**BAIS:3200**

**Final Project**

Cade Metzger; Dylan Tracy; Carson Stocker; Zack Lasek; Max Stevens

**Introduction:**

This project will break down data from college football teams and their performances to determine the relationships between wins/losses and various performance metrics. These can include, but are not limited to, rank of offense, defense, and football performance metrics such as yards, points, and turnover margin. The database application and analysis will be useful in displaying the key features to a team’s performance and how those metrics affect the success of the team in an individual season and across seasons as well.

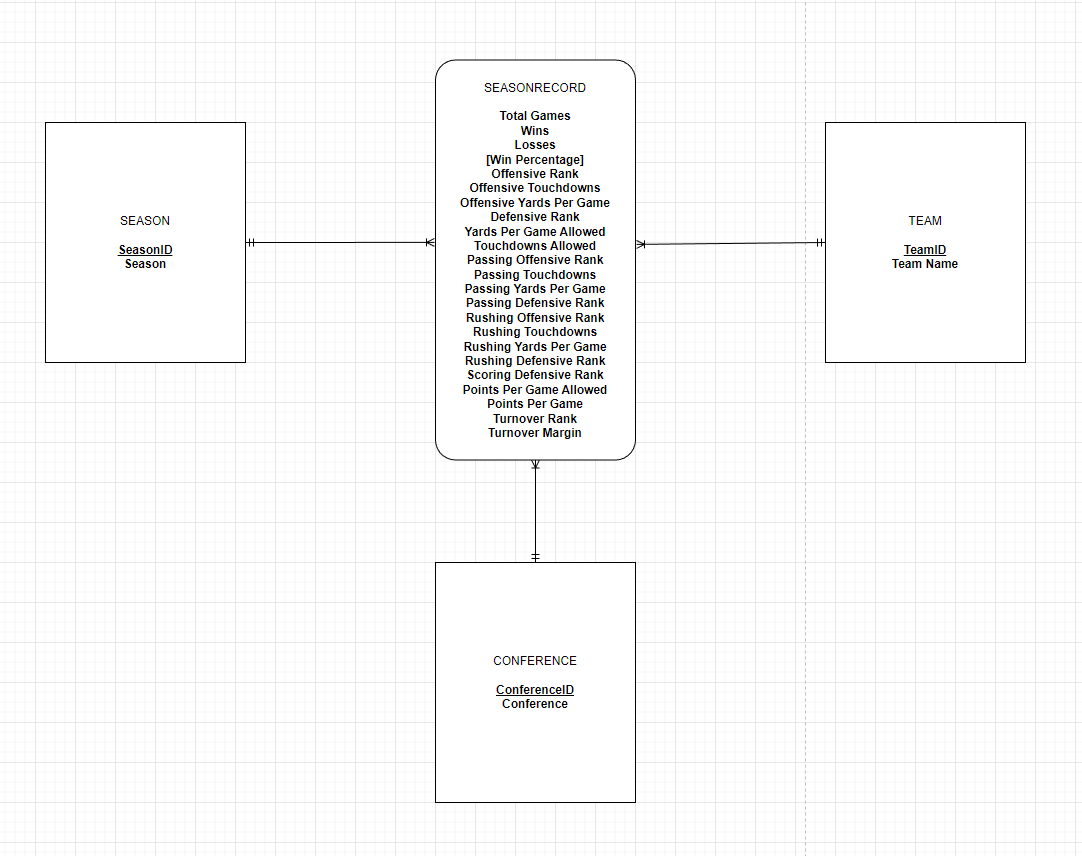
**Data:**

The data was collected from Kaggle based on statistics from the 2019 college football season (<https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019?select=cfb19.csv>). The original data includes data from all seasons from 2013 to 2021. We used data from 2018 and 2019 because 2021 had some inconsistent formatting and missing/illogical data in it and 2020 was the COVID-19 season, where a lot of teams played an unusual schedule with varying amounts of games and players missing because of it. We cleaned the 2018 and 2019 season data by removing some columns that we thought weren’t necessary to our project and repetitive data columns. Our data includes passing and rushing offensive statistics, defensive statistics, and performance metrics for each team. Additionally, we include some overall team information such as their conference, win-loss record, and turnover margin. We created 3 surrogate attributes/keys (TeamID, SeasonID, and ConferenceID) to easily identify the teams, what season the metrics are for, and what conference the team is a part of. Table 1 displays the descriptions of our data.

*Table 1 Data Dictionary:*

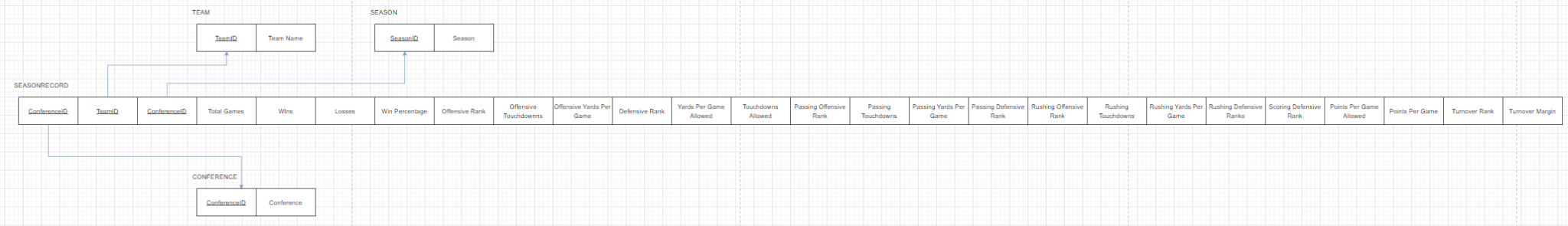
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| TeamID  TeamName  ConferenceID  Conference  SeasonID  Season  TotalGames  Wins  Losses  WinPercentage  Offensive Rank  Offensive Touchdowns  Offensive Yards Per Game  Rushing Offensive Rank  Rushing Touchdowns  Rushing Yards Per Game  Passing Offensive Rank  Passing Touchdowns  Pass Yards Per Game  Points Per Game  Defensive Rank  Yards Per Game Allowed  Offensive Touchdowns Allowed  Rushing Defense Rank  Scoring Defense Rank  Pass Defense Rank  Points Per Game Allowed  Turnover Rank  Turnover Margin | Text  Text  Text  Text  Text  Text  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric  Numeric | Unique ID for each team  School name of team  Unique ID for each conference  Conference the team belongs to  Unique ID for both seasons  The year of the season  Games played during the year  Wins for the year  Losses for the year  Percentage of wins/losses  National offensive team ranking  Total offensive touchdowns  Yards gained per game  National rushing ranking  Rush touchdowns for the year  Rush yards gained per game  National passing ranking  Passing touchdowns all year  Passing yards gained per game  Points scored per game  National defensive team ranking  Yards allowed per game  Touchdowns allowed all year  National rush defense rank  National ranking for pts allowed  National ranking vs pass  Points allowed per game  National ranking for turnovers  Margin of turnovers gained/lost |

There are four primary entities in the database. The first three are TEAM, SEASON, CONFERENCE, which are identified by TeamID, SeasonID, ConferenceID, respectively. The fourth primary entity is SEASONRECORD and that is identified by all the prior three primary keys as its composite primary key. The SEASONRECORD entity is an associative entity of the other 3 tables. All attributes involved in the database are mandatory. All the relationships between the entities to the associate entity are one to many mandatory. This is the case because the metrics for the teams in the SEASONRECORD table can only belong to one season, team, and conference, but all those metrics belong to many conferences, teams, and seasons. Figure 1 displays the ERD for this data.



*Fig. 1 Entity Relationship Diagram (ERD)*

Based on the ERD above, we had to normalize the data into 3NF. This resulted in 4 tables. There were no multivalued or composite attributes that needed to be resolved and there were no M:M relationships that required new tables to be made. We did have to separate the associate entity of SEASONRECORD into its own table (which is very long and has lots of columns) with all the other table’s primary keys as the composite primary key of the SEASONRECORD table. Other than that, there was not a large amount of normalization that was needed.

