**BAIS:3200**

**Final Project Report**

**Tyler Eshkenazi, Paul Gawne, Michelle Hopkins, Brant O’Shea, Alec Schwager**

**Introduction**

Betting on college basketball has been a trend for quite some time. Analyzing each team's statistics and rankings comes into play when money is on the line [(https://www.kaggle.com/andrewsundberg/college-basketball-dataset](https://www.kaggle.com/andrewsundberg/college-basketball-dataset)). In this project, we will use various data analysis techniques to better understand how one team's statistics in a particular year affect their overall ranking and season performance. By examining trends and applying knowledge backed by statistical evidence, we can assist those interested in understanding college basketball performance and making educated decisions.

**Data**

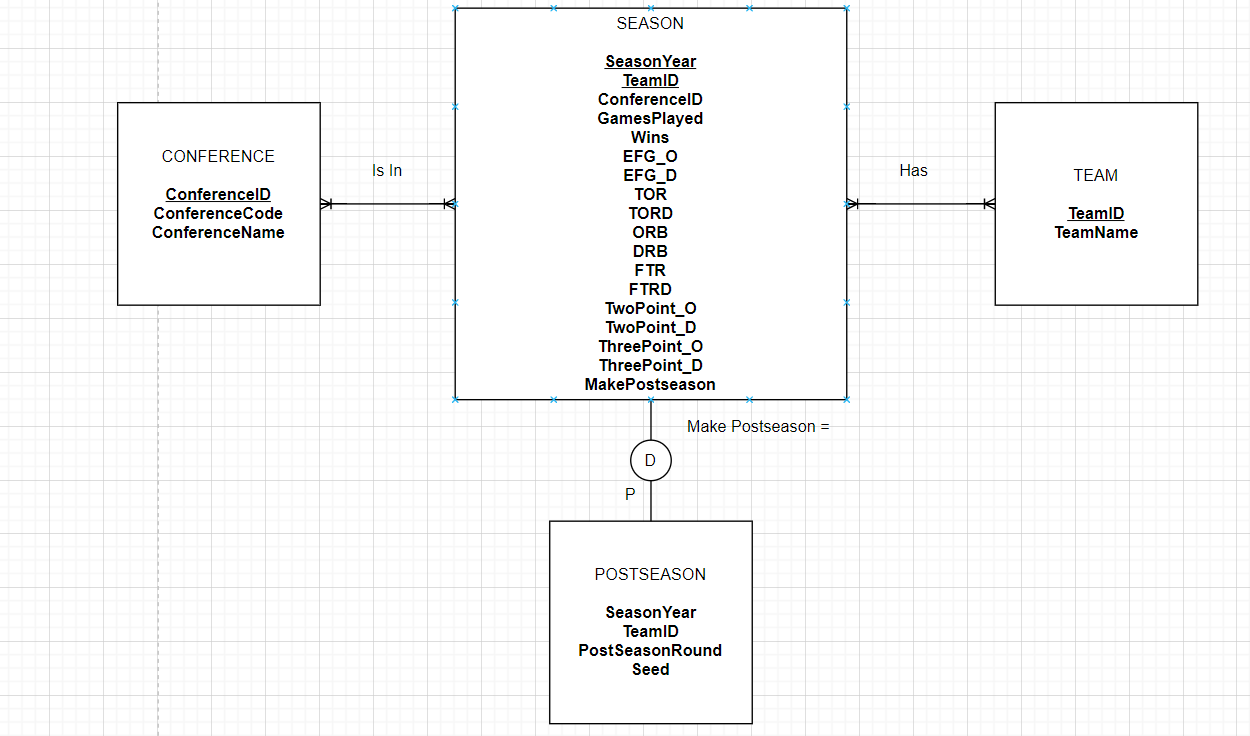
This project uses data gathered from 2015 to 2020 from a Kaggle database about college basketball statistics and tournament results (<https://www.kaggle.com/andrewsundberg/college-basketball-dataset>). For the sake of this project, we have condensed the dataset to include only those fields we feel are most pertinent to our analysis. These include fields describing individual teams, their conference affiliations, a wide range of offensive and defensive statistics sorted on a year-to-year basis, as well as the outcome of each individual team for any given year within the dataset. The fields representing the outcome of a season take the form of the teams’ records (i.e. wins and losses), their seed (a numerical ranking system with 1 being the highest value achievable), and the round they were eliminated in the postseason (if applicable).

*Table 1 Data Dictionary*

| **Field** | **Type** | **Description** |
| --- | --- | --- |
| TEAM | Text | The Division I college basketball school |
| CONF | Alphanumeric | The Athletic Conference in which the school participates in |
| G | Numeric | Number of games played |
| W | Numeric | Number of games won |
| EFG\_O | Numeric | Effective Field Goal Percentage Shot |
| EFG\_D | Numeric | Effective Field Goal Percentage Allowed |
| TOR | Numeric | Turnover Percentage Allowed (Turnover Rate) |
| TORD | Numeric | Turnover Percentage Committed (Steal Rate) |
| ORB | Numeric | Offensive Rebound Percentage |
| DRB | Numeric | Defensive Rebound Percentage |
| FTR | Numeric | Free Throw Rate (How often the given team shoots Free Throws) |
| FTRD | Numeric | Free Throw Rate Allowed |
| 2P\_O | Numeric | Two-Point Shooting Percentage |
| 2P\_D | Numeric | Two-Point Shooting Percentage Allowed |
| 3P\_O | Numeric | Three-Point Shooting Percentage |
| 3P\_D | Numeric | Three-Point Shooting Percentage Allowed |
| SEED | Numeric | Seed in the NCAA March Madness Tournament |
| YEAR | Numeric | Season |
| POSTSEASON | Alphanumeric | Round where the given team was eliminated or where their season ended |

**EERD Diagram**

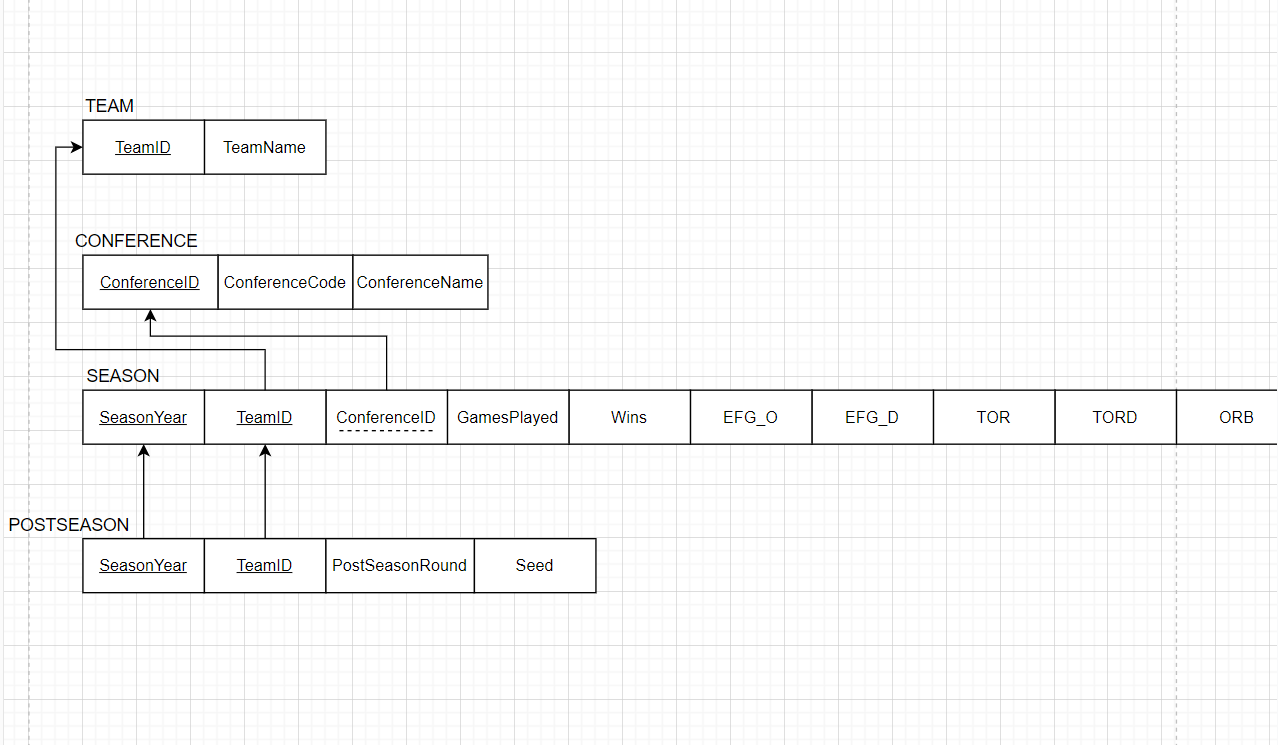
Figure 1 describes the two primary entities in our database as, Conference, which is identified by ConferenceID, and Team, which is identified by TeamID. All attributes are required in the Conference and Team tables. Both of these entities helped create the season entity. A conference is in one or many seasons, while a team also has one or many seasons. The attributes are all required within the season entity because it allows individuals to fully comprehend how successful one's season was. The season is also broken down into a subtype called Postseason. The Postseason table has partial specialization and is disjoint to the supertype.



