**BAIS:3200**

**NFL Data Final Project**

Jack Abel, Teagan Fletcher, Cole Ludwig, Trey Martin, Jake Nelson

**Introduction**

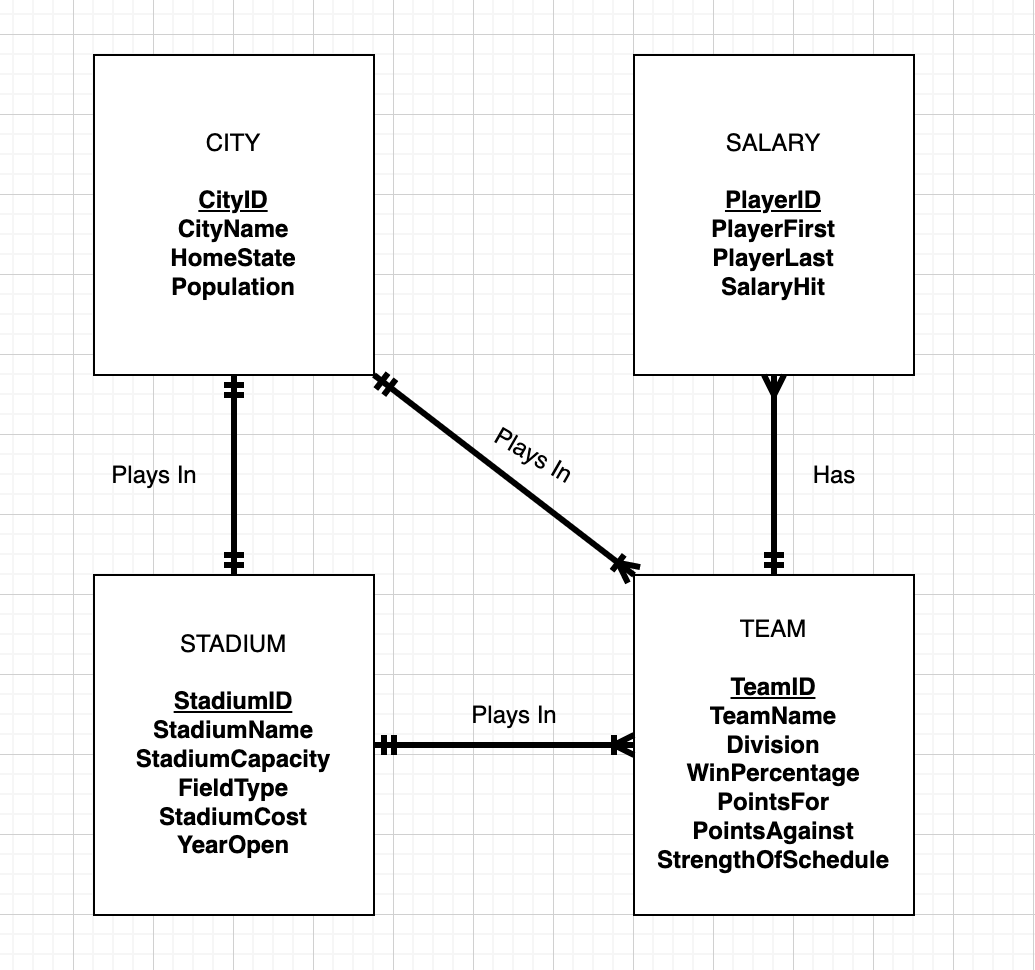
Since the first official game in 1920, football has been America’s beloved sport for over a century. In the United States, the NFL has provided viewers with an enticing outlet for entertainment, and in doing so, has made significant contributions to the nations GDP. Successful NFL franchises support the creation of new jobs, increase consumer spending, and increase tax revenue. To further their economic success, teams must increase their franchise value by having winning seasons. Winning seasons for NFL teams provide increased revenue streams, higher ticket revenue, higher value product sponsorships, and better broadcasting agreements. Our database application and analysis will allow users to comprehend how an NFL franchise may bolster its’ winning percentage.

**Data**

This project uses data from Pro Football Reference to break down each NFL team individually (<https://www.pro-football-reference.com/years/2021/>). The reference website contains data originating from the first game played in the NFL. To reduce the size of data, we collected data from the completed 2020-2021 season to maintain current and relevant statistics for our analysis. We collected the player data for each team by going through the team’s roster page and taking their top five highest-paid players. It must also be stated that the city population is included for the metro area of the city that the team plays in, not just the city alone.

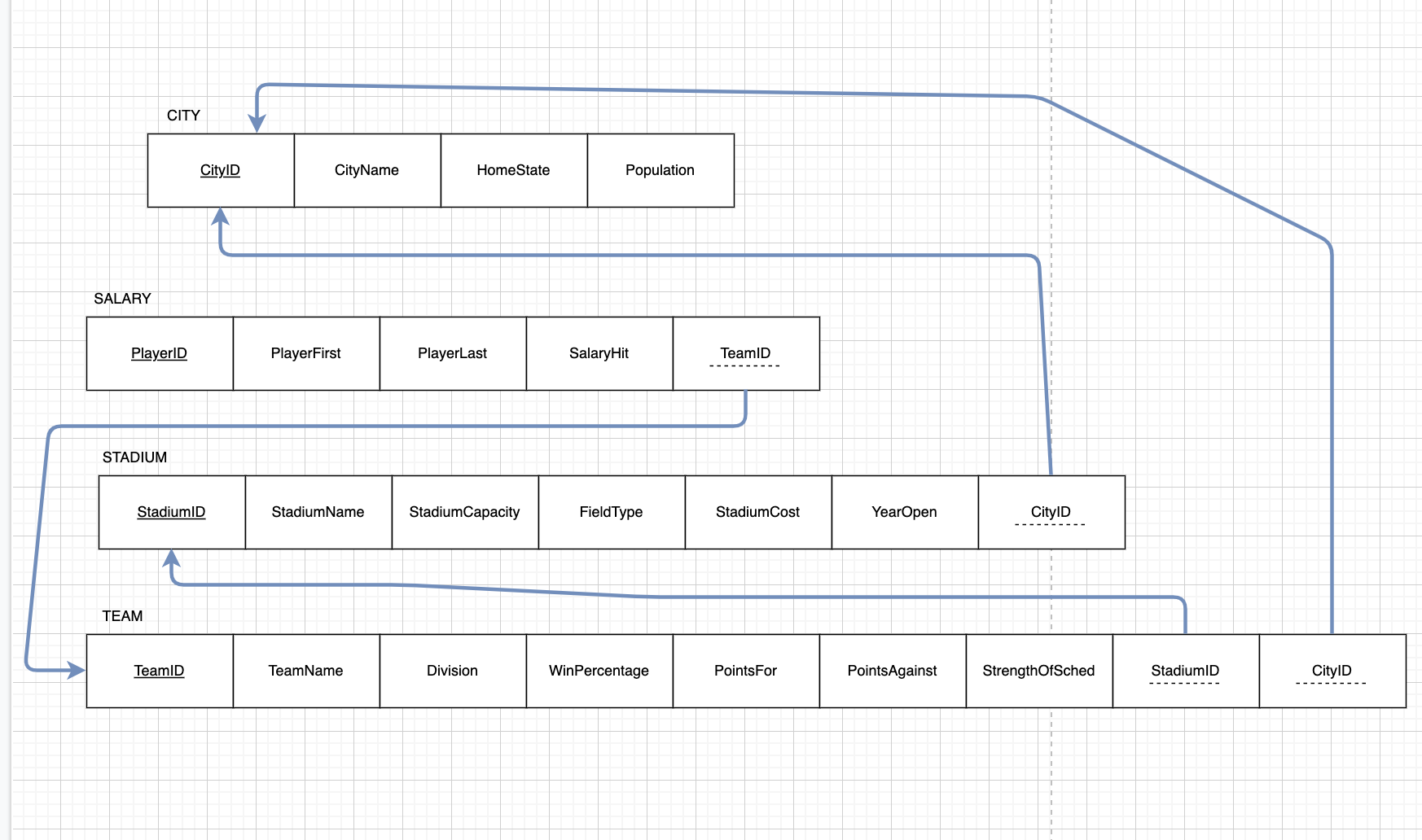
*Table 1: Data dictionary*





*Figure 1: Entity Relationship Diagram (ERD)*

The four primary entities in the database are CITY, SALARY, STADIUM, and TEAM, each identified with CityID, PlayerID, StadiumID, and TeamID respectively. For these identifiers, surrogate keys were created. The raw data has no optional attributes, as each attribute is filled in for each team, player, city and stadium. Each city can have one or many teams that play there and a team can play in only one city. Each city also has one and only one stadium and vice versa. Each team has one and only one stadium, but a stadium can have one or many teams that play in it. Lastly, a team will have many salaried players, and a player must be on one and only one team (players traded midseason will only be a member of the team that they played for at the beginning of the season).