*Fig. 2 Graphical Relational Schema*

**Database Implementation:**

We wrote 4 CREATE TABLE commands for the 4 tables in the relational schema in order to implement the data into Apex. Each table also has an example INSERT command for an entry/row in the respective tables.

The 3 tables that are connected to the associate entity are created before the associate entity (SEASONRECORD). The order that those 3 tables are created does not matter, but SEASONRECORD must be created last.

SEASON

CREATE TABLE Season(

SeasonID VARCHAR(2) NOT NULL,

SeasonYear CHAR(4) NOT NULL,

CONSTRAINT SeasonPK PRIMARY KEY (SeasonID));

INSERT INTO SEASON VALUES (‘Y1’, ‘2018’);

CONFERENCE

CREATE TABLE Conference(

ConferenceID VARCHAR(2) NOT NULL,

ConferenceName VARCHAR(50) NOT NULL,

CONSTRAINT ConferencePK PRIMARY KEY (ConferenceID));

INSERT INTO CONFERENCE VALUES (‘1’, ‘Mountain West’);

TEAM

CREATE TABLE Team(

TeamID VARCHAR(3) NOT NULL,

TeamName VARCHAR(100) NOT NULL,

CONSTRAINT TeamPK PRIMARY KEY (TeamID));

INSERT INTO TEAM VALUES (‘1’, ‘Air Force’);

SEASONRECORD

CREATE TABLE SeasonRecord(

ConferenceID VARCHAR2(3) NOT NULL,

TeamID VARCHAR2(3) NOT NULL,

SeasonID VARCHAR(2) NOT NULL,

Total\_Games NUMBER(2,0) NOT NULL,

Wins NUMBER(2,0) NOT NULL,

Losses NUMBER(2,0) NOT NULL,

WinPercentage NUMBER(3,2) NOT NULL,

OffensiveRank NUMBER(3,0) NOT NULL,

OffensiveTouchdowns NUMBER(3,0) NOT NULL,

PointsPerGame NUMBER(3,1) NOT NULL,

OffensiveYardsPerGame NUMBER(4,1) NOT NULL,

PassingOffensiveRank NUMBER(3,0) NOT NULL,

PassingTouchdowns NUMBER(3,0) NOT NULL,

PassingYardsPerGame NUMBER(4,1) NOT NULL,

RushingOffensiveRank NUMBER(3,0) NOT NULL,

RushingTouchdowns NUMBER(3,0) NOT NULL,

RushingYardsPerGame NUMBER(4,1) NOT NULL,

DefensiveRank NUMBER(3,0) NOT NULL,

ScoringDefensiveRank NUMBER(3,0) NOT NULL,

PassingDefensiveRank NUMBER(3,0) NOT NULL,

RushingDefensiveRank NUMBER(3,0) NOT NULL,

TouchdownsAllowed NUMBER(3,0) NOT NULL,

PointsPerGameAllowed NUMBER(3,1) NOT NULL,

YardsPerGameAllowed NUMBER(4,1) NOT NULL,

TurnoverRank NUMBER(3,0) NOT NULL,

TurnoverMargin NUMBER(2,0) NOT NULL,

CONSTRAINT SeasonRecordPK PRIMARY KEY (SeasonID, ConferenceID, TeamID),

CONSTRAINT SeasonFK FOREIGN KEY (SeasonID) REFERENCES Season (SeasonID),

CONSTRAINT TeamFK FOREIGN KEY (TeamID) REFERENCES Team (TeamID),

CONSTRAINT ConferenceFK FOREIGN KEY (ConferenceID) REFERENCES Conference (ConferenceID));

INSERT INTO SEASONRECORD VALUES (‘1’, ‘1’, ‘Y2’, 13, 11, 2, 0.85, 51, 55, 421.8, 17, 319.6, 32, 125, 14, 123.2, 51, 2, 41, 298.5, 7, 20, 19.8, 34.1, 78, -2);

**Analysis:**

The analysis presented in this project is useful for teams and team managers to determine the most important areas of a team and how they affect wins, as well as considerations such as conference success and overall team success. It is also useful for fans who want to know more about college football statistics and metrics and what effects they have on team success.

Question 1: Passing & Winning

Are higher average passing yards per game associated with better winning percentages? We wrote a simple query that returns the team name, average passing yards per game, and winning percentage for all teams across both years and grouped it by team. We then ordered the results from highest to lowest based on average passing yards per game.

SELECT TeamName as Team\_Name, AVG(PassingYardsPerGame) AS Average\_Passing\_Yards\_Per\_Game, AVG(WinPercentage \* 100)||'%' AS Win\_Percentage

FROM SeasonRecord JOIN Team ON Team.TeamID = SeasonRecord.TeamID

GROUP BY TeamName

ORDER BY Average\_Passing\_Yards\_Per\_Game DESC;

The results of the query are shown below (Figure 3). It returns 130 rows, but we are just showing the first 15 because of how long the results are. The teams with the most passing yards per game are Washington St, Texas Tech, and Alabama. Not all the top teams in terms of passing yards per game have the highest winning percentages, so there isn’t a strong correlation between the two. As we could not definitively decide the extent that average passing yards affect win percentage, we must also look to other factors in determining what leads to a solid win percentage.



*Fig. 3 Passing & Winning*

Question 2: Comparing Conferences

How do all the conferences compare in terms of their average win percentage for all the teams in that conference? We wrote a join query that returns each conference and the average winning percentage for all the teams in that conference. We grouped the averages by conference and ordered the results from highest to lowest based on the conference win percentage.

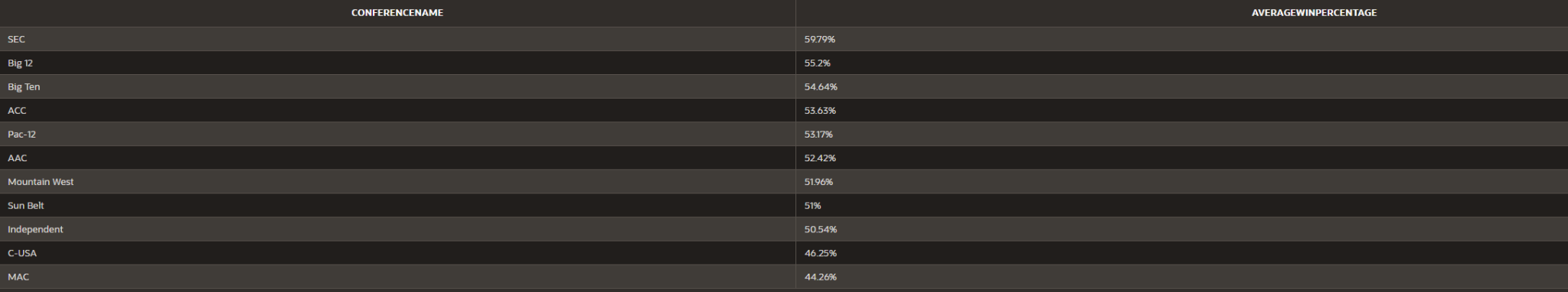
SELECT ConferenceName, (ROUND((AVG(WinPercentage)\*100),2) || '%') AS AverageWinPercentage

FROM SeasonRecord JOIN Conference ON Conference.ConferenceID = SeasonRecord.ConferenceID

GROUP BY ConferenceName

ORDER BY AVG(WinPercentage) DESC;

The results of the query are shown below (Figure 4). It returns 11 rows and each row is one of the 11 conferences (the 10 conferences plus Independents grouped together). The SEC, Big 12, and Big Ten were the three conferences with the highest winning percentages in that order. Conversely, the MAC, C-USA and Independent schools had the worst winning percentages in that order. There is not a huge discrepancy between all the conferences (outside of the very top and bottom ones), but that is expected since most of the games that are played for teams are between conference opponents. Also, in general, the best conferences are the "Power 5" conferences and the low-end ones are not a part of this group. This is expected since Power 5 conferences usually recruit better players and have more popular teams than the other conferences.



