BIG DATA MANAGEMENT

TRIMISTER – 3

A POST GADUATION DIPLOMA

IN DATA ENGINEERING

ASSIGNMENT - 1

SUBMITTED BY

NIRAJ BHAGCHANDANI [G23AI2087]

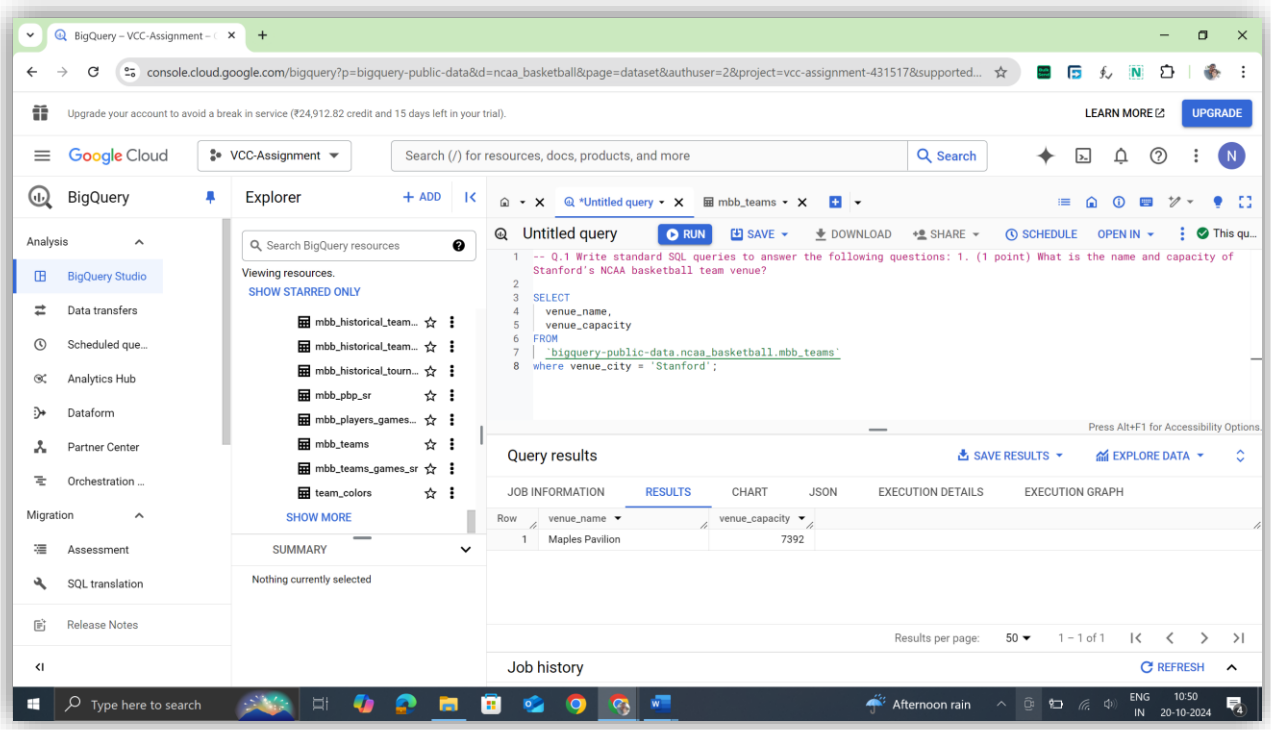
SUBMISSION DATE: *20TH OCTOBER, 2024*

DEPARTMENT OF ARTIFICIAL INTELLIGENCE DATA ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY - JODHPUR

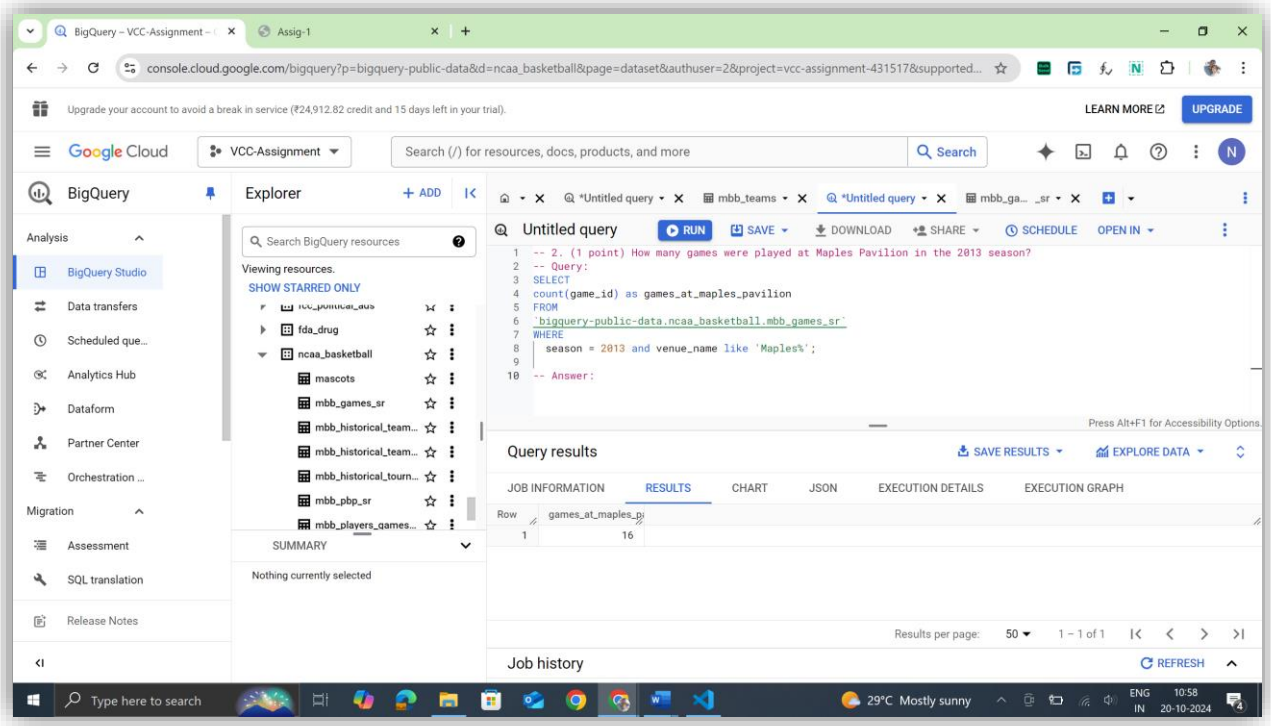
1. ***Write standard SQL queries to answer the following questions: 1. (1 point) What is the name and capacity of Stanford's NCAA basketball team venue?***

```sql
SELECT
  venue_name,
  venue_capacity
FROM
  `bigquery-public-data.ncaa_basketball.mbb_teams`
where venue_city = 'Stanford';
```



2. (1 point) How many games were played at Maples Pavilion in the 2013 season? Query: Answer:

```
SELECT
count(game_id) as games_at_maples_pavilion
FROM
`bigquery-public-data.ncaa_basketball.mbb_games_sr`
WHERE
  season = 2013 and venue_name like 'Maples%';
```