*Figure 2: Graphical Relational Schema*

Based on the ERD, we normalized the data and created a relational schema with 4 tables. Figure 2 displays the graphical relational schema of the database. For the Team table, TeamID acts as the primary key with StadiumID and CityID acting as the foreign keys. In the STADIUM table, StadiumID acts as the primary key with CityID acting as the foreign key. PlayerID acts as the primary key and TeamID acts as the foreign key for the SALARY table. Lastly, CityID acts as the primary key for the CITY table.

**Database Implementation**

To implement the data into APEX, we wrote CREATE TABLE commands for each of the tables in the relational schema.

TEAM

The TEAM table was created first:

CREATE TABLE "TEAM"   
 ( "TEAMID" VARCHAR2(2) NOT NULL ENABLE,   
 "TEAMNAME" VARCHAR2(18) NOT NULL ENABLE,   
 "DIVISION" VARCHAR2(15) NOT NULL ENABLE,   
 "WINPERCENTAGE" NUMBER(4,3) NOT NULL ENABLE,   
 "POINTSFOR" NUMBER NOT NULL ENABLE,   
 "POINTSAGAINST" NUMBER NOT NULL ENABLE,   
 "STRENGTHOFSCHEDULE" NUMBER NOT NULL ENABLE,   
 "STADIUMID" CHAR(3) NOT NULL ENABLE,   
 "CITYID" CHAR(3) NOT NULL ENABLE,   
 CONSTRAINT "TEAM\_PK" PRIMARY KEY ("TEAMID")  
 USING INDEX ENABLE  
 )  
/  
ALTER TABLE "TEAM" ADD CONSTRAINT "TEAM\_FK\_CITY" FOREIGN KEY ("CITYID")  
 REFERENCES "CITY" ("CITYID") ENABLE  
/  
ALTER TABLE "TEAM" ADD CONSTRAINT "TEAM\_FK\_STADIUM" FOREIGN KEY ("STADIUMID")  
 REFERENCES "STADIUM" ("STADIUMID") ENABLE  
/

SALARY

CREATE TABLE "SALARY"   
 ( "PLAYERID" CHAR(3) NOT NULL ENABLE,   
 "PLAYERFIRST" VARCHAR2(25) NOT NULL ENABLE,   
 "PLAYERLAST" VARCHAR2(25) NOT NULL ENABLE,   
 "SALARYHIT" NUMBER NOT NULL ENABLE,   
 "TEAMID" VARCHAR2(2) NOT NULL ENABLE,   
 CONSTRAINT "SALARY\_PK" PRIMARY KEY ("PLAYERID")  
 USING INDEX ENABLE  
 )  
/  
ALTER TABLE "SALARY" ADD CONSTRAINT "SALARY\_FK" FOREIGN KEY ("TEAMID")  
 REFERENCES "TEAM" ("TEAMID") ENABLE  
/

STADIUM

CREATE TABLE "STADIUM"   
 ( "STADIUMID" CHAR(3) NOT NULL ENABLE,   
 "STADIUMNAME" VARCHAR2(55) NOT NULL ENABLE,   
 "STADIUMCAPACITY" NUMBER NOT NULL ENABLE,   
 "FIELDTYPE" VARCHAR2(10) NOT NULL ENABLE,   
 "STADIUMCOST" NUMBER NOT NULL ENABLE,   
 "YEAROPEN" CHAR(4) NOT NULL ENABLE,   
 "CITYID" CHAR(3) NOT NULL ENABLE,   
 CONSTRAINT "STADIUM\_PK" PRIMARY KEY ("STADIUMID")  
 USING INDEX ENABLE  
 )  
/  
ALTER TABLE "STADIUM" ADD CONSTRAINT "STADIUM\_FK" FOREIGN KEY ("CITYID")  
 REFERENCES "CITY" ("CITYID") ENABLE  
/

CITY

CREATE TABLE "CITY"   
 ( "CITYID" CHAR(3) NOT NULL ENABLE,   
 "CITYNAME" VARCHAR2(25) NOT NULL ENABLE,   
 "HOMESTATE" VARCHAR2(15) NOT NULL ENABLE,   
 "POPULATION" VARCHAR2(15) NOT NULL ENABLE,   
 CONSTRAINT "CITY\_PK" PRIMARY KEY ("CITYID")  
 USING INDEX ENABLE  
 )  
/

**Analysis**

Our analysis is intended to inform football fans on the relationships between teams, stadiums, cities, and player salaries and how understanding these relationships may be able to translate into a more successful football team.

Question 1: Strength of schedule and win/loss record

How does a team’s strength of schedule (SoS) impact the team’s final WL record? To address this question, we created a simple query that collects each team’s name, strength of schedule, and win percentage from the team table and ordered them by their win percentage in descending order.

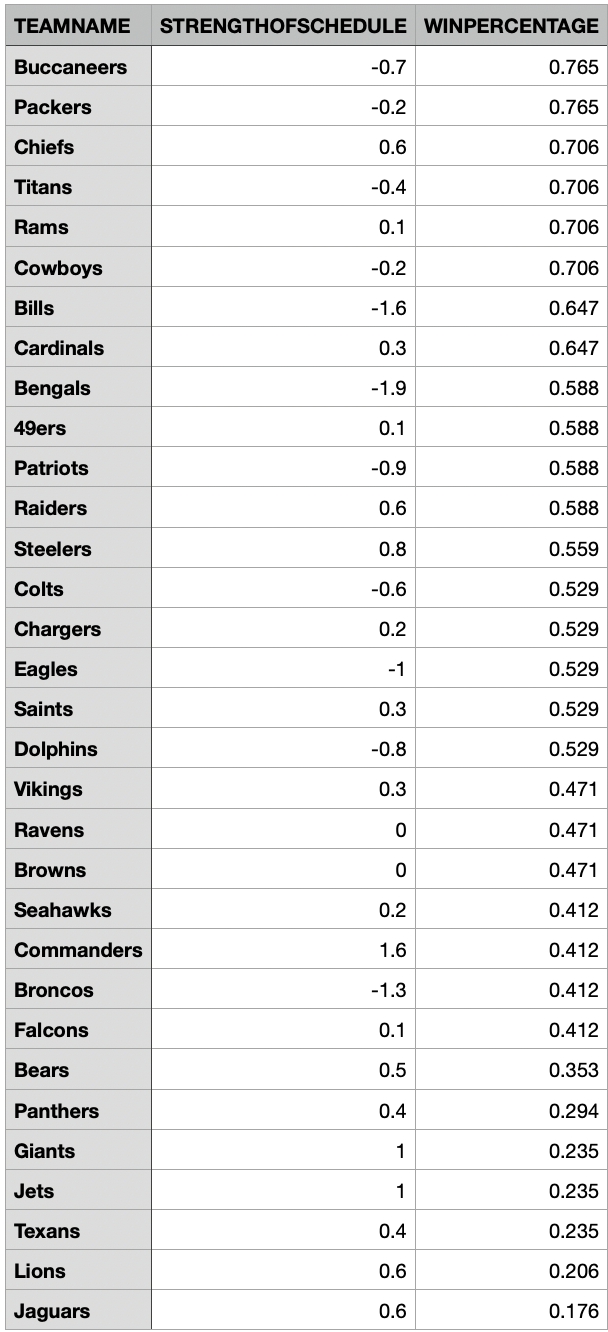
-- QUERY 1

SELECT TEAMNAME, STRENGTHOFSCHEDULE, WINPERCENTAGE

FROM TEAM

ORDER BY WINPERCENTAGE DESC;

The results of this query are shown below (Figure 3). Out of the top ten winning percentages, note that six of them have negative results, meaning, on average, their strength of schedule is weaker than average. For the bottom ten results, nine of the teams have a positive strength of schedule, meaning their schedules are, on average, more difficult.



