Premier League Standings on Boxing Day's Impact on End-of-Season Standings

Introduction

This report explores whether a team's performance in the English Premier League (EPL) after Boxing Day's round of matches serves as a predictor for its end-of-season outcome. The central questions being investigated are: Is there a strong positive association between points a team has after Boxing Day matches and their final point total? Does a team's standing after Boxing Day matches accurately predict their final standing?

Boxing Day, December 26th, is an important day in English Football (Soccer) Culture due to its position halfway through the EPL season (August to May). It's believed that the results after Boxing Day matches are predictors for how well the teams will perform till the end of the season. Boxing Day match results are significant because they will shape expectations for the rest of the season. The objective of this report is to assess whether this belief holds true by analyzing actual EPL performance data from the 2023–24 season. The goal is to provide data-driven insights to support or challenge this narrative for the upcoming ESPN article.

Data Summary and Discussion

The dataset consists of information on the 20 teams in the 2023-24 English Premier League season. It includes each team's point total and table position immediately after the Boxing Day matches and their final point total and final table position at the end of the season. The data was sourced from ESPN's published website, which posts the EPL's official statistics. The dataset includes the following variables (better seen in Table 1 in Appendix A): Club (Name of the team), Points after Boxing Day Matches (earned as of end of Boxing Day), Position after Boxing Day Matches (League table ranking on end of Boxing Day, with 1 being best), Final Points (as earned by end of the season), Final Positions (League table ranking by end of season), and Color-coded Categories (rankings categorized by color indicating Champions League Qualification).

It is important to acknowledge a few data limitations. Two teams, Everton and Nottingham Forest, were penalized with point deductions due to breaches of the Premier League's Profitability and Sustainability Rules. Everton was initially docked 10 points (later reduced to 8), while Nottingham Forest was penalized 4 points between Boxing Day and the season's end. These administrative deductions introduce non-performance-based changes to point totals and standings, which could bias the relationship between Boxing Day and final outcomes. Despite these exceptions, the dataset provides a robust overview of team performances across the season (from mid-point to end), and is well-suited to explore associations between mid-season and final outcomes.

Methodology and Results

To answer the first research question, the study investigates whether there is a strong positive association between the number of points a team earned through Boxing Day and the total number of points a team has at the end of the season. From the data, it is known that these two variables are not independent of each other. Since these variables are not independent, the analysis first assesses whether the difference between them is normally distributed. Figure 1 (Appendix A) shows a histogram of the differences, which appears left-skewed, indicating non-normality. Therefore, this data does not follow the Normal distribution assumption for parametric methods. Additionally, the sample size provided is quite small, at 20 teams, further establishing the necessity of a nonparametric method. The analysis examines a scatterplot (Figure 2 in Appendix A) between these two variables to determine if there is a linear relationship to further explore which test will fit best. The scatterplot shows a strong linear trend between final points and points after Boxing Day matches for each team. Additionally, the correlation between these two variables is very high, at 0.915. Therefore, the research has established that this research question requires a non-parametric method, the two variables (Boxing Day points and end-of-season points) are dependent, and that their relationship is linear. These observations meet the assumption of the Permutation Test for Linear Association. The null hypothesis is that there is no linear association between points after Boxing Day matches and total end-of-season points per team (slope is 0). The alternative hypothesis is that as Boxing Day match points increase, so do final points at the end of the season (slope is greater than 0). The analysis uses permutations of the data to gather a large number of observations to test the significance of the test. This means that the study checks if the slope obtained from the original data is unusual by randomly mixing up the data many times and seeing how often a slope that large is achieved. Since the data is paired, the observations are permuted across pairs of Boxing Day points and final season points. The test is conducted using R.

