Data Mining Project Report

**Executive Summary**

This report presents an analysis of data collected from the 52 super bowls, Super Bowl I in 1967 to Super Bowl LII in 2018. It is a time series data set containing 53 rows and 26 different variables. The variables we analyzed consisted of two types. First we looked at categorical variables, such as conference, broadcast network, winning team, etc. We analyzed these with numerical variables, which consisted of variables such as average US viewers, ad cost, point differential, etc. Exploratory analysis was performed to address questions about variable correlations, their statistical significance, and key trends over time. There were 12 queries that were created based off of our data set, which include visualizations such as bar charts, scatter plots, box plots, and line graphs to portray insights into several key factors of the data set. Key findings include the dominance of quarterbacks in MVP awards, a strong positive correlation between average viewership and ad costs, and a weak negative relationship between ticket prices and attendance. Additional statistical tests, such as t-tests and correlation analyses, were conducted to validate observations, with results included in the visual outputs.

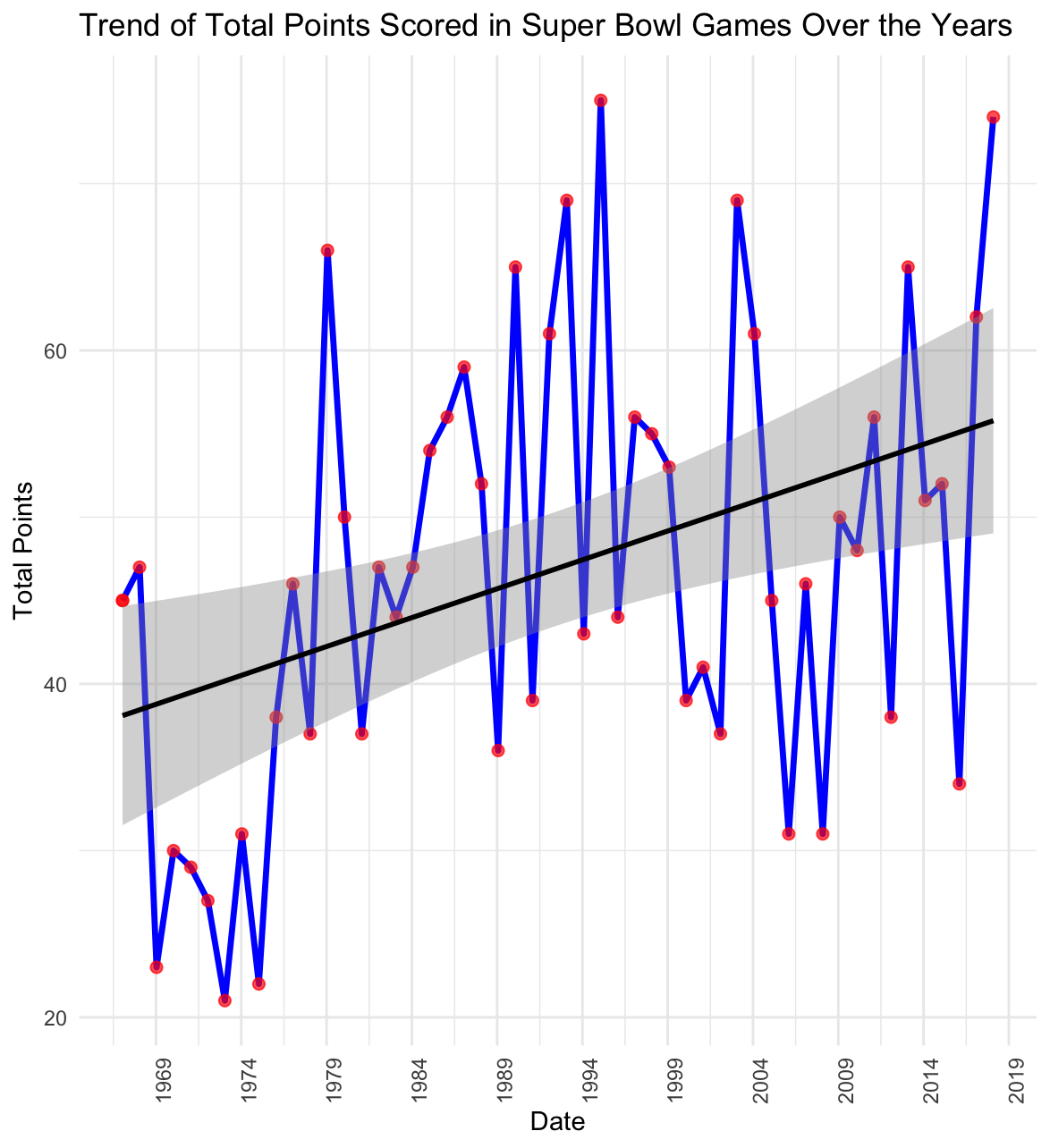
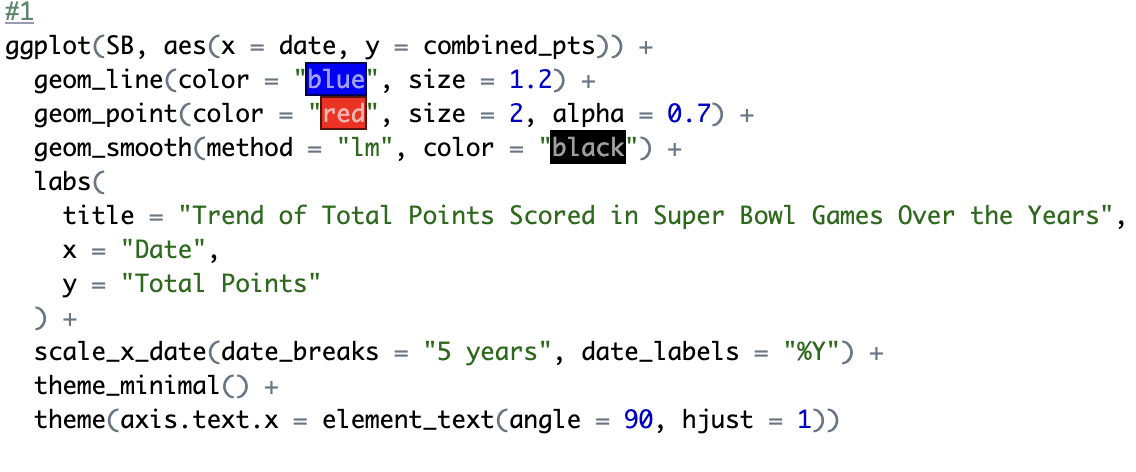
**Data Cleaning**

Once we decided on what data to use we had five datasets that we wanted to use to conduct our analysis. Four were from Kaggle and had minimal null observations or invalid values that needed cleaning. The other one we manually made in Excel. We first used select functions to get rid of unwanted variables like qb\_loser\_2 which gave us the backup quarterback of the losing team. We deemed this and a handful of other variables useless. Then we cleaned some of the entries by removing “+” from some names which indicated players that have passed away. We also needed to change all data sets' Super Bowl numbers to the same normal number format, so that they were cohesive in order to merge all the datasets (i.e. one dataset had the classic Roman numerals). Following these steps, we merged all data sets. Then we ran into a few issues with NFL teams that either changed names (i.e. the Washington Redskins to the Commanders) or changed location (i.e. the Baltimore Colts moving to Indianapolis). These issues resulted in NA values for our conference and division variable. We had to manually adjust each team that changed. Lastly, we converted variables like date from character to date or Super Bowl number to numeric in order for R to provide us with better visualizations.