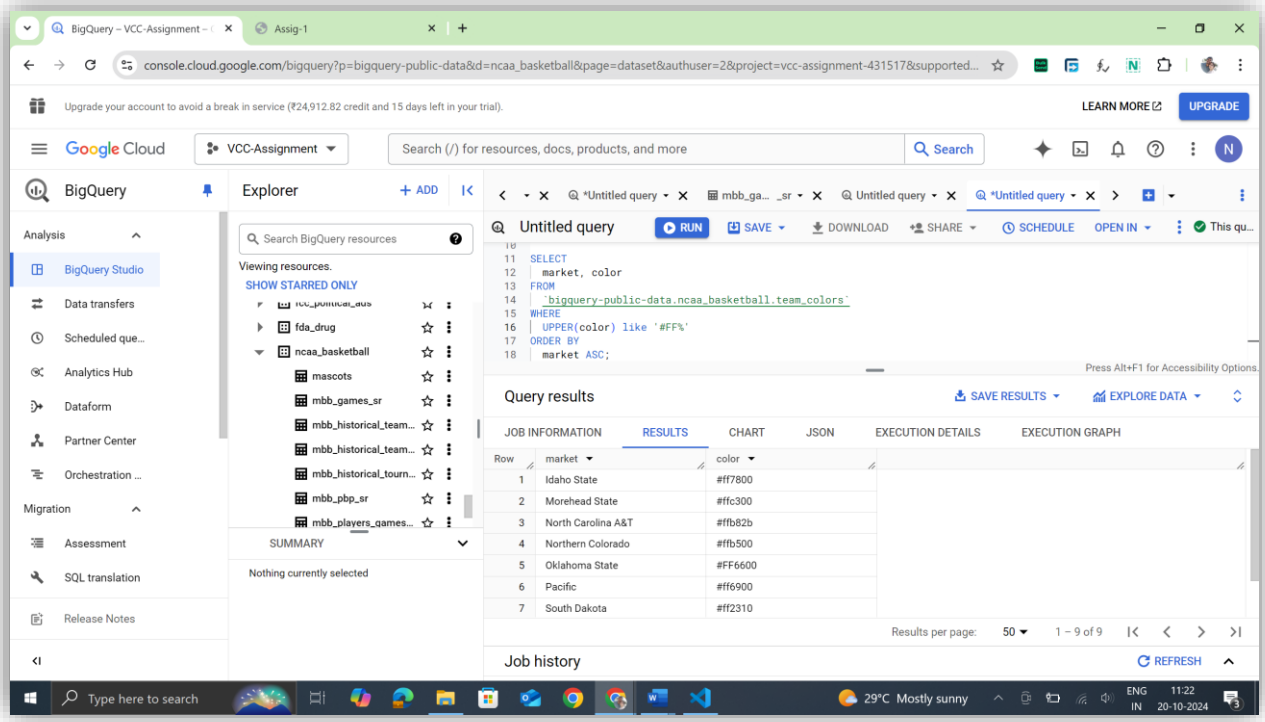


3. ***(1 point) Hexadecimal colors codes are a way of representing color on a computer. Hex color codes are of form #AABBCC, where AA, BB, and CC are hexadecimal numbers (00, 01, ... , FE, FF) indicating the intensity of red, green, and blue in the color, respectively.***

Hint: be careful with the case of the colors in the dataset -- some use lower case characters and some use upper case characters. Note that in the expected answer below, the original case from the dataset is kept. What teams have the maximum possible red intensity in their color? Give (team market, color) as your answer. Order your results alphabetically by the team market.

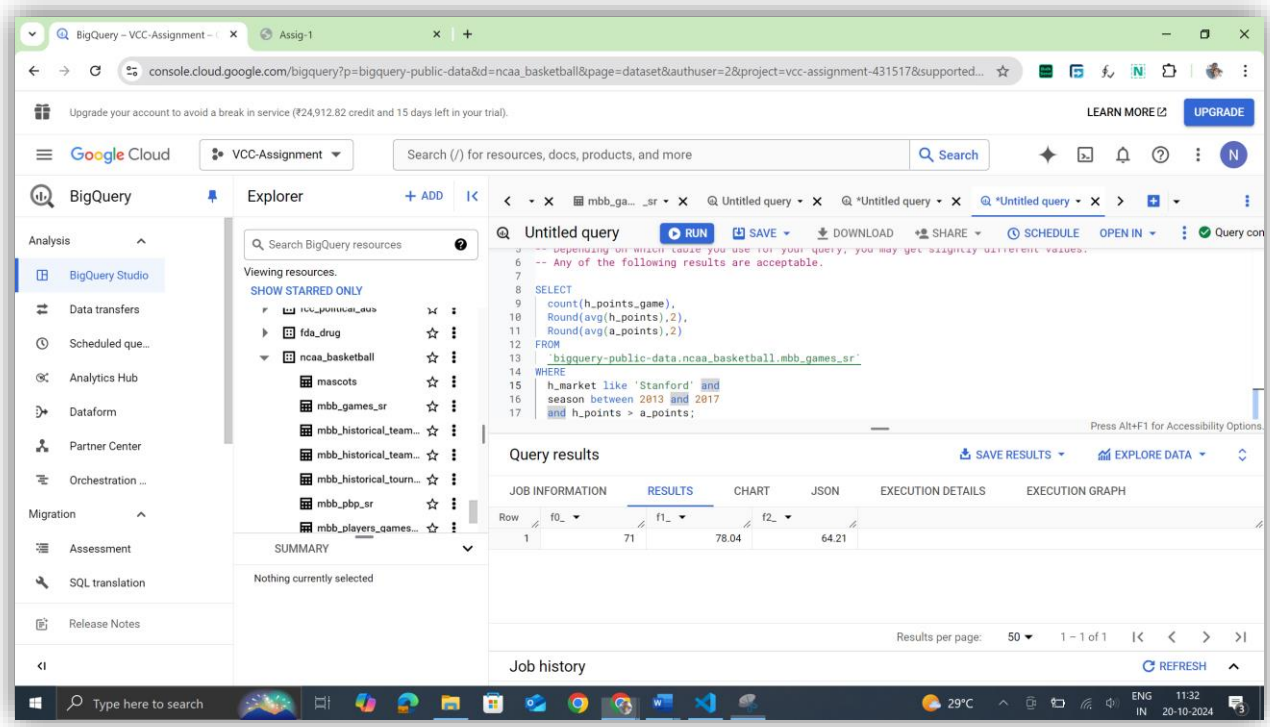
Answer:

```
SELECT
  market, color
FROM
  `bigquery-public-data.ncaa_basketball.team_colors`
WHERE
  UPPER(color) like '#FF%'
ORDER BY
  market ASC;
```



