

Project Part 1

CJ Onibala and Ray Santiago

2025-11-04

Introduction

UVA Football. The pride and joy of many students and something that countless students look forward to each, especially when it's a home game. Our project plans to analyze three questions using the data from the 2024 UVA football season:

Does the UVA football's team's performance have more variation in points scored during away games compared to home matches?

Does attendance influence the amount of average points scored by both teams?

Does UVA convert a lower proportion of 3rd downs than their opponents?

The initial question we will explore is whether the team shows greater variability in points scored during home games compared to away games, which indicates whether the team performs consistently or shows more frequent fluctuations in scoring depending on where they play. This brings us to our next question of whether attendance influences the number of points scored by both teams, to see if there is any trend in fan support. This helps identify if fans cheering for the teams truly does hype up the players and allow for more points to be scored. Lastly, we want to explore the potential weakness of our team and see if our 3rd down conversion rate is lower than our opponent's. This may help determine where our offense crumbles.

Data Summary:

- (a) If the data represents a population, explain how the data were collected.

This data is representative of a population of all UVA Football games from the 2024 season. It was collected during each game and uploaded to the UVA Athletics website. The third-

down results were collected from the official UVA Athletics official NCAA game statistics and box scores. As NCAA tracks play-by-play outcomes our data is recorded directly from the in-game tracking systems, which generate a full population of games for the season.

- (c) Explain any data modifications after collection and the reasons behind these changes.

After downloading the data from the UVA athletics webpage no alterations to the original play-by-play data were made. The data were organized into a simplified formats for the total game statistics, third-down attempts, and conversions. The clean game to game summaries make comparisons easier when attempting to analyze the seasons results.

- (d) Discuss potential issues with the data and their possible impact.

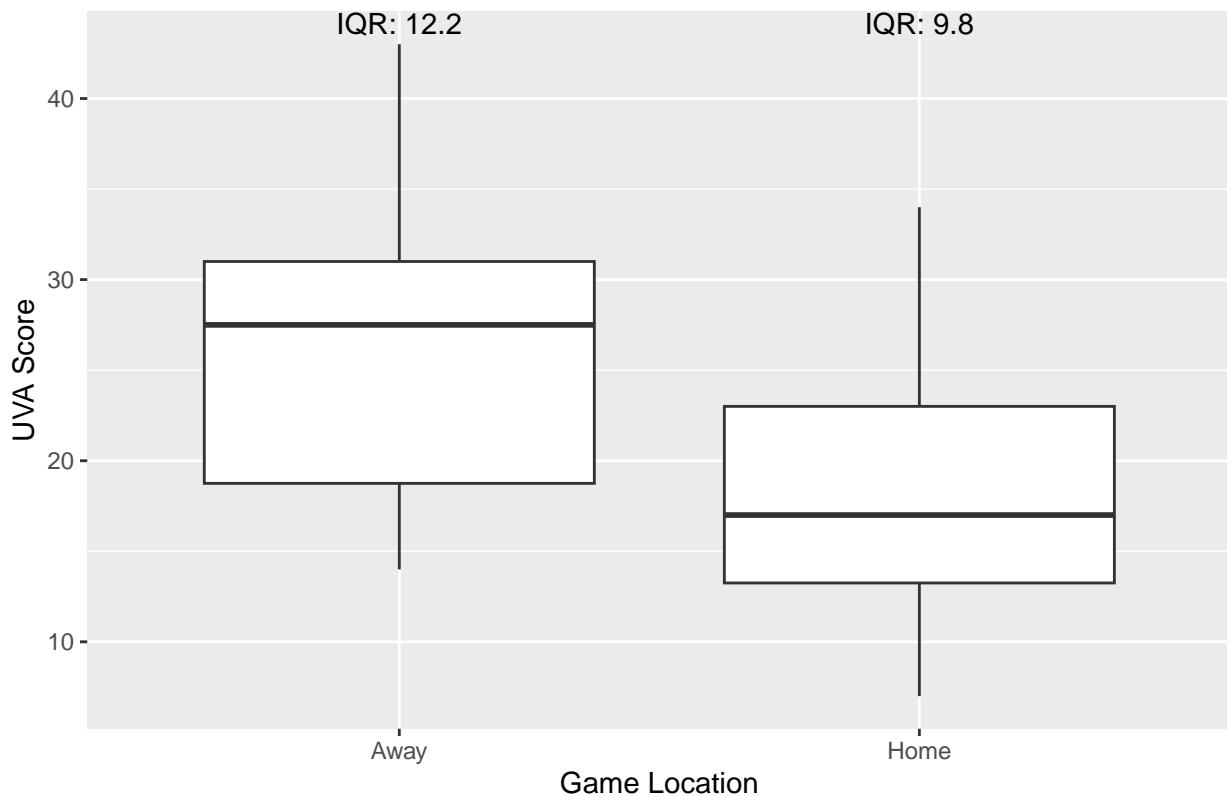
Although the data is descriptive there are still uncaptured details of each game. First, injuries, weather, and field conditions are not included or accounted for. These confounding variables may play a role in the teams performance, but our data has no way of showing that. Secondly the data this project uses is only of the 2024 season. Meaning the results cannot be generalized to team data of other years.