To answer the second research question, the study tests whether a high placement on Boxing Day indicates a high placement at the end of the season. Therefore, the explanatory variable is the team's placement at Boxing Day, and the response variable is the team's placement at the end of the season. Both of these are ordinal, meaning that there is a clear ordering to the variable (4 placement types). Based on these stipulations, the Jonckheere-Terpstra test is the best test to use to investigate this relationship. First, numeric labels are assigned to each of the placement groups: 1 for Champions League, 2 for Europa League or Conference League, 3 for safety teams, and 4 for teams relegated to England's second division. A contingency table (Table 2 in Appendix A) is created where the rows are the 4 placements for teams after the Boxing Day matches, and the columns are the 4 placements for teams at the end of the season. The 16 values in the table represent the frequency of teams for each combination of placements. The JT-Test is then performed on this data. The null hypothesis is that there is no difference between the distributions of placements after Boxing Day at the end of the season. The alternative hypothesis is that teams with higher placements after Boxing Day (lower numeric values, like 1 for Champions League) tend to also have higher placements (lower numeric values) at the end of the season. This is the increasing alternative hypothesis. The test is conducted using R.

Results

For the permutation test for linear association, after fitting a linear regression line to the data, with the x-variable being Boxing Day points and the y-variable being final season points, the slope of the line is 1.8867. After permuting the data and performing the test, the analysis finds that the p-value is 0.

Additionally, the 95% confidence interval is 1.474 and 2.231, which means that there is 95% confidence that the slope of Boxing Day match points and final season points is between 1.474 and 2.231.

After performing the Jonkheere-Terpstra test with 5000 permutations (for the same reason permutations are performed for the first research question), the p-value of this test is 0.0002, which is well below the alpha threshold of 0.05.

Discussion

This analysis examines whether a team's performance at the end of the Boxing Day rounds can reliably predict their outcomes by the end of the season. Two core questions were addressed: Is there a strong positive association between points a team has after Boxing Day matches and their final point total? Does a team's standing after Boxing Day matches accurately predict their final standing?

The analysis began by examining the distribution of differences between Boxing Day and final point totals. The left-skewed histogram and small sample size of 20 indicated that traditional parametric methods like linear regression were not suitable. Instead, the permutation test for linear regression was applied, which avoids the assumption of normality. The results supported a strong linear association, with the correlation between Boxing Day points and final points being 0.915, and the slope of the fitted line was 1.8867. Therefore, for every additional point earned by Boxing Day, a team gained nearly two more by season's end. A 95% bootstrap confidence interval for the slope of [1.474,2,231] provided further evidence of the strength and reliability of the relationship. These findings are actionable as the point totals after Boxing Day rounds are not just reflective of a team's performance in the first half of the season, but predictive of the final outcomes, validating the belief that Boxing Day marks a strategic point in the season.

The analysis then examined placement categories, with teams grouped into four performance tiers: Champions League qualification (Top 4), Europa/Conference League qualification (5th-7th), Safety

(8th-17th), and Relegation (bottom 3). Using these ordered categories, a Jonckheere-Terpstra test was conducted, which is designed to find trends across ordinal groups. This nonparametric method allowed the determination of whether these placement categories stayed consistent. The results showed a p-value of 0.0002, which provides strong statistical evidence that higher placement at Boxing Day is associated with higher final placement. Even though individual placement may shift, the league's general structure remained consistent from Boxing Day all the way to the end of the season in May.

Although nonparametric methods are less common in sports reporting, they are more appropriate and provide more robust results for non-normal data. The use of these methods ensures that the findings are reliable and not dependent on assumptions that the data violates. As with any analysis, there are limitations. This study only examines one season of data, so broader generalizations should be made cautiously. External factors like administrative point deductions (Everton and Nottingham Forest) influenced standings in ways unrelated to general season performance. While these events are not common, they create noise in the data that can affect predictive models. Outcomes can be affected by injuries, transfers, and general sports dynamics, which create noise that cannot be captured in this dataset.