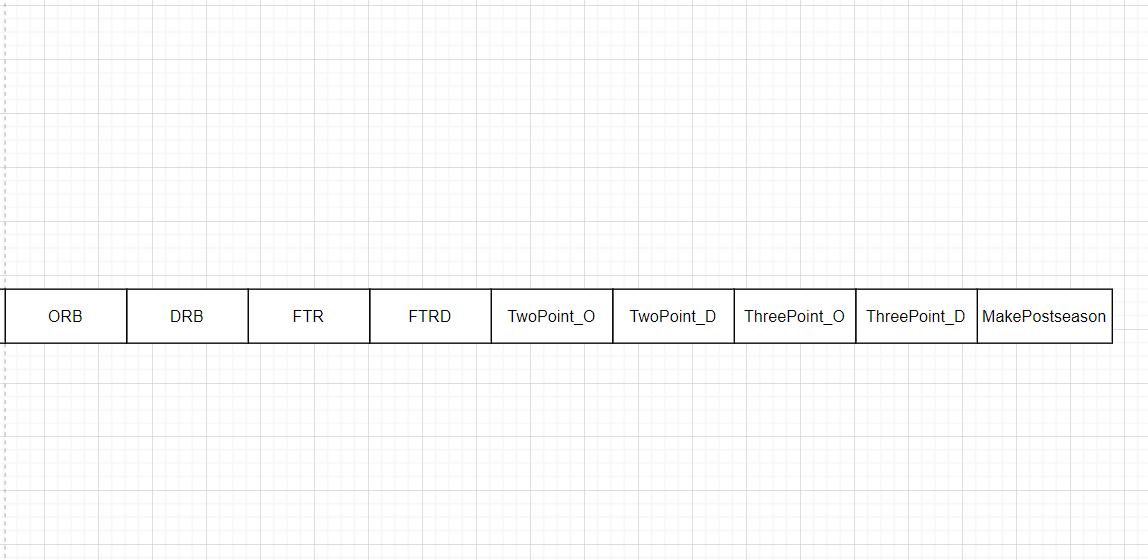
*Fig 1 Entity Relationship Diagram (ERD)*

**Normalized Relational Schema**

Based on the EERD above, we normalized the data into four different tables. Figure 2 describes the four tables, Team, Conference, Season, and Postseason and all of the attributes correlated with the tables. The Team and Conference table act as the parent table for the Season table. Then the Season table acts as the parent table for the Postseason table. Team and Conference have one primary key. Season and Postseason both have composite primary keys to help identify the results. Season also has a foreign key of ConferenceID within its table.



*Fig 2A Graphical Relational Schema*

*Fig 2B Graphical Relational Schema*

**Database Implementation**

In order to create the database within APEX, we wrote CREATE TABLE commands for all of the tables in our relational schema.

TEAM

*This table was created first as it serves as a parent table to SEASON*

CREATE TABLE TEAM(

TeamID CHAR(4) NOT NULL,

TeamName VARCHAR(30) NOT NULL,

CONSTRAINT TEAM\_PK PRIMARY KEY (TeamID));

INSERT INTO TEAM VALUES ('T101', 'University of Iowa');

CONFERENCE

*This table was created second as it also serves as a parent table to SEASON*

CREATE TABLE CONFERENCE(

ConferenceID CHAR(4) NOT NULL,

ConferenceCode VARCHAR(5) NOT NULL,

ConferenceName VARCHAR(35) NOT NULL,

CONSTRAINT CONFERENCE\_PK PRIMARY KEY (ConferenceID));

INSERT INTO CONFERENCE VALUES ('C011', 'B10', 'Big Ten');

SEASON

CREATE TABLE SEASON(

SeasonYear NUMBER NOT NULL,

TeamID CHAR(4) NOT NULL,

ConferenceID CHAR(4) NOT NULL,

GamesPlayed NUMBER NOT NULL,

Wins NUMBER NOT NULL,

EFG\_O NUMBER(4,2) NOT NULL,

EFG\_D NUMBER(4,2) NOT NULL,

TOR NUMBER(4,2) NOT NULL,

TORD NUMBER(4,2) NOT NULL,

ORB NUMBER(4,2) NOT NULL,

DRB NUMBER(4,2) NOT NULL,

FTR NUMBER(4,2) NOT NULL,

FTRD NUMBER(4,2) NOT NULL,

TwoPoint\_O NUMBER(4,2) NOT NULL,

TwoPoint\_D NUMBER(4,2) NOT NULL,

ThreePoint\_O NUMBER(4,2) NOT NULL,

ThreePoint\_D NUMBER(4,2) NOT NULL,

MakePostseason NUMBER(1,0) CHECK(MakePostseason IN(0,1)) NOT NULL,

CONSTRAINT SEASON\_PK PRIMARY KEY (SeasonYear, TeamID),

CONSTRAINT SEASON\_FK FOREIGN KEY (ConferenceID) REFERENCES CONFERENCE(ConferenceID));

INSERT INTO SEASON VALUES ('2019', 'T101', 'C011', '35', '20', '50', '45', '20', '25', '30', '33', '47', '20', '60', '37', '45', '32', '1');

POSTSEASON

CREATE TABLE POSTSEASON(

SeasonYear NUMBER NOT NULL,

TeamID CHAR(4) NOT NULL,

PostseasonRound VARCHAR(10) NOT NULL,

Seed NUMBER NOT NULL,

CONSTRAINT POSTSEASON\_PK PRIMARY KEY (SeasonYear, TeamID),

CONSTRAINT POSTSEAON\_FK FOREIGN KEY (SeasonYear, TeamID) REFERENCES SEASON (SeasonYear, TeamID));

INSERT INTO POSTSEASON VALUES ('2019', 'T101', 'F4', '10');

**Analysis**

The analysis of this data is intended to utilize trends and apply knowledge backed by statistical evidence to assist those who want to better understand or predict the performance of college basketball teams.

Question 1: Free Throw Rate Average

Which teams made the postseason with a below-average FTR for each year? For this question, we wrote a query returning the year, team name, and their free throw rate for the given year. We nested a subquery in the WHERE clause that returns the average free throw rate across all seasons to compare individual teams’ rates with the average.

SELECT SeasonYear, TeamName, FTR

FROM SEASON JOIN TEAM

ON SEASON.TeamID = TEAM.TeamID

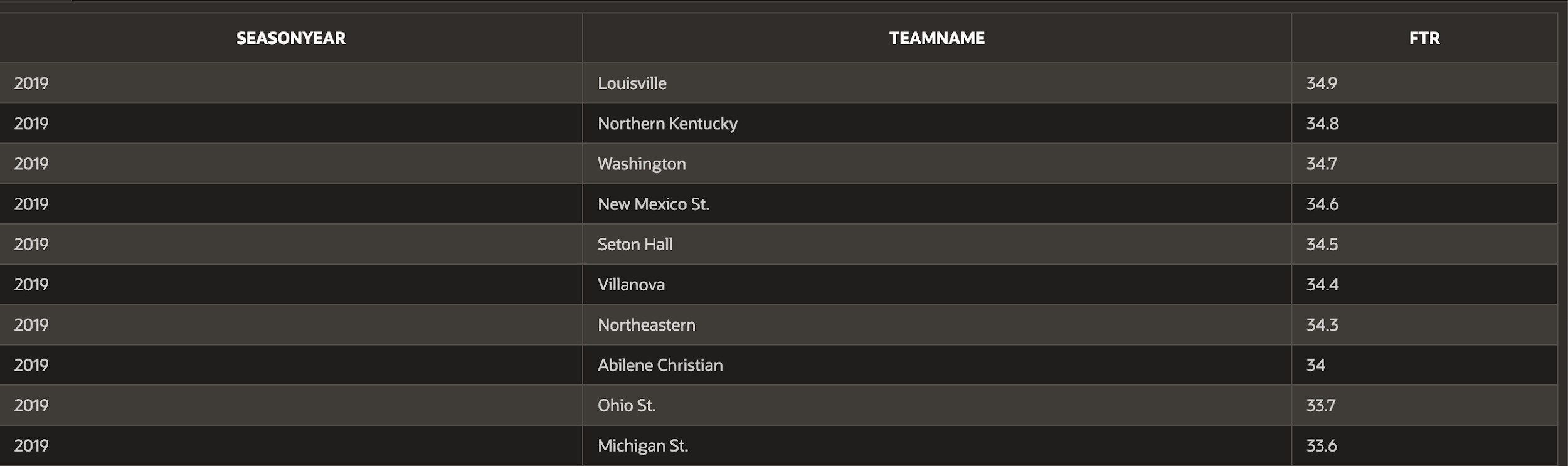
WHERE MakePostSeason = '1' and FTR <

(select AVG(ftr)

from season)

ORDER BY SEASON.SeasonYear desc, FTR desc;

The results of this query are shown below (Figure 3). Based on the results, it appears that having a low free throw rating does not necessarily mean a team is less likely to make the postseason. For example, in 2019, 46 teams (roughly 67 percent of all teams in the postseason) made the postseason while having a below-average free throw rating.



*Figure 3 Made postseason with below-average FTR*

Question 2: Three-Point Percentages

What was the three-point shooting percentage for teams with a win percentage of at least 80%? For this question, we wrote a query that returns each team each year with this three-point percentage and their calculated win percentage.

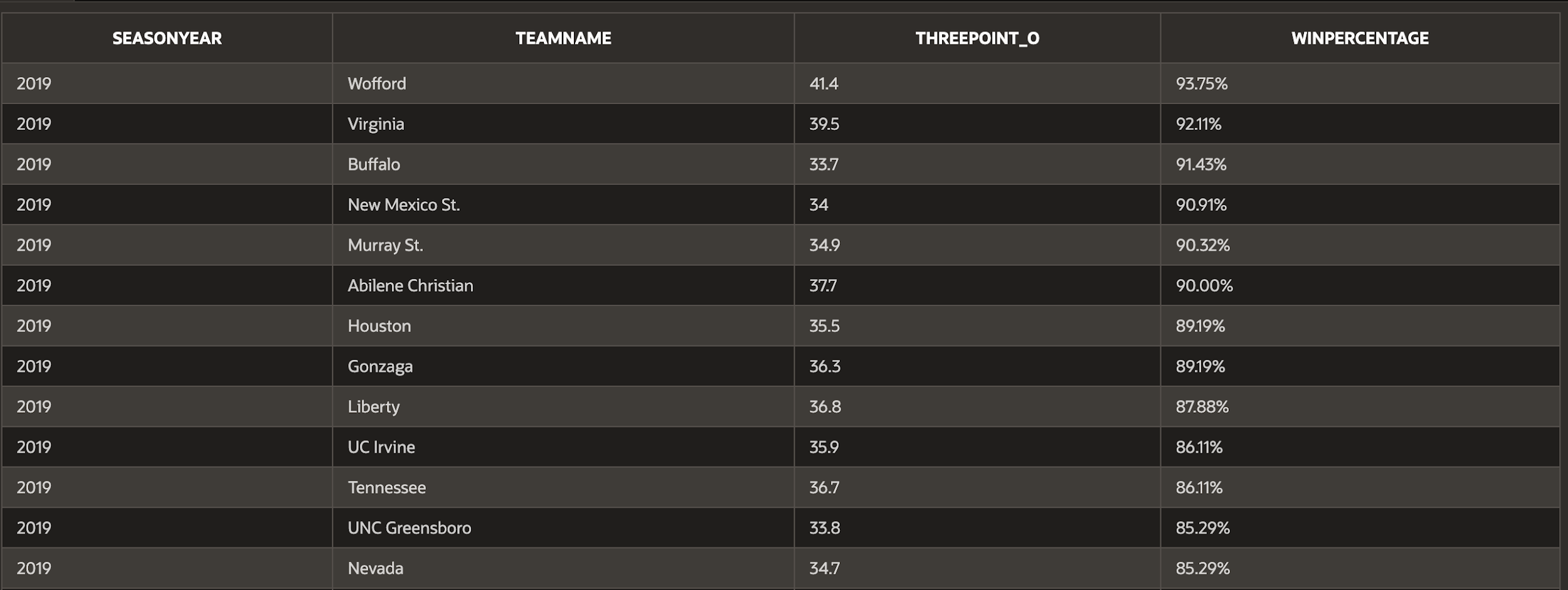
SELECT SeasonYear, TeamName, ThreePoint\_O, TO\_CHAR(ROUND((Wins/GamesPlayed)\*100, 2), '99.99') || '%' AS WinPercentage