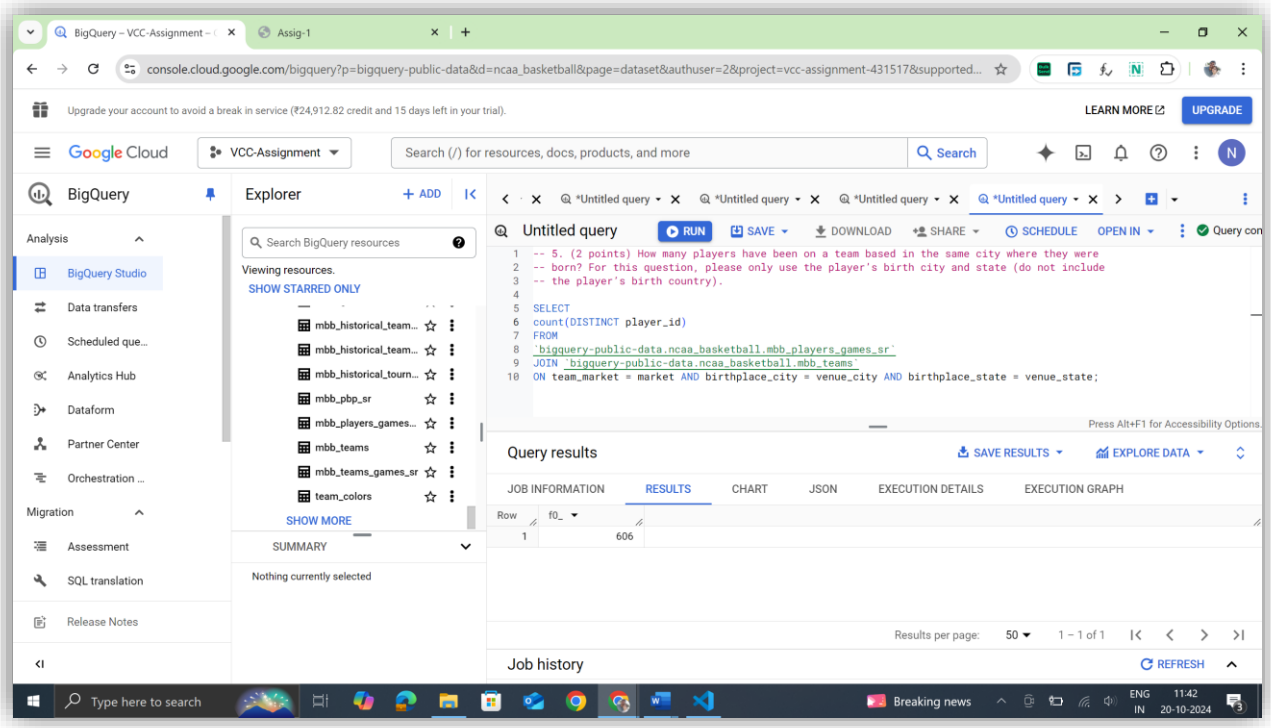
4. ***(1 point) How many home games has Stanford won in seasons 2013 to 2017 (inclusive)? Give (number of games won, average score for Stanford in those games, average score of the opponents in those games) as your answer. Round any decimal values to two places. Answer: Depending on which table you use for your query, you may get slightly different values. Any of the following results are acceptable.***

```
SELECT
  count(h_points_game),
  Round(avg(h_points),2),
  Round(avg(a_points),2)
FROM
  `bigquery-public-data.ncaa_basketball.mbb_games_sr`
WHERE
  h_market like 'Stanford' and
  season between 2013 and 2017
  and h_points > a_points;
```



