Compile your workbook!

Now that you have submitted all of your answers use this opportunity to proof read your work before compiling them into a workbook. The workbook will be submitted in the next section, so take your time to double-check your answers!

Please note, when reviewing your workbook before submitting, any documents you have submitted as part of the task will be shown as a link, appearing as the document name. In order to review any graphs, diagrams or presentations you have created, please click on the link to view your document.

Click 'next' to begin proofing your work.

Preparation:

Step one:

Research:

In the 2023 season, the San Francisco 49ers demonstrated a balanced and highly efficient performance across both offense and defense. Offensively, they averaged 28.9 points per game, with strong efficiency in both passing (8.3 yards per attempt) and rushing (4.7 yards per attempt), producing a total of 387.5 yards per game compared to 329.1 for their opponents. Defensively, they allowed just 18.8 points per game, well below the league average, while maintaining a positive turnover differential of +0.6 per game.

Research

Step one:

Data:

From the following website: https://www.pro-football-reference.com/teams/sfo/2025.htm

I have found the following stats for each game of the 2023/2024 season:

Week -- Week number in season
Rec -- Team's record following this game
Score
Tm -- Points scored
Opp -- Points allowed
Offense
TotYd -- Total Yards Gained on Offense
PassY -- Total Yards Gained by Passing (includes lost sack yardage)
RushY -- Total Yards Gained by Rushing
TO -- Team Turnovers Lost
Defense
TotYd -- Total Yards Allowed by Defense
PassY -- Total Passing Yards Allowed by Defense
(includes lost sack yardage)
RushY -- Total Rushing Yards Allowed by Defense
TO -- Turnovers Gained by Defense
Expected Points

Analysis:

Step one:

Clean and organise data:

Before conducting the analysis, the dataset was thoroughly cleaned and organised to ensure accuracy and consistency. Column names were standardised, with clear prefixes (e.g. offence_ and defence_) applied to distinguish different categories of statistics. Non-regular fixtures, such as the Super Bowl, were removed to focus solely on the regular season. Data types were corrected so that all numerical values were stored as integers or floats rather than strings. Additional derived attributes were created, including point difference and turnover difference, to provide clearer measures of performance. Finally, binary indicator columns were added to identify match outcomes (win or loss) and venue (home or away), ensuring the dataset was structured for both exploratory analysis and subsequent modelling.

Step two:

Data analysis:

In 2023, the 49ers' offence improved in the second half of the season. They averaged 31.63 points per game in weeks 10-18 vs. 27.25 in weeks 1-8, while the defence allowed slightly fewer points (17.13 vs 17.50). As a result, the average point differential widened from +9.75 to +14.50. Under the hood, total offence rose (432.5 vs 376.6 yards per game) and yards allowed fell (298.6 vs 315.50), although turnovers lost increased (2.67 vs 1.80) alongside a bump in takeaways (2.33 vs 1.86).

The highest scoring game was week 15 at Arizona (away), a 45-29 win with 406 total yards, 24 first downs, and two defensive takeaways; with the opponent having had 3 wins at this point.
The lowest scoring game was week 6 at Cleveland (away), a 17-19 loss with just 215 total yards, 15 first downs, and one giveaway; the opponent had 1 win at that time.

Together, this indicates that opponent quality, with yardage and turnovers, help explain the scoring extremes, while the second-half offence was notable more productive and the defence remained consistently strong.

Create:

Step one:

Predictive analysis:

Based on historical game data from 2016–2024, predictive models were developed to estimate the 49ers' likelihood of winning future games. A Random Forest classifier, trained on rolling offensive and defensive performance metrics, turnover differential, and contextual features such as opponent win percentage and home/away status, achieved strong results with an accuracy of 75% and an ROC AUC of 0.90 on unseen test data. This indicates that the model is able to distinguish wins from losses with a high degree of reliability.

Analysis of feature importance revealed that opponent strength (opponent win percentage) and the 49ers' own pre-game win percentage were the most influential predictors of outcomes, followed by offensive yardage, points scored in recent games, and turnover differential. Defensive metrics, such as yards allowed, also contributed meaningfully, while home advantage was comparatively less important. These findings highlight that sustained offensive production, controlling turnovers, and facing weaker opponents are the strongest indicators of success.

At the season level, historical win rates show considerable fluctuation, from lows of 12.5% in 2016 to a high of 81% in 2019. More recent years indicate consistent playoff-level performance between 2021 and 2023, followed by a downturn in 2024 (37.5%). A simple linear projection fitted to the 2016–2024 trend suggests a projected 2025 win rate of approximately 73.5%, equivalent to 12–13 wins in a 17-game schedule. This projection, combined with the model's emphasis on offensive yardage and turnover control, points towards a return to strong performance levels if current long-term patterns are maintained.

Step two:

Data visualisations (document submission):

win_rate_trend.png
77 KB

Document and Present

Step one:

Report (document submission):

San Francisco 49ers Performance Analysis.pdf
424 KB

Reflection

Step one:

What did you learn through completing these data analytics tasks?

Through completing these tasks, I learned how to take a dataset from its raw form through to a cleaned, structured version ready for analysis, and how to engineer new features that capture meaningful patterns such as recent form and opponent strength. I also deepened my understanding of exploratory data analysis by identifying trends within and across seasons, and learned how to link these to context (such as opponent quality) to explain performance outcomes.

On the modelling side, I gained experience in evaluating different approaches, recognising the limitations of regression in predicting volatile outcomes like exact points, and seeing how classification models can provide stronger and more interpretable insights. Importantly, I developed a clearer appreciation of how to communicate results effectively - using visualisations, trend projections, and feature importance to make findings accessible to diverse audiences.

How did you consider data limitations throughout your analysis?

From the outset, I was aware that the dataset had important limitations. Initially, I only had access to one or two seasons of data, which restricted the ability of predictive models to generalise. To address this, I extended the analysis to cover more than 10 seasons (2014–2024), which provided a larger sample size and helped capture longer-term trends rather than season-specific anomalies.

The raw data also contained inconsistencies — for example, stacked offensive and defensive columns, missing values, and non-standard formats for categorical fields. I dealt with these by restructuring and renaming columns, correcting data types, removing non-regular fixtures, and creating derived variables such as point differential and turnover differential. In some cases, missing values in rolling features were handled by dropping the earliest games of each season to avoid introducing bias.

How would you enhance your sports data analysis in the future?

In the future, I would enhance this sports data analysis by expanding both the scope and depth of the dataset. Incorporating player-level statistics (such as quarterback ratings, rushing efficiency, or defensive pressures) and drive-level data would allow more detailed modelling of performance drivers than team-level aggregates alone. I would also integrate external contextual factors such as injuries, weather conditions, and betting odds, which are known to affect outcomes but were not captured here.

On the modelling side, I would explore more advanced approaches such as gradient boosting methods or time-series forecasting models that could account for evolving trends within and across seasons. Finally, I would focus on building more interactive dashboards or visual reports, enabling stakeholders to explore the data dynamically and apply insights directly to strategy and decision-making.

Acknowledgement

Please check this box to acknowledge that you are aware that you can only download your workbook and upload your assignment once.

☑ Agree