*Figure 3: Strength of schedule and win percentage*

Question 2: Grass v. turf fields in the NFL

How does the field type impact a team’s final WL record? Specifically, does a team playing on grass have a higher win percentage or vice versa? The style of play can be changed drastically from one playing surface to another, and certain situations such as weather can play a large role in the outcome of a game. To address this question, we ran a join query using the TEAM table and the STADIUM table to determine the winning percentage of teams playing on grass v. teams playing on turf.

-- QUERY 2

SELECT FIELDTYPE, ROUND(AVG(WINPERCENTAGE),3) AS WINPERECNTAGE

FROM TEAM JOIN STADIUM

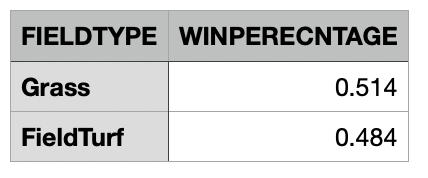
ON TEAM.STADIUMID = STADIUM.STADIUMID

GROUP BY FIELDTYPE

ORDER BY ROUND(AVG(WINPERCENTAGE),3) DESC

;

The results of this query are shown below (Figure 4). Overall, the playing surface of the stadium does not seem to have a large impact on a team's final win/loss record.



*Figure 4: Grass v. field turf winning percentages*

The results from this data are also represented in graphical form on the web application using a scatter chart with a trend line within the data.

Question 3: Player salary and a team's final winning percentage

How does a player’s salary affect a team’s overall win percentage? Specifically, what impact do the highest-paid players have on the result of their team's winning percentage in a season? To address this question, we created a query that took each team from the table and grouped them by win percentage while also calculating the sum of the top five highest paid players on their respective roster.

--QUERY3

SELECT TEAM.TEAMID, TEAM.TEAMNAME, WINPERCENTAGE, TO\_CHAR(SUM(SALARYHIT),'999,999,999') AS TOP\_5\_SALARY

FROM SALARY JOIN TEAM

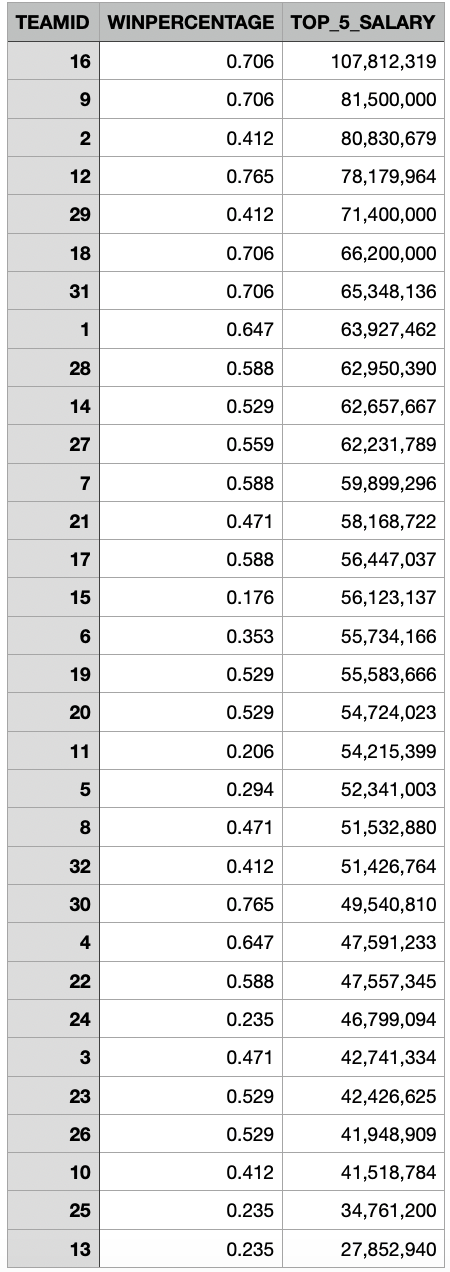
ON SALARY.TEAMID = TEAM.TEAMID

GROUP BY TEAM.TEAMID, TEAM.TEAMNAME, WINPERCENTAGE

ORDER BY TOP\_5\_SALARY DESC

;

The results are shown below (Figure 5). Based on just the top five highest paid players on each team, there is a moderate correlation between the two variables. When running an analysis report for this data, there was a 0.50 R-value associated with the results of this query. Of course, the top five highest players are not a good representation of the rest of the team because you need at least 25 starters per team. However, with more all-pro and pro bowl players, a team should have more firepower.



*Figure 5: Top 5 salaries on each team v. overall win percentage*

Question 4: Stadium capacity and city population

Do cities with a higher population host stadiums that have a higher seating capacity? Cities that have a higher population are able to market their brand to a higher number of people that live close to their respective stadiums and may be able to make more money off their brand. With more capital, a team may be able to play in bigger stadiums. To address this question, we ran a join query using the TEAM table, CITY table, and STADIUM table.

-- QUERY 4

SELECT

TEAMNAME,CITYNAME AS CITY, HOMESTATE AS STATE , STADIUMNAME AS STADIUM,

TO\_CHAR(STADIUMCAPACITY,'999,999') AS STADIUMCAPACITY, TO\_CHAR(POPULATION, '999,999,999') AS POPULATION

FROM TEAM JOIN CITY ON TEAM.CITYID = CITY.CITYID

JOIN STADIUM ON CITY.CITYID = STADIUM.CITYID

ORDER BY STADIUMCAPACITY DESC

;

The results are shown below (Figure 6). Although MetLife Field in New Jersey, home of the Jets and Giants, has both the highest capacity and highest population, there seems to be little correlation beyond that. Outside factors, such as history and age of stadium, difficulty of finding new land for a new stadium, and devoted fan bases may play into this outcome.



*Figure 6: Stadium capacity v. city population*

Question 5: Teams that are above the league average winning percentage

During the season when this data was collected, which teams were above the average points for (offensive points scored) of all teams? To answer this question, we ran a subquery to find the average points for all teams and filtered out all teams below that mark.

--QUERY 5

SELECT TeamID, TeamName, AVG(PointsFor) AS AveragePointsFor

FROM TEAM GROUP BY TeamID, TeamName

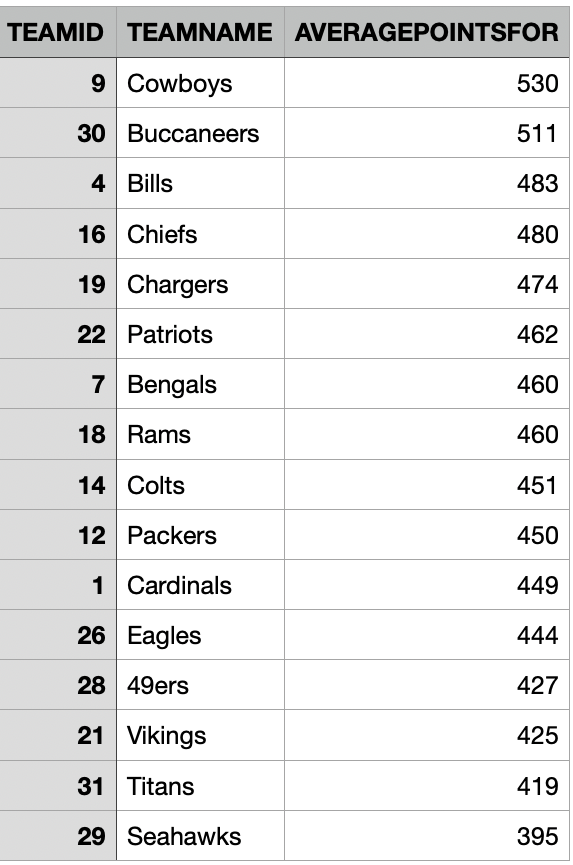
HAVING AVG(PointsFor) >

(SELECT AVG(PointsFor) FROM TEAM)

ORDER BY AveragePointsFor DESC

;

The results are shown below (Figure 7). There were 16 teams that were above the league average in offensive points scored. These teams were statistically more offensively dominate than teams not included in the query results.