Conclusion

This analysis confirms that Boxing Day is a significant milestone in the EPL season with a strong, positive relationship between a team's points on Boxing Day and their final point total. Every Boxing Day point estimated approximately 1.9 final points, showing a statistically significant association. Team placement tiers at mid-season are also predictive of final placement. Teams in higher tiers at Boxing Day are more likely to finish in those same tiers, as confirmed by the Jonckheere-Terpstra test. From a reporting perspective, this analysis supports the belief that Boxing Day standings are meaningful, forecasting final standings. For stakeholders, the mid-season checkpoint provides a reliable prediction of where teams will be at the end of the season. While this report covers a single season, the methods and findings lay the foundation for more analysis with multiple seasons. Extending this analysis to more seasons would offer an even more robust foundation for predictions and commentary.

Appendix A: Graphs and Tables

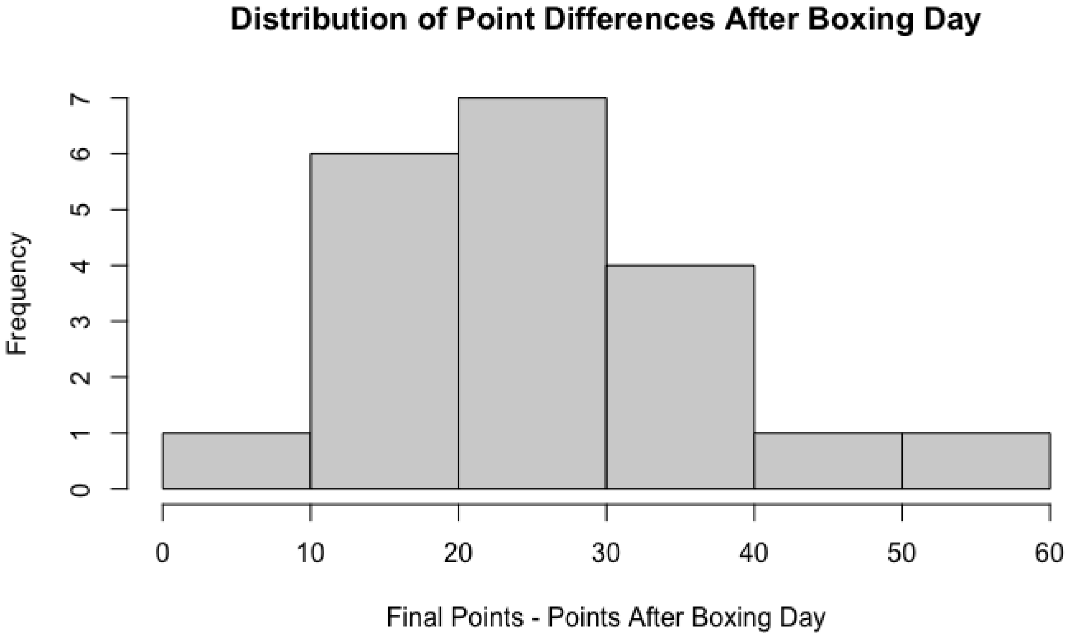


Figure 1: Histogram of the differences between a team's points after Boxing Day and at the end of the season