*Fig. 4 Comparing Conferences*

Question 3: Best Offenses

Which 15 teams had the most total offensive yards, and what are their respective winning percentages? We wrote a join query that returns the team name, average offensive yards in a season, and average winning percentage for all teams across both seasons. We grouped the results by team and ordered it from highest to lowest by the season yards. We decided we only wanted to return in the final results (FETCH) the top 15 teams in terms of total offensive yards and compare them.

SELECT TeamName AS Team\_Name, AVG(Total\_Games \* OffensiveYardsPerGame) AS Season\_Yards, AVG(WinPercentage \* 100)||'%' AS Win\_Percentage

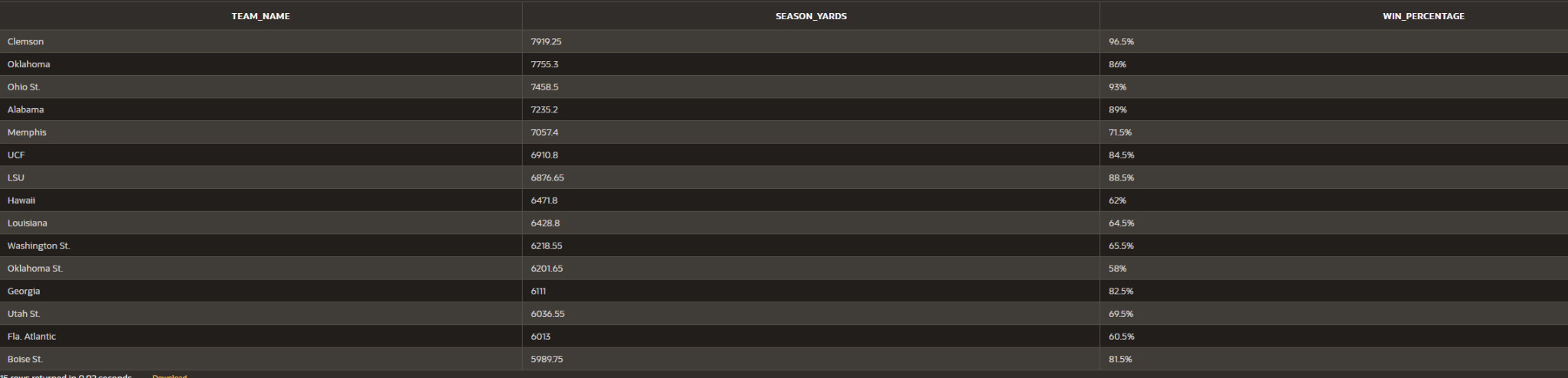
FROM SeasonRecord JOIN Team ON Team.TeamID = SeasonRecord.TeamID

GROUP BY TeamName

ORDER BY Season\_Yards DESC

FETCH FIRST 15 ROWS ONLY;

The results of the query are shown below (Figure 5). It returns 15 rows, which are the top 15 teams in terms of average total offensive yards in a season. Clemson not only had the best total offense, but the highest winning percentage as well. It seems that gaining more yards on offense leads to a better win percentage; however, it is not definitive since some of the teams with the most yards gained do not have great winning percentages. This is most likely due to those teams having poor defenses, but it could be other factors as well.



*Fig. 5 Best Offenses*

Question 4: Team Success

Which teams had the most success across both years? We wrote a CASE query that returns the team name, average winning percentage across both years, and then it grouped the teams based on their winning percentage. The groups go from high to low based on success (winning percentage) and the specific groups and cutoffs can be seen in the query below. We grouped the results by team name and then ordered the results from highest to lowest based on the average winning percentage.

SELECT TeamName AS Team\_Name, ROUND((AVG(WinPercentage)\*100),1)||'%' AS Win\_Percentage , CASE

WHEN AVG(WinPercentage) >= .8 THEN 'Contending'

WHEN AVG(winPercentage) >= .7 THEN 'Great'

WHEN AVG(WinPercentage) >= .6 THEN 'Above Average'

WHEN AVG(WinPercentage) >= .5 THEN 'Average'

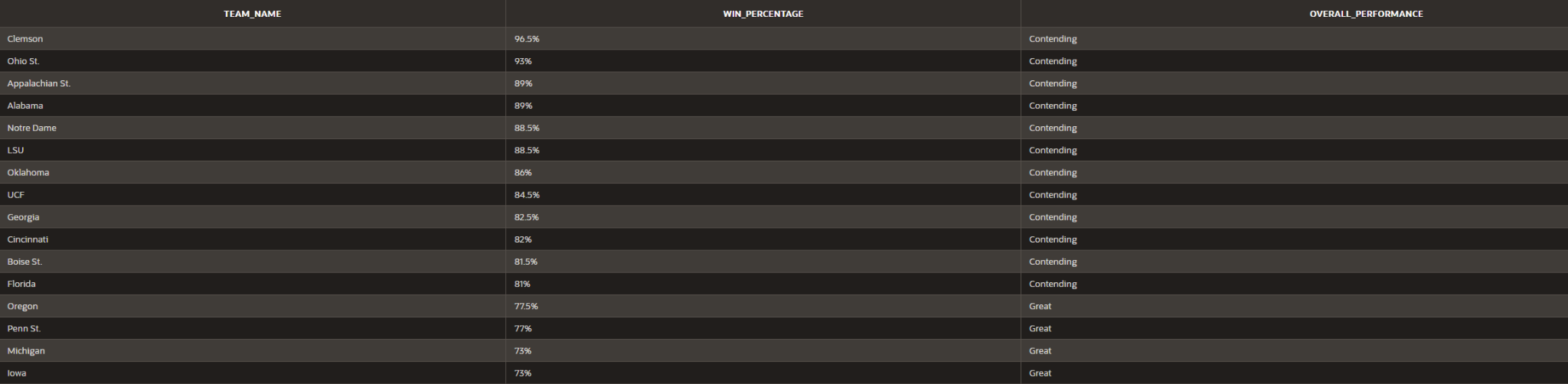
ELSE 'Below Average' END AS Overall\_Performance

FROM SeasonRecord JOIN Team ON Team.TeamID = SeasonRecord.TeamID

GROUP BY TeamName

ORDER BY AVG(WinPercentage) DESC;

The results of the query are shown below (Figure 6). It returns 130 rows, but we are only showing the first 16 as a representation. The teams with the highest degree of success across both years (winning percentage) are considered "Contending" and the group is led by a top 3 consisting of Clemson, Ohio State, and Appalachian State in that order. There are not many teams that fall into that "Contending" category, and it is clearly an exclusive group. On the other hand, most teams fall under the other categories, and more and more are in each group as the list goes down. Most of the teams in Division 1 FBS fall under the "Average" or "Below Average" categories.



*Fig. 6 Team Success*

Question 5: Turnovers & Winning

Is a higher turnover margin associated with a higher win percentage? We wrote a join query that returns the team name, the average turnover margin across both years, and the average winning percentage across both years. The results were grouped by the team name and the results were ordered from highest to lowest based on the average turnover margin.

SELECT TeamName AS Team\_Name, ROUND(AVG(TurnoverMargin),1) AS Turnover\_Margin, ROUND(AVG(WinPercentage)\*100,1)||'%' AS Win\_Percentage

FROM SeasonRecord JOIN Team ON SeasonRecord.TeamID = Team.TeamID

GROUP BY TeamName

ORDER BY Turnover\_Margin DESC;

The results of the query are shown below (Figure 7). The query returns 130 rows, but again we are only displaying the first 15 rows as a representation. There appears to be correlation between turnover margin and higher win percentage as demonstrated by results. The results clearly show that the teams at the top of the results (the teams with the highest turnover margin) have relatively high or the highest winning percentages. While there are outliers (like the top two teams in the results only having mediocre to good winning percentages), the general winning percentage of the top 10 in turnover margin is significantly better than that of the bottom 10 performers based on turnover margin.