FROM SEASON JOIN TEAM ON SEASON.TeamID = TEAM.TeamID

WHERE (Wins/GamesPlayed)\*100 >= 80.00

ORDER BY SeasonYear DESC, WinPercentage DESC;

The results of this query are shown below (Figure 4). We calculated each team’s win percentage by dividing the games won by the games played and displayed it as a percentage. Furthermore, we used a WHERE clause to only show teams whose win percentage was above 80%. Based on the current data, a team’s three-point percentage does not appear to be correlated to the team’s win percentage, even though 3 points is the highest a team can score on a single shot made.



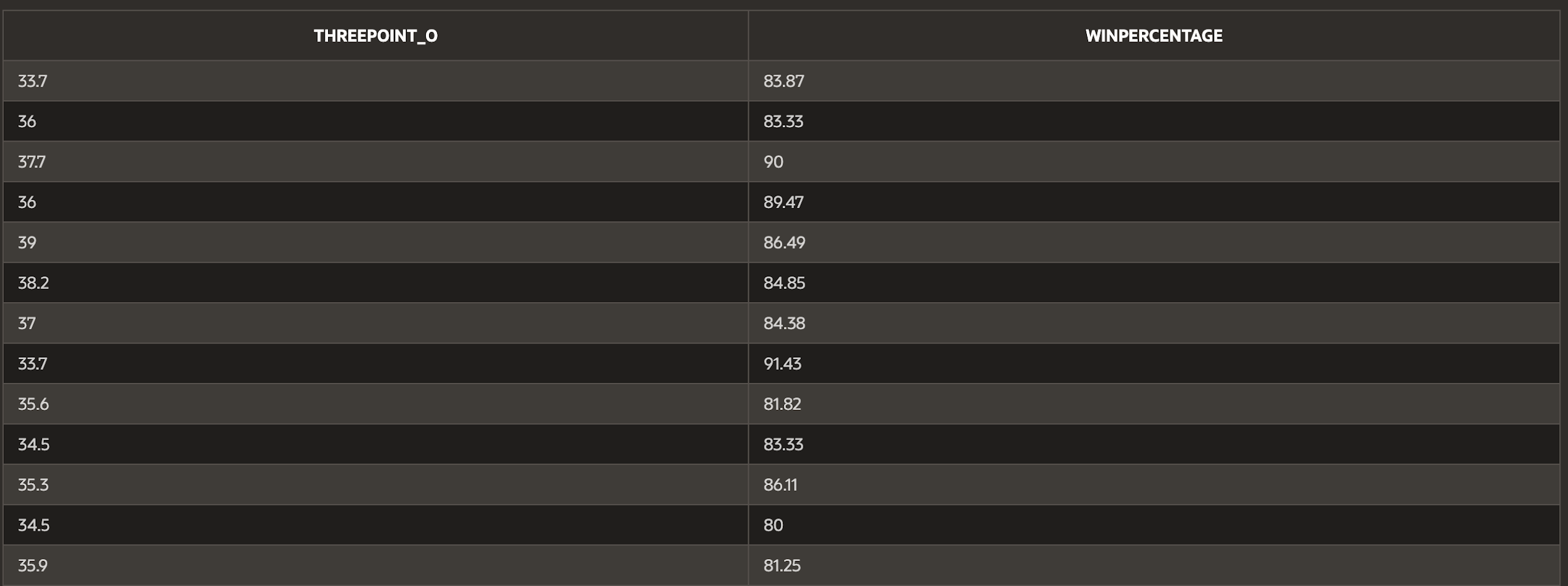
*Figure 4 Three-point shooting for each team each year*

Is there a correlation between three-point shooting percentage and win percentage for team’s with a win percentage over 80%? To answer this question, we wrote a simple query that returned each team's three-point shooting percentage and their win percentage.

SELECT ThreePoint\_O, ROUND((Wins/GamesPlayed)\*100) AS WinPercentage

FROM SEASON

WHERE (Wins/GamesPlayed)\*100 >= 80.00;

The results of this query are shown below (Figure 5). Given each team’s three-point shooting percentage and their win percentage, there does not appear to be a correlation. When considering a team’s three-point shooting percentage, there is no way of knowing what their win percentage is. 

*Figure 5 Three-point shooting percentage and win percentage*

*This query is also presented as a scatter chart in the web application* (Figure 17)

Question 3: Team Grade

Based on the grading scale below, when examining a team's win percentage, how many teams fall into each grade category for the 2019 season? To answer this question, we created a grading scale and calculated each team’s win percentage, and assigned a grade using a CASE query.

Grading Scale: A: 100%-90%; B: 90%-80%, C: 80%-70%, D: 70%-50%, F: <50%.

SELECT CASE

WHEN (Wins/GamesPlayed)\*100 >= 0.00 AND (Wins/GamesPlayed)\*100 < 50.00 THEN 'F'

WHEN (Wins/GamesPlayed)\*100 >= 50.00 AND (Wins/GamesPlayed)\*100 < 70.00 THEN 'D'

WHEN (Wins/GamesPlayed)\*100 >= 70.00 AND (Wins/GamesPlayed)\*100 < 80.00 THEN 'C'

WHEN (Wins/GamesPlayed)\*100 >= 80.00 AND (Wins/GamesPlayed)\*100 < 90.00 THEN 'B'

ELSE 'A'

END AS Grade, COUNT(TeamID) AS TOTAL

FROM SEASON

WHERE SeasonYear = '2019'

GROUP BY CASE

WHEN (Wins/GamesPlayed)\*100 >= 0.00 AND (Wins/GamesPlayed)\*100 < 50.00 THEN 'F'

WHEN (Wins/GamesPlayed)\*100 >= 50.00 AND (Wins/GamesPlayed)\*100 < 70.00 THEN 'D'

WHEN (Wins/GamesPlayed)\*100 >= 70.00 AND (Wins/GamesPlayed)\*100 < 80.00 THEN 'C'

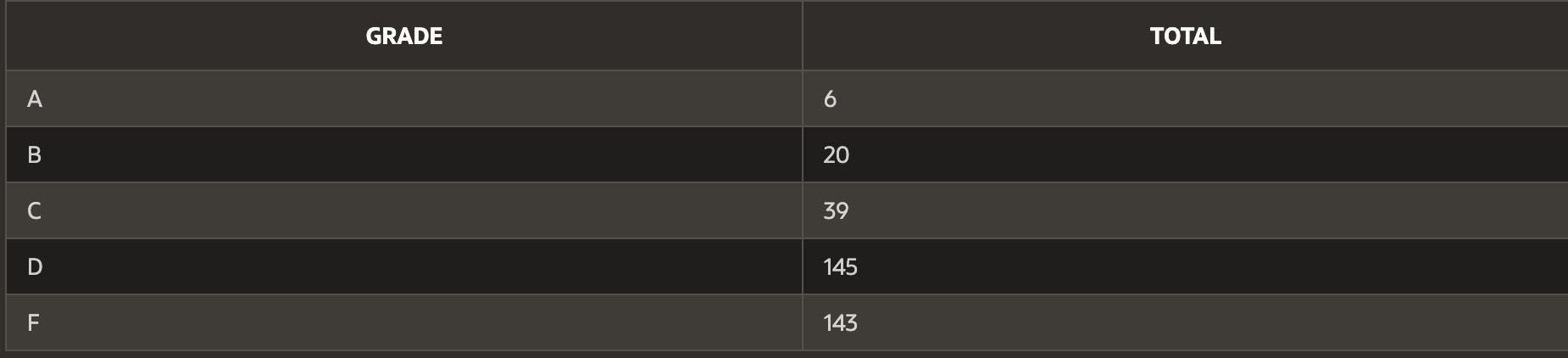
WHEN (Wins/GamesPlayed)\*100 >= 80.00 AND (Wins/GamesPlayed)\*100 < 90.00 THEN 'B'

ELSE 'A'

END

ORDER BY Grade ASC;

The results of this query are shown below (Figure 6). Grades “A” and “B” have the smallest number of teams which is expected because not every team can have a winning record. Based on the data sample, there is a much greater chance of a team having a losing or close to losing record and not making the postseason and attempting to be crowned champions. Further analysis could be done to determine the teams who made the postseason with grade “A” or “B”.



*Figure 6 Team grade*

Question 4: 2019 Conference Size

How many teams were in each conference in the year 2019? Which conference had the greatest amount of teams? To answer this question, we created a query that counted how many teams were in each specific conference, only in the year 2019.

