Final Project: SQL Queries

DS 230

Introduction

This project will give you the opportunity to practice more complex queries than we have seen in class so far. Working through these query statements is probably the best way to learn the SQL language. Learning what it can and cannot do, learning how to group certain things, but not others. While this is a “project”, it is not a group project - this is an **individual** assignment. Discussion with peers is encouraged, but by no means should one copy files, share code, or otherwise provide animate assistance.

This assignment will explore the NCAA basketball data set on Google BigQuery. This data set is fairly robust, with quite a large amount of helpful documentation. I strongly recommend you spend some time reading over it to familiarize yourself with the different tables and their attributes. Just a quick note: the historical data is split up into regular season games and tournament games, and “pbp” stands for “play-by-play” and represents the most fine grained information about the game.

Submission

Instead of submitting your queries in raw text, you need to submit them as a Python script. This enables your SQL commands to be run directly, and the results compared to mine. This is to help me account for the **many different** possible queries that all generate the exact same results. Please use the example python script included on the Moodle page and *do not* change the function names. The edited .py file is what should be finally uploaded for your completed project.

Here is an example of what your python script might look like for a single query:

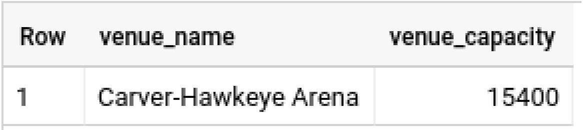
```
def problem_1():
    my_SQL_command = """
       SELECT P.Product
       FROM Table P
       WHERE P.price > 10.00
    """
    return my_SQL_command
```

Questions

Write the SQL query that would answer the following questions. None of these queries should take longer than 10 seconds, with most happening in less than 2. All of these queries (and your much experimentation!) is very likely to even exceed the free 1TB monthly limit for BigQuery.

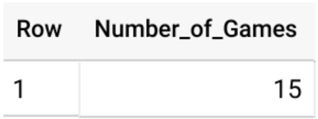
As requested, the resulting table answers have been included for you to double check your query statements. Good luck!

1. (10 points) What is the name and capacity of Iowa's NCAA basketball team venue?



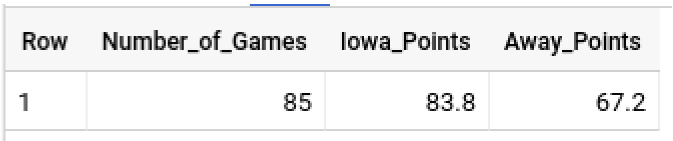
Row	venue_name	venue_capacity
1	Carver-Hawkeye Arena	15400

2. (10 points) How many games were played in Iowa's venue in the 2017 season?



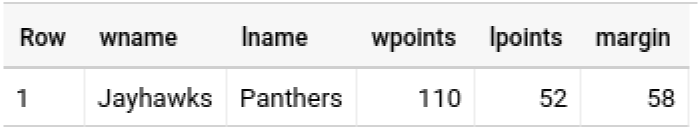
Row	Number_of_Games
1	15

3. (10 points) How many home games has Iowa won in the 2013 to 2017 seasons (inclusive)? Include as your answer the number of games won, the average score for Iowa, and the average score for the opponent in those games. Round your averages to 1 decimal place. Yes, SQL has a round function.



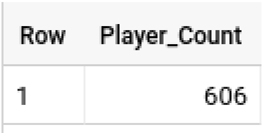
Row	Number_of_Games	Iowa_Points	Away_Points
1	85	83.8	67.2

4. (20 points) What is the largest margin of victory recorded for the historical tournament data? Include the name of the winning and losing team, their individual scores, and the win margin for that game.



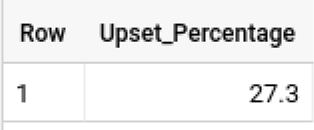
Row	wname	lname	wpoints	lpoints	margin
1	Jayhawks	Panthers	110	52	58

5. (20 points) How many players have played for a team based in the *same city* they were born in? Only include the birth city and state in your conditions.