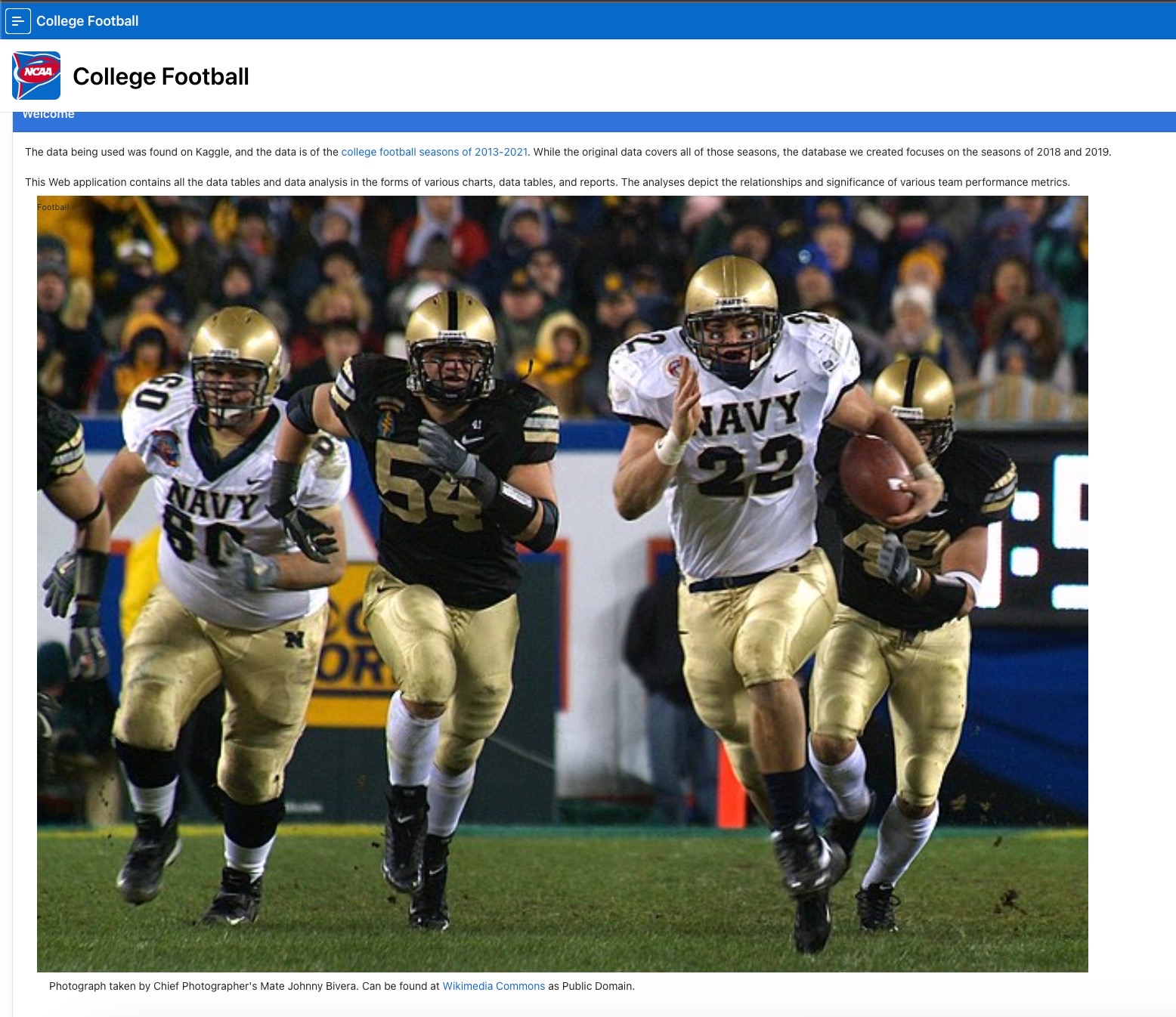
*Fig. 7 Turnovers & Winning*

**Web Design:**

<https://apex.oracle.com/pls/apex/r/shared_database_workspace/college-football/>

Home Page

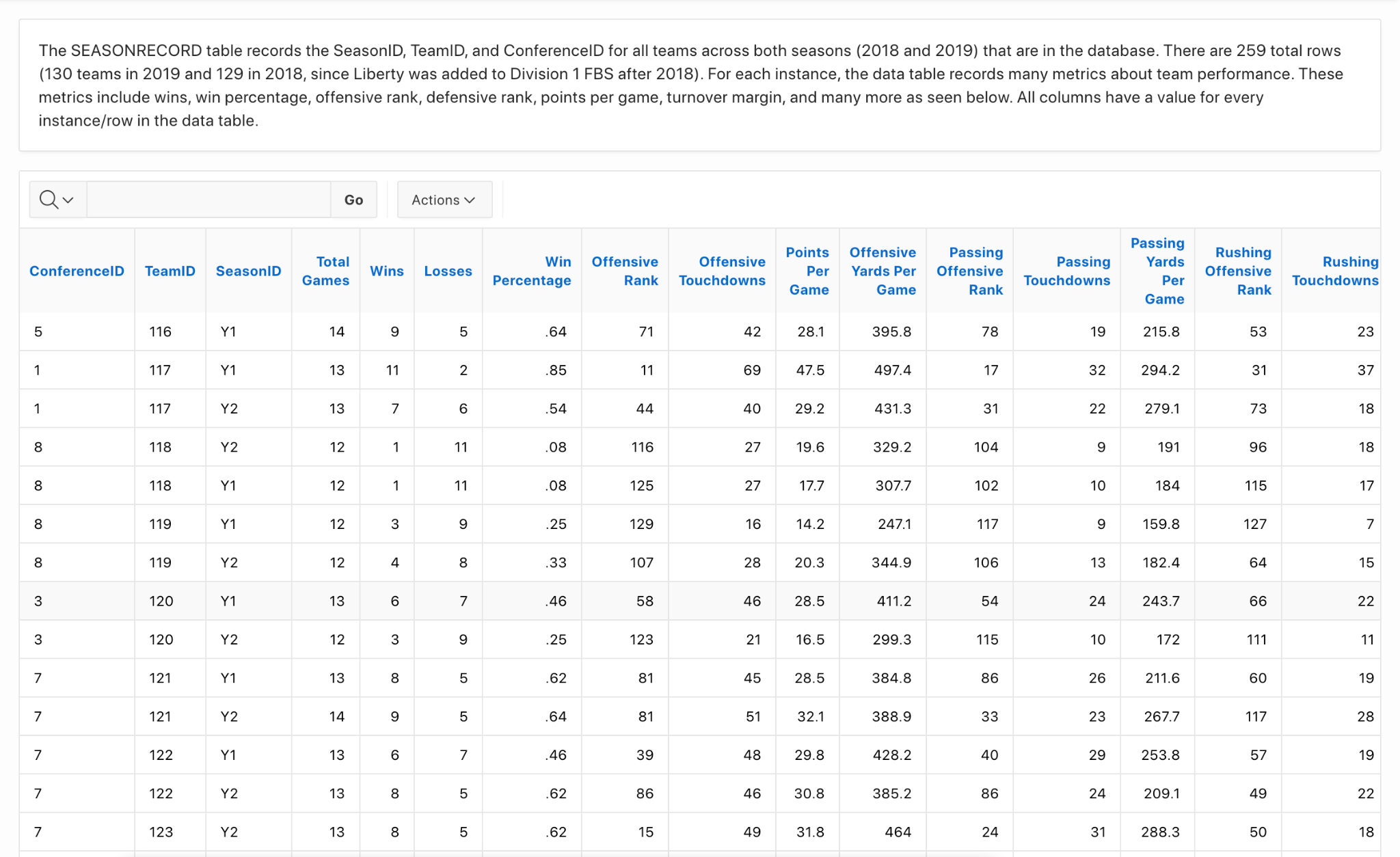
The home page of our web application includes a brief description and introduction to the rest of the application. It also includes a link to where we got the original dataset (Kaggle) and a link to where we found the photo that is being displayed. The photo we used is from Wikimedia Commons, and it is of football players from what appears to be an Army vs. Navy game. We used consistent accent colors here and throughout the application to draw attention to the main features and to create a better, nicer-looking overall page. The main color is blue because that is the color of the NCAA logo. Figure 8 shows a screenshot of the home page.



