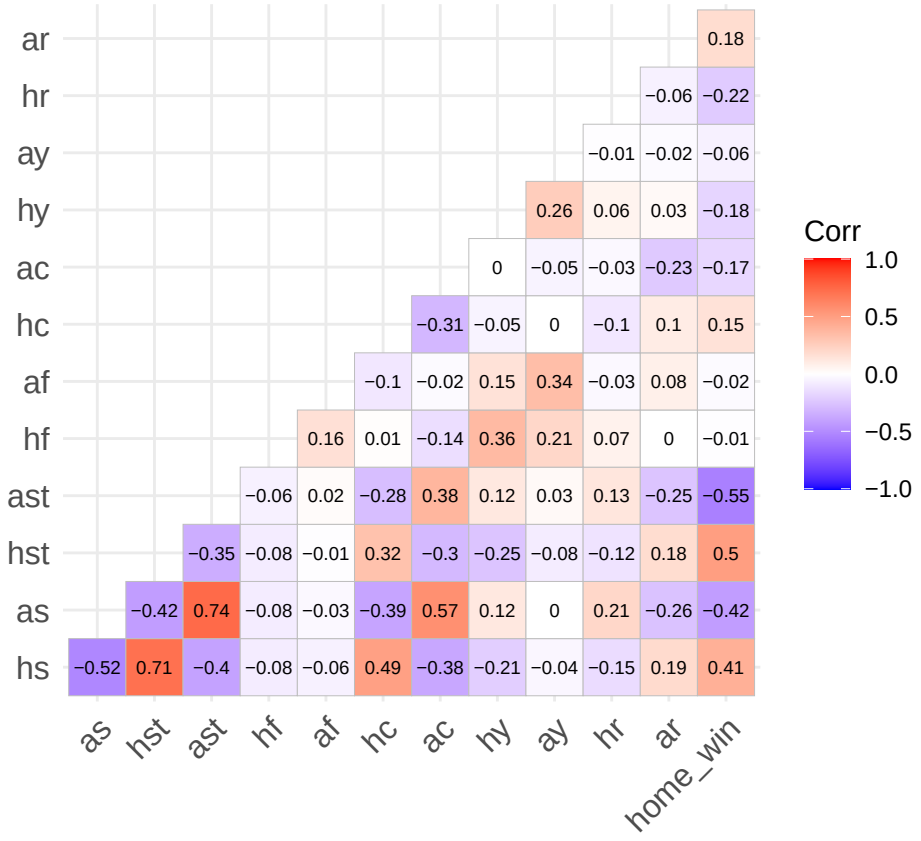
Insights:

- Attendance rates have shown a general decline over the years, with a marked drop between 2021 and 2022.
- Remote areas initially showed an increase in attendance from 2011 to 2014, but then displayed a declining trend.
- Larger urban areas, such as major cities and inner regional areas, consistently demonstrate higher median attendance rates than remote regions.
- The northern suburbs, typified by North Sydney and Chatswood, are attendance hotspots, whereas southwestern regions, particularly less urbanized ones, record lower attendance rates in 2022.

Question 2

a)

```
# Loading the data
soccer <- read.csv("soccer.csv")
# Computing correlation matrix
cor_matrix <- cor(soccer)
# Visualizing the correlation matrix
ggcorrplot(cor_matrix, method="square", type="lower", lab = TRUE, lab_size = 2.5)
```



