Introduction

The project involves the creation of a database called nfl_statistics, which is designed to hold extensive data on NFL player statistics. The data is sourced from Kaggle and is composed of various files that include player basic stats and career statistics across different performance types. The primary goal is to organize this data within a MySQL database, allowing for complex queries to be executed to answer business-driven questions.

1. Database Creation and Population

1. 1 Database: nfl_statistics

1.2 Entities:

- Player
- Team
- College
- High_School
- Yearly_Statistics
- Passing_Statistics
- Receiving_Statistics
- Rushing_Statistics
- Defensive_Statistics

1.3 Data Population

To create and populate the nfl_statistics database, I cleaned and organized the raw data into a structured format suitable for a relational database using Python scripts. I then created the database and the tables using SQL commands and populated the tables using Python scripts.

College and High_School Tables

Starting with the Basic_Stats.csv file, which contains player-specific information, I ran Python scripts to:

- identify and extract unique colleges and high schools,
- save them into separate CSV datasets using Python scripts, and
- load these datasets into College and High_School tables.

Other details related to high schools such as Location were extracted directly from Basic_Stats.csv based on the high school name.

Team Table

Using Basic_Stats.csv and Career Statistics files (which are directly related to player careers), I ran Python scripts to:

- identify and extract unique teams,
- save them into a separate CSV dataset, and
- populate the Teams table using this dataset.

Player Table

Using Basic_Stats.csv, I ran Python scripts to:

- ensure that there are no duplicate records based on PlayerID,
- re-format the player names in the database to standardize and maintain consistency,
- format birthday column from ‘dd/mm/yyyy’ to ‘yyyy-mm-dd’,
- match the names of teams, colleges, and high schools listed in the player's records with the unique IDs from the Team, College, and High School tables,
- link players to their respective TeamID, CollegeID, and HighSchoolID in the database,
- extract and combine with other player details from Basic_Stats.csv, and
- populate Player table using the organized dataset.

Yearly_Statistics Table

Updating Performance Files:

- Updated each performance-specific career file with a new column named ‘StatsType’. This column serves to identify the type of statistics contained in each record.

Python Script for Data Integration:

- Combined data from Player and Team tables and various performance-specific files based on unique PlayerID, TeamID, Year and StatsType combination.
- Loaded the organized dataset into the database to populate the table.

Each record in the YEARLY_STATISTICS table includes:

- StatID: Unique identifier for the record,
- PlayerID and TeamID: Linked to the PLAYER and TEAM tables, connecting each stat to the right player and team,
- Year, GamesPlayed, and StatsType extracted from the performance files.

Performance-Specific Tables

- Passing_Statistics
- Receiving_Statistics
- Rushing_Statistics
- Defensive_Statistics

Loading Data into Performance Tables:

- Ran a Python script to populate these tables.
- For each record in the performance files, it first replaced '--' with blanks wherever applicable.
- Checked for a matching PlayerID, TeamID, Year, and StatsType in the YEARLY_STATISTICS table.
- Once a match was found, the script extracted the corresponding StatID from YEARLY_STATISTICS.
- Using StatID, along with other performance-specific details, the script loaded the data into the relevant performance table.

The resulting database provides an overall picture of NFL player career statistics.

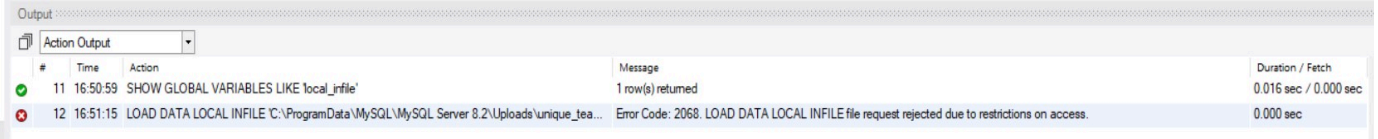
2. Data Importation Challenges

2.1 Challenge:

SQL Workbench Local File Access Issue

- Encountered persistent access restrictions when attempting to load data using SQL Workbench.
- Explored various troubleshooting methods, such as permissions review and software configuration checks, but the issue remained unresolved.