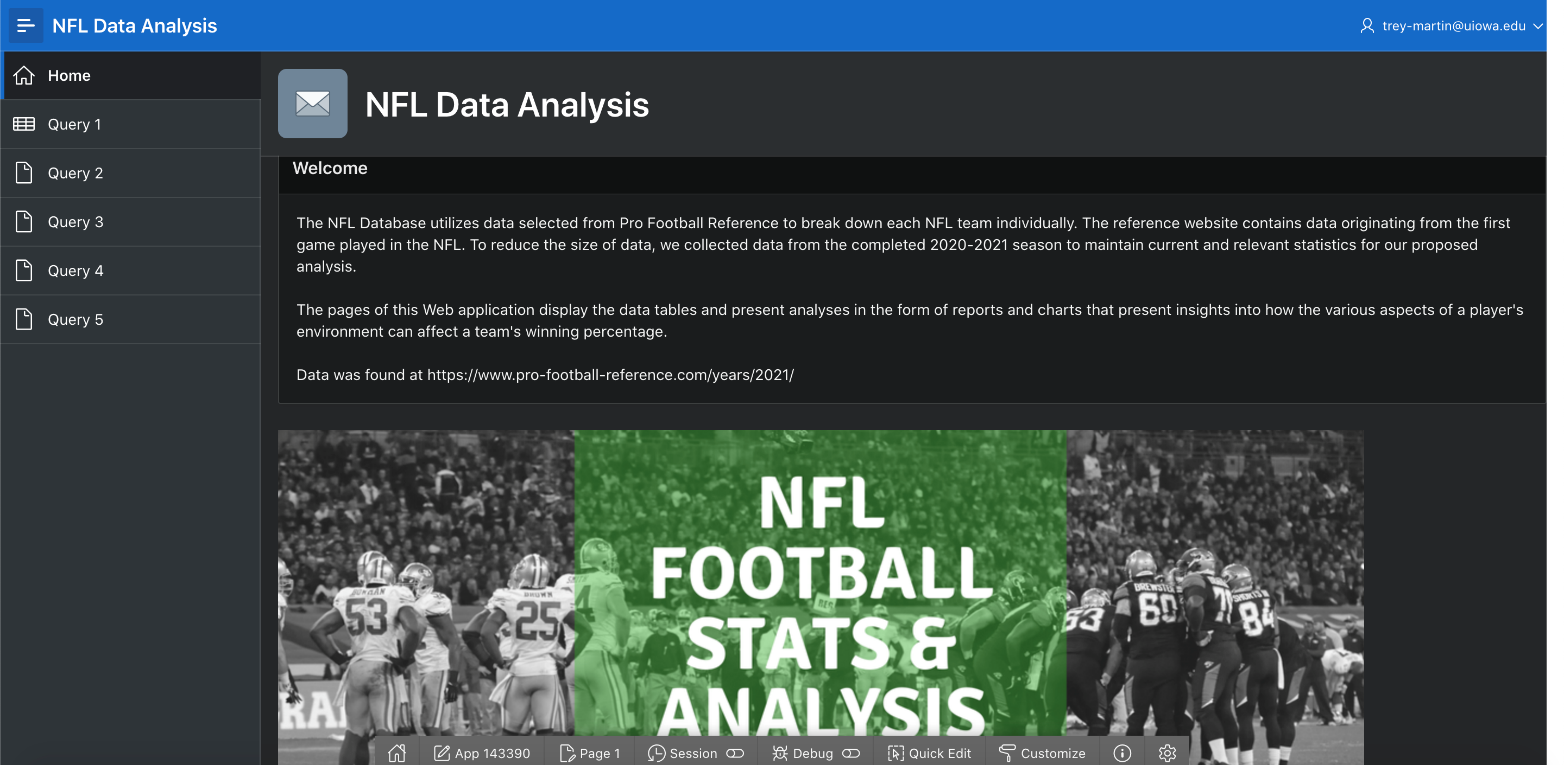
*Figure 7: Teams above the league average offensive points scored*

**Web Design**

[Web Application](https://apex.oracle.com/pls/apex/r/group_10_final_project/nfl-data-analysis/home?session=3035027818224)