**Analysis**

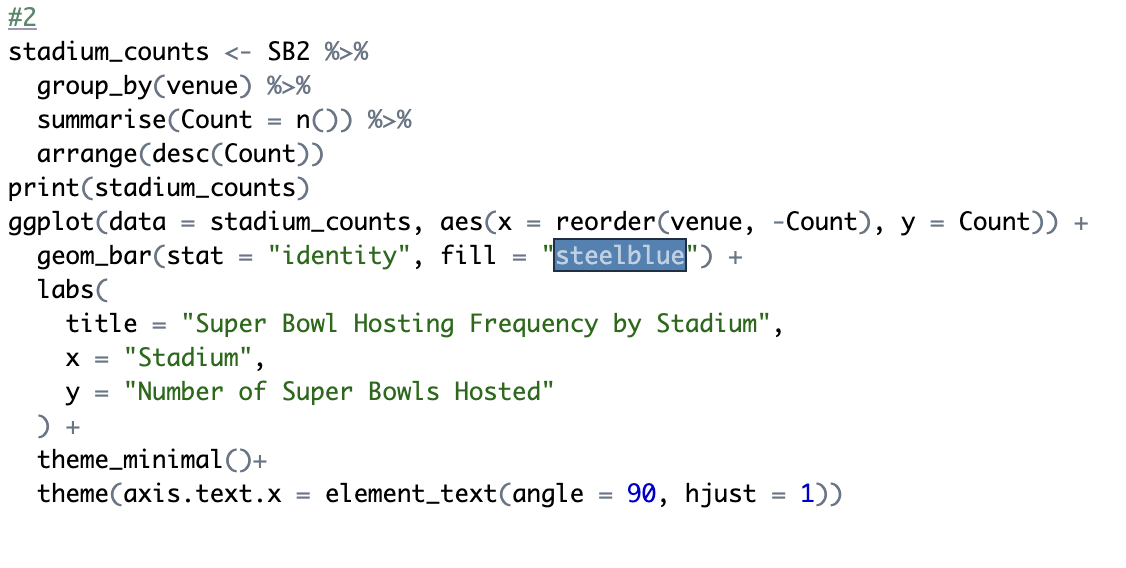
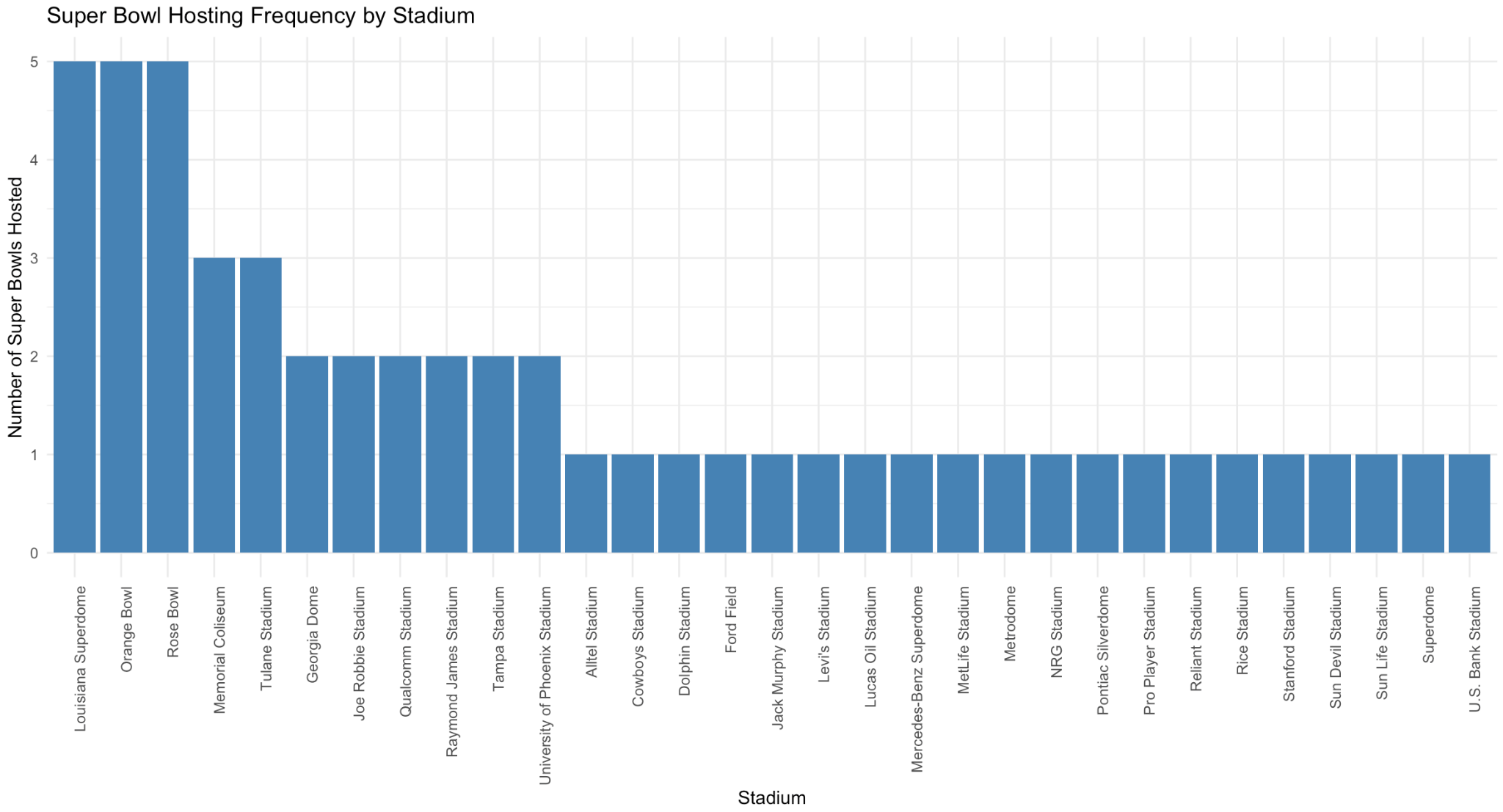
Query 1: Trend of total points scored

Query 1 analyzes the trend of the total points scored in each superbowl starting in 1967. We used ggplot, geom\_line, geom\_point, and geom\_smooth to create a line graph with a line of best fit to analyze how total points scored have changed over time. The graph shows a gradual increase over time, which can be attributed to the evolution of the sport. In the beginning, the league saw a run heavy offensive approach that would waste clock time and result in lower scoring games. As time went on, we saw historically great quarterbacks enter the league, leading to a more pass heavy offensive approach, resulting in higher scoring games as we can see on the graph.



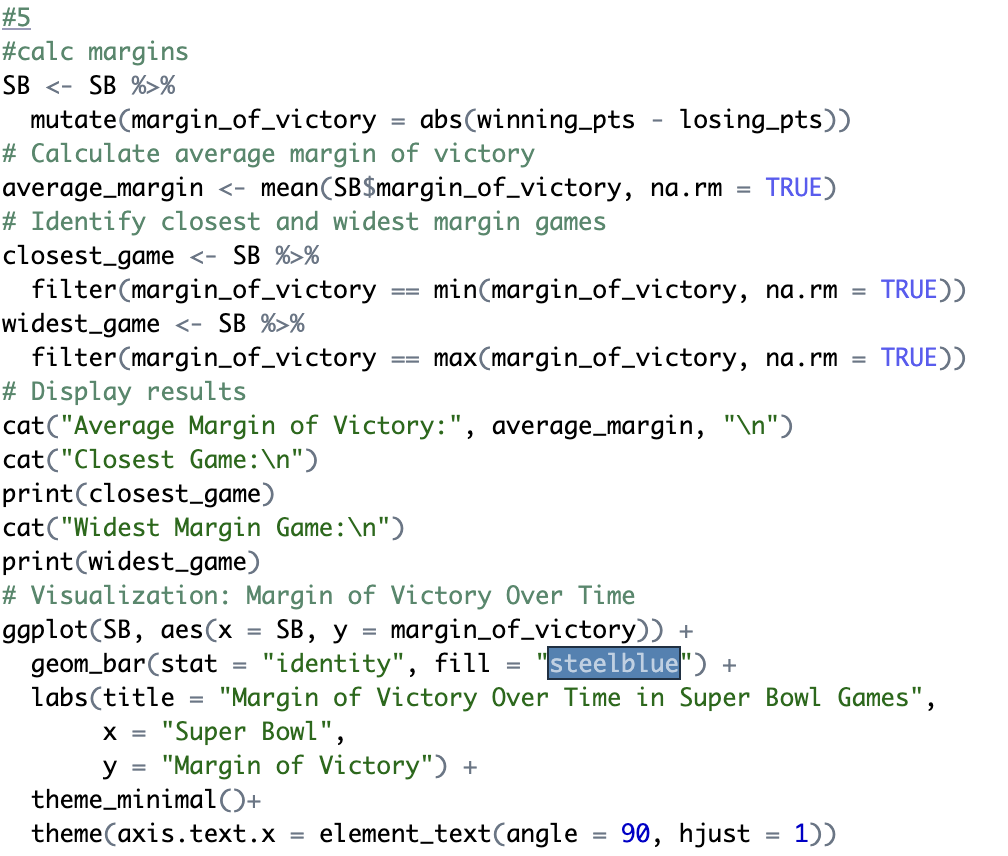
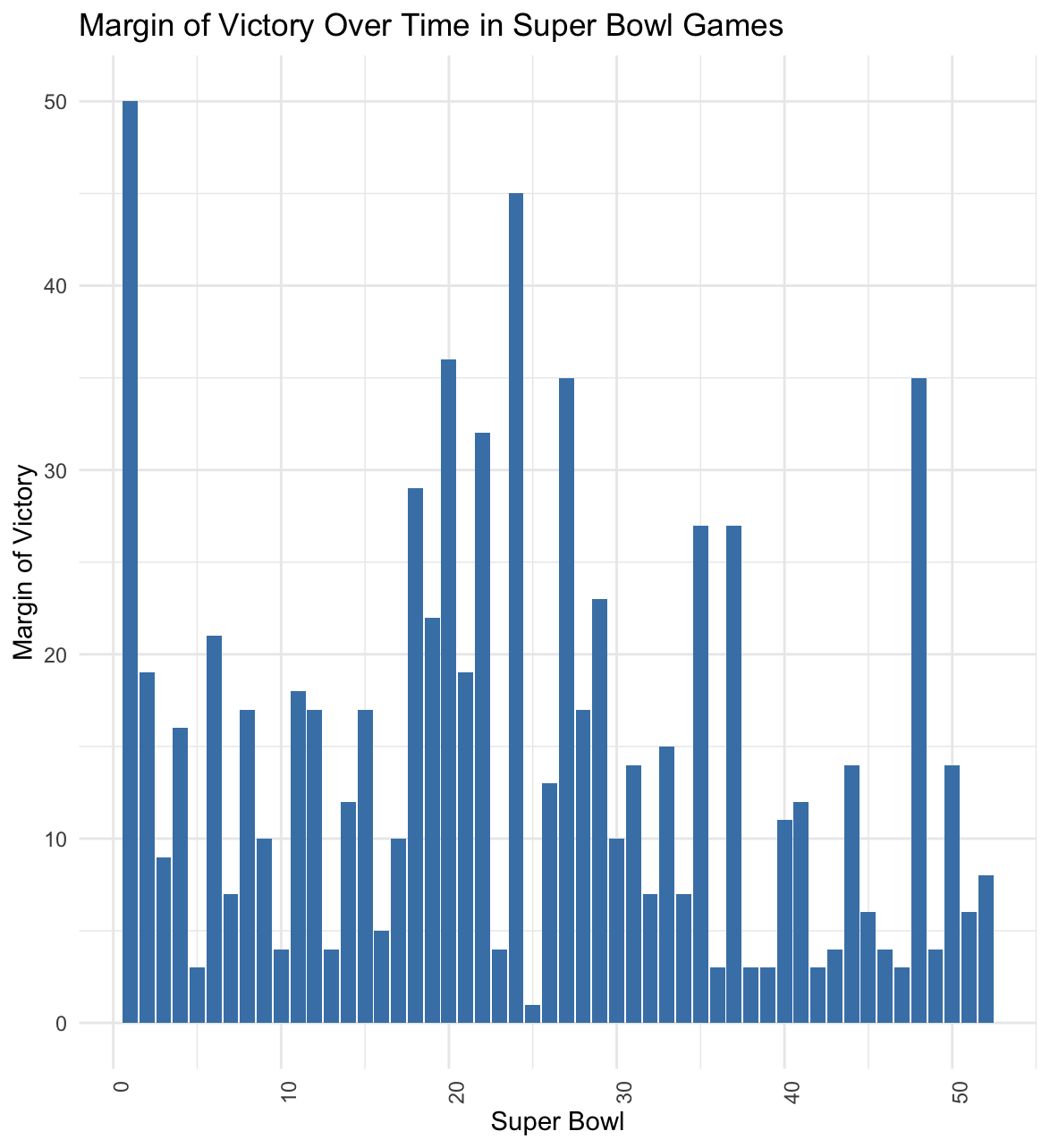
Query 2: Stadiums with the most Super Bowls hosted:

Query 2 highlights the frequency of Super Bowl games hosted by different stadiums, using a bar graph to display the number of games hosted for each stadium. The code groups the data by venue and calculates the count of Super Bowls hosted at each stadium. The results are sorted in descending order, and the graph is plotted with ggplot, where we used the reorder function to make sure the bars are displayed from most to least games hosted. The graph reveals that the Louisiana Superdome and Orange Bowl hosted the most Super Bowls, with five each, followed by the Rose Bowl and Memorial Coliseum.