Error:



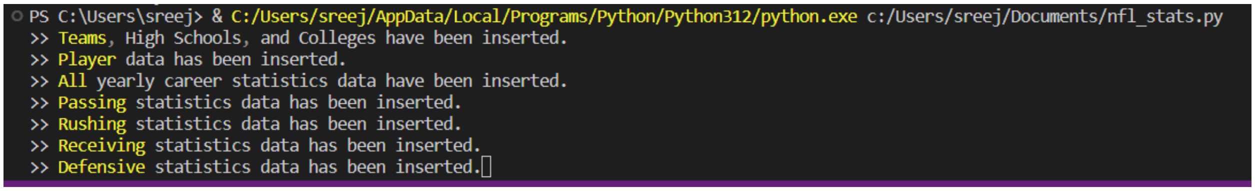
	#	Time	Action	Message	Duration / Fetch
✓	11	16:50:59	SHOW GLOBAL VARIABLES LIKE 'local_infile'	1 row(s) returned	0.016 sec / 0.000 sec
✕	12	16:51:15	LOAD DATA LOCAL INFILE 'C:\ProgramData\MySQL\MySQL Server 8.2\Uploads\unique_tea...	Error Code: 2068. LOAD DATA LOCAL INFILE file request rejected due to restrictions on access.	0.000 sec

2.2 Resolution:

Utilizing Python Script

- Cleaned and prepared datasets using Python to ensure data quality.
- Implemented a Python script to interact with the MySQL database, bypassing the access issues in SQL Workbench and successfully populating the tables with the data

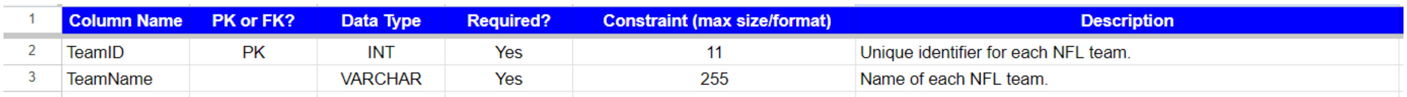
Result:



```
PS C:\Users\sreej> & C:/Users/sreej/AppData/Local/Programs/Python/Python312/python.exe c:/Users/sreej/Documents/nfl_stats.py
>> Teams, High Schools, and Colleges have been inserted.
>> Player data has been inserted.
>> All yearly career statistics data have been inserted.
>> Passing statistics data has been inserted.
>> Rushing statistics data has been inserted.
>> Receiving statistics data has been inserted.
>> Defensive statistics data has been inserted.
```

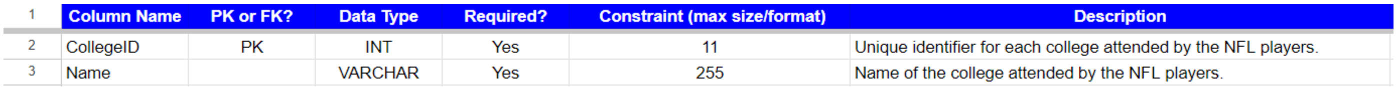
3. Data Dictionary

# Team Table



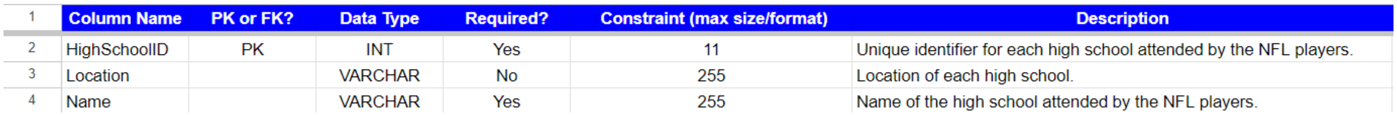
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format)	Description
TeamID	PK	INT	Yes	11	Unique identifier for each NFL team.
TeamName		VARCHAR	Yes	255	Name of each NFL team.