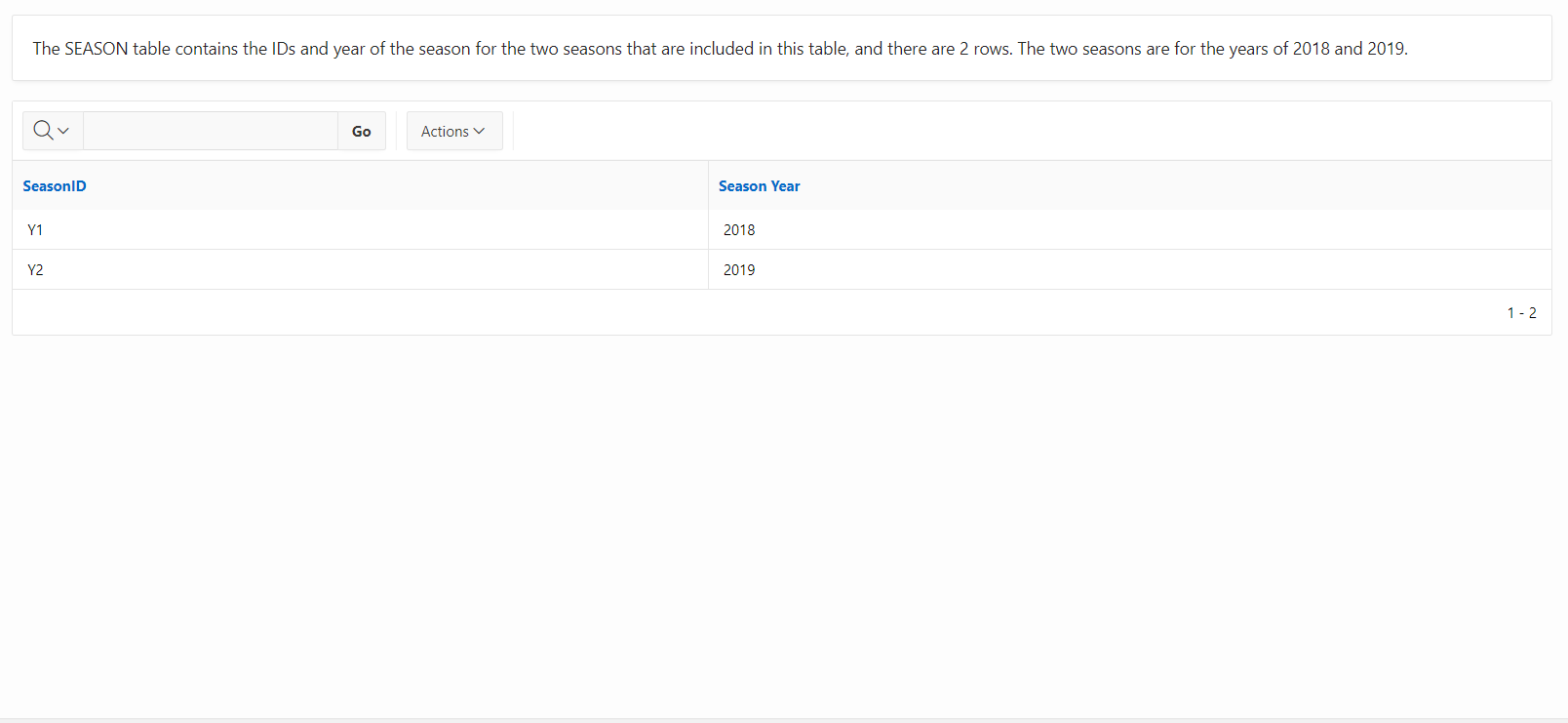
*Fig. 8 Home Page*

Tables

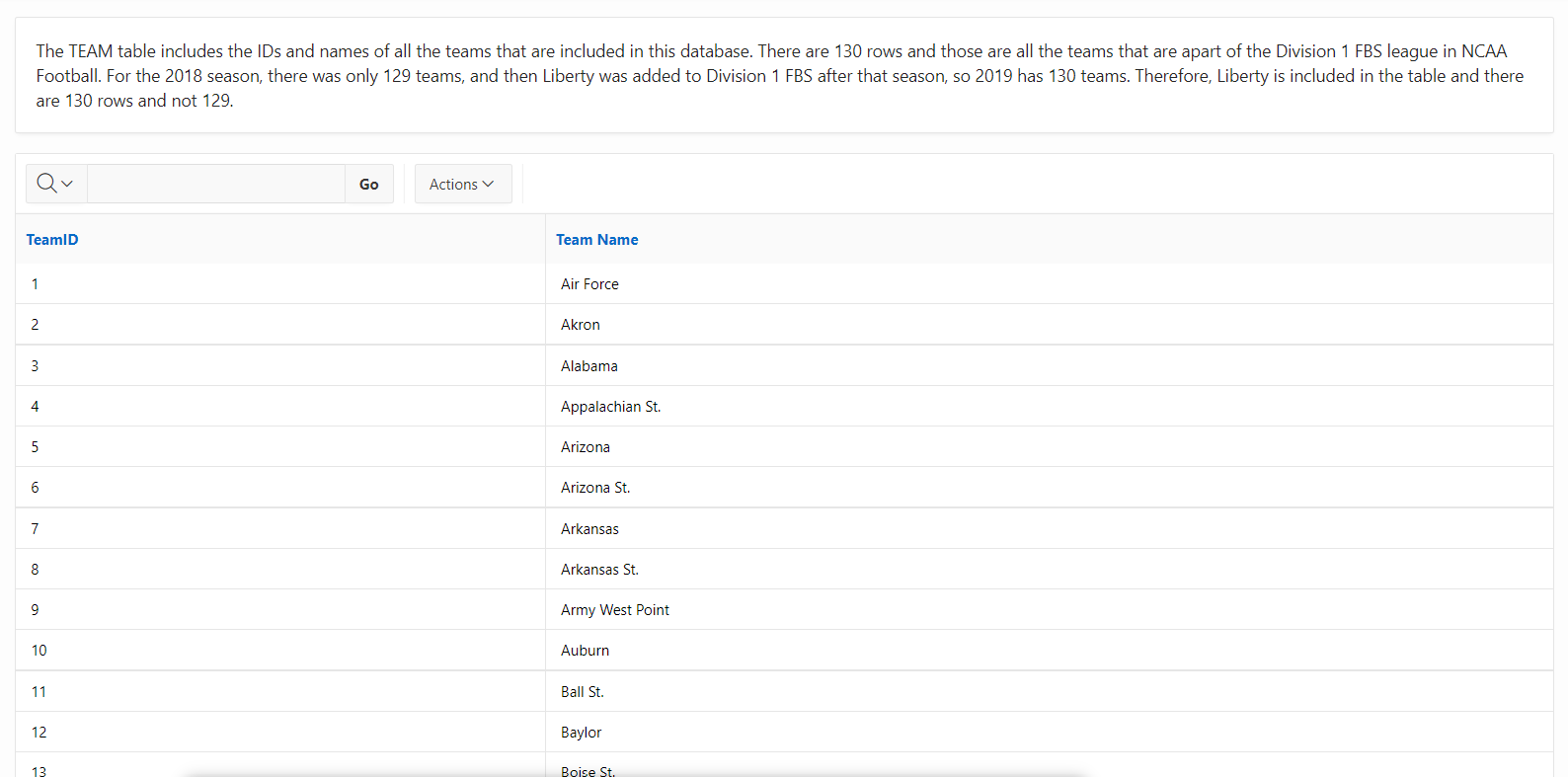
We created interactive reports for each of the tables we utilized in our database (Figures 9-12). Each table allows the user to search and filter through the data within it. We also formatted it to be easy on the eye with differentiated formatting between the header row and the actual data. Each table also has an overhead textbox describing the data to make it more clear if needed. Every column header and data format are made to be readable, easy to consume, and correctly formatted.