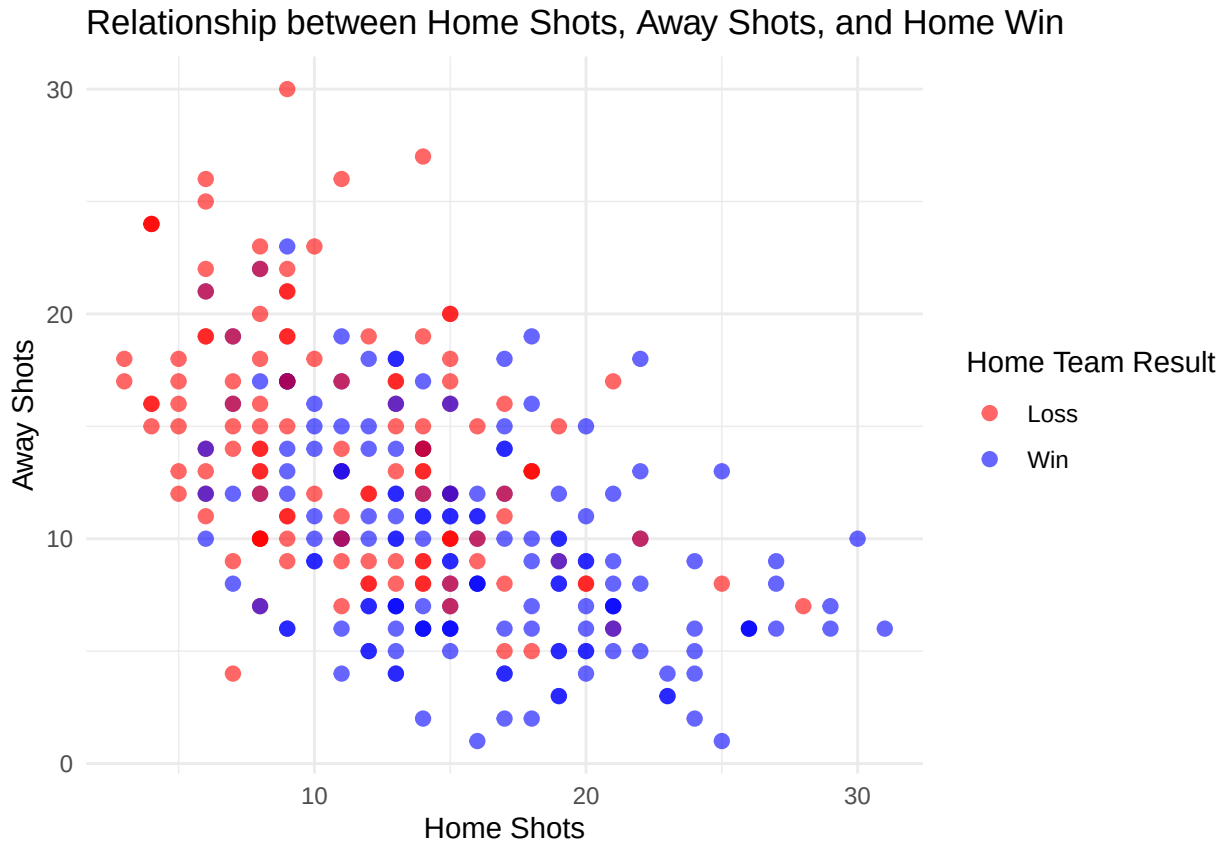
Interpretation:

- The strong positive correlation between 'hs' and 'hst' indicates that when the home team takes more shots, more of those shots are on target.
- 'as' and 'ast' show a similar relationship, suggesting the away team's shots on target also increase with the total number of shots.
- Yellow cards ('hy' and 'ay') have a noticeable positive correlation with fouls ('hf' and 'af'), suggesting more fouls can lead to more yellow cards.
- Negative correlations, like between 'hs' and 'as', imply that when the home team takes more shots, the away team takes fewer shots and vice versa.

b)

(i) A Scatter plot

```
# Plot
ggplot(data = soccer, aes(x=hs, y=as, color=factor(home_win))) +
  geom_point(alpha=0.6, size= 2.2) +
  scale_color_manual(values=c("red", "blue"), labels=c("Loss", "Win"),
                     name="Home Team Result") +
  labs(title="Relationship between Home Shots, Away Shots, and Home Win",
       x="Home Shots",
       y="Away Shots") +
  theme_minimal()
```