SELECT ConferenceName, COUNT(SEASON.ConferenceID) AS TeamsInConference\_2019

FROM CONFERENCE JOIN SEASON

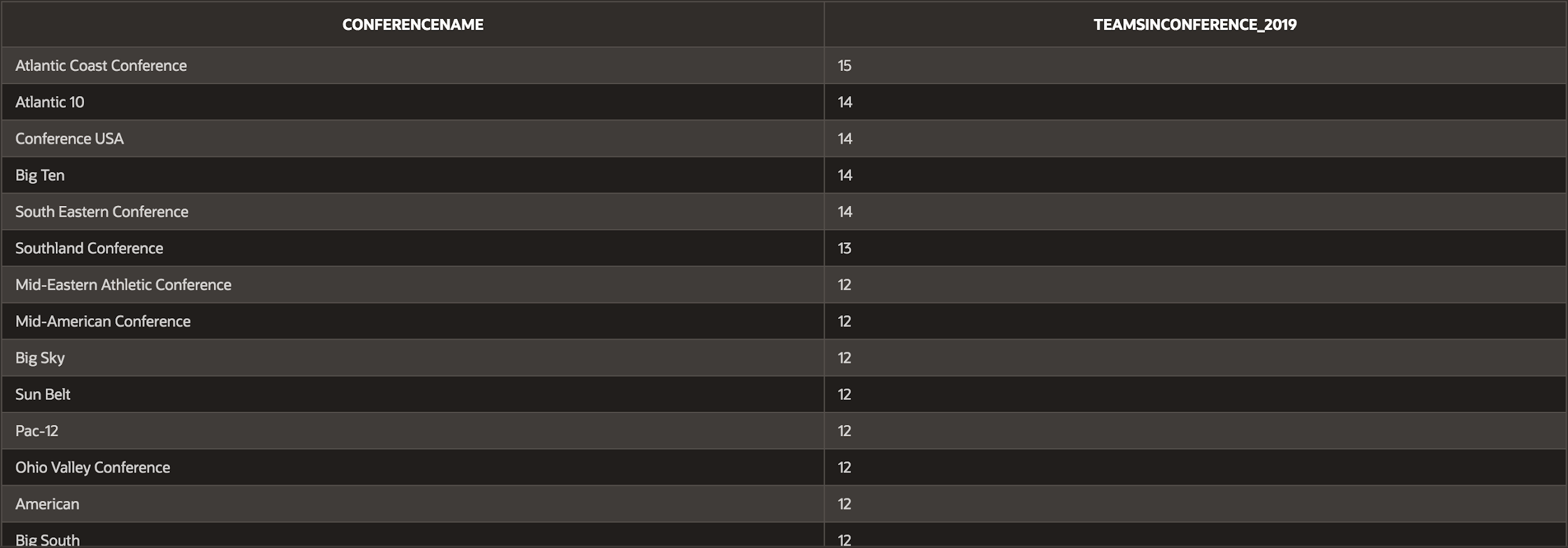
ON CONFERENCE.ConferenceID = SEASON.ConferenceID

WHERE SeasonYear = 2019

GROUP BY ConferenceName

ORDER BY TeamsInConference\_2019 DESC;

The results of this query are shown below (Figure 7). Based on the current data, the largest conference that we found is the ACC at 15 teams, followed by Atlantic 10, Conference USA, Big 10, and SEC at 14 teams. The smallest conferences are Summit League, Ivy League, and ASUN at 8 teams each. This information is useful because the teams in each conference are changing relatively often. This query also gives context for other questions that we answer. For example, when determining how many teams a conference had in the postseason, it is useful to know how many teams are in each conference to understand the chances of having more or fewer teams make the postseason.



*Figure 7 Conference size*

Question 5: Conference Appearance in Postseason

Given there are a limited number of openings for the postseason, which conference has the most teams in the postseason per year, across all years? To answer this question we created a join query with a count function to count the number of teams in each conference while checking to make sure the team had actually made the postseason. We also made sure to exclude conferences with 2 teams or less who made the postseason as we are looking for the most appearances.

SELECT ConferenceName, SeasonYear, COUNT(TeamID) AS TotalTeams

FROM CONFERENCE JOIN SEASON

ON CONFERENCE.ConferenceID = SEASON.ConferenceID

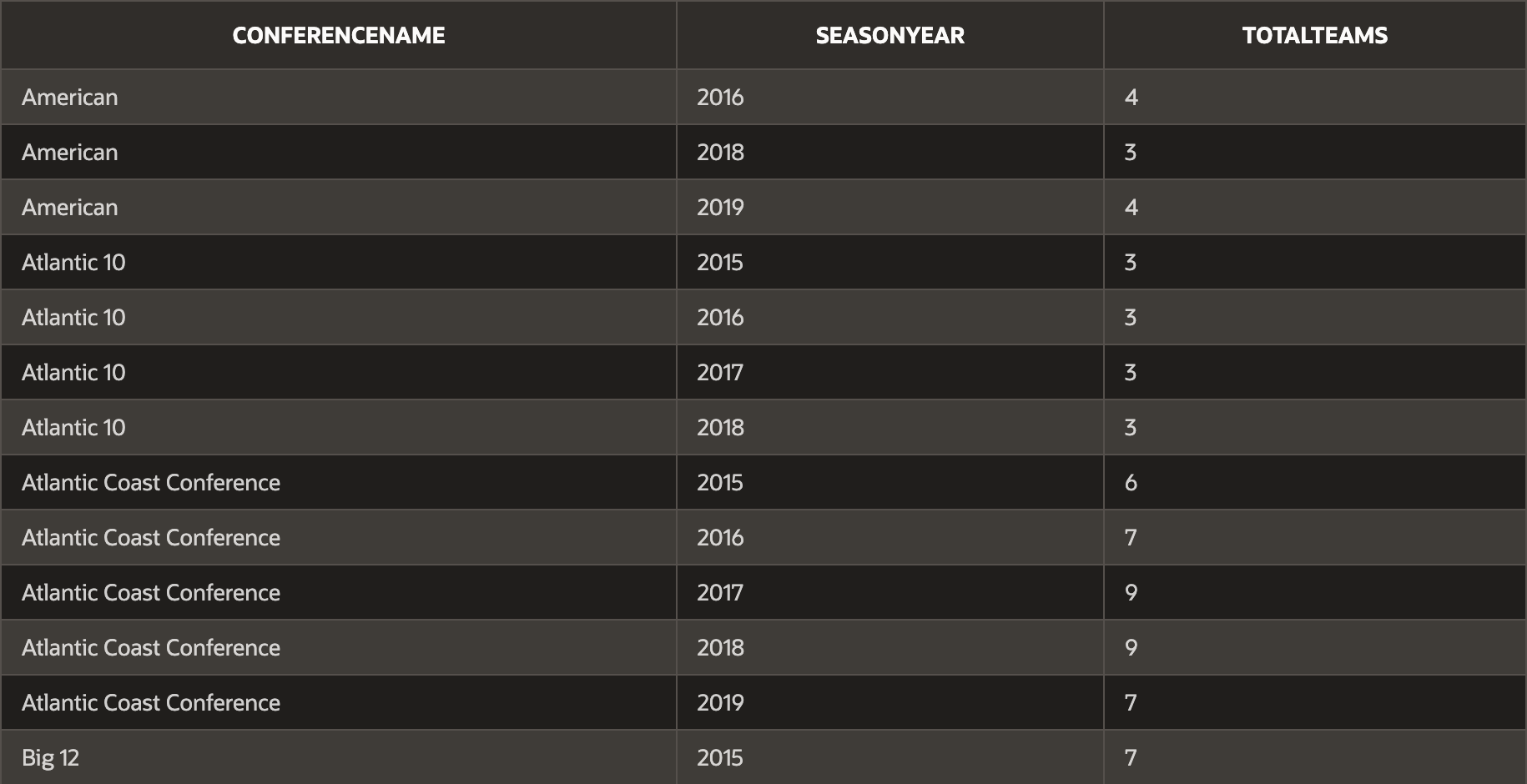
WHERE MakePostseason = '1'

GROUP BY SeasonYear, ConferenceName

HAVING COUNT(TeamID) > 2

ORDER BY ConferenceName ASC, SeasonYear ASC, TotalTeams DESC;

The results of this query are shown below (Figure 8). Across all years the top 3 conferences with the most teams that made the postseason include the Atlantic Coast, Big 10, and Big 12 Conferences. The Atlantic Coast Conference had the highest number of teams in the postseason with 9 in both 2017 and 2018. These results are supported by the fact that the Power 5 conferences (Atlantic Coast, Big Ten, Big 12, Pac-12, and Southeastern) have the most championships across all conferences. (<https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_men%27s_basketball_champions#Champions_by_conference_status:~:text=2018-,Champions,-by%20conference%20status>) Furthermore, the number of teams that make the postseason per conference is generally consistent across the five years, except for the Mountain West Conference which only had 3 teams make the postseason in 2015.



*Figure 8 Conference appearances in postseason*

Question 6: Postseason Appearances

How many times did the team appear in the postseason between 2015-2019? Also, which teams had the most appearances? In order to answer these questions, we created a join query that returned each college team and the number of times they played in the postseason ranked in descending order.