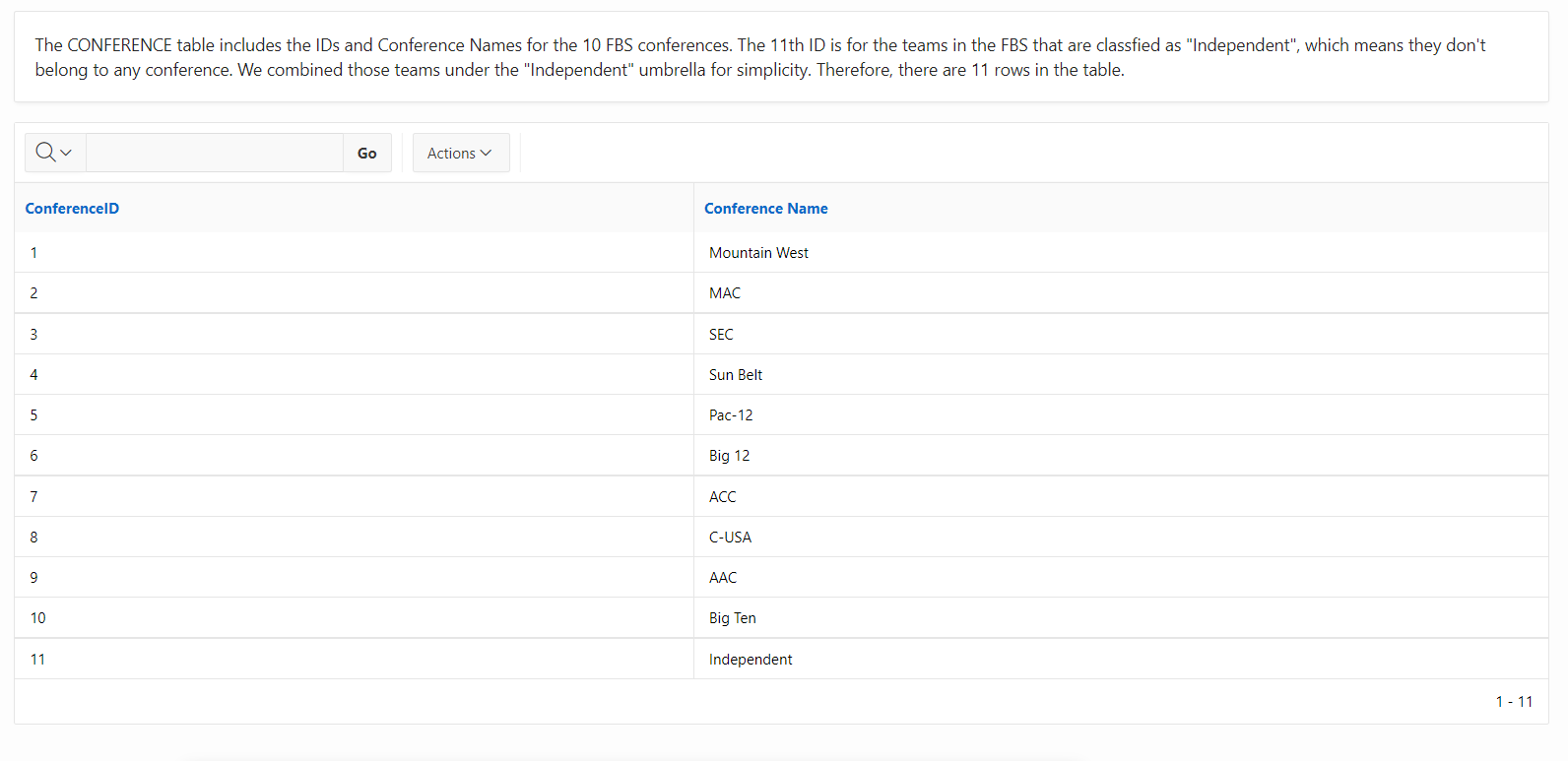
*Fig. 9 Season Record*

**

*Fig. 10 Season*

**

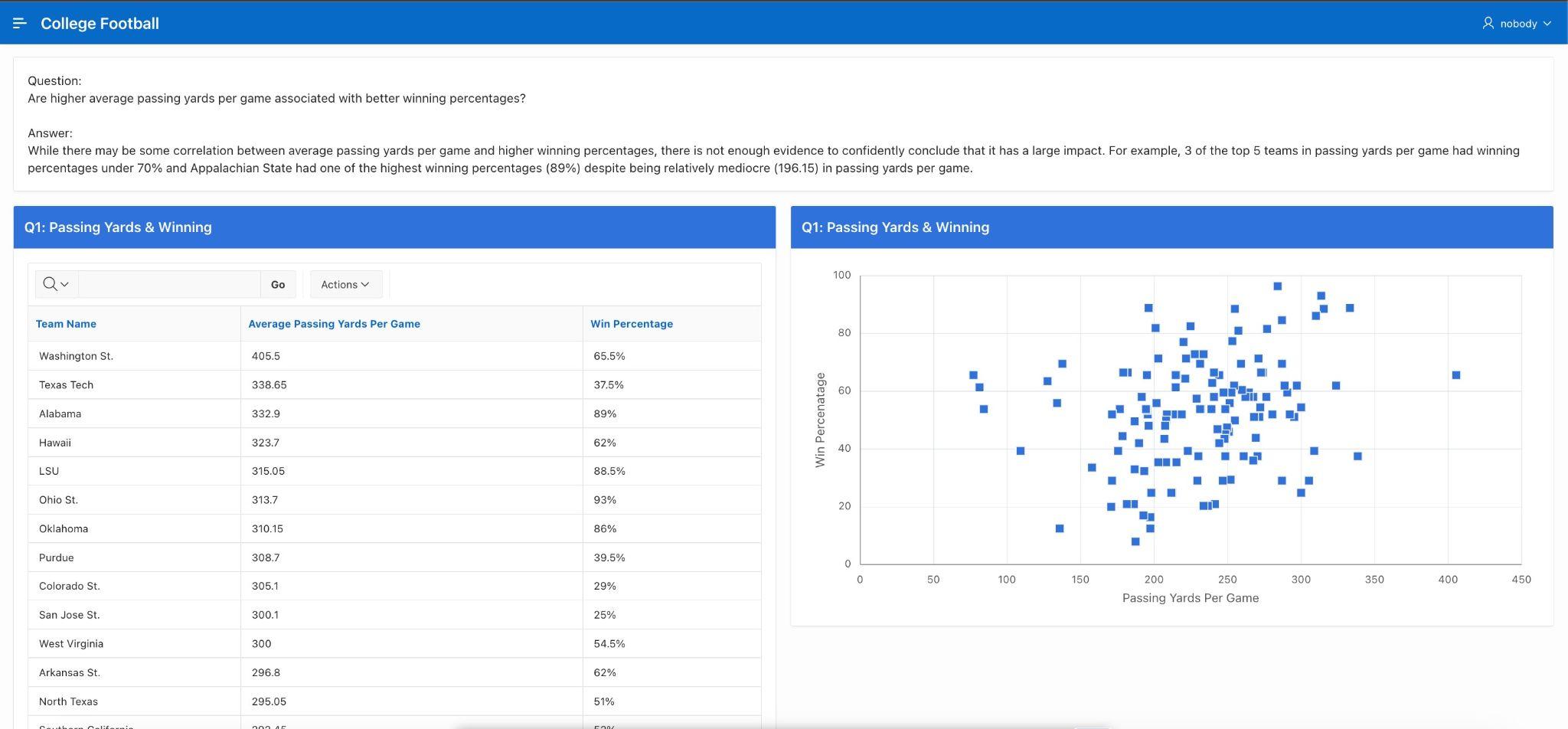
*Fig. 11 Team*

**

*Fig. 12 Conference*

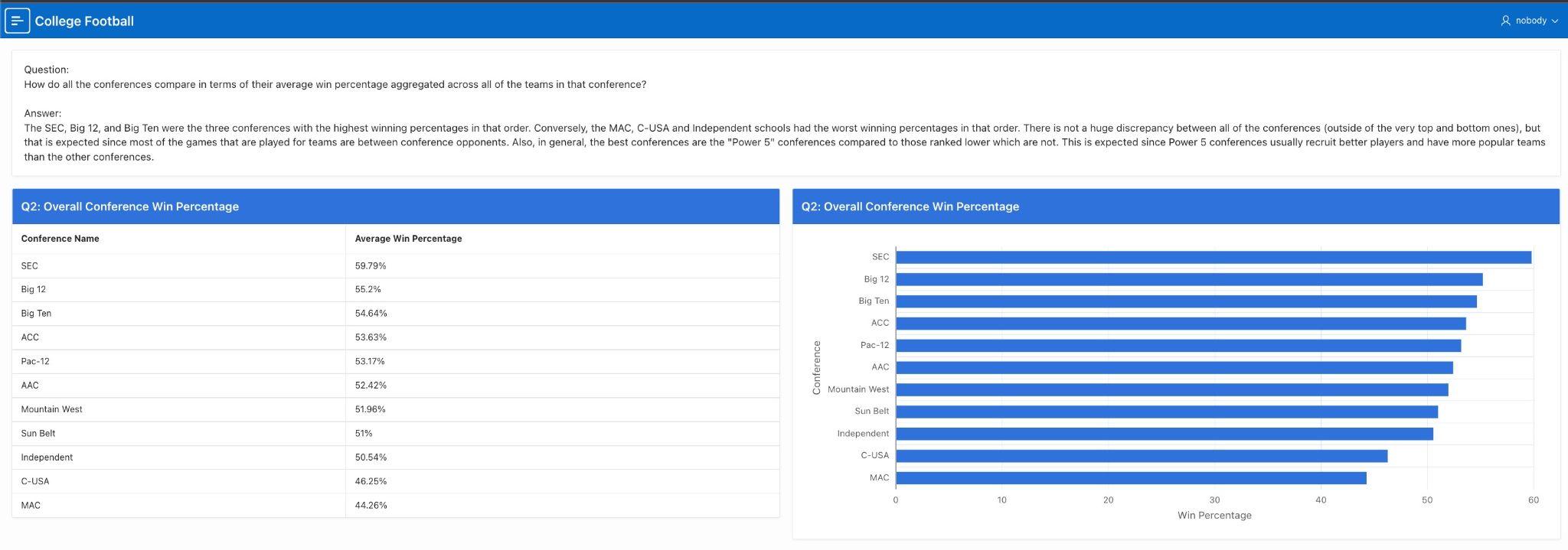
Queries

We first wanted to answer whether or not a team's passing yards per game affected their winning % over the two years with an interactive table with and a scatter plot (Figure 13). We added a text box above to show our question with a detailed response for the results. You can hover over each dot on the scatter plot, and it will show the team name, its average passing yards per game, and average winning percentage aggregated across both seasons.