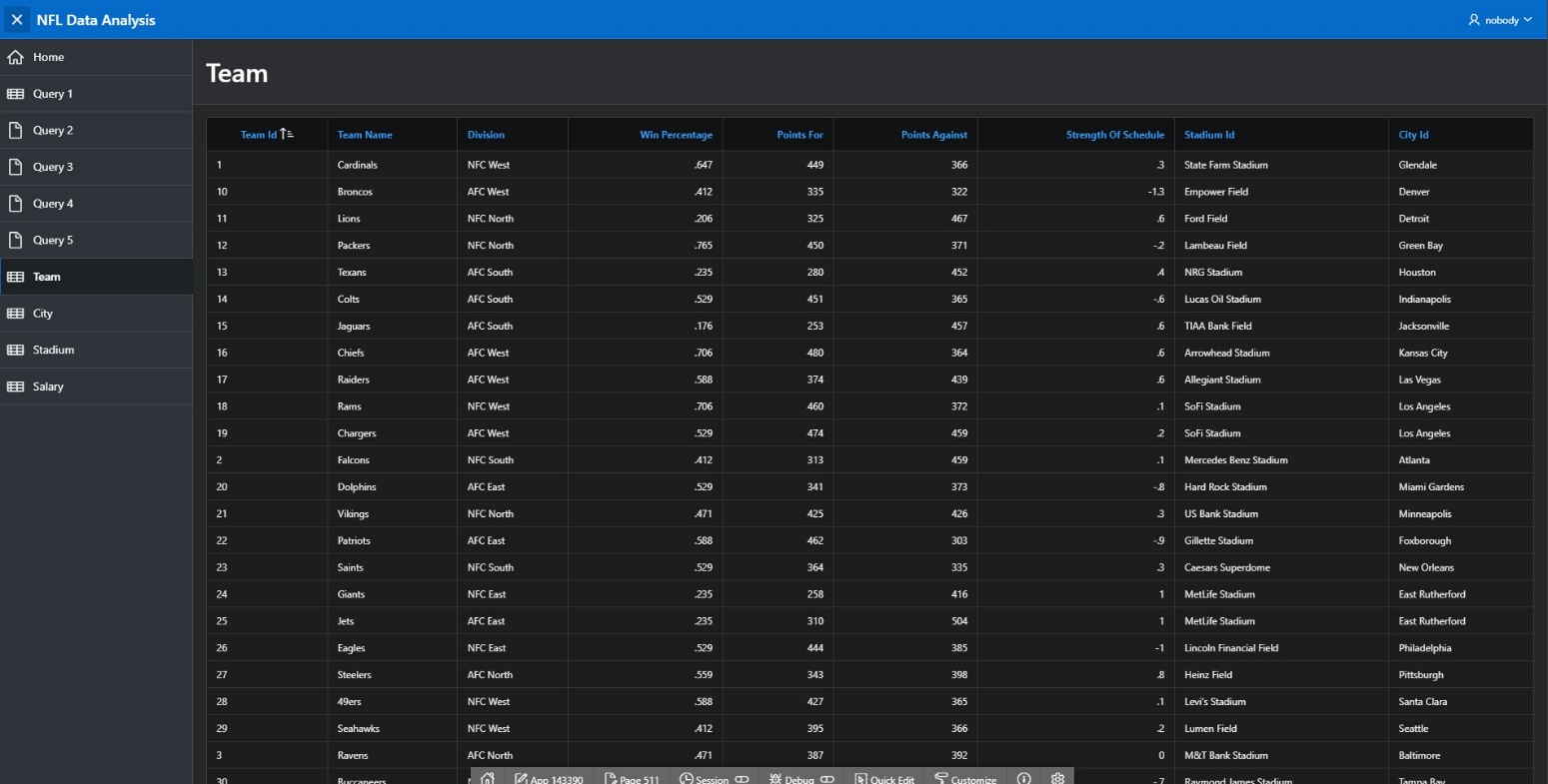
Home Page

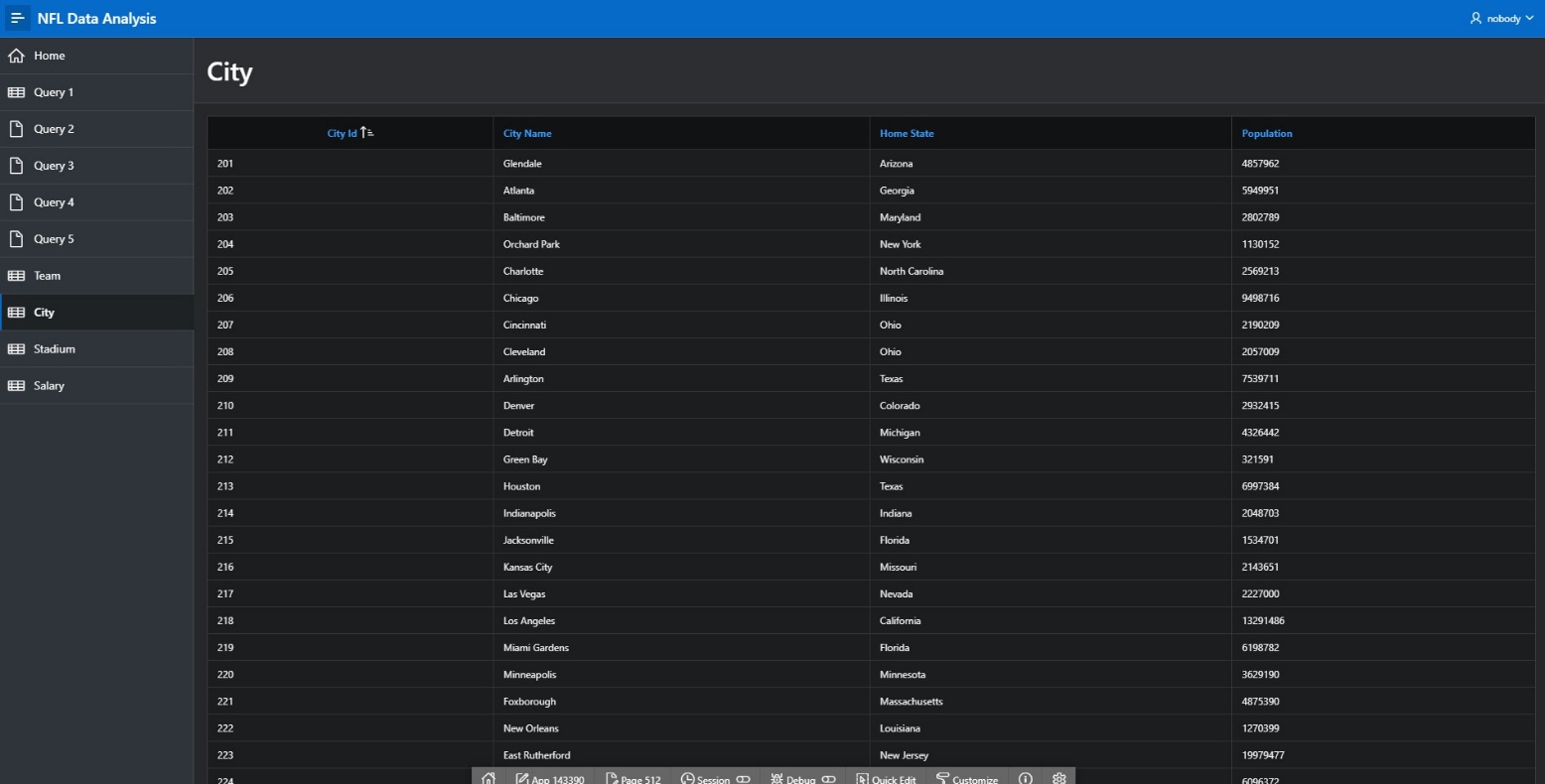
The home page of the web application includes a brief description of the project, as well as a link to the source where data was collected. There is also a main navigation pane where users can find the results for each query we ran related to the data. Also included on the home page is an image (<https://sqldusty.com/tag/nfl-statistics/> ) of an NFL game with text that says, “NFL Football Stats and Analysis.”