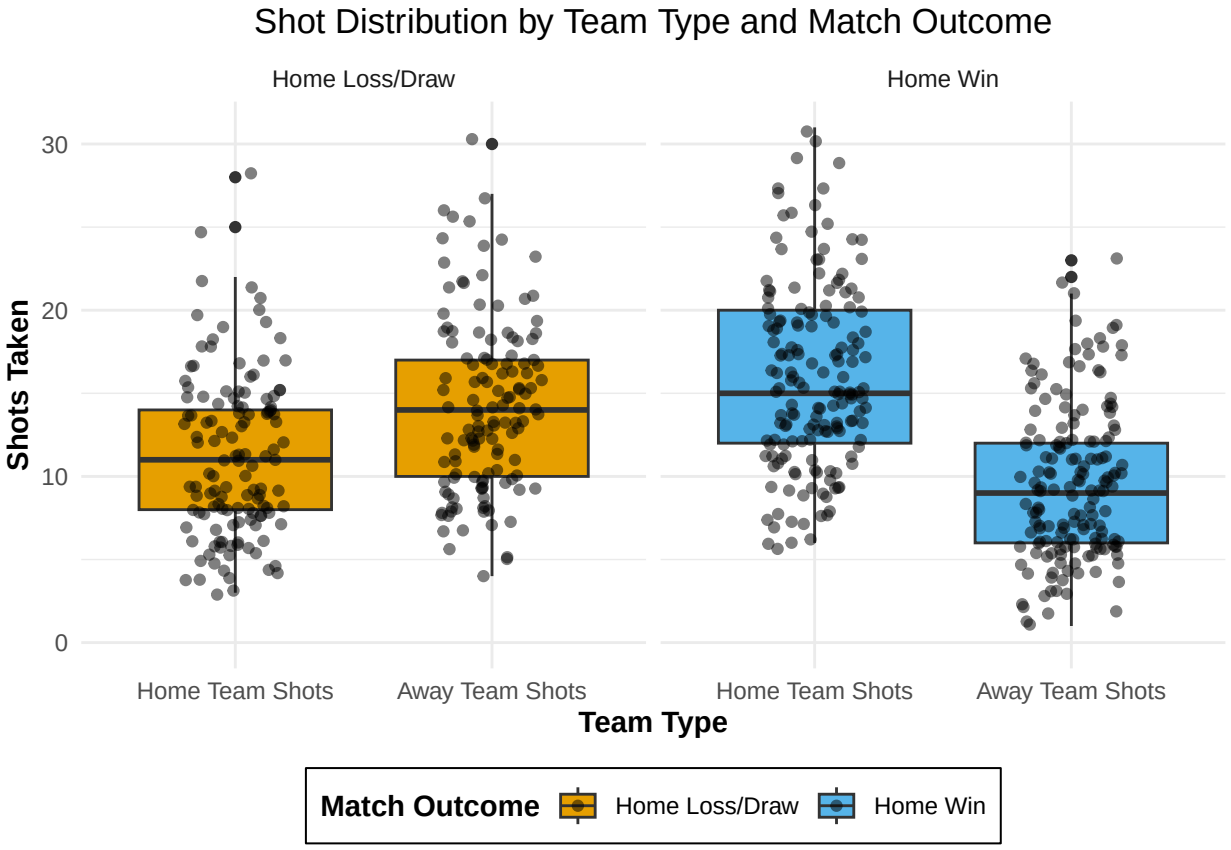
This scatter plot displays the relationship between shots taken by home teams (hs) and away teams (as), with colors indicating if the home team won.

Key Insights:

Blue points (home wins) cluster around higher shot counts for home team(Lower right). Similarly, red points (home losses/draws) cluster around higher shot counts for Away team(Upper left). A denser concentration of both points suggest a shot advantage often leads to win. However, many exceptions indicate that shot count alone doesn't ensure victory. Additionally, No points for both Home and Away shots higher than 20 are observed (on the top right). In essence, while more shots by the home team can indicate a higher chance of winning, it's not a guaranteed predictor of match outcomes.

(ii) A box plot

```r
# Melt the data to long format for easy plotting
long_data <- melt(soccer, id.vars = "home_win", measure.vars = c("hs", "as"))

# Create a ggplot object for the boxplot
ggplot_boxplot <- ggplot(long_data, aes(x = variable,
                                        y = value, fill = factor(home_win))) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.5) +
  facet_wrap(~ home_win, scales = "free_x", labeller = labeller(home_win =
                                  c('0' = "Home Loss/Draw", '1' = "Home Win"))) +
  scale_fill_manual(values = c("0" = "#E69F00", "1" = "#56B4E9"),
                    name = "Match Outcome",
                    labels = c("0" = "Home Loss/Draw", "1" = "Home Win")) +
  scale_x_discrete(labels = c("hs" = "Home Team Shots", "as" = "Away Team Shots")) +
  labs(x = "Team Type", y = "Shots Taken",
       title = "Shot Distribution by Team Type and Match Outcome") +
  theme_minimal() +
  theme(legend.title = element_text(face = "bold"),
        legend.position = "bottom",
        plot.title = element_text(hjust = 0.5),
        axis.title = element_text(face = "bold"),
        legend.background = element_blank(),
        legend.box.background = element_rect(color = "black", fill = NA)) +
  expand_limits(y = 0) # Ensuring the y-axis starts at 0 for a better comperison

# Print the ggplot object
print(ggplot_boxplot)
```