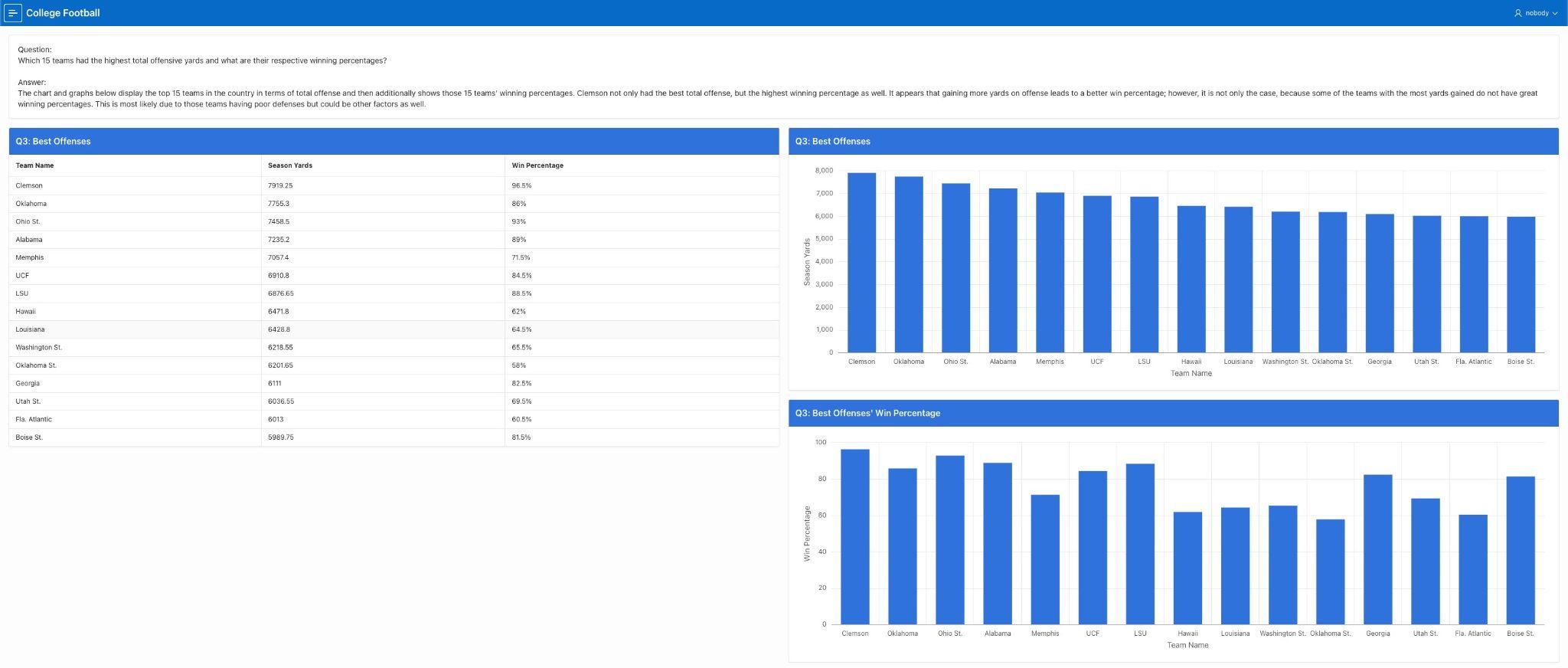
*Fig. 13 Passing Yards & Winning Percentage*

We also investigated each conference and ran a query to see which conferences had the highest average win percentage. We presented the data as both a static table and a bar chart (Figure 14). You can hover over the bar associated with each conference to view their average win percentage value numerically. At the top of the page, we added a text box with our question and our analysis of the query results for clarity.