SELECT TeamName, COUNT(POSTSEASON.TeamID) AS PostseasonAppearances

FROM TEAM JOIN POSTSEASON

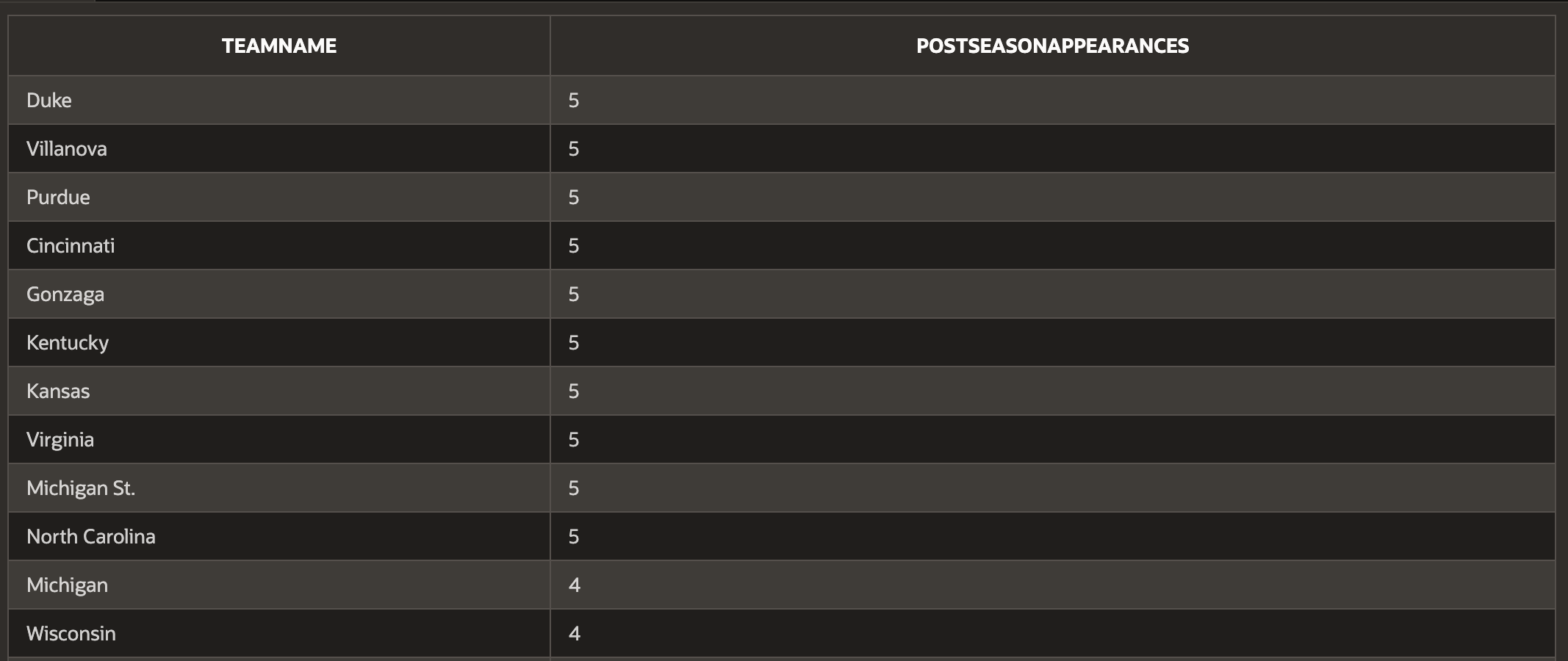
ON TEAM.TeamID = POSTSEASON.TeamID

GROUP BY TeamName

ORDER BY PostseasonAppearances DESC

FETCH FIRST 15 ROWS ONLY;

The results of this query are shown below (Figure 9). From 2015-2019, there were 10 teams within the top 15 who made the postseason each year, and 5 teams within the top 15 that missed the postseason only one time across that span. These results are helpful to understand which teams have been high performers consistently over the past five seasons. For further analysis, we could consider these teams, and evaluate their individual performances within the postseason, to determine whether they are consistent championship contenders, or often appear as early-round exits.

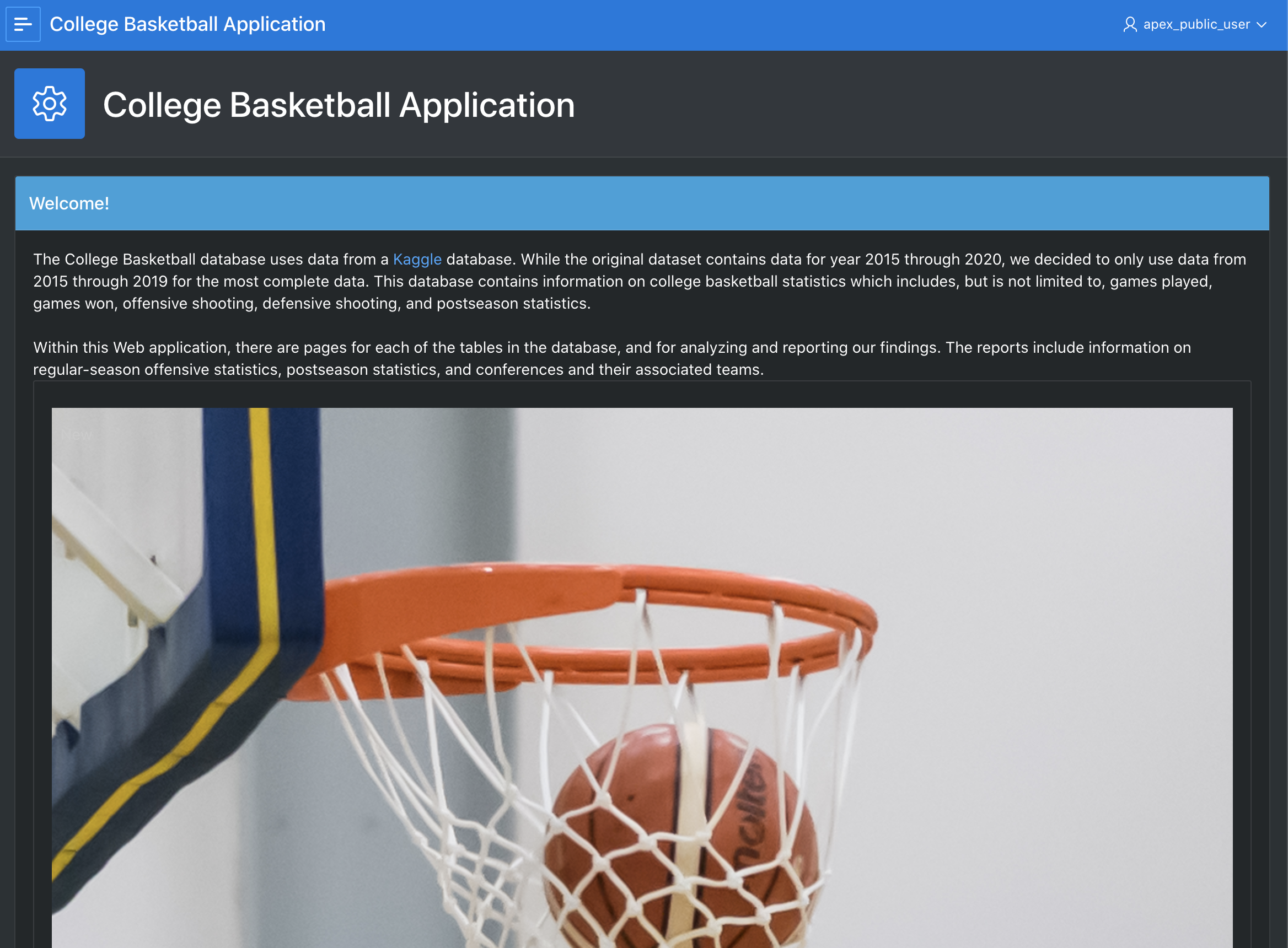
*Figure 9 Postseason appearances*

**Web Application**

[https://apex.oracle.com/pls/apex/teshkenazi-personal/r/college-basketball-application/home](https://apex.oracle.com/pls/apex/teshkenazi-personal/r/college-basketball-application/home?session=707153646099254)