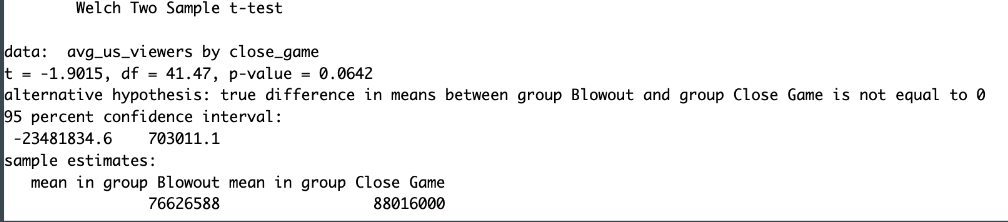


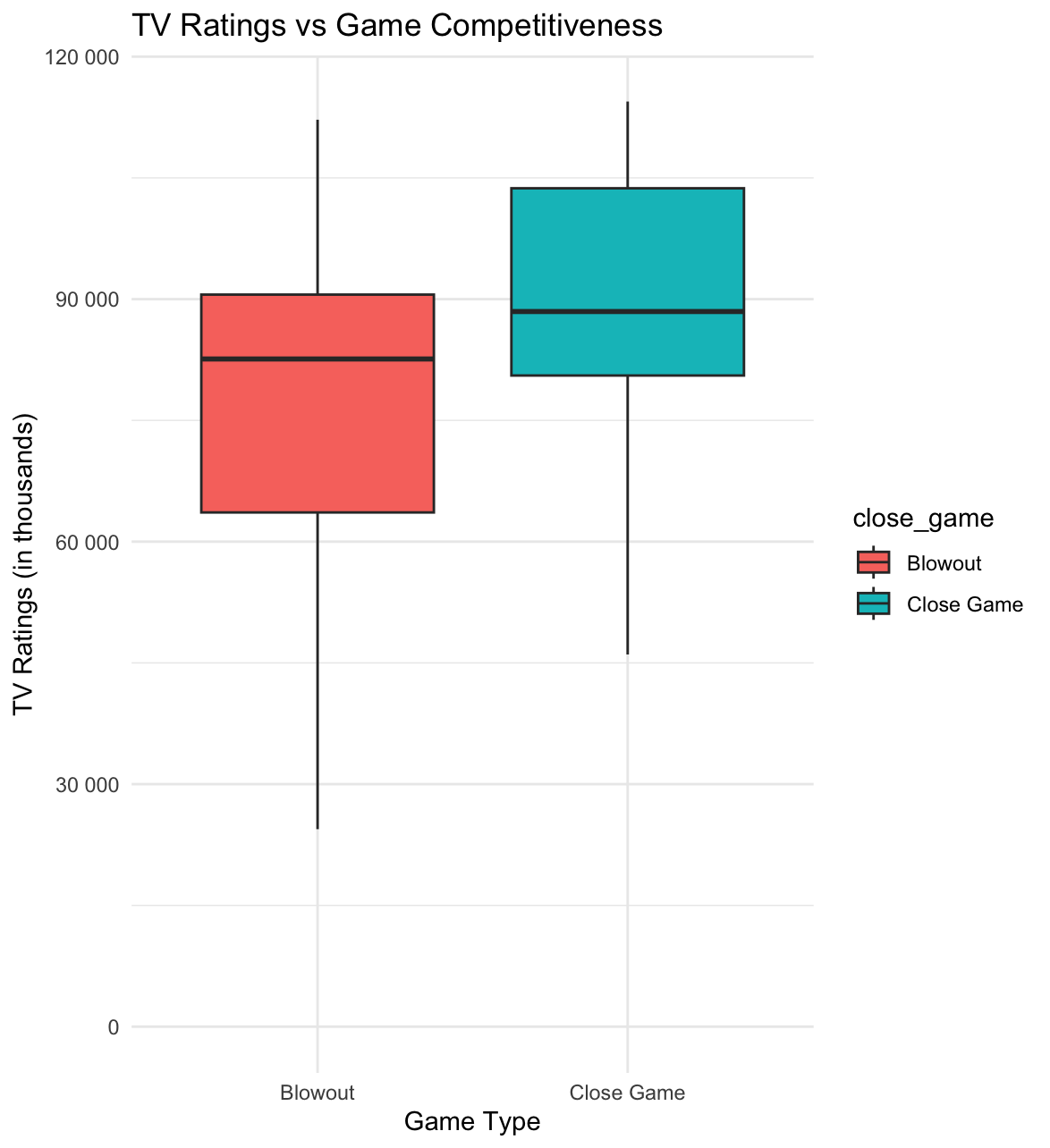
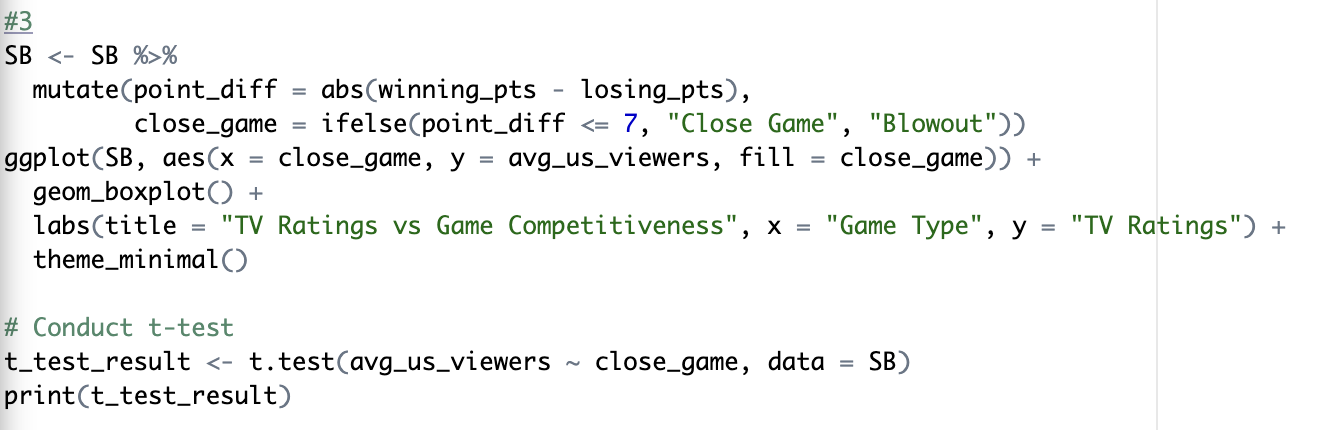
Query 3: Average margin of victory

Query 3 displays the margin of victory for each Super Bowl game over time in a bar graph. The code calculates the absolute point difference between winning and losing teams, represented as "margin of victory." Key insights include the closest and widest margins of victory, which are determined using minimum and maximum values. The average margin is also computed for context. The graph, generated with ggplot, clearly shows that some games had significantly higher margins, indicating less competitive matchups, while others were closer in score. The x-axis labels are rotated to make the graph look neat, and the bars are filled in "steelblue" for the visual appeal. This margin of victory helps identify trends in game competitiveness throughout Super Bowl history, however we can see that it is random, which makes sense as there is a different ebb and flow to every game.