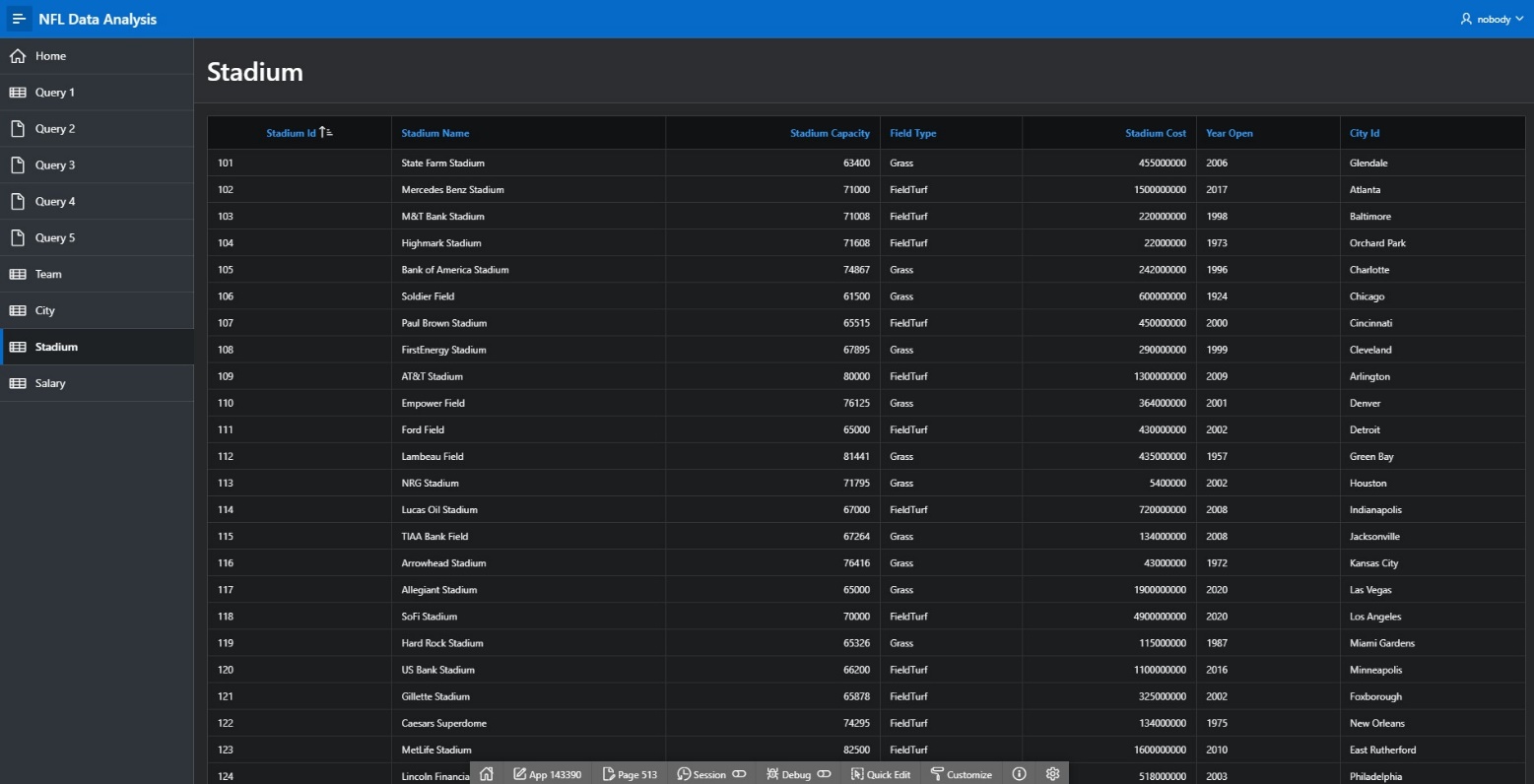
*Figure 8: Web application homepage*

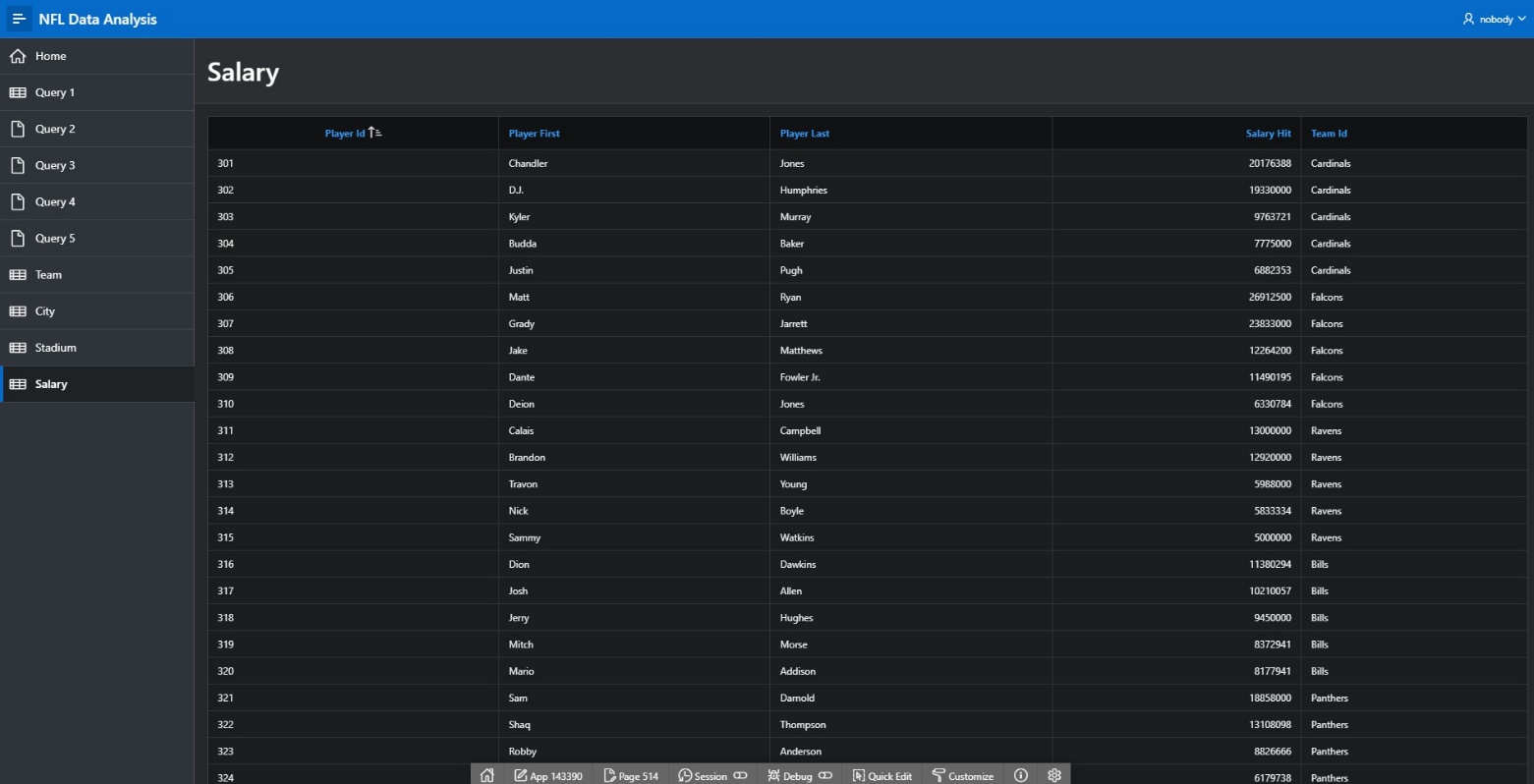
Tables

Included in our web application are the tables of data used to find our query results (Figures 9-12). Column headings and number formats have been customized to ensure that the data is easily interpretable.