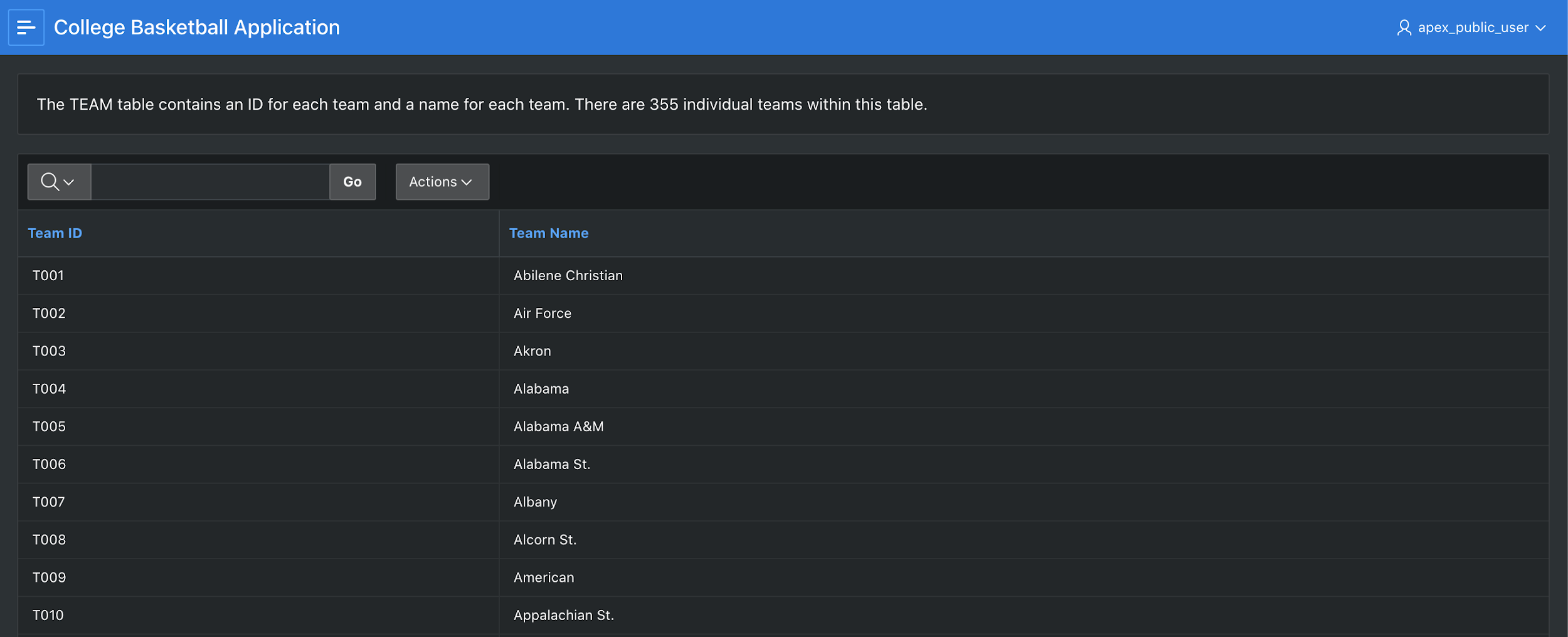
Home Page

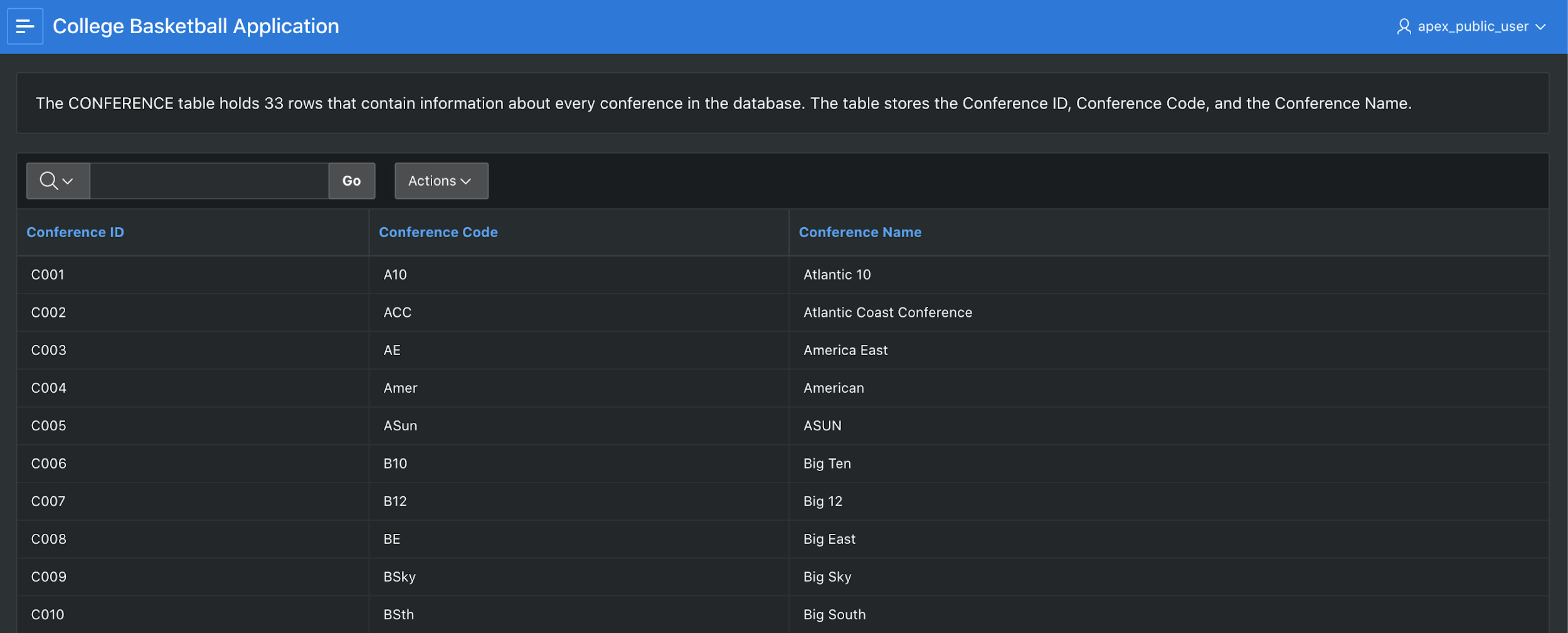
The home page of the web application has a short introduction and description of the application and provides a hyperlink to the original dataset. Under the description we added a picture, with a creative commons license, of a basketball to keep up with our theme. There is also a nested navigation menu and we used the same accent colors to keep users attention. Figure 10 shows a screenshot of the home page.

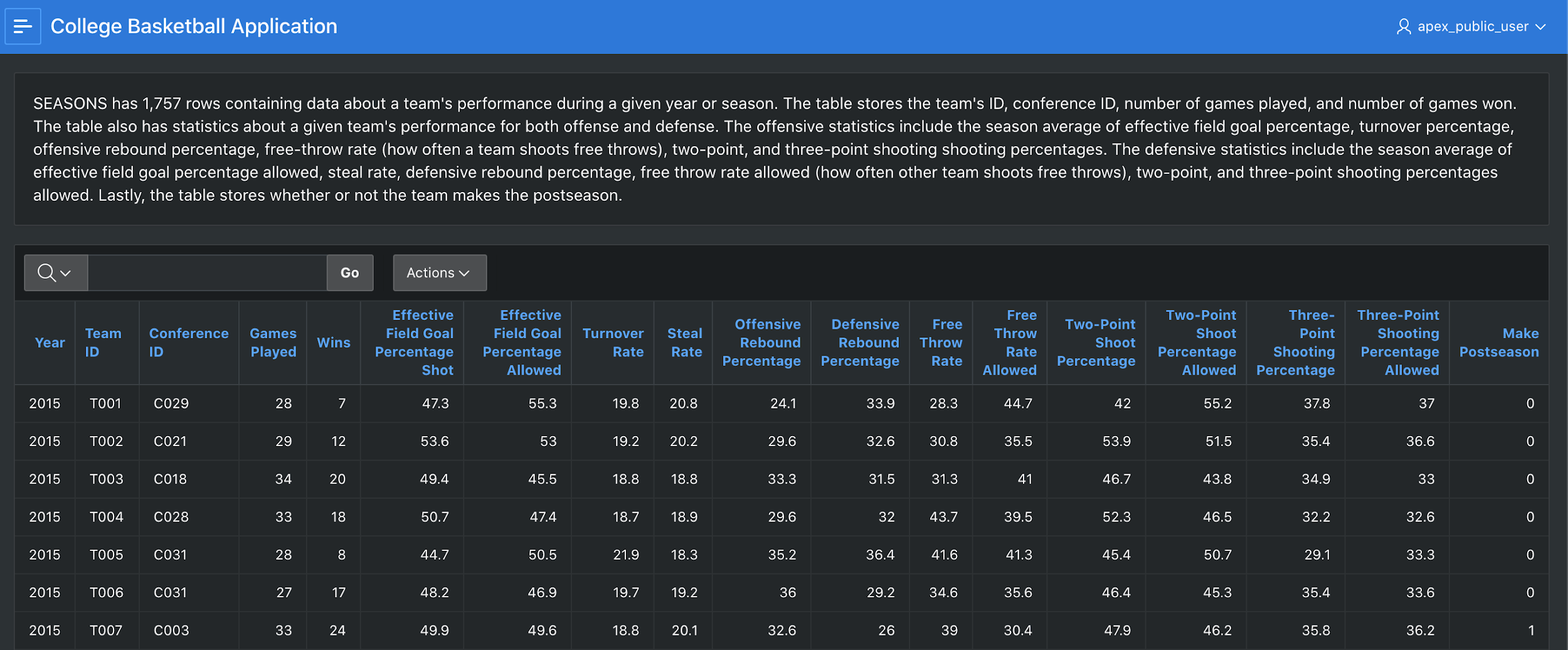
*Figure 10 Home page*

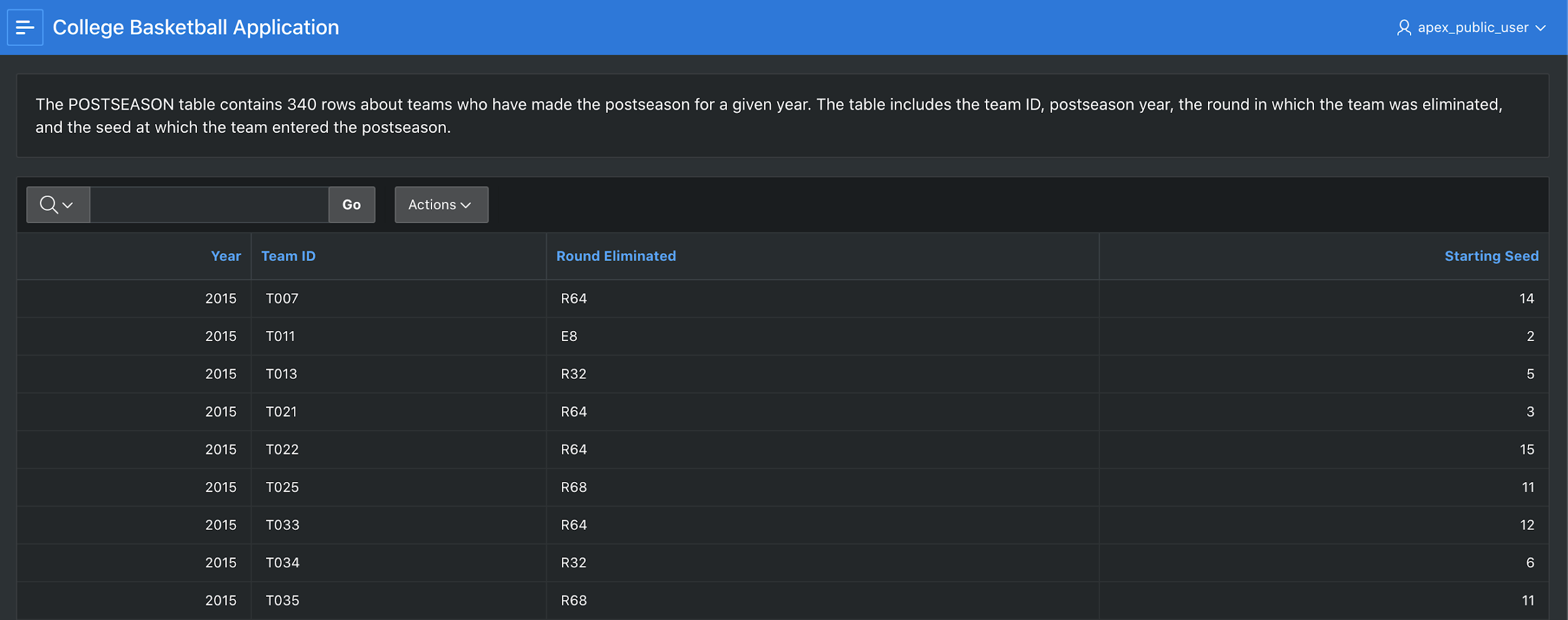
Tables

Our group created interactive reports for each of the four tables in our database (Figures 11-14). Presenting these tables as interactive reports allows users to search, filter, and group data. At the top of each page, we included a small text box that gives some detail on the data in the table. Lastly, we renamed column headers to clarify their meaning.