# College Table



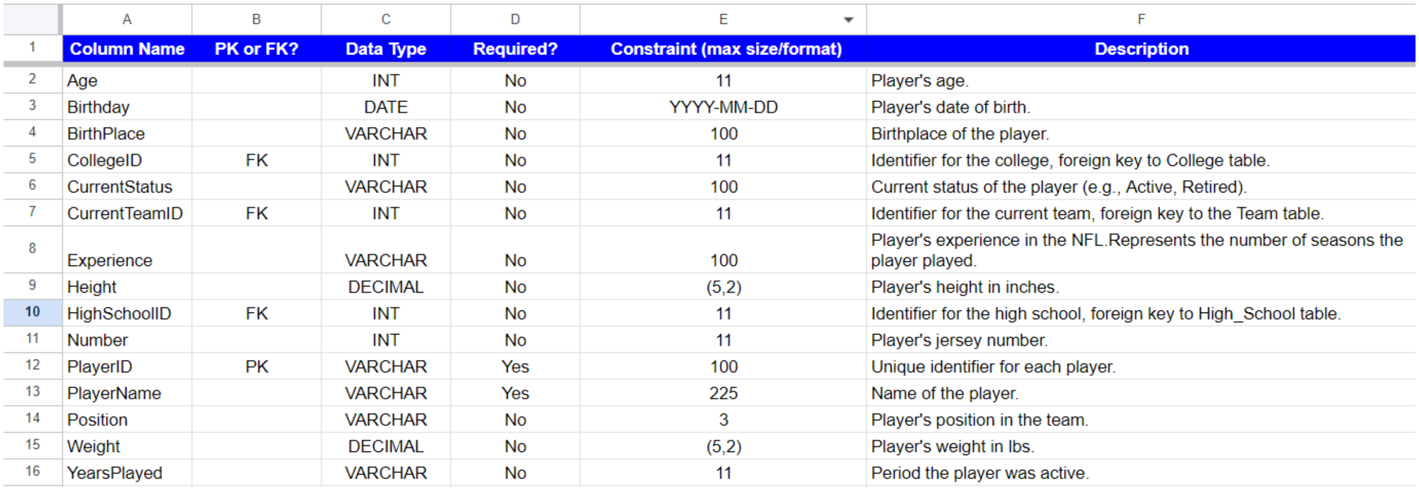
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format)	Description
CollegeID	PK	INT	Yes	11	Unique identifier for each college attended by the NFL players.
Name		VARCHAR	Yes	255	Name of the college attended by the NFL players.

# High_School Table



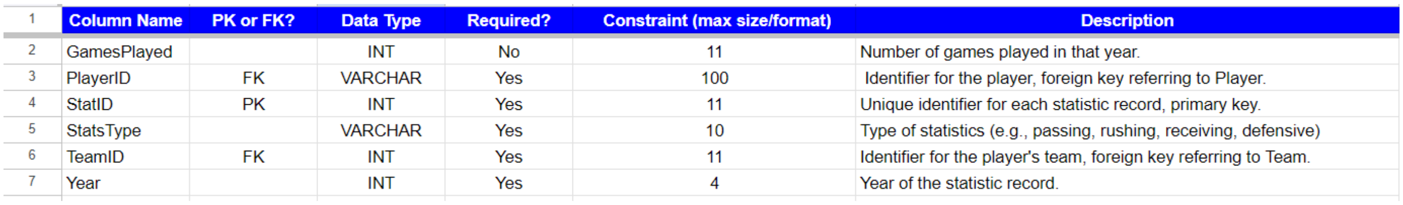
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format)	Description
HighSchoolID	PK	INT	Yes	11	Unique identifier for each high school attended by the NFL players.
Location		VARCHAR	No	255	Location of each high school.
Name		VARCHAR	Yes	255	Name of the high school attended by the NFL players.