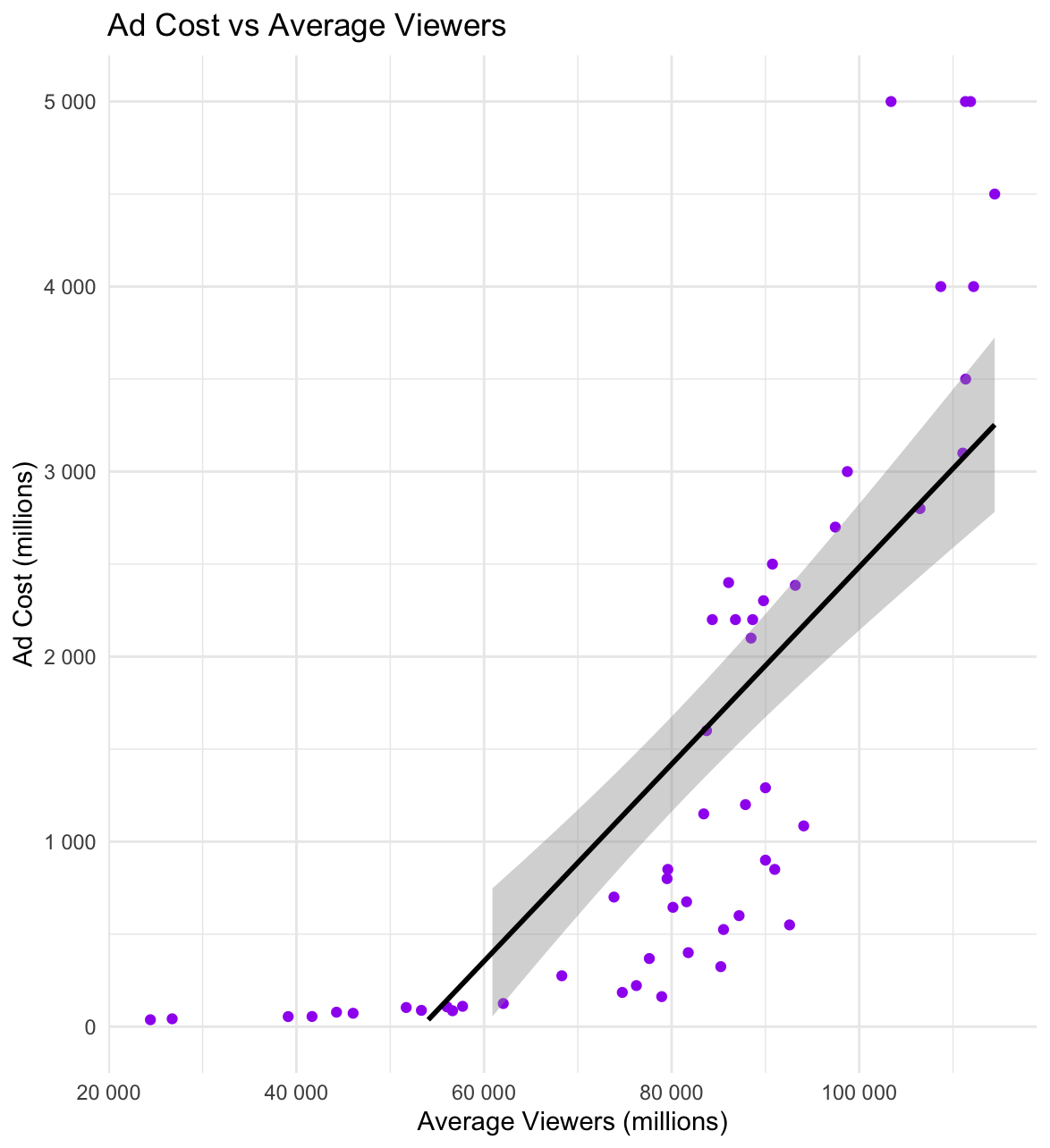
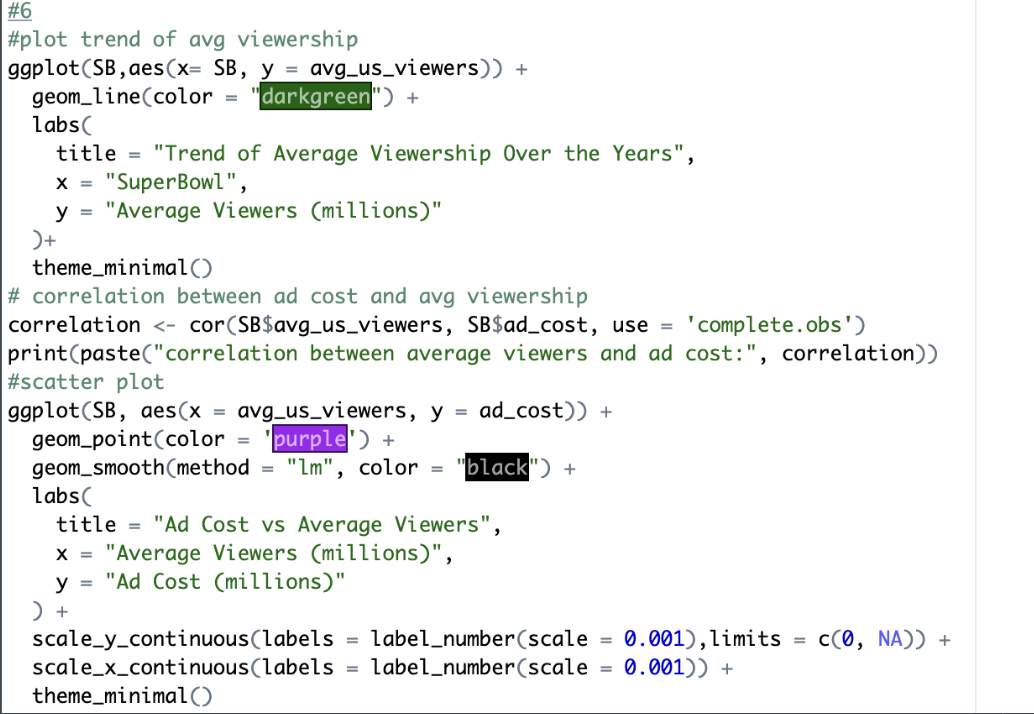
Query 4: Relationship between TV ratings and Super Bowl outcome

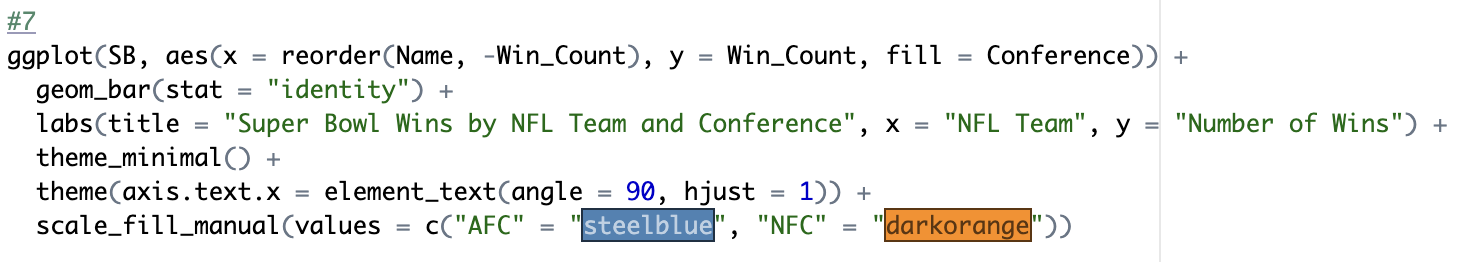
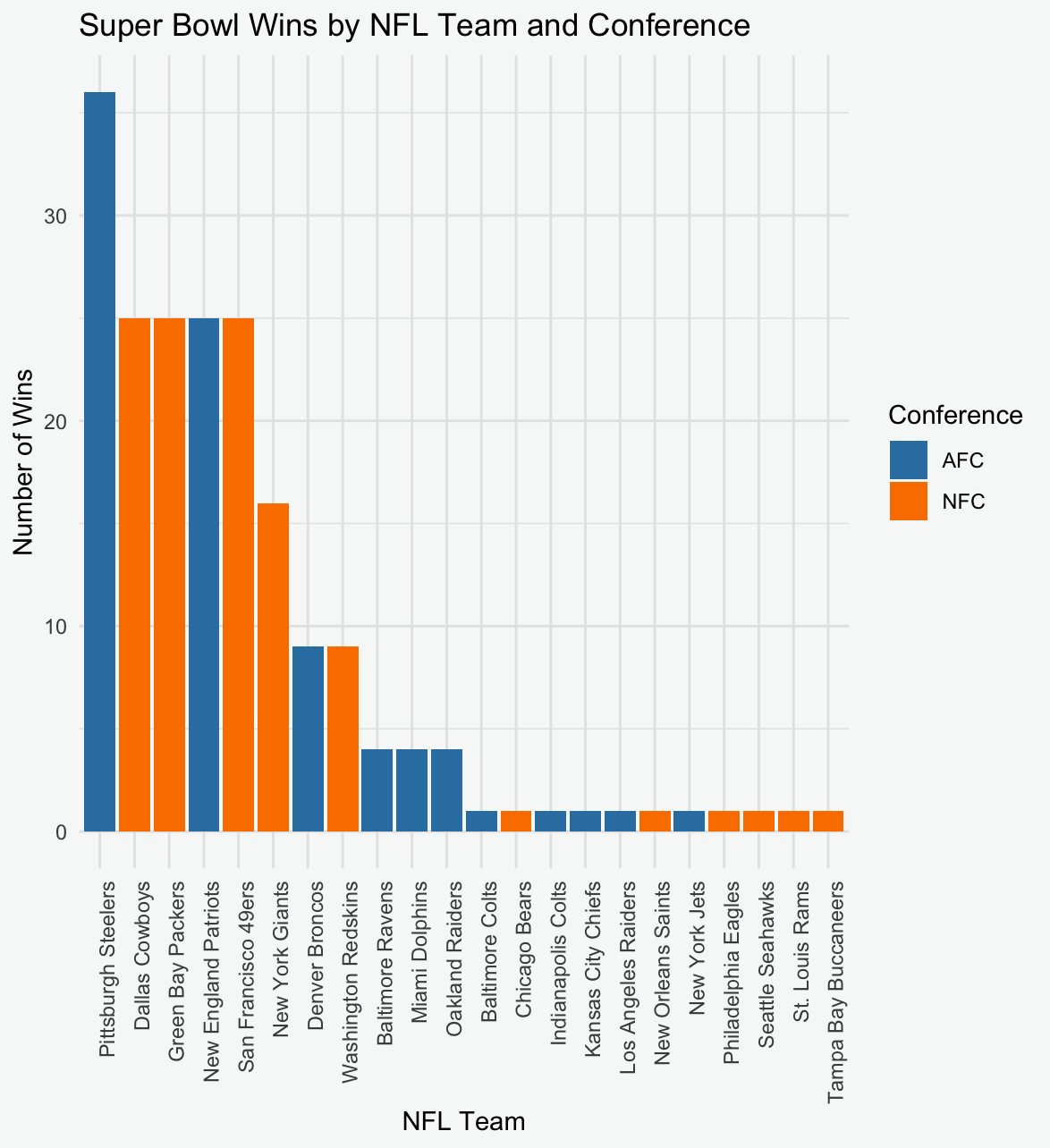
Query 4 analyzes the relationship between Super Bowl competitiveness and TV ratings, categorizing games as a "Close Game" (point difference ≤ 7) or a "Blowout." The box plot shows that close games generally attract higher median viewership, with a mean of 88,016,000 compared to 76,626,588 for blowouts. A Welch t-test was conducted to assess the difference, resulting in a p-value of 0.0642, which is slightly above the statistically significant level of 0.05. While not statistically significant, the trend suggests that closer games may engage larger audiences, highlighting the potential impact of game competitiveness on viewership.



Query 5: Correlation between TV viewership and ad cost

Query 5 shows the relationship between Super Bowl average viewership and ad costs. We calculated the correlation between the two variables using the cor () function, and used ggplot, geom\_point, and geom\_smooth to plot the data. The results show a strong positive correlation, indicating that higher viewership is associated with increased ad costs, demonstrating how audience size influences advertising expenses for the Super Bowl.