****

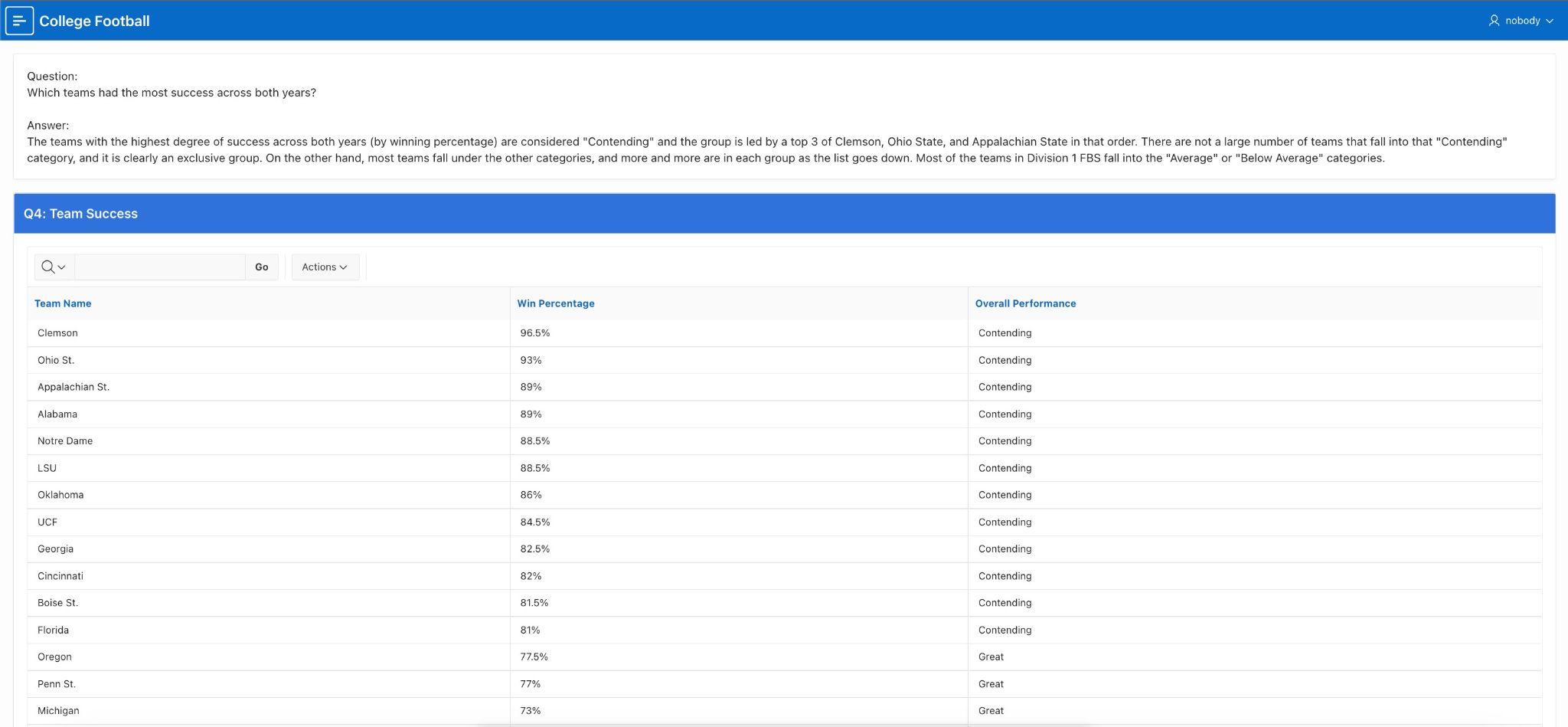
*Fig. 14 Comparing Conferences*

We also researched which 15 teams had the highest total yards and overall winning percentage over the two years. We exhibit this with a static table with the 15 teams, their season yards, and their average win percentage. We also made separate bar charts for total yards and winning percentage (Figure 15). At the top of the page, we added a text box presenting the question and a response to that question in the form of an analysis of the table and both graphs.



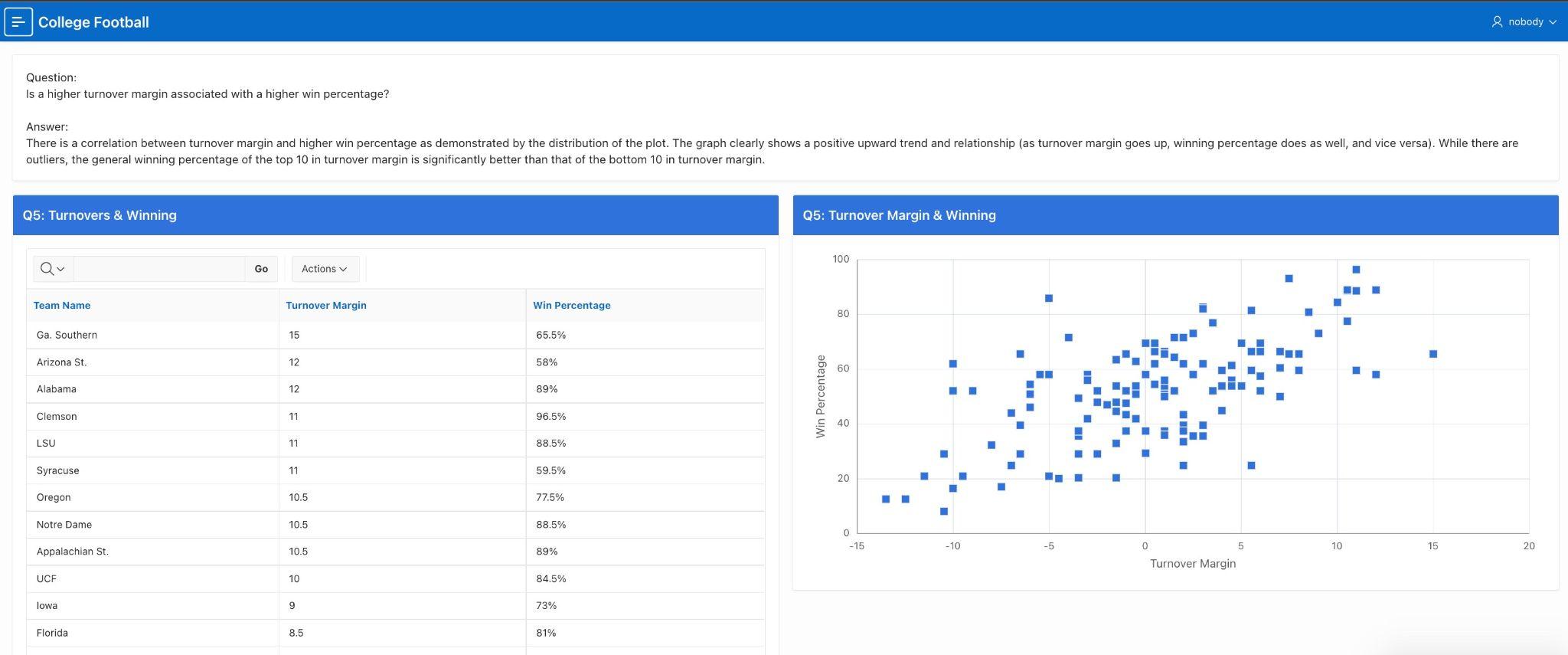
*Fig. 15 Best Offenses*

Question four of our web application included investigating which teams had the most success across both years. Below the question, we gave our answer to the question along with describing the figure below. We revealed the best teams by winning percentage and then discussed what category most teams fall into. The table contains team name, win percentage, and then overall performance. The table is sorted by win percentage with Clemson, Ohio State, and Appalachian State being the first three teams listed. Overall performance is based on win percentage and can fall into contending, great, above average, average, and below average.



*Fig. 16 Team Success*

Our final query and table/graph explored how turnover margin and winning percentage correlate as the question was “Is a higher turnover margin associated with a higher win percentage?”. Our findings were presented with our data in sortable columns in addition to a scatter plot that clearly affirms our findings which can be found in the textbox above. The text box states the question and an analysis of the output (table and graph) as an answer.



*Fig. 17 Turnovers & Winning*