The boxplot interpretation:

When Home Loss/Draw:

The median number of shots taken by the Home Team when they either lost or drew is slightly higher than 10. The range of shots is broad, with a few outliers above 25. For the Away Team, the median shots are a bit higher than the home team, indicating that away teams tend to shoot more when the home team doesn't win.

When Home Win:

The median number of shots taken by the Home Team is approximately 15 when they win, which is noticeably higher than when they lose or draw. Interestingly, the median number of shots by the Away Team during home team victories is slightly lower than the home team's shots during losses or draws. The range of shots is also narrower in this situation, suggesting more consistent performance.

Overall Observations:

Teams tend to take more shots when they win, which is intuitive. The home team, especially, has a noticeable increase in the number of shots when they win compared to when they lose or draw.

c)

```
# Initially, fit a model with all the predictors:
# Adding interaction terms and some transformations
full_model <- glm(home_win ~ hs + hr + as +hst + ast + hc + ac + I(log(hst+1))
                  + I(log(ast+1)) + hf + af
                  + hy + ay+ ar ,
                  data=soccer, family="binomial")
# Backward Elimination
final_model <- step(full_model, direction="backward")
# Summary of final model
summary(final_model)
# Results are hidden for aesthetic reason
```

```
# Tidy up the model
tidy_model <- tidy(final_model, conf.int = TRUE)
# Create a separate d9ata frame for diagnostic statistics
diagnostic_df <-glance(final_model)

# The tables display above the code chunk for some reason that I could not figure out.
# So leaving this description as requested
# Print the model coefficients table
kable(tidy_model, caption = "Model Coefficients", align = 'c', format = "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 2: Model Coefficients

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.7198033	1.1606255	0.6201857	0.5351356	-1.5535844	3.0210934
hr	-2.2201580	1.1444528	-1.9399297	0.0523882	-5.2303407	-0.3335560
hc	-0.1483868	0.0645584	-2.2984899	0.0215339	-0.2782320	-0.0239353
ac	0.1830467	0.0746488	2.4521046	0.0142023	0.0404071	0.3343594
I(log(hst + 1))	3.2321451	0.5181982	6.2372760	0.0000000	2.2876362	4.3259482
I(log(ast + 1))	-3.6418780	0.5228094	-6.9659769	0.0000000	-4.7413432	-2.6834711