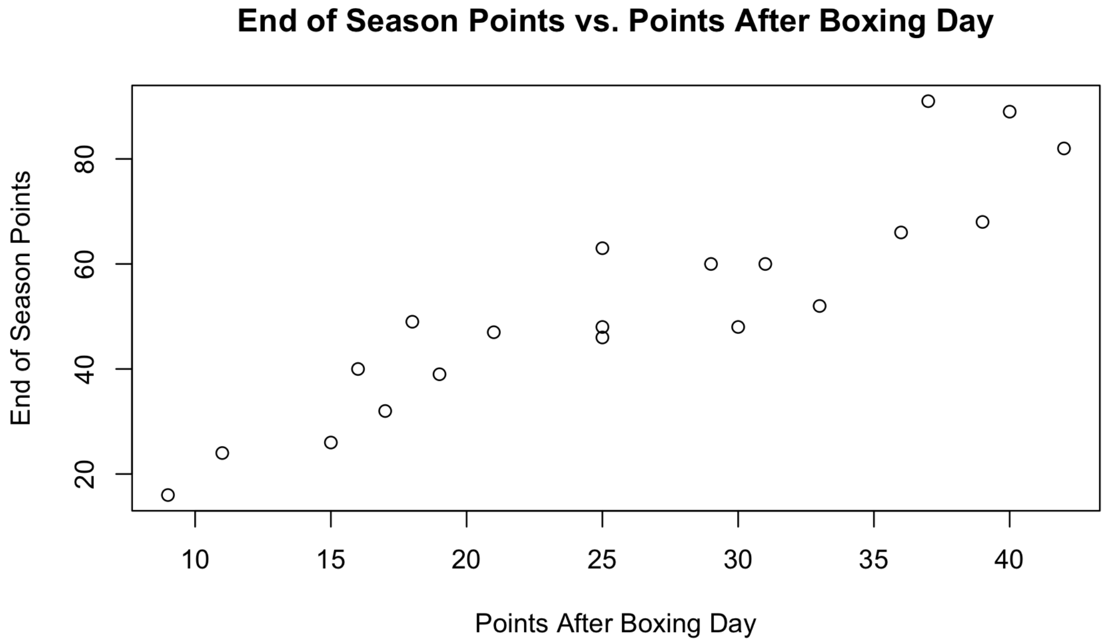


Figure 2: Scatterplot of the relationship between the teams' points after Boxing Day and at the end of the season

Variable	**Description**
Club	Name of the team
Points after Boxing Day Matches	Total Points earned as of end of Boxing Day Matches
Position after Boxing Day Matches	League Table ranking at end of Boxing Day Matches (1 being best)
Final Points	Total Points earned by the end of the season
Final Position	Final ranking in the league table (1-20)
Color-coded categories	Each position is also categorized by color to indicate Champions League qualification (green), Europa/Conference League (Blue), Safe from Relegation (Yellow), and Relegation (Red)

Table 1: Data Dictionary

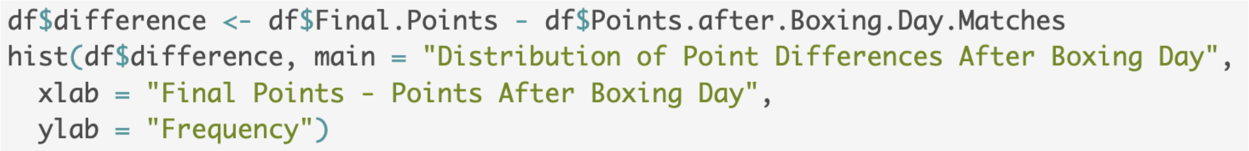
	End of Season Placement				
Placement After Boxing Day		Champion’s League	Europa or Conference League	Safety	Relegation
	Champion’s League	4	0	0	0
	Europa or Conference League	0	1	2	0
	Safety	0	2	8	0
	Relegation	0	0	0	3

Table 2: Contingency table on the frequency of teams at each combination of placements after Boxing Day and at the end of the season

Appendix B: R Code

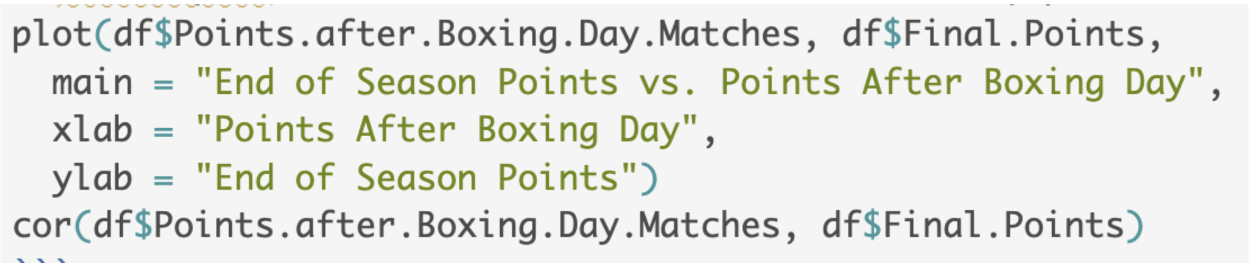
Part A: Research Question 1

1: Creating the histogram of the differences in points after Boxing Day and at the end of the season to check for normality.