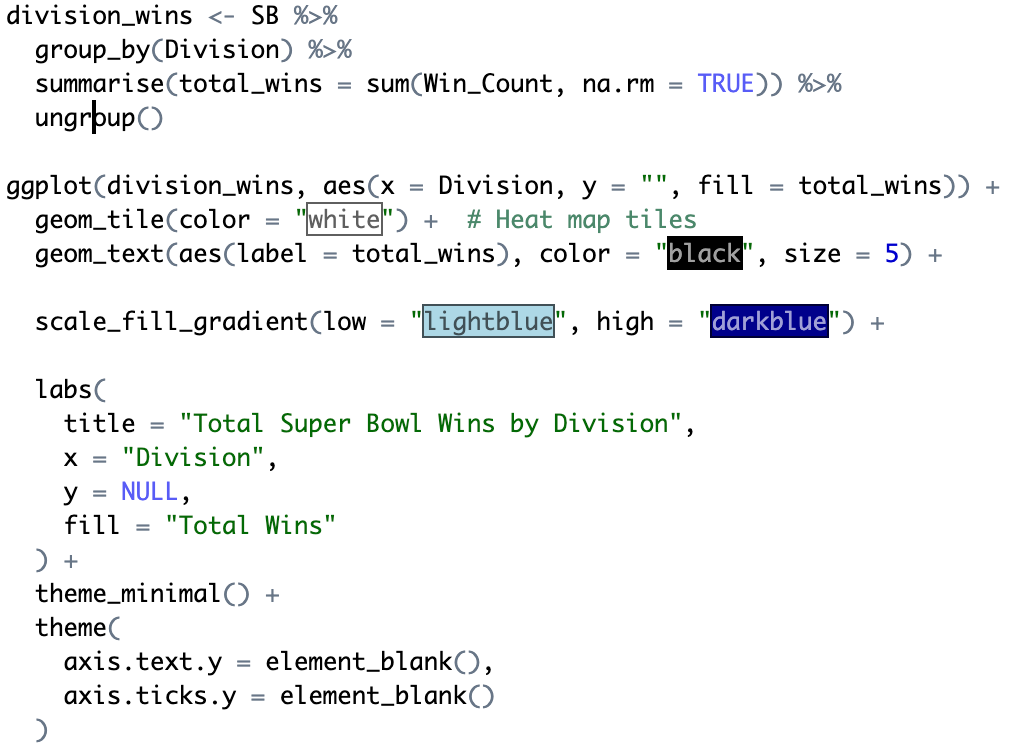
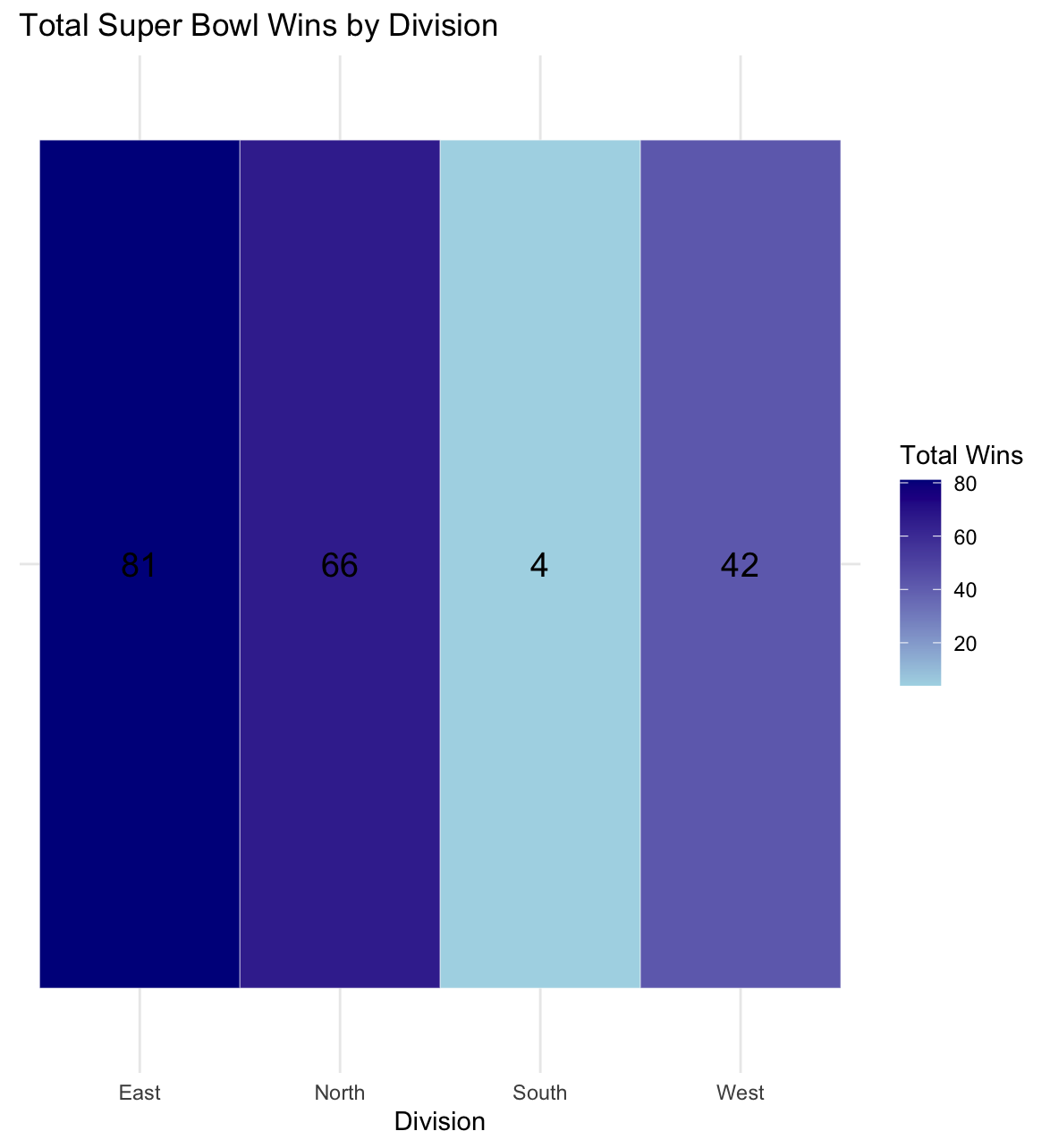


Query 6: Super Bowl wins by conference (AFC/NFC)

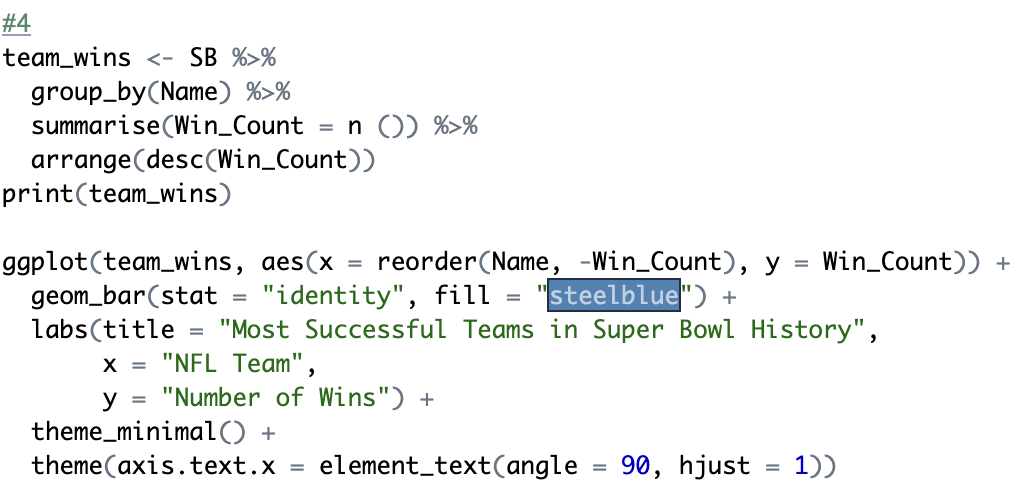
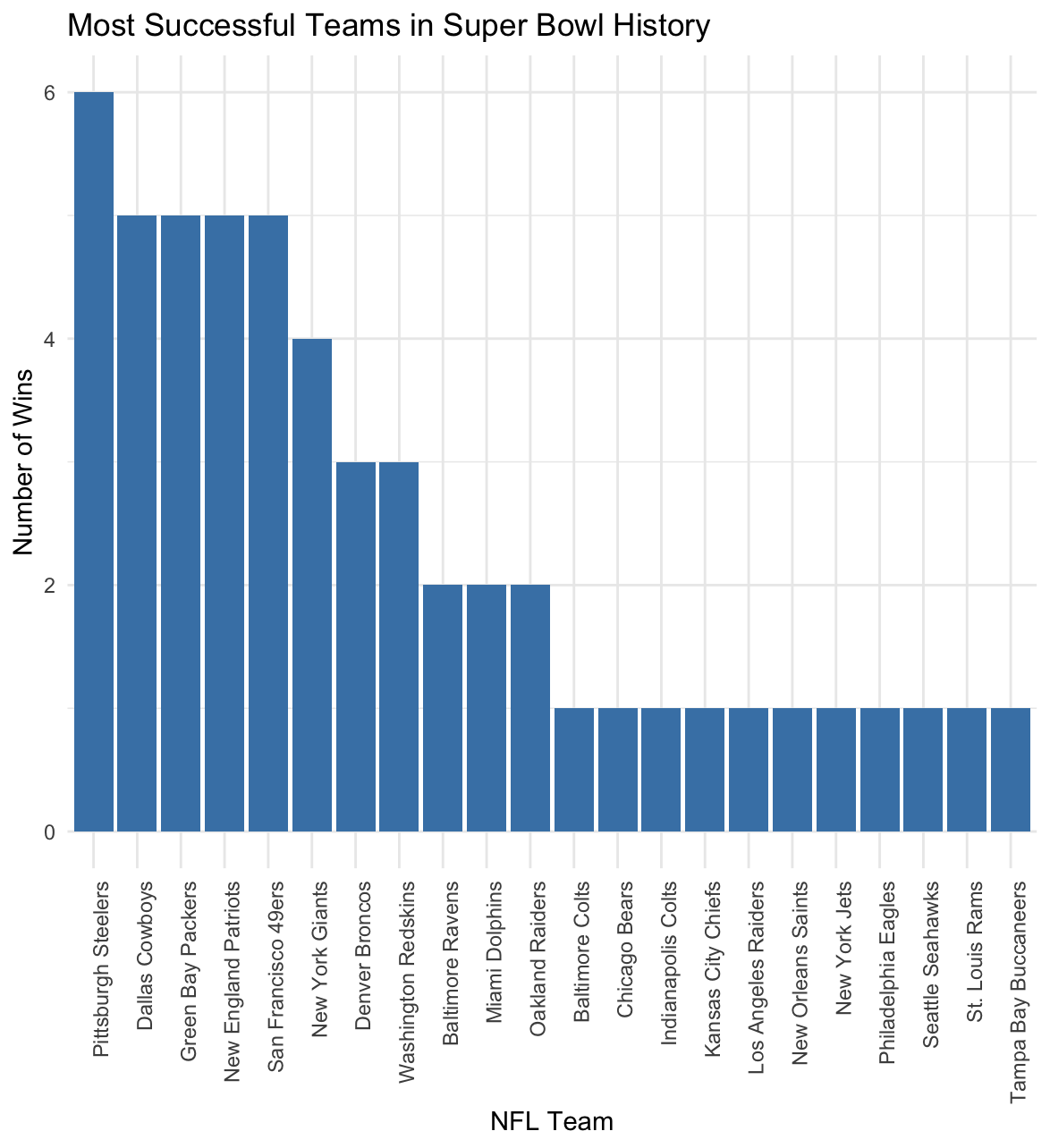
In Query 6 we used ggplot(), geom\_bar(), and scale\_fill\_manual(), to create a visualization where we could analyze which conference has the most superbowl win. We could conclude while the AFC has the team with the most superbowl wins the NFC has the second and third teams. In the top 5 the NFC holds 3 of those spots. Based on the results, we can conclude that the NFC is more likely to win the superbowl over the AFC.