- (e) Explain why the data are appropriate to answer the research question.

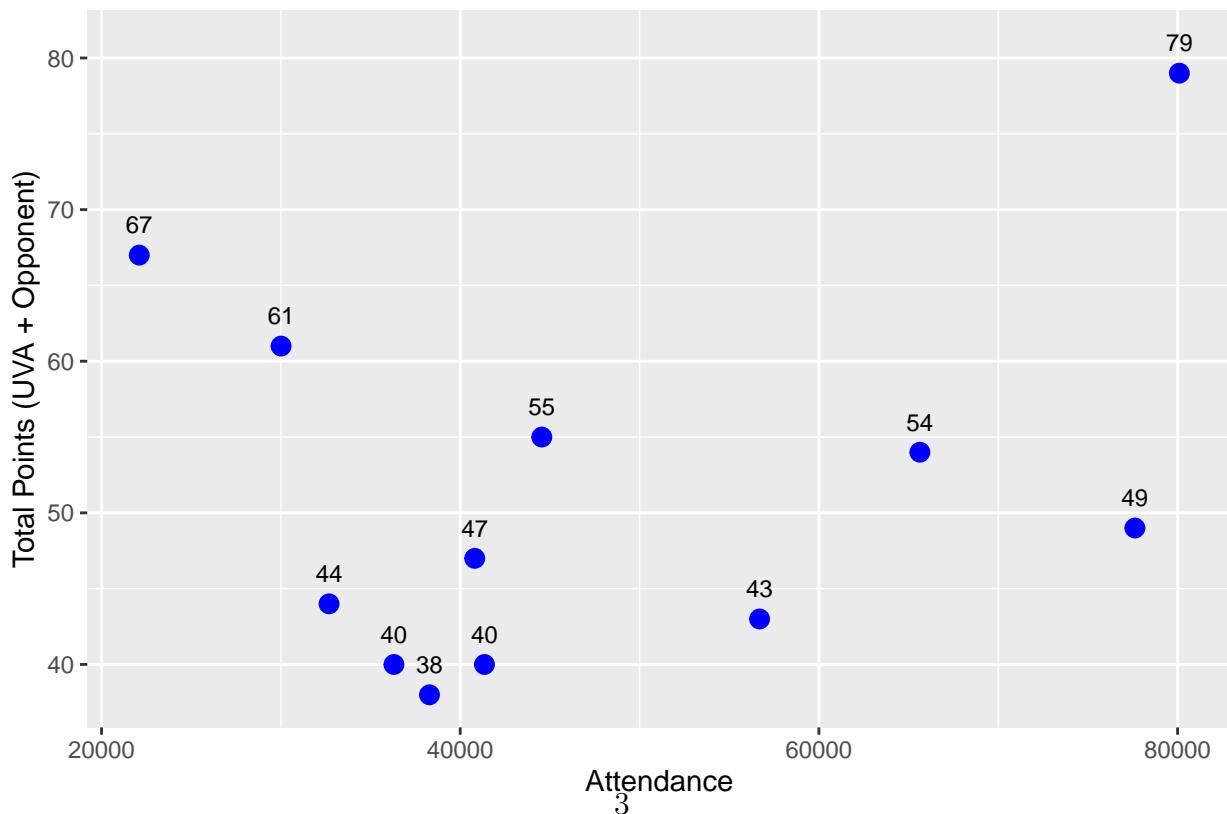
This dataset includes the research question because they measure the variables in the research question. Third-down attempts within the data allow us to calculate conversion percentages so that we can compare performance across all games, and game outcomes. The metric of third-down efficiency is a meaningful metric, because drive plays a large role in football success. Since the data is complete, unbiased, and recorded using legitimate NCAA measurements, it is reliable for statistical analysis.