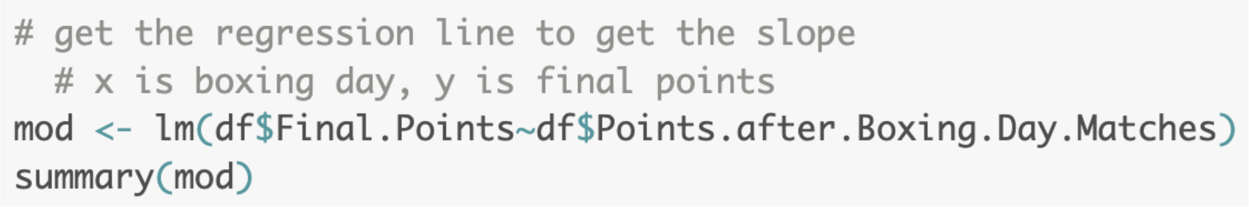
```r
df$difference <- df$Final.Points - df$Points.after.Boxing.Day.Matches
hist(df$difference, main = "Distribution of Point Differences After Boxing Day",
  xlab = "Final Points - Points After Boxing Day",
  ylab = "Frequency")
```

2: Creating a scatterplot of the differences in points to check for a linear relationship, and calculating the correlation between the two variables.



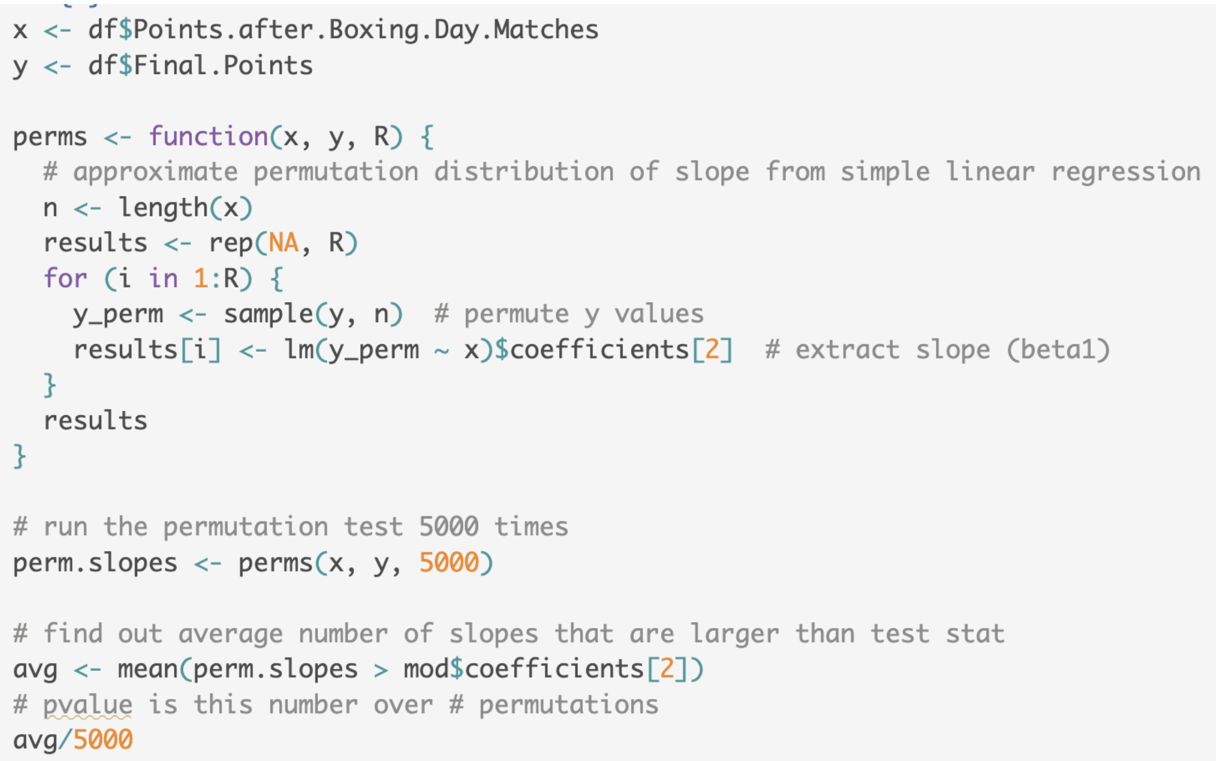
```r
plot(df$Points.after.Boxing.Day.Matches, df$Final.Points,
  main = "End of Season Points vs. Points After Boxing Day",
  xlab = "Points After Boxing Day",
  ylab = "End of Season Points")
cor(df$Points.after.Boxing.Day.Matches, df$Final.Points)
```