Query 7: Super Bowl wins by division

For Query 7 we conducted a similar analysis to query 6 but instead of conference we analyzed divisions. We used a heat map to demonstrate the wins. We created a new variable to group wins by division and then used ggplot() to create the visualization below. Based on the graph we can see that South teams have a very slim chance of winning the superbowl. While teams in the east have the highest chance of winning the superbowl.

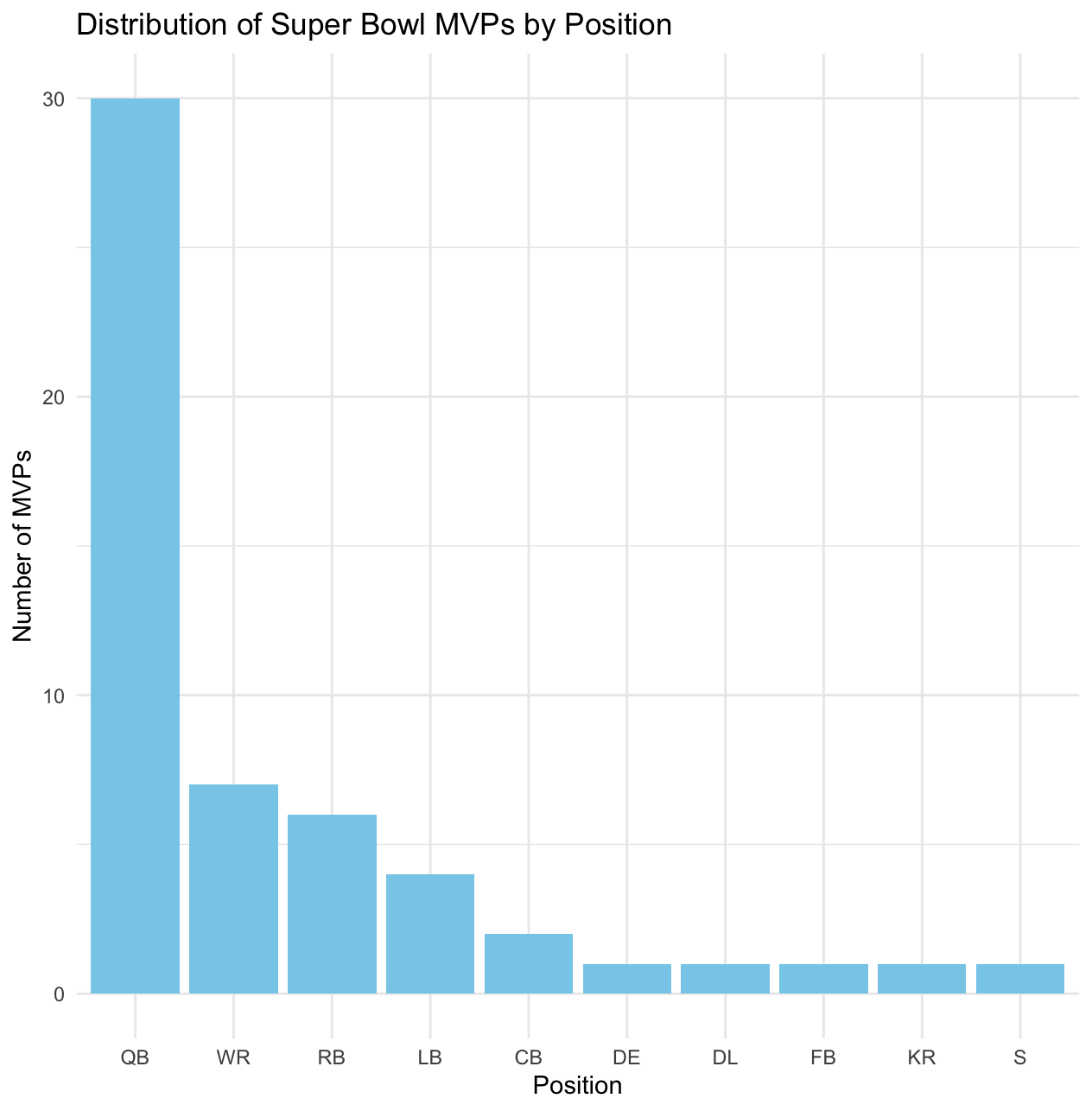
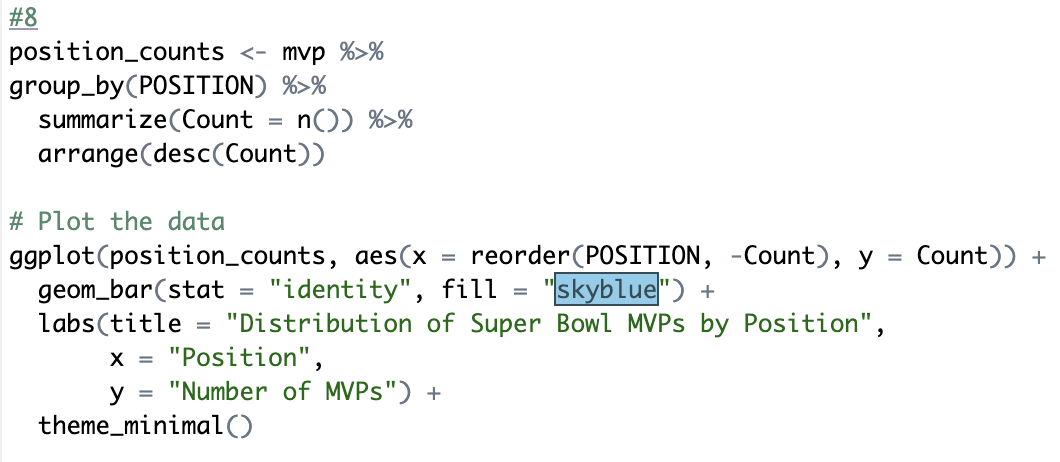


Query 8: Number of titles per team

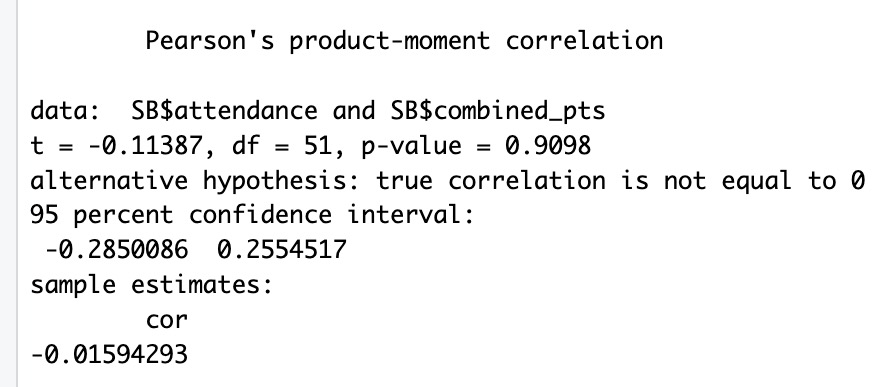
In query 8 we analyzed superbowl wins by team. We created a new variable to count the number of wins by each team. Then using the ggplot() function we created this bar graph to show wins by team. Based off of the visualization we can see that the Steelers have the most wins at 6 and the Cowboys, Packers, Patriots, 49ers are all tied for second with 5 superbowl wins each. We can conclude that the Steelers are the most successful team in the NFL currently with the four other teams lurking close by being only one superbowl away from matching them. 