Exploratory Analysis:

Boxplot of UVA Scores by Game Location



Attendance vs Total Points Scored

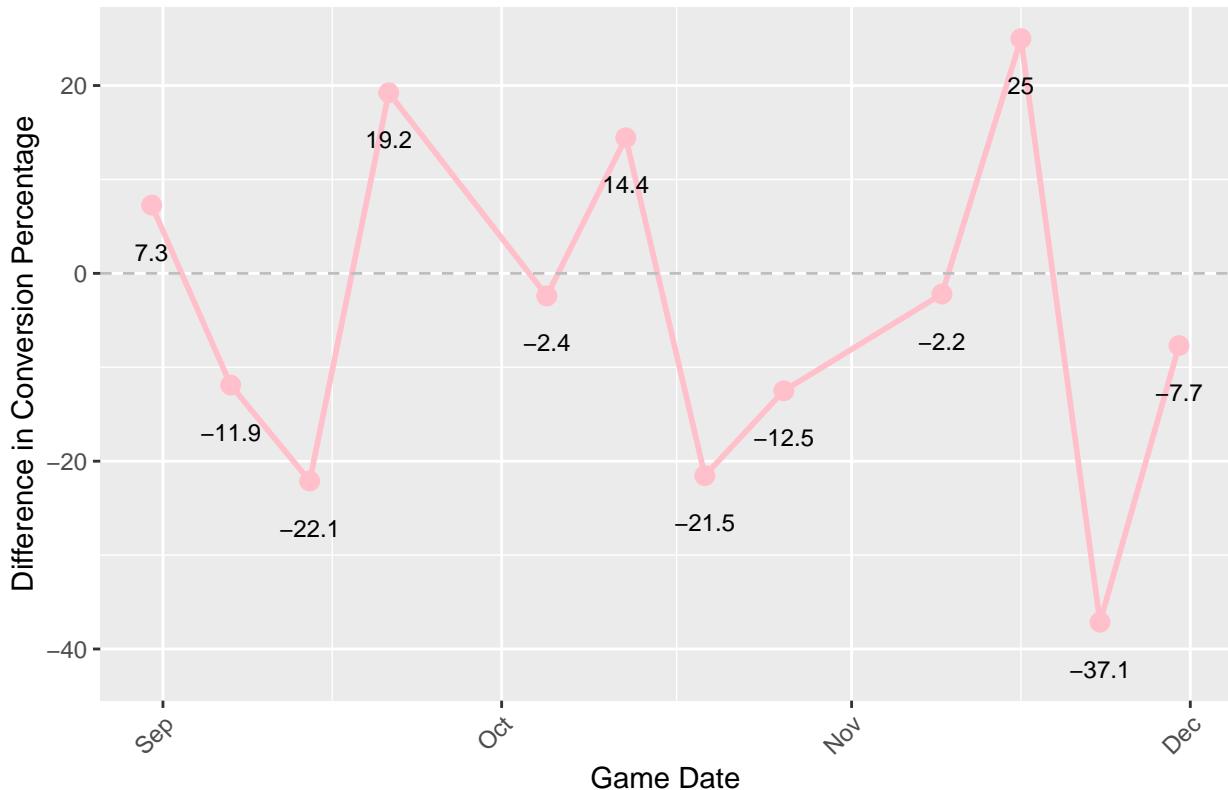


```

## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```

Difference in 3rd Down Conversion %: UVA minus Opponents



Conclusions: (a) Interpret the chosen summaries and make initial conclusions based on the information displayed in them.

Interestingly, the away games have a higher scoring average, whereas the home games score less, as evident by the box plot. The higher variability indicates that we are more unpredictable in how much we score in away games. Lastly, our home games seem to be more tightly clustered around an overall lower score count in games.

In the plot of attendance vs total points scored, games with low and high attendance saw higher point totals, with the middle of the pack attendance counts clustered around lower total points. This may indicate the type of game to expect and how the matchup is expected to turn out.

In the 3rd down conversions, 4 games out of 12 do we have a better rate of 3rd down

conversions compared to the other time. Of those 4, only 2 were above 15%. Of the 8 games where we had a worse conversion percent, 3 of the games were by more than 15%, with one being by 37.1%. There seems to be no improvement/worsening in conversion percent the later the season went.

- (b) Interpret the values, trends, and patterns displayed in the chosen summaries statistically.

Notably the medians are over 10 points apart, which is a touchdown and a field goal worth of points. But, the variability in away games is slightly more, as depicted by a roughly 2.4 difference in IQR between the two. Lastly, the whiskers in the box plot are higher and lower for away games, showing that away games had a higher amount of points and the highest amount of points in their lowest scoring game.

The 80,000+ attendee game may be an outlier and skew a perception that higher attendance always equals more points scored. The scatter plot suggests a weak correlation and unlikely for there to be a trend between the two, but potentially a different correlation made by team relevance/difficulty that will be talked about later.

The range in percentages is from -37.1 to 25, signifying a high volatility and variance. There was a decent streak of 3 games between September and October, starting with the 19.2 percent, which may signify a reason for good performance in those games. The standard deviation here appears to be very high.

- (c) Explain what the values, trends, and patterns displayed in the chosen summaries may mean or imply in the context of the question.

Our initial question ended up being that there is more variability within away games, but surprisingly also higher scoring. This may show that there is more motivation for the players and a “nothing to lose” mindset, alongside that our players do not experience a home field advantage energy.

In the attendees versus points scored, a potential reason for how the relationship turned out could be accounted for by other factors such as strength of opponent, football legacy, rivalries, and weather. If looking purely at strength of opponent, you can find a pattern that there is a lot of points scored in less attended events because the other team is a lot worse, which means less fans, and in the more attended events you can expect more points because the other team is a lot better, meaning more fans. This data does not exactly answer

our question, but rather shows that many factors account for higher points and attendance amounts.