3: Obtain the test statistic for the permutation test for linear association by fitting a linear regression model to the data.



```r
# get the regression line to get the slope
  # x is boxing day, y is final points
mod <- lm(df$Final.Points~df$Points.after.Boxing.Day.Matches)
summary(mod)
```

4: Permutation test for Linear Association



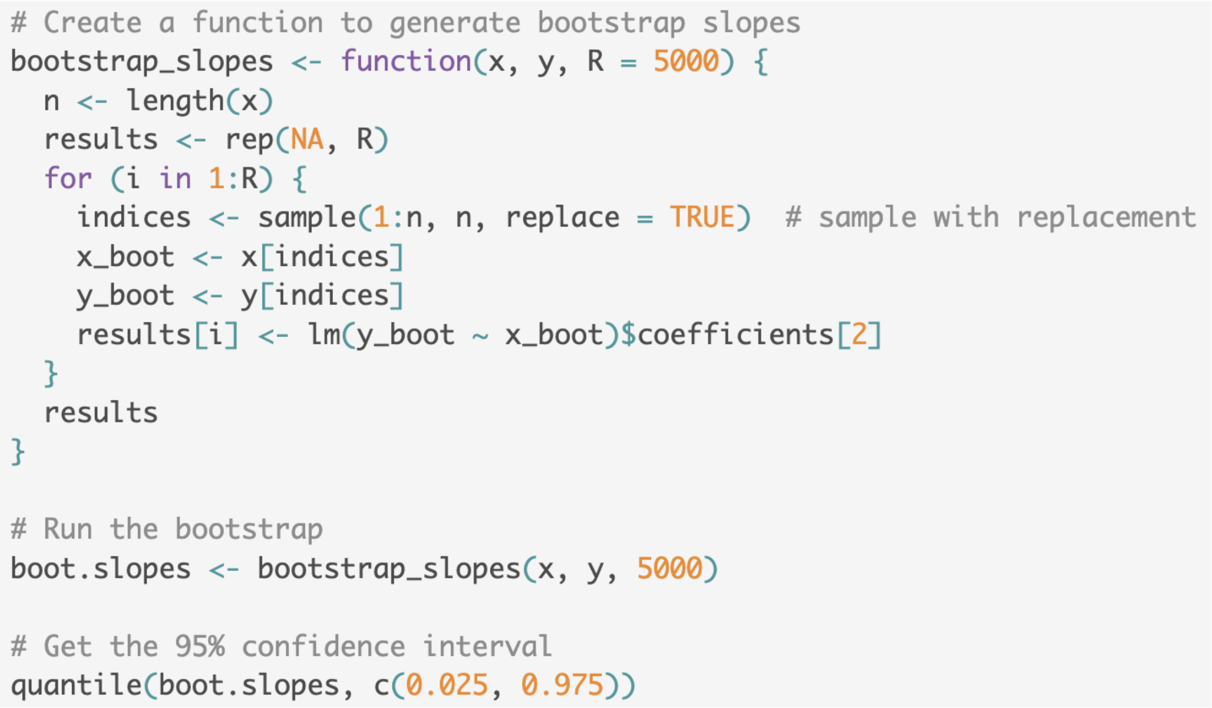
```
x <- df$Points.after.Boxing.Day.Matches
y <- df$Final.Points

perms <- function(x, y, R) {
  # approximate permutation distribution of slope from simple linear regression
  n <- length(x)
  results <- rep(NA, R)
  for (i in 1:R) {
    y_perm <- sample(y, n)  # permute y values
    results[i] <- lm(y_perm ~ x)$coefficients[2]  # extract slope (beta1)
  }
  results
}

# run the permutation test 5000 times
perm.slopes <- perms(x, y, 5000)

# find out average number of slopes that are larger than test stat
avg <- mean(perm.slopes > mod$coefficients[2])
# pvalue is this number over # permutations
avg/5000
```