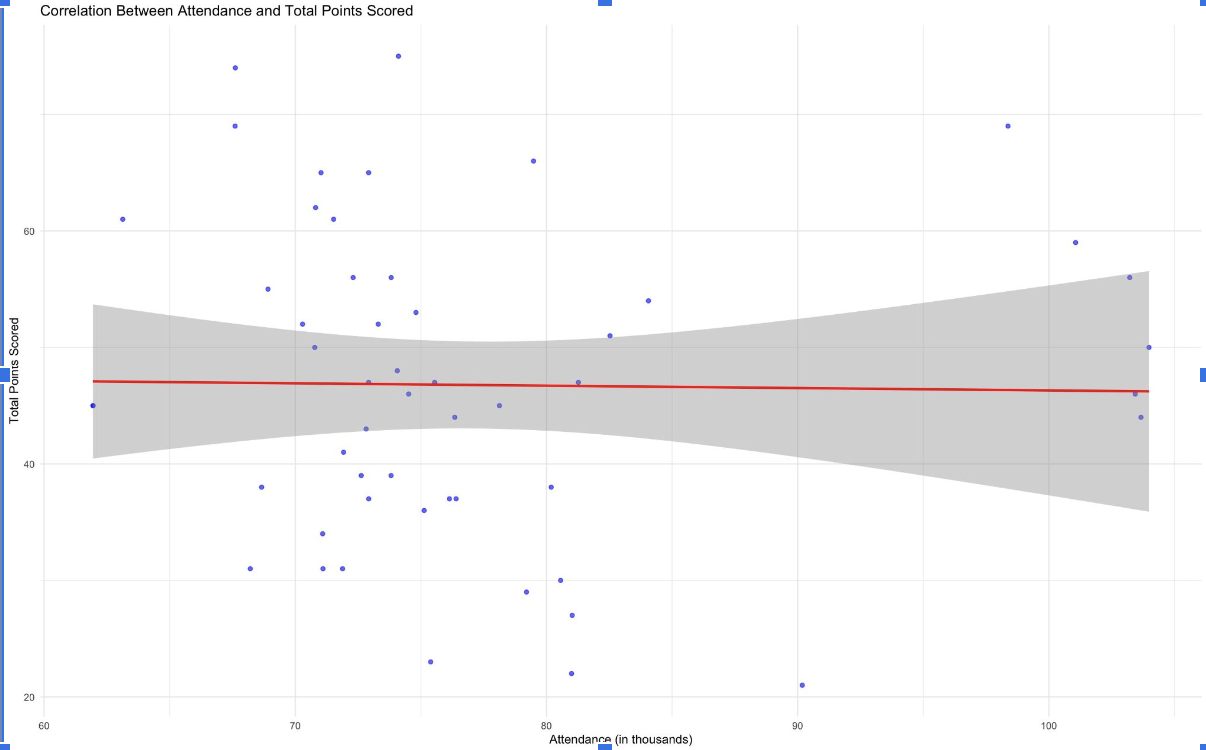
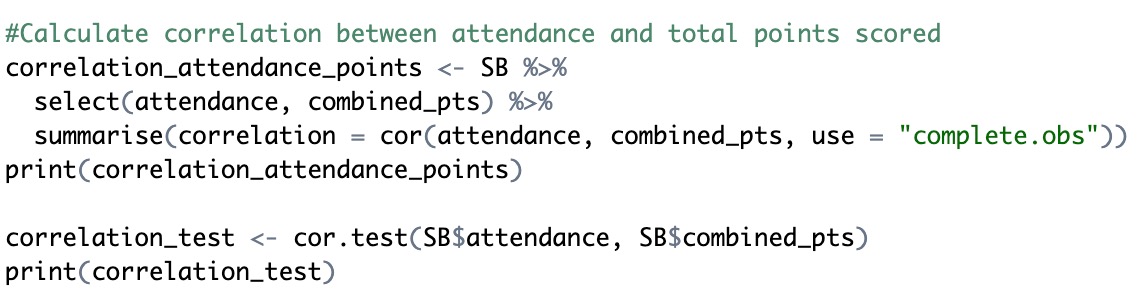
Query 9: Super Bowl MVP by position

Query 9 contains a bar chart that illustrates the distribution of Super Bowl MVPs by player position, with quarterbacks (QBs) dominating the awards. The code uses the group\_by () and summarize () functions to calculate the count of MVPs for each position, and the arrange () function sorts the results in descending order. Using ggplot, we created a chart that visualizes the data with sky blue-colored bars, making the dominance of QBs, with over 30 MVP awards, visually clear. Wide receivers (WR) and running backs (RB) follow as the next most common MVP positions. This visualization portrays the significant role of quarterbacks in determining Super Bowl outcomes and their tendency to receive the MVP award.



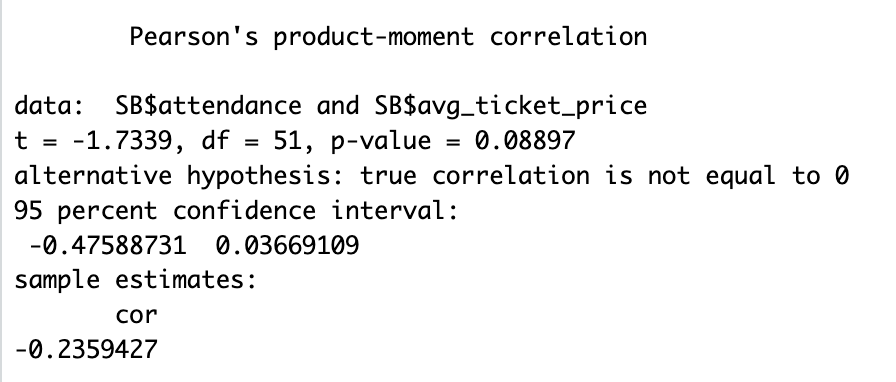
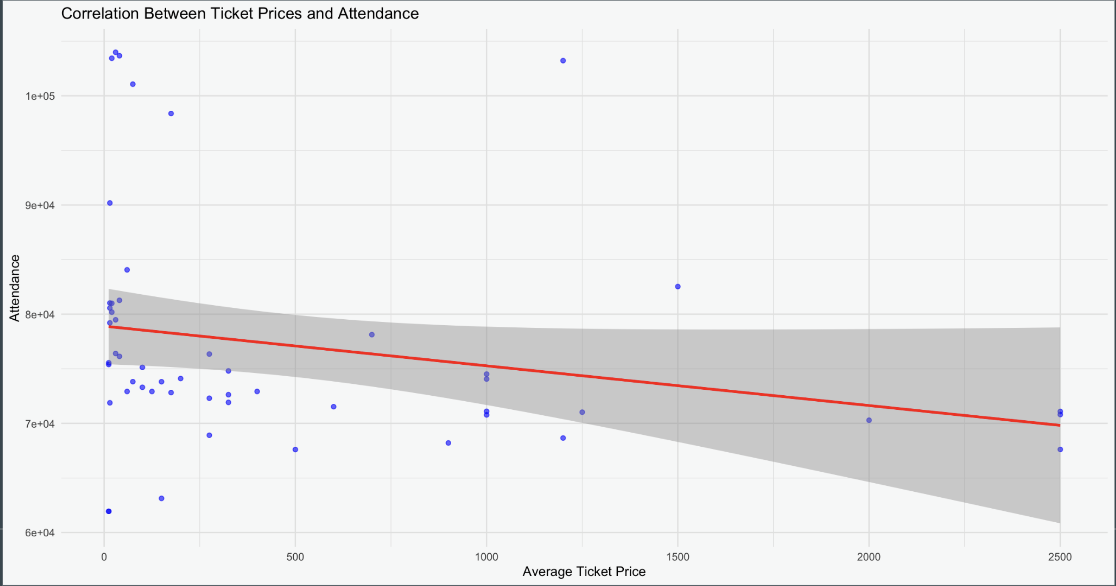
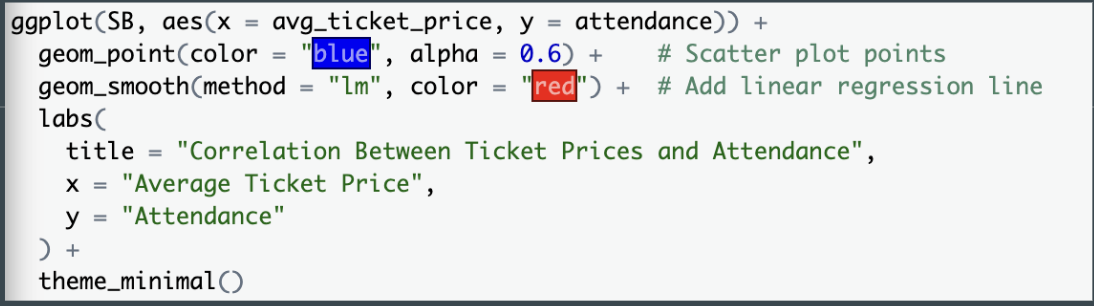
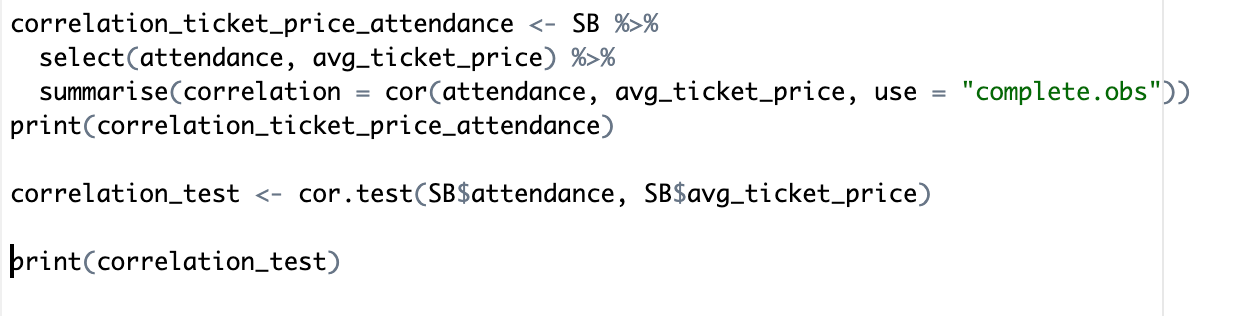
Query 10: Correlation between attendance and total points scored

Query 10 analyzes the correlation between Super Bowl attendance and total points scored. The code first calculates the Pearson correlation coefficient using the cor() function, followed by a hypothesis test with cor.test() to assess the statistical significance of the relationship. The resulting correlation coefficient is -0.016, with a p-value of 0.9098, allowing us to determine that there is no meaningful correlation between attendance and total points scored. This result is displayed in the scatter plot, where the red regression line remains relatively flat, confirming the lack of correlation. 