# Player Table



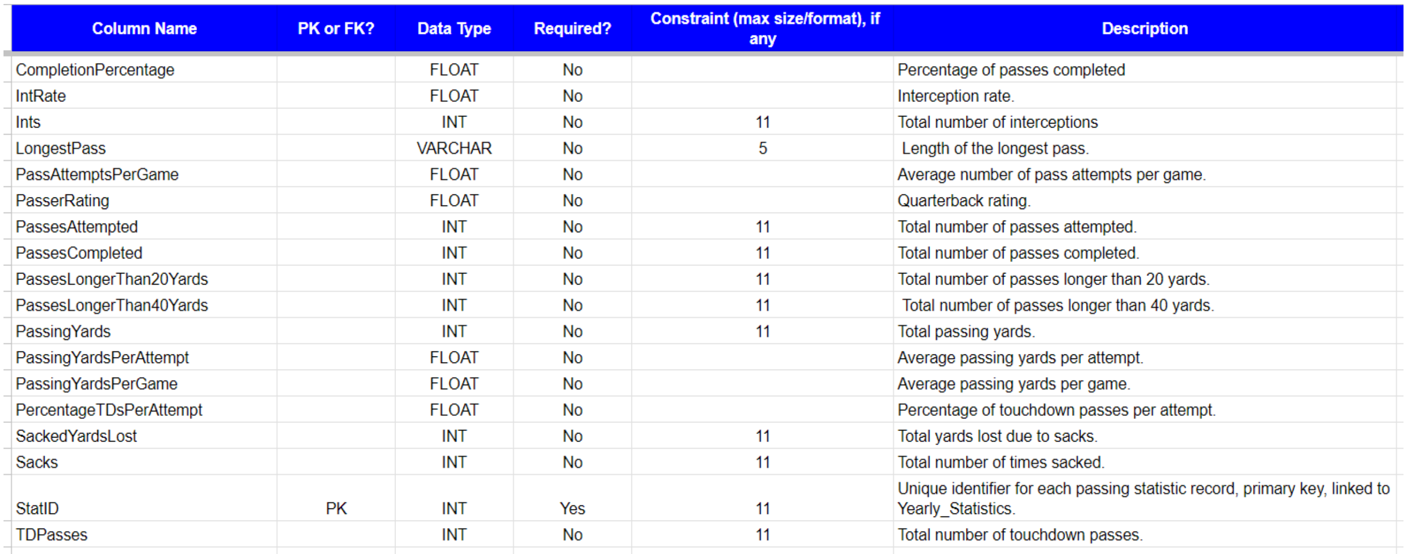
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format)	Description
Age		INT	No	11	Player's age.
Birthday		DATE	No	YYYY-MM-DD	Player's date of birth.
BirthPlace		VARCHAR	No	100	Birthplace of the player.
CollegeID	FK	INT	No	11	Identifier for the college, foreign key to College table.
CurrentStatus		VARCHAR	No	100	Current status of the player (e.g., Active, Retired).
CurrentTeamID	FK	INT	No	11	Identifier for the current team, foreign key to the Team table.
Experience		VARCHAR	No	100	Player's experience in the NFL.Represents the number of seasons the player played.
Height		DECIMAL	No	(5,2)	Player's height in inches.
HighSchoolID	FK	INT	No	11	Identifier for the high school, foreign key to High_School table.
Number		INT	No	11	Player's jersey number.
PlayerID	PK	VARCHAR	Yes	100	Unique identifier for each player.
PlayerName		VARCHAR	Yes	225	Name of the player.
Position		VARCHAR	No	3	Player's position in the team.
Weight		DECIMAL	No	(5,2)	Player's weight in lbs.
YearsPlayed		VARCHAR	No	11	Period the player was active.

# Yearly_Statistics Table



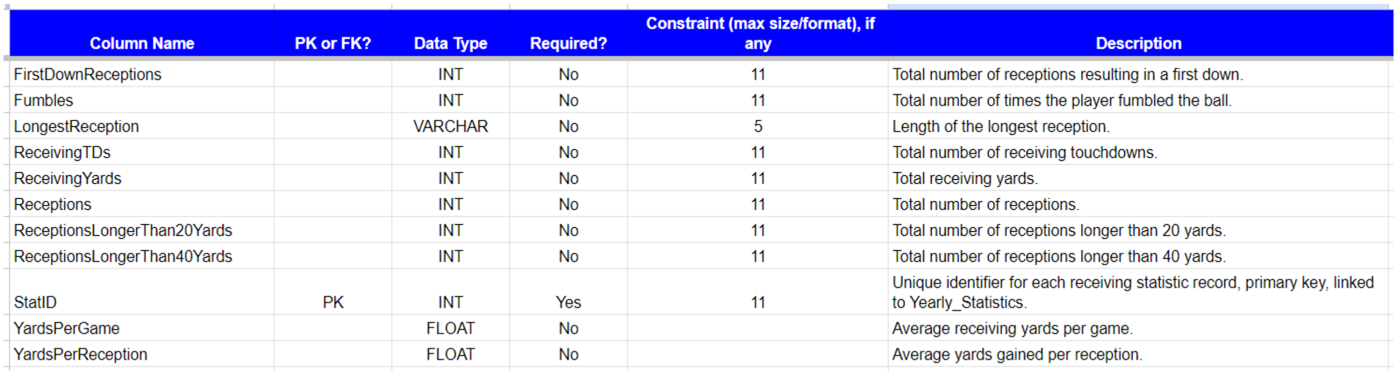
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format)	Description
GamesPlayed		INT	No	11	Number of games played in that year.
PlayerID	FK	VARCHAR	Yes	100	Identifier for the player, foreign key referring to Player.
StatID	PK	INT	Yes	11	Unique identifier for each statistic record, primary key.
StatsType		VARCHAR	Yes	10	Type of statistics (e.g., passing, rushing, receiving, defensive)
TeamID	FK	INT	Yes	11	Identifier for the player's team, foreign key referring to Team.
Year		INT	Yes	4	Year of the statistic record.

