Andres Melendez

DSC 350

9 / 21 / 2024

**List of Sources and Description of Data**

**1. Flat File: National Football League Franchise Histories**

* **Data Description**: This flat file likely contains historical data on NFL franchises. The data may include the team’s name, years of operation, key milestones (e.g., Super Bowl wins), and location changes.
* **Fields**:
  + **Team Name**: The name of the NFL franchise (e.g., Green Bay Packers).
  + **Franchise ID**: A unique identifier for each NFL team.
  + **Years Active**: The years the franchise has been active in the league.
  + **Championships Won**: Number of Super Bowls or NFL Championships won.
  + **City/State**: The geographic location of the franchise.

**2. API: Fantasy Football Data Pros**

* **Data Description**: This API provides real-time or historical performance data for NFL players, which is used in fantasy football. It may include player statistics for rushing yards, passing yards, touchdowns, and other performance metrics.
* **Fields**:
  + **Player Name**: The name of the NFL player.
  + **Player ID**: A unique identifier for each player.
  + **Team Name**: The team to which the player is currently or historically affiliated.
  + **Season Stats**: Detailed performance metrics for each player per season, including touchdowns, yards, interceptions, etc.
  + **Position**: The position that the player plays (e.g., quarterback, wide receiver).

**3. Website: NFL Combine Results**

* **Data Description**: This website lists player performance statistics from the NFL Combine, including speed, agility, and strength tests.
* **Fields**:
  + **Player Name**: The name of the player who participated in the combine.
  + **Player ID**: A unique identifier for each player (if available).
  + **Position**: The position the player was tested for (e.g., wide receiver).
  + **Combine Metrics**: Performance statistics, such as 40-yard dash time, bench press, vertical jump, etc.
  + **Draft Year**: The year the player was drafted into the NFL.

**Relationships Between the Data Sources**

* **Player Name/Player ID**: The primary relationship between the **Fantasy Football API** and the **NFL Combine Website** is through player names or player IDs. Player performance data from the API can be tied to their NFL Combine performance via these identifiers.
* **Team Name**: The **NFL Franchise History Flat File** and the **Fantasy Football API** can be linked through the team name. Team-based historical data from the flat file can be associated with current player performance from the API.
* **Position**: While the position field may not be unique enough to create direct relationships between the sources, it can be used to filter or group players across the combine data and the fantasy API.

If no player ID exists in both the API and website, I may need to create a **synthetic player ID** by matching player names and draft years to unify the data. Similarly, if team IDs between the flat file and the API don’t match, I can manually map them or create fake IDs for alignment.

**Data Preparation and Transformation for Milestones**

1. **Data Cleaning and Standardization**: I’ll need to clean all datasets by standardizing player names, team names, and dates (e.g., standardizing year formats). Some fields, such as player names, may have different variations (e.g., John Doe vs. J. Doe), requiring me to merge records using fuzzy matching techniques.
2. **Handling Missing Data**: For any missing values in combined data or player performance metrics, I’ll need to either fill them with estimates or flag them as null for analysis. Missing franchise IDs or player IDs can be created where necessary.
3. **Merging Data**: Once standardized, I’ll merge the data by:
   * **Player Name/ID**: To link the **Fantasy Football API** with the **NFL Combine** data, player name or ID will be the primary key. If this data doesn’t exist in both, I’ll create matching IDs based on names, draft years, and positions.
   * **Team Name**: To link the **NFL Franchise History** with the **Fantasy Football API**, I will use team names or franchise IDs. This will allow me to analyze how a franchise’s historical performance correlates with the players' current or past performances.
4. **Create Relationships**: If no direct relationships exist between the datasets (such as team or player IDs), I’ll generate **synthetic IDs** to match records and fill in gaps where needed.
5. **Data Dictionary or Summary**:
   * **Player Name**: Shared across the API and Combine results.
   * **Player ID**: A unique identifier for each player (potentially generated).
   * **Team Name/Franchise ID**: Links the flat file and API for team-based analysis.
   * **Performance Metrics**: Include both season statistics from the API and combine metrics from the website.
   * **Position**: Helps categorize and filter data across all datasets.

In conclusion, the relationships between these datasets are primarily built around player names/IDs and team names. The integration will involve matching fields where possible and generating synthetic data where necessary to ensure consistency across the datasets. The final goal is to create a unified dataset that combines franchise history, player combine results, and ongoing performance metrics to derive meaningful insights.