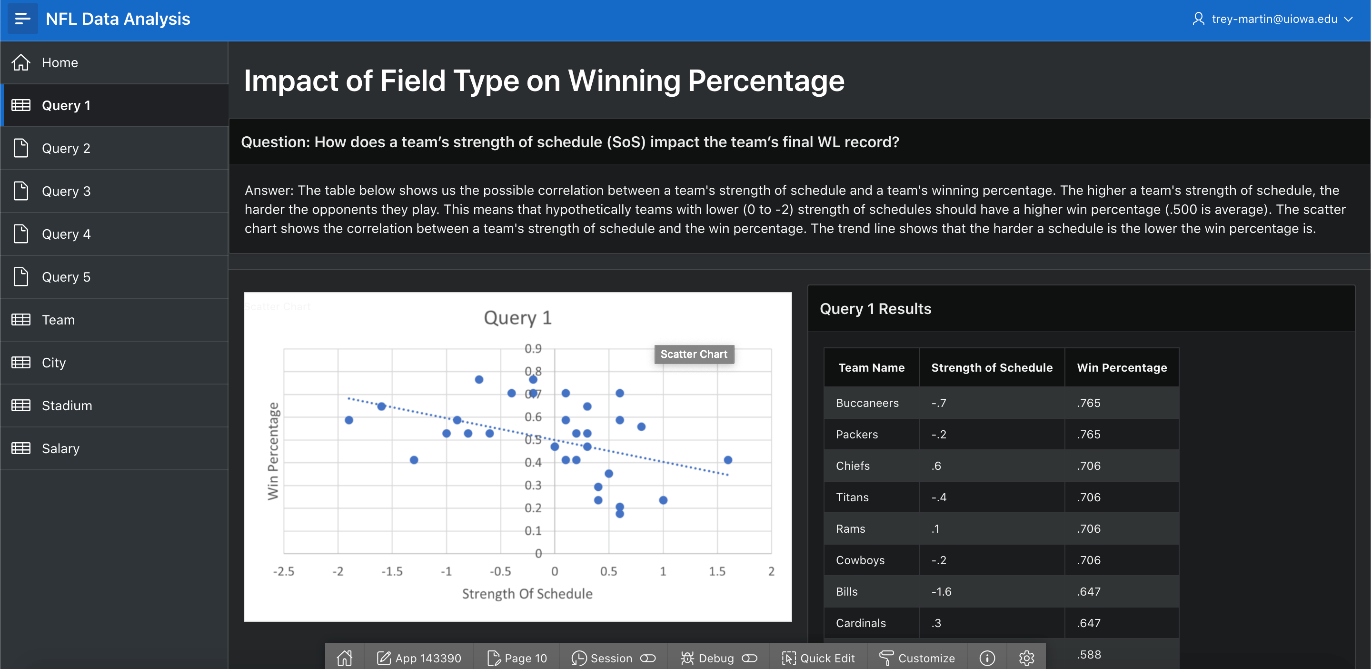
*Figure 9: TEAM*

*Figure 10: CITY*

*Figure 11: STADIUM*

*Figure 12: SALARY*

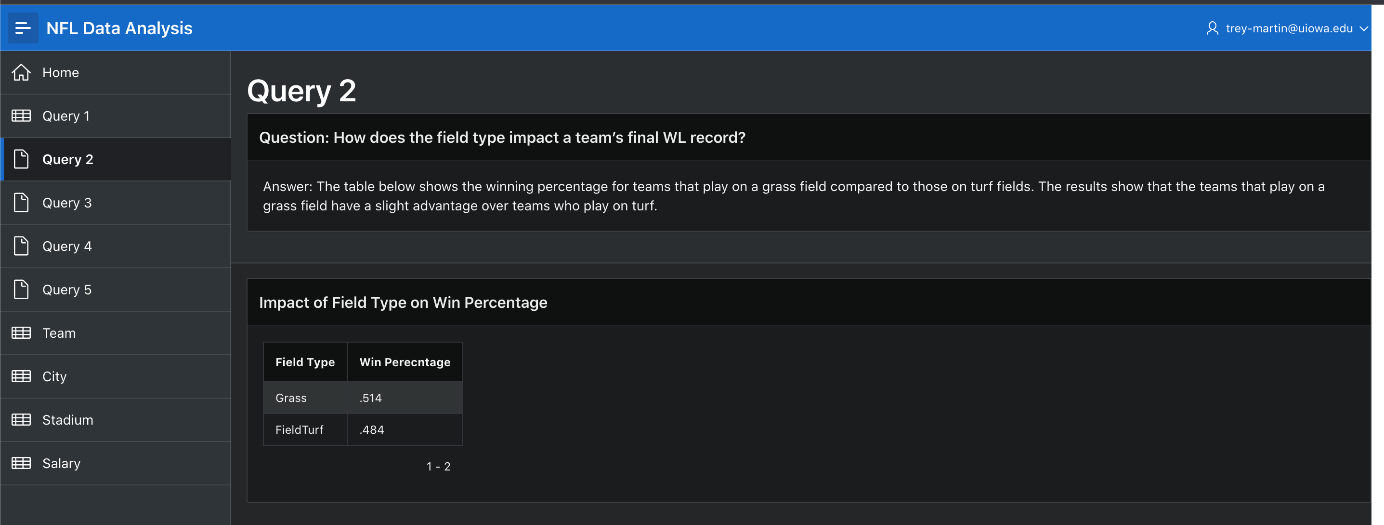
Queries

We presented the results of the first query as both a scatter chart and as a simple table (Figure 13). This query was focused on the strength of each team's schedule and how that relates to a team's overall win percentage. Included in the web application was the question as well the query results in the form of a table and a scatter plot. 

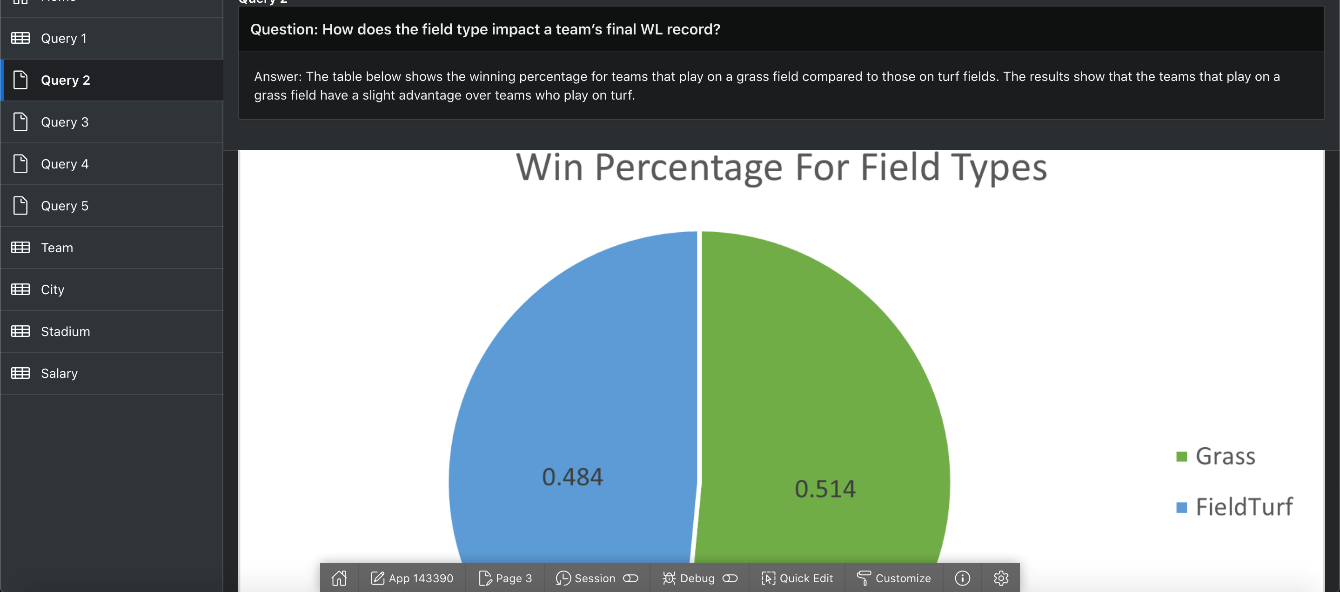
*Figure 13: Query 1 results*

*\*Please disregard the heading. We could not remove it.*

The next question we focused on and answered using a query was related to the field type that each team plays on and how that relates to their overall win percentage (Figures 14 and 15). Included in the web application is the question and the query results both in table form as a pie chart.