# Passing_Statistics Table



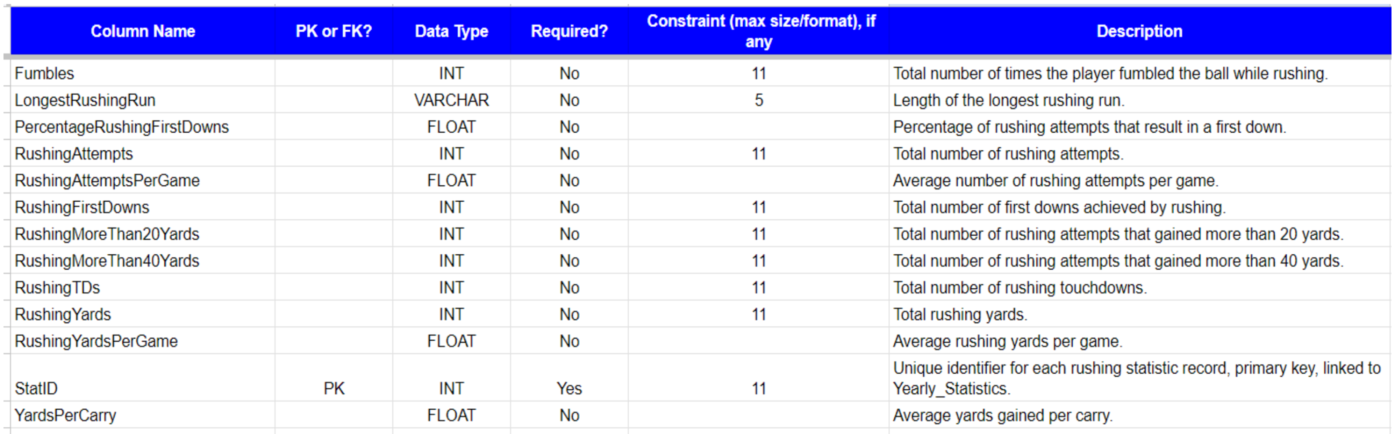
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format), if any	Description
CompletionPercentage		FLOAT	No		Percentage of passes completed
IntRate		FLOAT	No		Interception rate.
Ints		INT	No	11	Total number of interceptions
LongestPass		VARCHAR	No	5	Length of the longest pass.
PassAttemptsPerGame		FLOAT	No		Average number of pass attempts per game.
PasserRating		FLOAT	No		Quarterback rating.
PassesAttempted		INT	No	11	Total number of passes attempted.
PassesCompleted		INT	No	11	Total number of passes completed.
PassesLongerThan20Yards		INT	No	11	Total number of passes longer than 20 yards.
PassesLongerThan40Yards		INT	No	11	Total number of passes longer than 40 yards.
PassingYards		INT	No	11	Total passing yards.
PassingYardsPerAttempt		FLOAT	No		Average passing yards per attempt.
PassingYardsPerGame		FLOAT	No		Average passing yards per game.
PercentageTDsPerAttempt		FLOAT	No		Percentage of touchdown passes per attempt.
SackedYardsLost		INT	No	11	Total yards lost due to sacks.
Sacks		INT	No	11	Total number of times sacked.
StatID	PK	INT	Yes	11	Unique identifier for each passing statistic record, primary key, linked to Yearly_Statistics.
TDPasses		INT	No	11	Total number of touchdown passes.