5: 95% confidence interval for the slope.



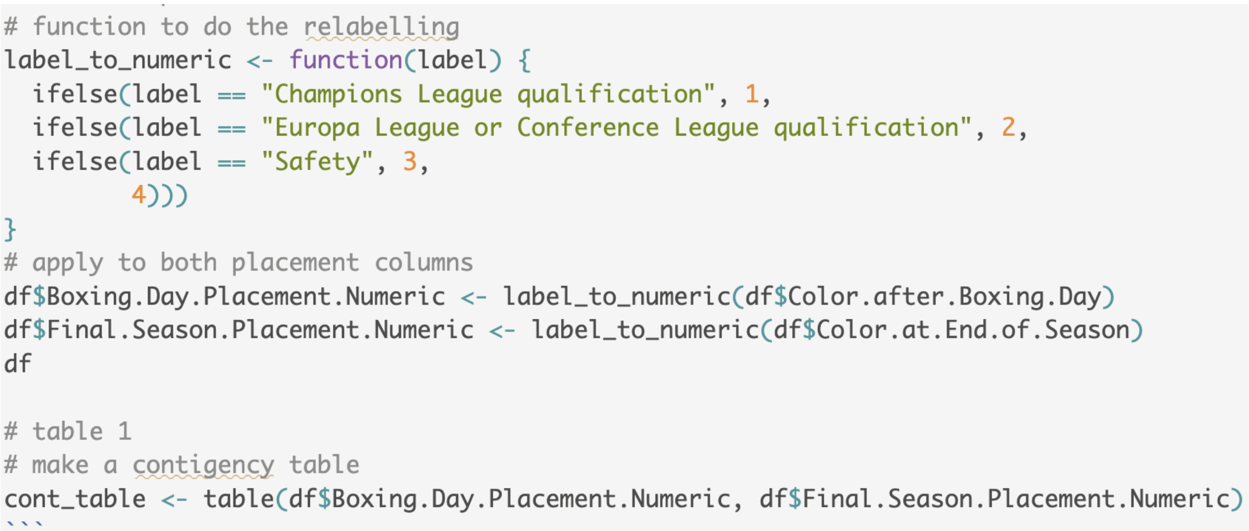
```
# Create a function to generate bootstrap slopes
bootstrap_slopes <- function(x, y, R = 5000) {
  n <- length(x)
  results <- rep(NA, R)
  for (i in 1:R) {
    indices <- sample(1:n, n, replace = TRUE)  # sample with replacement
    x_boot <- x[indices]
    y_boot <- y[indices]
    results[i] <- lm(y_boot ~ x_boot)$coefficients[2]
  }
  results
}

# Run the bootstrap
boot.slopes <- bootstrap_slopes(x, y, 5000)

# Get the 95% confidence interval
quantile(boot.slopes, c(0.025, 0.975))
```