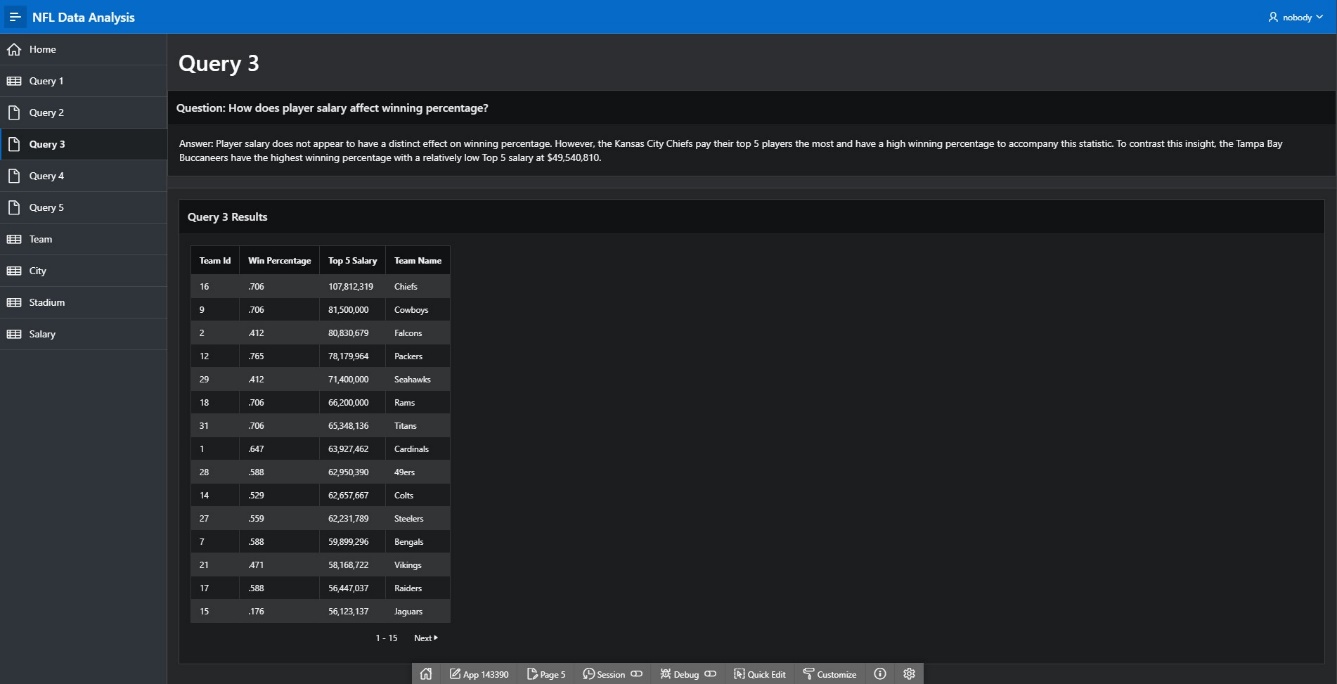


*Figure 14: Query 2 results part 1*



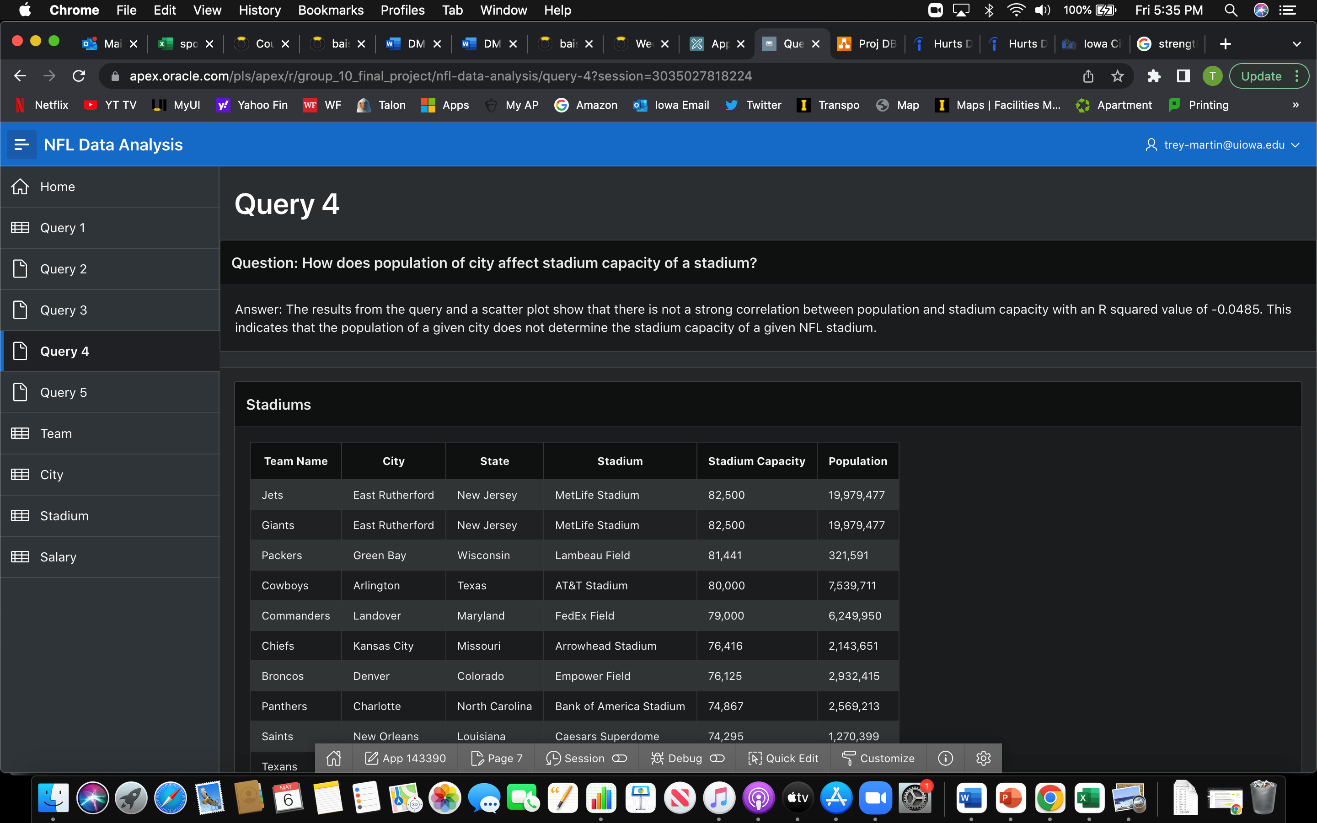
*Figure 15: Query 2 results part 2*

As a follow up question, we investigated whether the total amount for top 5 player salary has a relationship with the winning percentage of the team (Figure 16). Included in the web application is the question and the query results in table form.



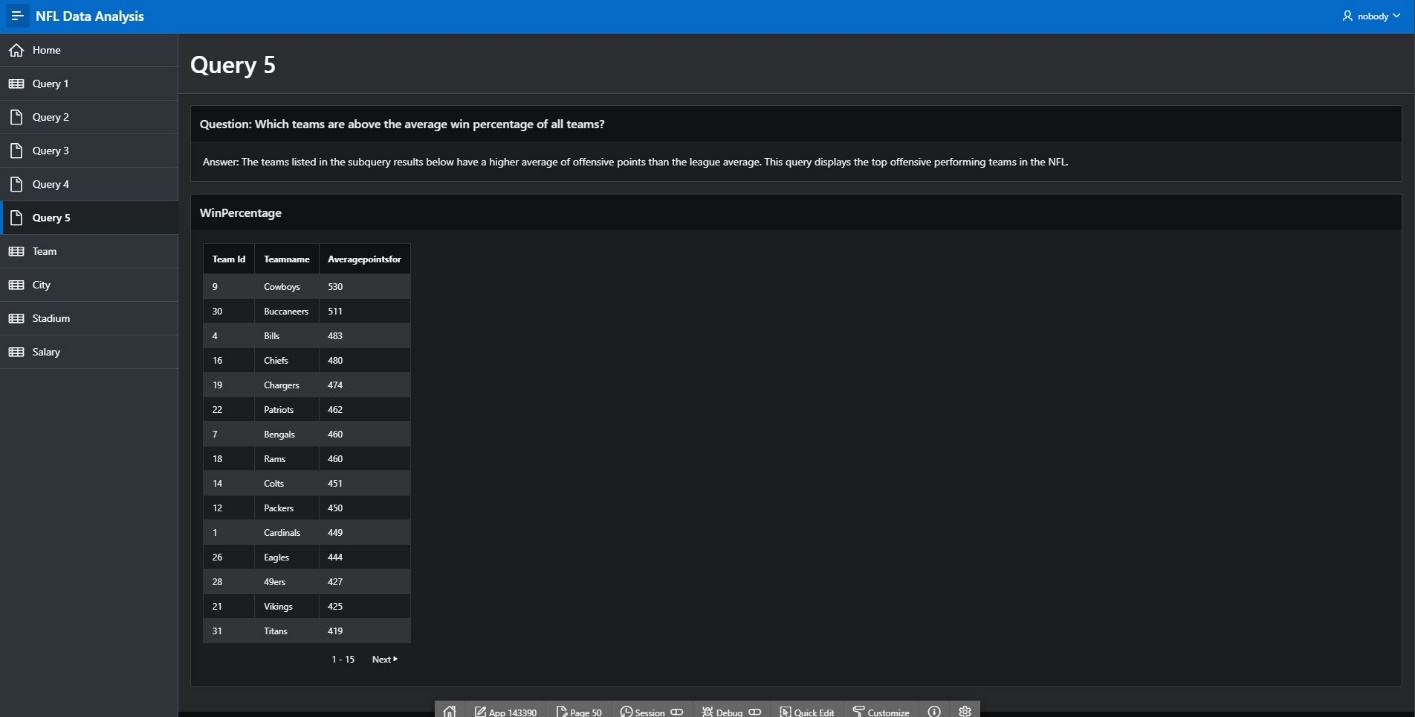
*Figure 16: Query 3 results*

The fourth query we ran covered the relationship between the population of a city and capacity of a stadium. Included in the web application are the question and query results in the form of a table.



*Figure 17: Query 4 results*

The fifth and final query we ran looked at which teams were more offensively driven than other teams in the league by writing a subquery that returned all teams that had a higher points for than the average across the league. (Figure 18). Included in the application are the question and the query results in table form.

*Figure 18: Query 5 results*