*Figure 11 TEAMS*

*Figure 12 CONFERENCES*

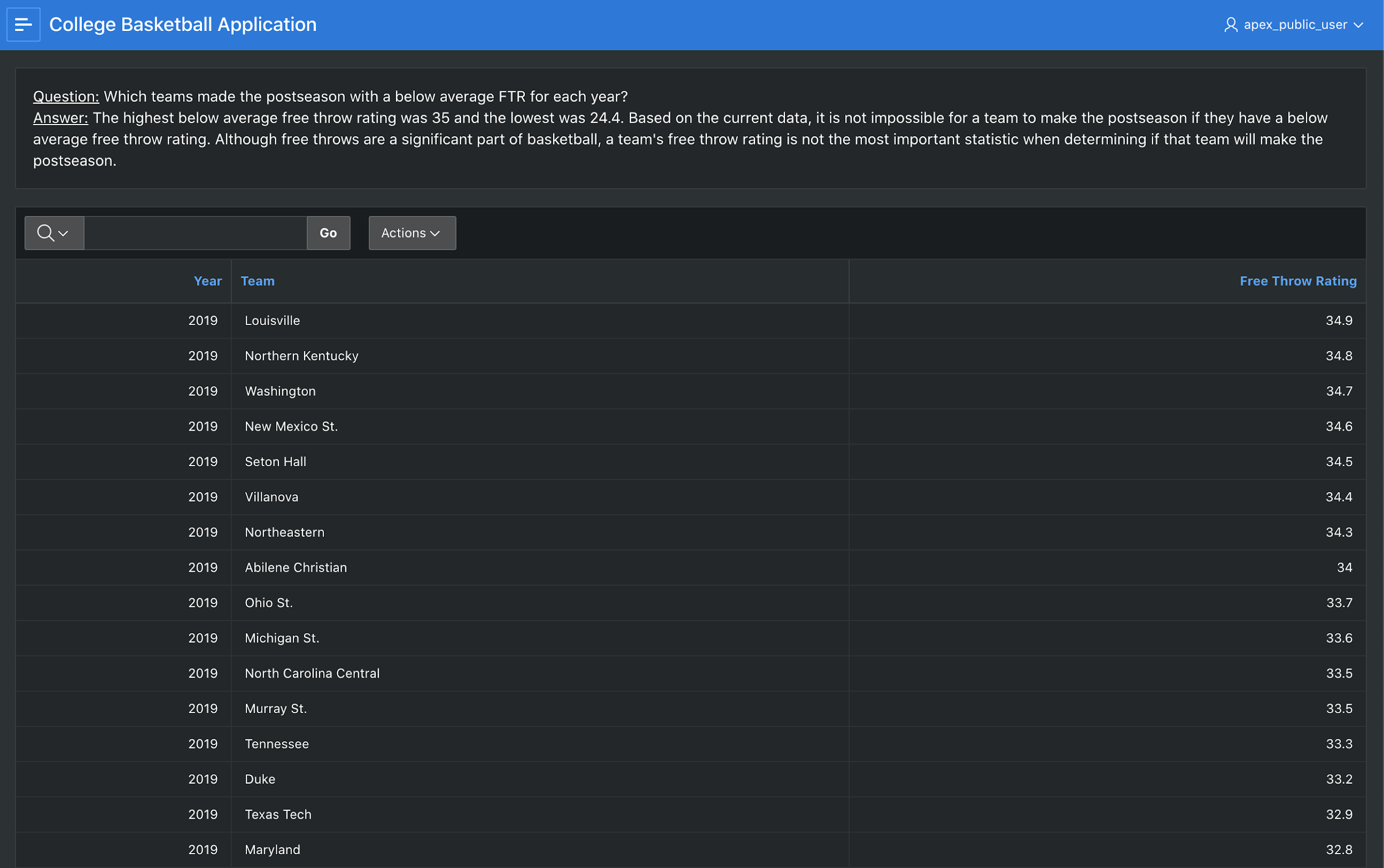
*Figure 13 SEASON*

*Figure 14 POSTSEASON*

Queries

Query 1

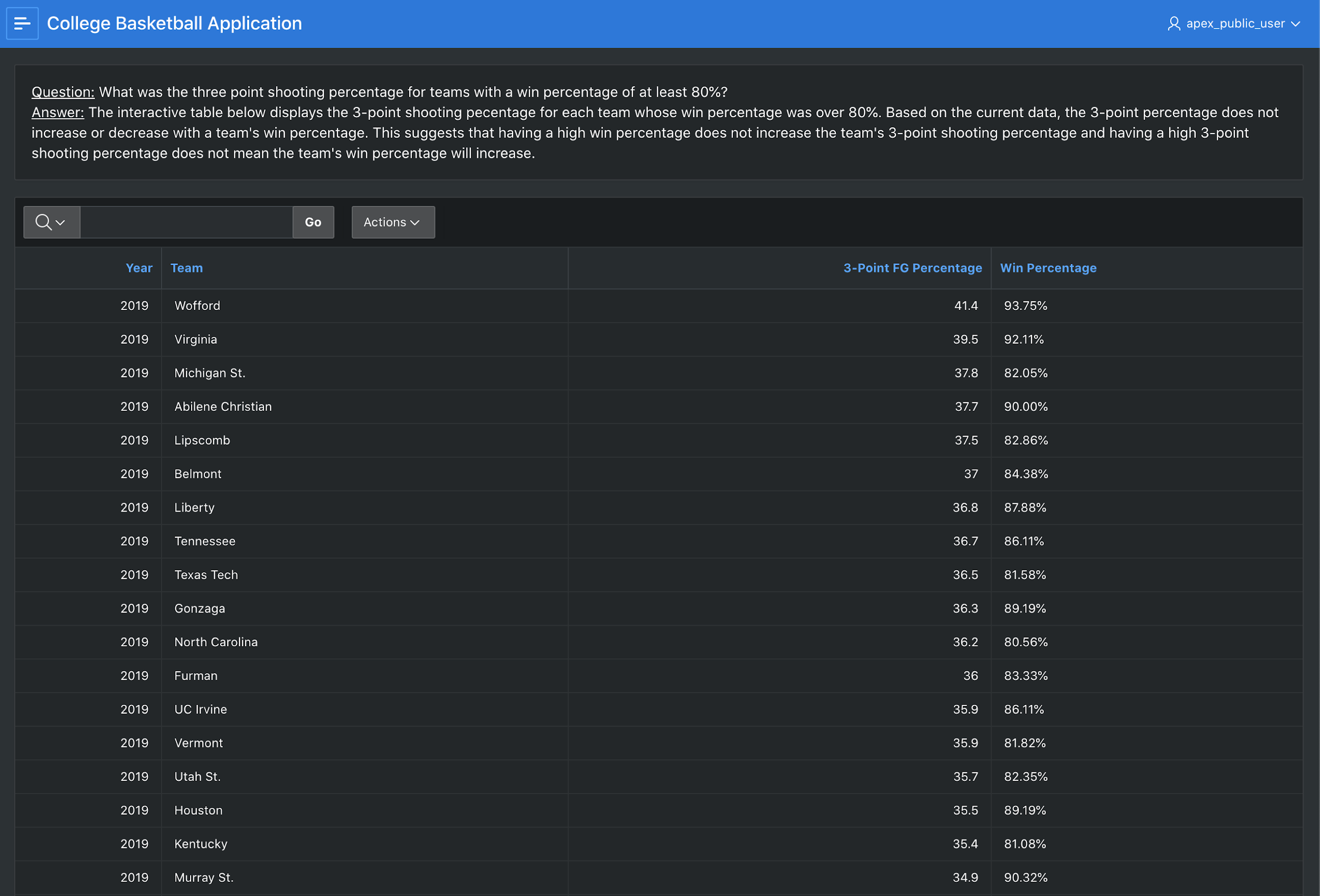
For the first question’s results we used an interactive report as the query returned more than 25 rows (Figure 15). Column headers were changed for clarity purposes. This table is helpful in understanding that teams can still make the postseason without having a free throw rating that is above average.



*Figure 15 Teams with below average free throw rating in the postseason*

Query 2

The results of the second query are shown as an interactive table and column headers were changed for clarity (Figure 16). This table presents the year, team, and the respective three point and win percentages.



*Figure 16 Three point shooting percentage and win percentage by team and year*

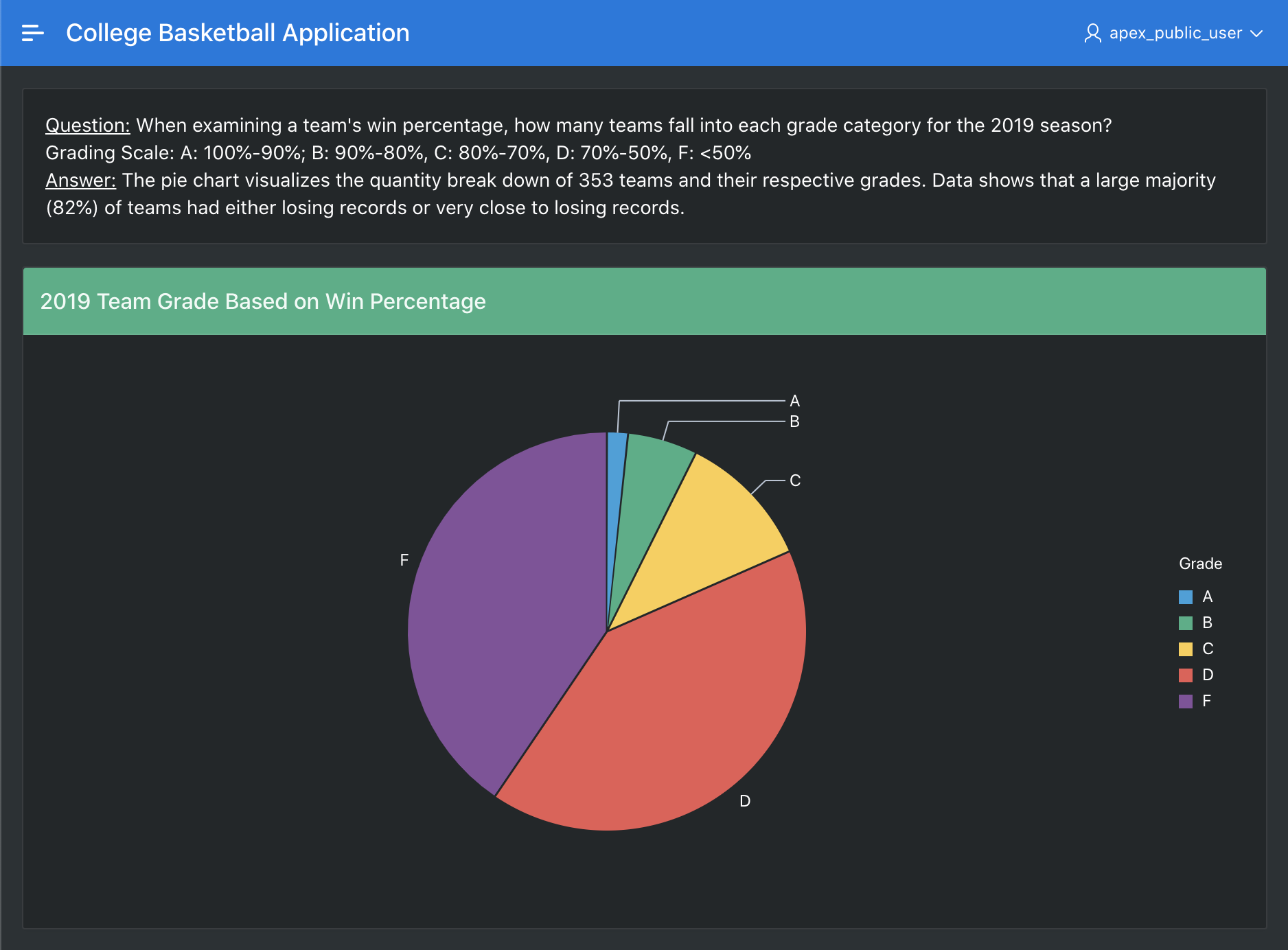
As a follow-up to the second question, we wrote a simple query for the three point shooting and win percentages, these results were presented as a scatter chart (Figure 17). The scatter chart shows that the team's win percentages ranged widely irrespective of the team’s three point shooting percentage.