```
# Print the Diagnostic Statistics
kable(diagnostic_df, caption = "Diagnostic Statistics", align = 'c', format = "latex", booktabs = TRUE)
  kable_styling(latex_options = c("hold_position"))
```

Table 3: Diagnostic Statistics

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
400.8301	291	-104.8424	221.6848	243.7454	209.6848	286	292

Interpretation:

This model predicts the probability of a home team's win in soccer based on match statistics:

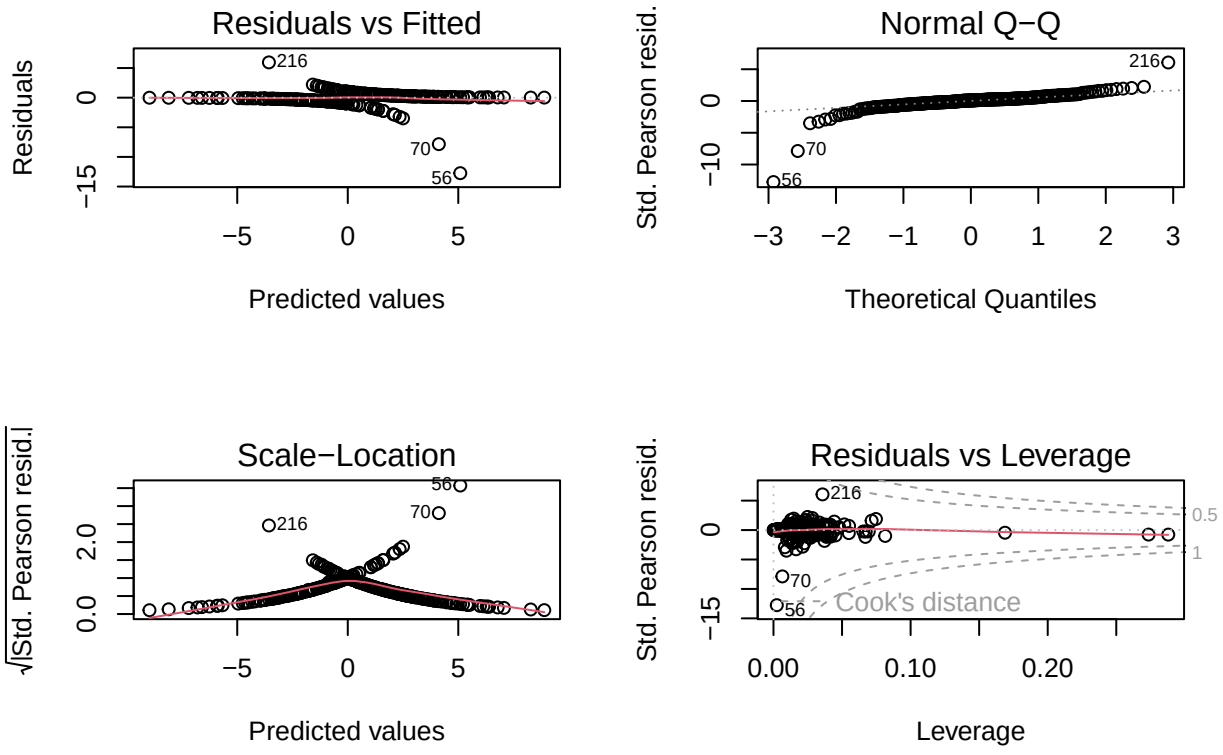
- **Intercept**: The estimated log-odds of a home win when all predictor variables are 0 is 0.71980, which is not statistically significant ($p = 0.5351$).

- **Red Cards (hr):** Each additional red card for the home team is associated with a decrease in the odds of winning, which is near significant ($p = 0.0524$).
- **Home Corners (hc):** Surprisingly, more corners for the home team are linked with a slight decrease in winning odds ($p = 0.0215$).
- **Away Corners (ac):** More corners for the away team are correlated with a slight increase in home team winning odds ($p = 0.0142$).
- **Home Shots on Target (log-transformed hst):** Increasing the number of home shots on target is strongly related to increased winning odds ($p < 0.001$).
- **Away Shots on Target (log-transformed ast):** More shots on target for the away team significantly decrease the home team's odds of winning ($p < 0.001$).

Comment on the Model: The model's AIC (221.68) indicates a good fit relative to the complexity of the model.The model, optimized through backward elimination, maintains significant predictors and leverages log transformations to reduce skewness and enhance model accuracy, as reflected by the reduced AIC and BIC scores.

Diagnostic Plots:

```
par(mfrow = c(2,2))
plot(final_model)
```

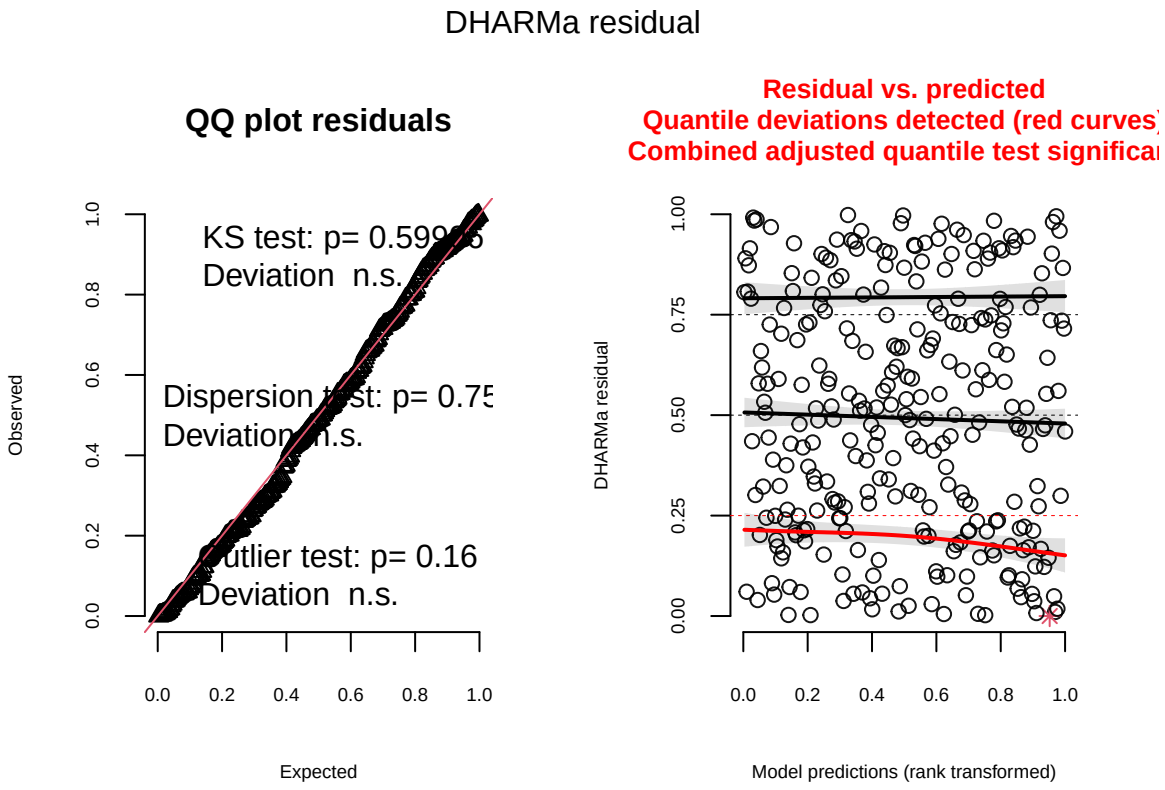


These plots above are designed with the assumption of continuous response data and are most suitable for models where residuals can be reasonably assumed to be normally distributed, such as linear regression models. However, for a binomial model (e.g., logistic regression), the residuals do not follow a normal distribution; instead, they are binomially distributed. The Normal Q-Q plot, which is used to check normality of

residuals, is not appropriate for residuals from a binomial model because these residuals will not be normally distributed. Additionally, the concept of homoscedasticity (constant variance across levels of predictors) doesn't apply in the same way to logistic regression, making the Scale-Location plot less relevant. The residuals vs fitted plot also becomes less interpretable due to the discrete nature of the response variable in a binomial distribution.

For binomial GLMs, alternative diagnostic approaches are more suitable. One such approach involves using diagnostic plots from the DHARMa package in R, which creates standardized residuals that are simulated from a uniform distribution under the model assumptions. This allows for the creation of diagnostic plots that are more interpretable for binomial and other non-normal models.

