# Datasheet for NFLFastR*

Vandan Patel

Invalid Date

This datasheet provides a comprehensive overview of the dataset used to analyze the relationship between quarterback performance metrics and team success in the NFL's 2023 season. It details data provenance, variable descriptions, processing steps, and potential limitations to ensure transparency and reproducibility for future research and applications.

Extract of the questions from @gebru2021datasheets.

## 0.1 Motivation

The dataset was created to analyze the impact of quarterback performance on team success in the 2023 NFL season, focusing on metrics such as wins, playoff qualification, and point differential.

The dataset was compiled from publicly available NFL play-by-play data using the `nflfastR` package [@nflfastR]. The authors aimed to explore the hypothesis that quarterbacks are the primary drivers of team success.

The dataset aims to answer the following questions:

- How do quarterback statistics correlate with team success metrics like wins and playoffs?
- Are turnovers (e.g., interceptions) more detrimental than touchdowns are beneficial?

## 0.2 Composition

Dataset containts the following data:

- **Quarterback dataset**: Aggregated statistics such as passing yards, touchdowns, interceptions, and average metrics per game.

*Code and data are available at: https://github.com/vandanppatel/impact_qb_nfl/.

- **Team dataset**: Metrics including wins, points scored, points allowed, and playoff qualification.

No significant missing values were identified after cleaning. However, ties in games were excluded from win counts, and some metrics rely on aggregations that may introduce bias.

The following preprocessing was done: * Data was filtered for the 2023 NFL regular season. * Key metrics were aggregated for starting quarterbacks and teams. * Playoff qualification was manually assigned based on publicly available postseason data.

## 0.3 Use Cases

This dataset is intended for researchers, analysts, and sports enthusiasts interested in NFL statistics, player evaluation, and team performance analysis.

The intended uses of this dataset are:

- Statistical modeling of NFL team success.
- Exploratory data analysis for insights into quarterback contributions.
- Benchmarking for machine learning models.

## 0.4 Collection Process

The dataset was sourced from the NFL's public play-by-play data, processed using the `nflfastR` package. Efforts were made to ensure transparency in data cleaning and processing. No personally identifiable information about players or staff was included.

## 0.5 Limitation

The limitations of the dataset are

- **Data granularity**: Only aggregated metrics were used; situational data (e.g., specific game contexts) was excluded.
- **External factors**: The dataset does not account for coaching decisions, injuries, or schedule difficulty.
- **Season scope**: Focused solely on the 2023 regular season, limiting longitudinal analysis.

Potential sources of bias are:

- Strong quarterbacks on weak teams may have skewed stats due to playing from behind.
- Playoff qualification was manually added, introducing potential human error.

## 0.6 Ethics and Legal

The dataset reflects public data and does not include sensitive or personally identifiable information. However, its usage should avoid harm, such as misrepresentation of player performance. The dataset leverages publicly available data but adheres to NFL data-sharing policies.

## 0.7 Future Directions

This dataset can potentially be improved by doing the following:

- Incorporating situational data, such as fourth-quarter performance or playoff-specific stats.
- Expanding the scope to include multiple seasons for longitudinal analysis.
- Adding advanced metrics from platforms like NFL Next Gen Stats to capture nuanced player performance.