*Figure 17 Three point shooting percentage vs win percentage scatter chart*

Query 3

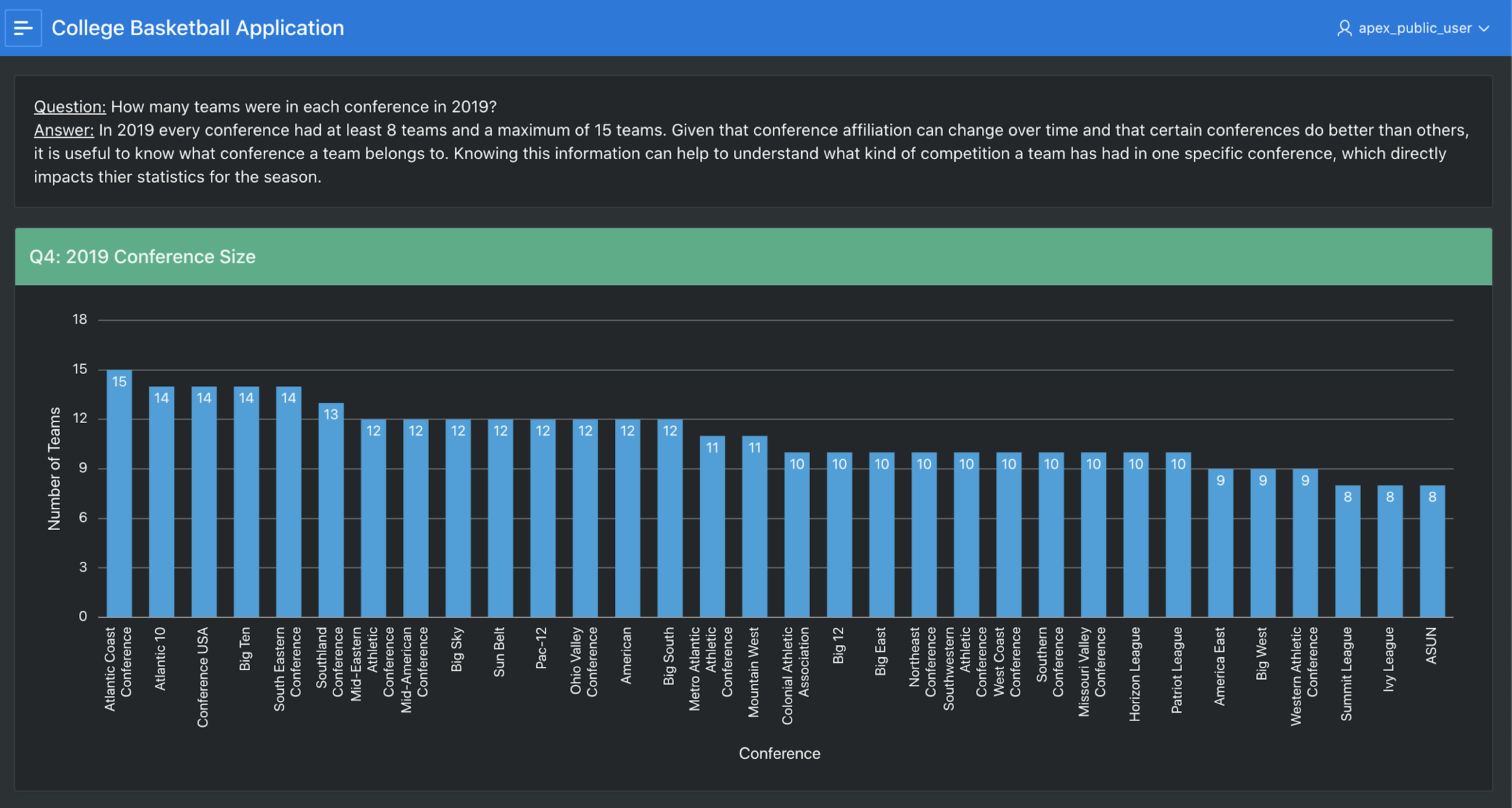
For the results of the third question which assigned each team a grade, we used a pie chart to display our results. Each slice represents the number of teams with that grade designation (Figure 18).



*Figure 18 Team grade designation*

Query 4

For our fourth query’s results, the number of teams in each conference, we decided to represent the data with bar charts (Figure 19). We found this representation useful for visualizing the amount of teams and the size difference between each conference. By seeing the conferences side by side, one can tell with a quick scan, how one conference compares in size to another.



*Figure 19 Teams per conference 2019*

Query 5

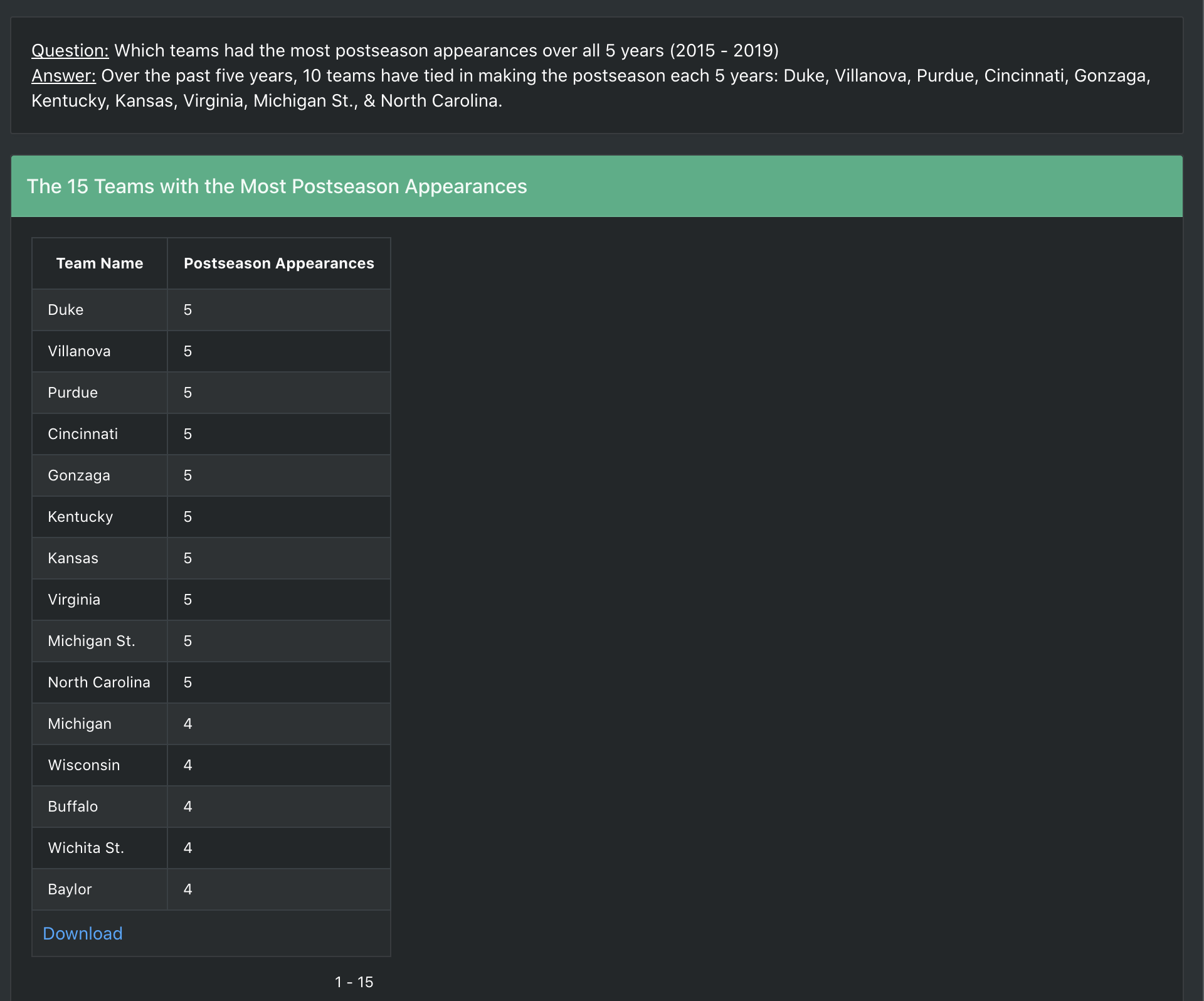
For our fifth question, which breaks down the teams who made the postseason by conference, we displayed our results as a clustered or grouped bar chart (Figure 20). By utilizing a clustered bar chart, we were easily able to see which conferences had the most teams in the postseason each year. In addition, we were able to analyze how the number of teams that made the postseason changed by conference.



*Figure 20 Number of postseason teams by conference*

Query 6

The last question we found the 15 teams with the most postseason appearances throughout the 5 years. These results are displayed as a classic table (Figure 21).



*Figure 21 Top 15 teams with the most postseason appearances*