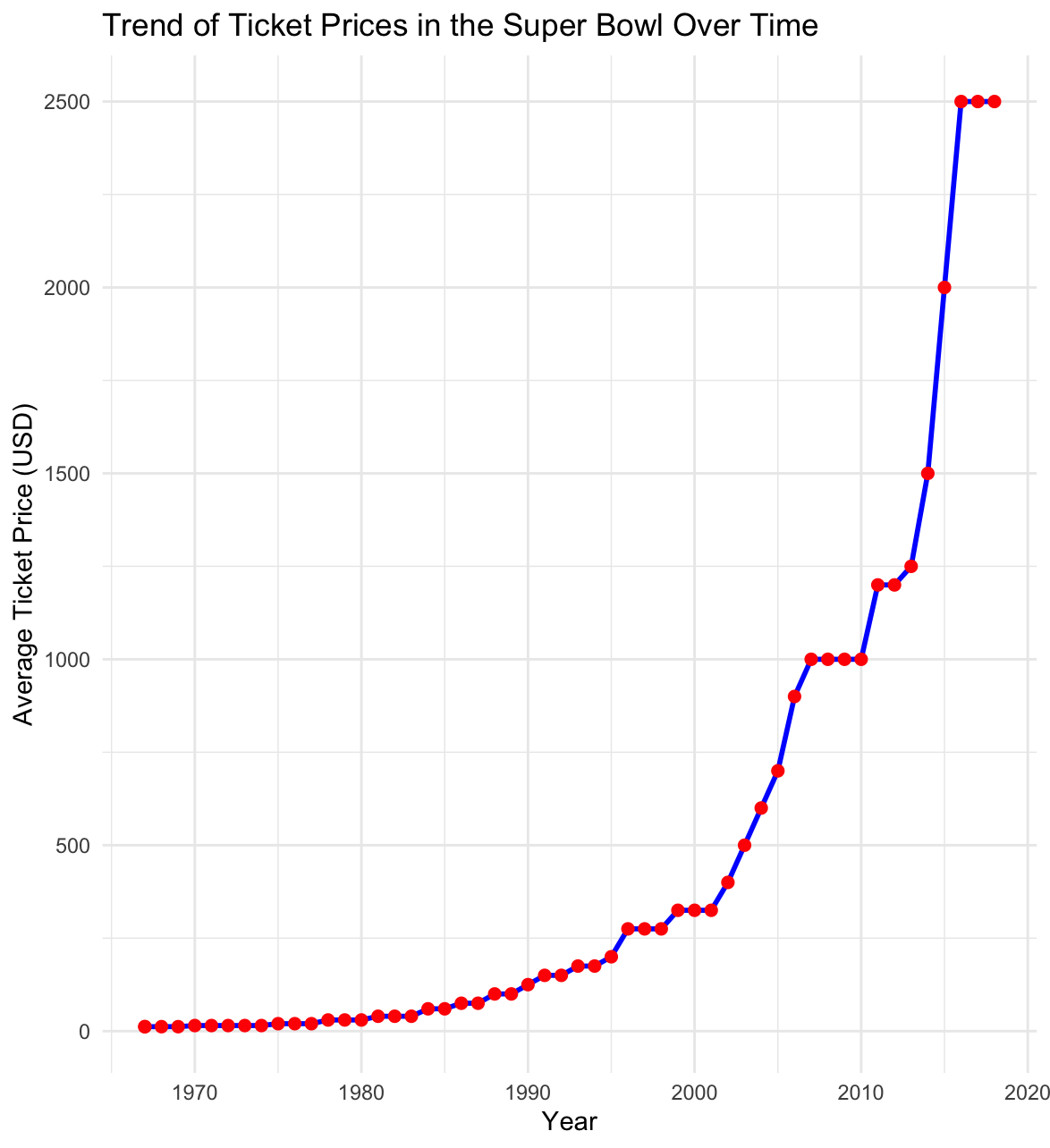
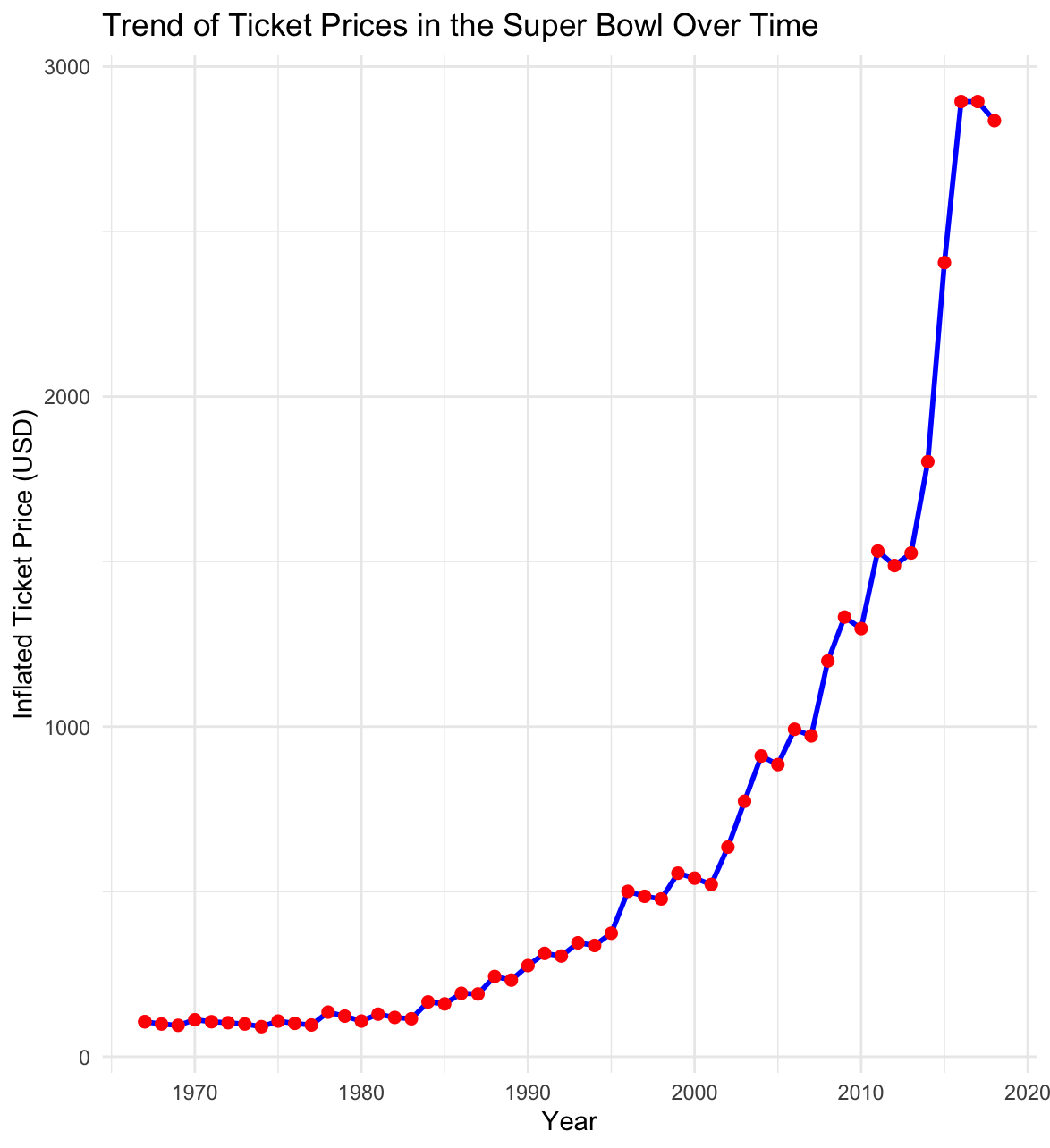
Query 11: Correlation between ticket prices and attendance

Since we were not able to determine that points scored had any correlation with attendance, we wanted to discover what did correlate with attendance. Therefore, query 11 examines the relationship between Super Bowl ticket prices and attendance. The code calculates the correlation coefficient using the cor() function and tests its statistical significance with cor.test(). The Pearson correlation result is -0.236, with a p-value of 0.08897, showing there is a weak negative correlation that is not statistically significant. The scatter plot visualizes this relationship, with blue points representing the data and a red regression line showing a slight downward trend. This suggests that higher average ticket prices may be weakly associated with lower attendance, but the relationship is not strong or significant.



Query 12:

Since we were able to determine ticket prices do have a slight correlation with attendance, we wanted to analyze the trend of ticket prices in general. The two line graphs display the trend of Super Bowl ticket prices over time, both adjusted for inflation and as average ticket prices. Each graph plots the years on the x-axis and ticket prices on the y-axis, with red points representing individual data values and a blue line connecting them to illustrate the trend. Both visualizations reveal a sharp increase in ticket prices starting around the early 2000s, with a large increase after 2010. The graphs highlight how ticket costs have escalated significantly over the decades, emphasizing the increasing demand and market value of attending the Super Bowl.



**Conclusion**

Our exploratory analysis of the Super Bowl dataset provided a comprehensive understanding of the games historical trends and its economic and competitive significance. By utilizing RStudio, we generated a series of queries and visualizations that offered both financial and performance-related insights into the Super Bowl. Our analysis highlighted key economic trends, such as the strong positive correlation between average viewership and advertisement costs, demonstrating the Super Bowl’s growing economic dominance. Additionally, the exponential increase in ticket prices over time further reflected the event’s rising market value and demand.

Other than financial analysis, our queries analyzed trends that showcased the evolution of the game itself. For example, the analysis of total points scored over time revealed a gradual shift away from low-scoring, defense-dominated games, aligning with the NFL’s move toward a passing-oriented style of play. Visualizations such as the distribution of Super Bowl MVPs by position demonstrates the consistent importance of the quarterback. Additionally, the bar chart of Super Bowl wins by team and conference reveals the historically dominant franchises and shows the underperformance of certain divisions in our heatmap analysis.

Overall, our analysis provided in depth analysis into the Super Bowl’s legacy, combining financial trends with on-field performance and historical patterns. By generating these visualizations and conducting statistical tests, we were able to draw a more data-driven understanding of the Super Bowl’s evolution.