**Group Members:**

Braden Bullon, Dylan Bland, Justin Roe, Trey Wagner

**Project Overview:**

**Objective**: Identify which quarterback and team statistics are most predictive of a successful NFL season, as measured by Points Scored (Points Scored).

This aligns with the Data Mining process by emphasizing:

* Business understanding: Framing a real-world problem.
* Data understanding and preparation: Cleaning and merging datasets.
* Modeling: Preparing data for regression/classification in Orange.
* Evaluation: Measuring and interpreting prediction performance.

**Data Cleaning Process:**

To begin the project, the team imported two datasets into Python using pandas: one containing team-level statistics for each NFL season (nfl\_data), and another containing quarterback passing statistics (qb\_data). This was an essential first step in gaining a foundational understanding of the available data. Previewing the data with functions like .head() allowed the team to quickly examine the structure, check for missing or misaligned values, and identify what information might be useful for modeling.

One of the biggest issues we noticed after our first time cleaning and merging the data, was that qb\_data from 2023 was copied into the 2007 rows. We assumed this was an error in the original data source’s scraping process, so we decided to drop the 2007 season. This ensures the data will not be skewed

Next, the data was cleaned to ensure alignment between the two sources. The nfl\_data included the 2000 season, while qb\_data extended through the 2022 and 2023 seasons. Because the goal was to merge each team-season with its corresponding primary quarterback, seasons that did not appear in both datasets were removed. This ensured that every row in the merged dataset would represent a complete and accurate snapshot of a single team-season.

One key challenge was that team names in the two datasets were formatted differently. The team data used full franchise names (e.g., "Atlanta Falcons"), while the quarterback data used three-letter abbreviations (e.g., "ATL"). To resolve this, the team created a mapping dictionary to convert all abbreviations in the quarterback data into full team names. This allowed for a clean, consistent merge later in the process.

Another important data cleaning step was the removal of quarterbacks who played for multiple teams in a given season. These players were marked with codes like "2TM" or "3TM", indicating split seasons. Since the purpose of the project was to analyze team-level performance, and these rows could not be cleanly assigned to a single franchise, they were excluded to avoid introducing ambiguity or noise into the analysis.

With those rows removed, the team needed to ensure that only one quarterback was represented per team-season. This was necessary to match the one-row-per-team-per-year format of the nfl\_data. The team identified the primary quarterback for each team by selecting the player with the most passing attempts, a strong proxy for starter status and a meaningful measure for determining who most influenced the team’s performance that year. This ensured a one-to-one alignment between quarterback and team records.

Finally, the cleaned and filtered datasets were merged using both Team and Year as keys. This resulted in a combined dataset where each row contained both team-level outcomes (such as win percentage, points scored, and net points) and quarterback-specific metrics (such as completion percentage, passer rating, and sacks taken). The resulting dataset was exported to a CSV file to be used in Orange, a data mining and modeling tool covered in the course.

**Analytical Procedure**:

* Originally, we started by having “win%” and “winning season” as our target variables, with the goal of measuring how quarterback statistics impact a team’s success. The original linear regression we ran for “win%” returned an R^2 value of .21, this was too low for our liking. Due to “win%” only being from 0 to 1, it did not make for a strong linear regression; we needed a better continuous variable. We chose “points scored” as a new target variable which provided an R^2 of 0.67 and mean absolute error of 32 points. The independent variables used for this linear regression were completions (Cmp), attempts (Att), adj net yards per attempt (ANY/A), touchdown % (td%), yards (yds), and interception% (int%).