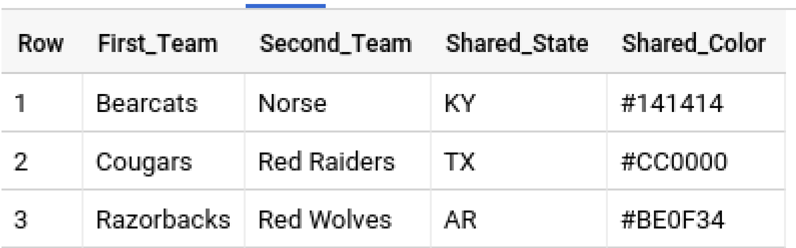
Row	Player_Count
1	606

6. (30 points) An **upset** occurs whenever a lower ranked team (which corresponds to a *higher* seed) beats a strictly higher ranked team (with a *lower* seed). What percentage of historical tournament games are upsets? Round to 1 decimal place.



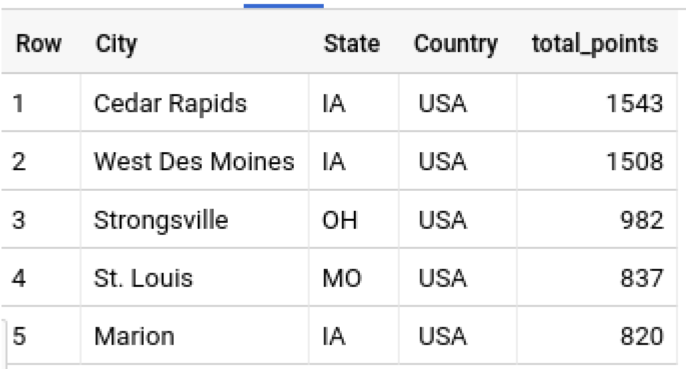
Row	Upset_Percentage
1	27.3

7. (30 points) Which pairs of teams have **both** the same state and team color? Your answer should list the name of both teams, their shared state, and their shared color. Because there are multiple teams that satisfy this condition, put them in alphabetical order (for each pair) and then alphabetical order for the rows.



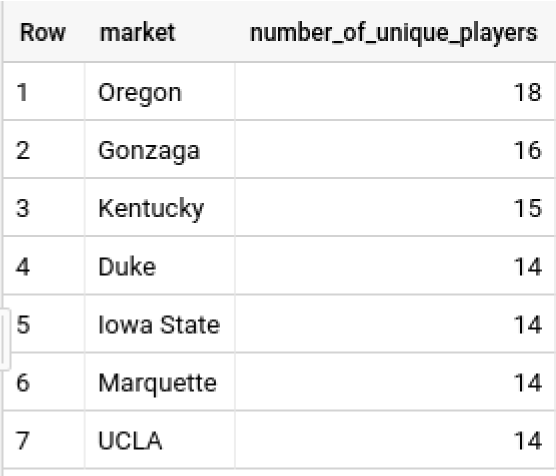
Row	First_Team	Second_Team	Shared_State	Shared_Color
1	Bearcats	Norse	KY	#141414
2	Cougars	Red Raiders	TX	#CC0000
3	Razorbacks	Red Wolves	AR	#BE0F34

8. (30 points) What are the top 5 scoring “locations” for the University of Iowa starting from their 2013 season? Here, a location is the city, state, and country of their *player*.



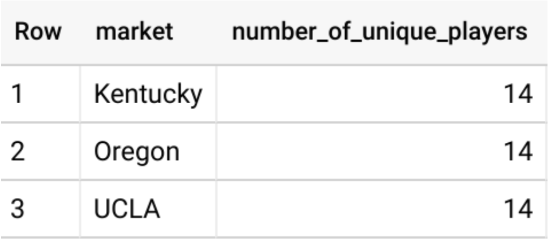
Row	City	State	Country	total_points
1	Cedar Rapids	IA	USA	1543
2	West Des Moines	IA	USA	1508
3	Strongsville	OH	USA	982
4	St. Louis	MO	USA	837
5	Marion	IA	USA	820

9. (40 points) Since the start of the 2013 season, which teams have had 14 or more players score 15 or more points in the *first half* of a game? Your answer should display the team market and the number of unique players they had satisfying the above condition. Order by the number of unique players, then by team market.



Row	market	number_of_unique_players
1	Oregon	18
2	Gonzaga	16
3	Kentucky	15
4	Duke	14
5	Iowa State	14
6	Marquette	14
7	UCLA	14

or



Row	market	number_of_unique_players
1	Kentucky	14
2	Oregon	14
3	UCLA	14