The large negative percentages in 3rd down conversions may highlight major losses, as the game with the highest negative of 37.1, we lost 7-33, which may indicate the degree of loss that the 3rd down conversion may effect. The high volatility may be a sign of inconsistent play-calling or coaching, showing that they are trying different solutions.

- (d) Summarize the conclusions in the context of the topic of interest.

UVA actually scores more points in away games, with a higher variability in scoring compared to home games, making potentially more “exciting” games to be not at home. Attendance does not effect scoring for teams, showing fan presence is not a major enough factor for offensive forces. Lastly, our 3rd down conversion was lower than the enemy 8/12 games played, which shows that it can impact the outcome of a game. We lost 5/8 of those games when we had a lower conversion rate and is 5/7 of the games we lost. This may show that 3rd down conversion rates can impact UVA’s winning chances in games. These insights could inform coaching adjustments, game preparation strategies, and future performance evaluations.

References

cfbstats.com - 2024 Virginia Cavaliers Third Down Conversion Defense Game Log. (2024).
Cfbstats.com. <https://cfbstats.com/2024/team/746/thirddown/defense/gamelog.html>

cfbstats.com - 2024 Virginia Cavaliers Third Down Conversion Offense Game Log. (2024).
Cfbstats.com. <https://cfbstats.com/2024/team/746/thirddown/offense/gamelog.html>

Football Stats.” Virginia Cavaliers Official Athletic Site, 24 June 2020, virginiaspports.com/football-stats/. Accessed 3 Nov. 2025

github URL:<https://github.com/wng8zw/stat3080-project-part1.git>

CJ Onibala query: chatGPT 11/3 11:57 PM “where to find conversions for each game” - “You can find the third down conversion data for each UVA game in the 2024 season on external specialized sports statistics websites, for example, cfbstats.com provides a detailed game log with third down attempts and conversions per game for the Virginia Cavaliers (2024 season). The link to their game log is here, for example:

<https://www.cfbstats.com/2024/team/746/thirddown/offense/gamelog.html>”

CJ Onibala query: chatGPT 11/4 12:20 AM “the merge isn’t working, is it cause the date format in bothtxts are different” - “Yes, if the merge isn’t working, it could very well be due to differences in date formats between your two TXT files.

To fix this, you should convert the Date columns in both data frames to the same format before merging.

Here is how you can do that in R using `as.Date()` with the appropriate format strings:

Convert Date in `game_stats` to Date class (assuming format YYYY-MM-DD) `game_statsDate <- as.Date(game_statsDate, format = "%Y-%m-%d")`

Cj Onibala query: chatGPT 11/4 1:00 AM “the plot isn’t working, do i need anything?” - “To plot Attendance vs both UVA_Score and Opponent_Score on the same scatter plot with different colors for easy comparison, you can reshape the data into a”long” format and then use `ggplot2`.”

Reshape the data to long format with scores stacked `long_data <- full_data %>% select(Date, Opponent, Attendance, UVA_Score, Opponent_Score) %>% pivot_longer(cols = c(UVA_Score, Opponent_Score), names_to = “Team”, values_to = “Score”)`

- (d) If you used AI to produce any portion of any file in your repository, you must provide attribution in accordance with UVA guidelines

```
## R code to conduct the test in part 2.  
## This R code will be printed in the PDF.  
## Remove the ## before all code.
```