# Receiving_Statistics Table



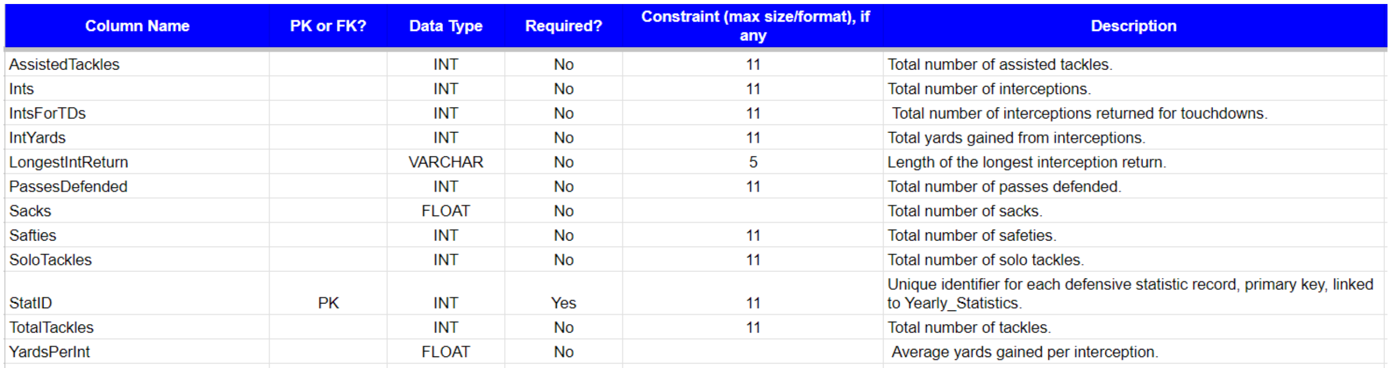
Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format), if any	Description
FirstDownReceptions		INT	No	11	Total number of receptions resulting in a first down.
Fumbles		INT	No	11	Total number of times the player fumbled the ball.
LongestReception		VARCHAR	No	5	Length of the longest reception.
ReceivingTDs		INT	No	11	Total number of receiving touchdowns.
ReceivingYards		INT	No	11	Total receiving yards.
Receptions		INT	No	11	Total number of receptions.
ReceptionsLongerThan20Yards		INT	No	11	Total number of receptions longer than 20 yards.
ReceptionsLongerThan40Yards		INT	No	11	Total number of receptions longer than 40 yards.
StatID	PK	INT	Yes	11	Unique identifier for each receiving statistic record, primary key, linked to Yearly_Statistics.
YardsPerGame		FLOAT	No		Average receiving yards per game.
YardsPerReception		FLOAT	No		Average yards gained per reception.

# Rushing_Statistics Table



Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format), if any	Description
Fumbles		INT	No	11	Total number of times the player fumbled the ball while rushing.
LongestRushingRun		VARCHAR	No	5	Length of the longest rushing run.
PercentageRushingFirstDowns		FLOAT	No		Percentage of rushing attempts that result in a first down.
RushingAttempts		INT	No	11	Total number of rushing attempts.
RushingAttemptsPerGame		FLOAT	No		Average number of rushing attempts per game.
RushingFirstDowns		INT	No	11	Total number of first downs achieved by rushing.
RushingMoreThan20Yards		INT	No	11	Total number of rushing attempts that gained more than 20 yards.
RushingMoreThan40Yards		INT	No	11	Total number of rushing attempts that gained more than 40 yards.
RushingTDs		INT	No	11	Total number of rushing touchdowns.
RushingYards		INT	No	11	Total rushing yards.
RushingYardsPerGame		FLOAT	No		Average rushing yards per game.
StatID	PK	INT	Yes	11	Unique identifier for each rushing statistic record, primary key, linked to Yearly_Statistics.
YardsPerCarry		FLOAT	No		Average yards gained per carry.

# Defensive Table



Column Name	PK or FK?	Data Type	Required?	Constraint (max size/format), if any	Description
AssistedTackles		INT	No	11	Total number of assisted tackles.
Ints		INT	No	11	Total number of interceptions.
IntsForTDs		INT	No	11	Total number of interceptions returned for touchdowns.
IntYards		INT	No	11	Total yards gained from interceptions.
LongestIntReturn		VARCHAR	No	5	Length of the longest interception return.
PassesDefended		INT	No	11	Total number of passes defended.
Sacks		FLOAT	No		Total number of sacks.
Safties		INT	No	11	Total number of safeties.
SoloTackles		INT	No	11	Total number of solo tackles.
StatID	PK	INT	Yes	11	Unique identifier for each defensive statistic record, primary key, linked to Yearly_Statistics.
TotalTackles		INT	No	11	Total number of tackles.
YardsPerInt		FLOAT	No		Average yards gained per interception.

4. Business Questions

1. Is there a correlation between the weight of defensive players and the number of sacks they make per season?

2. Which are the top 5 colleges in terms of producing players with the highest average rushing yards per game?

3. How has the average completion percentage changed for quarterbacks over the past decade?

4. How does the average rushing yards per attempt compare between players who attended college versus those who came directly from high school?

5. Which quarterbacks who have at least 15 seasons of recorded statistics demonstrate the greatest consistency in their passing performance, as evidenced by their career passer rating?

6. Which running back players, who have participated in at least 15 games, have experienced a decrease in rushing yards but an increase in rushing touchdowns compared to the previous year, focusing on the last five years?

7. Who are the top 1% of NFL rookie players in terms of total rushing yards who debuted in the latest season?

8. Which top 5 teams have shown a trend of improving defense over the last three seasons?