```
par(cex.main=0.6,
    cex.lab=0.6,
    cex.axis=0.6)
q <- simulateResiduals(fittedModel = final_model)
plot(q)
```

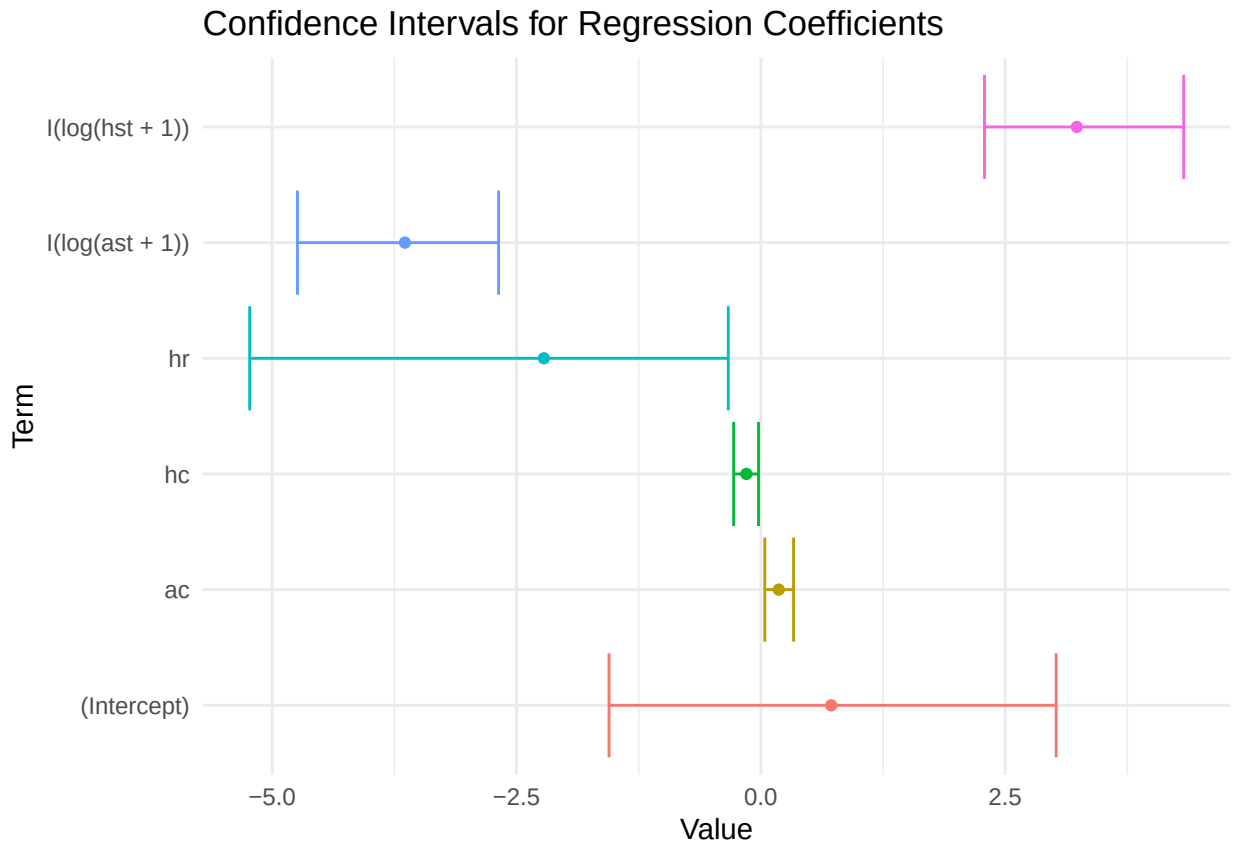


Diagnostic summary:

- QQ plot: Residuals closely follow the expected line, indicating they're well-distributed.
- Residual vs. predicted: Random distribution around the center, but red curves show some unexplained patterns.
- Tests: - KS test: Residuals fit expected uniform distribution (p=0.59906). - Dispersion test: Residuals' variance is consistent (p=0.752). - Outlier test: No significant outliers (p=0.14).
- In summary: The model is fairly good, but the red curve in the right plot suggests some patterns aren't captured.

d)

```
ggplot(tidy_model, aes(estimate, term, color = term)) +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Confidence Intervals for Regression Coefficients",
       x = "Value", y = "Term", color = "Term") +
  theme(legend.position = "none")
```



The plot showcases the confidence intervals for regression coefficients of various predictors. The terms I(log(hst+1)) and I(log(ast+1)) possess intervals that are entirely positive and negative respectively, signaling their clear statistical significance in influencing the model's outcome. Meanwhile, terms like 'hr', 'hc', and 'ac' display intervals spanning both sides of the zero line or closely hugging it, indicating potential ambiguity in their influence on the dependent variable. Notably, the intercept's confidence interval, which crosses the zero mark, indicating it's not statistically significant at the chosen confidence level.

Note:

The code chunk for loading libraries is hidden, and also the results are hidden in Question 2 part c(model selection) due to aesthetic reasons. It can be found in the R Markdown file attached.

AI Use Acknowledgement

I have used ChatGPT to revise my writing and debugging codes. It helped make the report more professional and concise.