Part B: Research Question 2

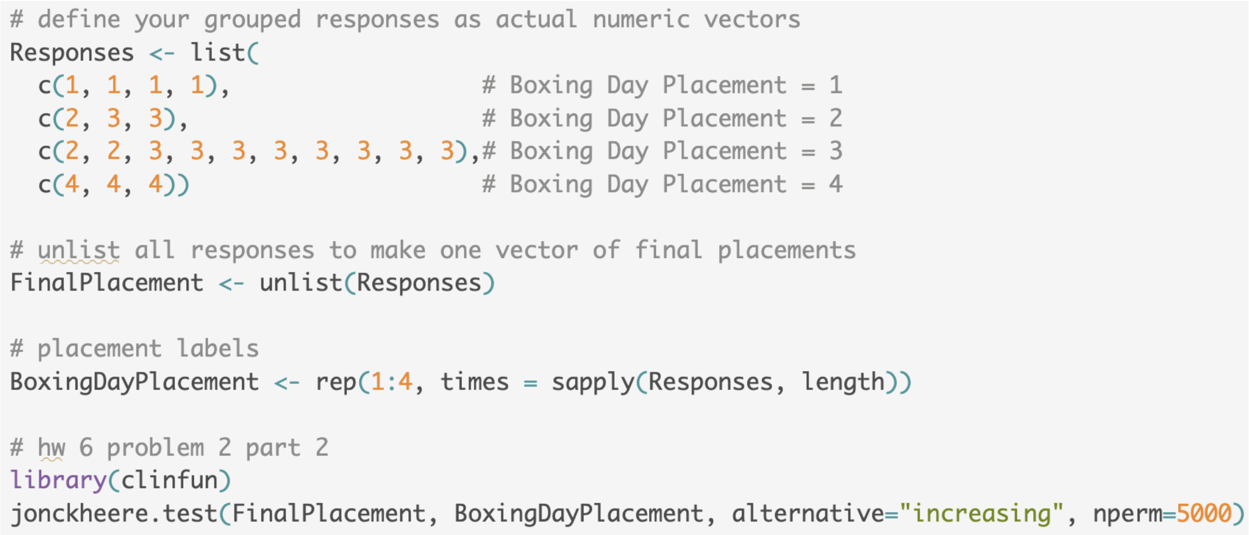
6: Manipulating the data into a format fit for the contingency table, and creating the contingency table.



~~~
# function to do the relabelling
label_to_numeric <- function(label) {
  ifelse(label == "Champions League qualification", 1,
  ifelse(label == "Europa League or Conference League qualification", 2,
  ifelse(label == "Safety", 3,
         4)))
}
# apply to both placement columns
df$Boxing.Day.Placement.Numeric <- label_to_numeric(df$Color.after.Boxing.Day)
df$Final.Season.Placement.Numeric <- label_to_numeric(df$Color.at.End.of.Season)
df

# table 1
# make a contigency table
cont_table <- table(df$Boxing.Day.Placement.Numeric, df$Final.Season.Placement.Numeric)
```
~~~

7: Cleaning and manipulating the data for the Jonckheere-Terpstra Test, then performing the test



```
# define your grouped responses as actual numeric vectors
Responses <- list(
  c(1, 1, 1, 1),                    # Boxing Day Placement = 1
  c(2, 3, 3),                       # Boxing Day Placement = 2
  c(2, 2, 3, 3, 3, 3, 3, 3, 3, 3),# Boxing Day Placement = 3
  c(4, 4, 4))                       # Boxing Day Placement = 4

# unlist all responses to make one vector of final placements
FinalPlacement <- unlist(Responses)

# placement labels
BoxingDayPlacement <- rep(1:4, times = sapply(Responses, length))

# hw 6 problem 2 part 2
library(clinfun)
jonckheere.test(FinalPlacement, BoxingDayPlacement, alternative="increasing", nperm=5000)
```