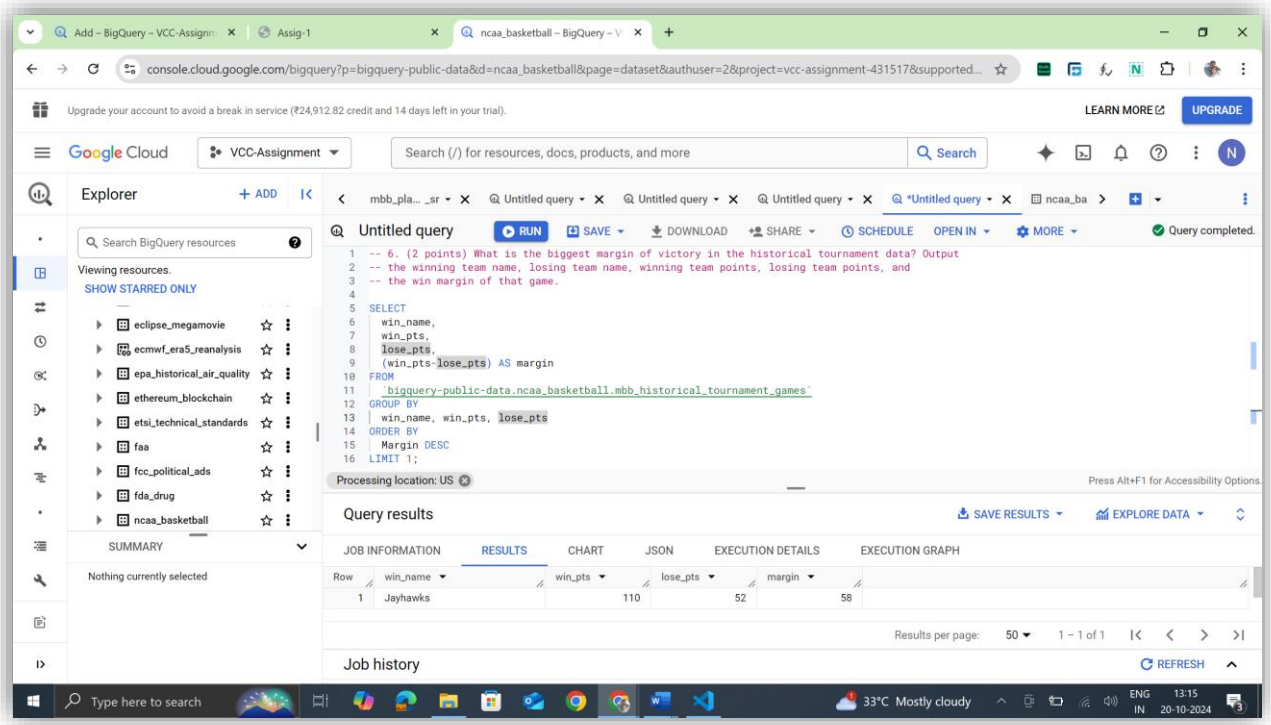
5. ***(2 points) How many players have been on a team based in the same city where they were born? For this question, please only use the player's birth city and state (do not include the player's birth country).***

```
SELECT
count(DISTINCT player_id)
FROM
`bigquery-public-data.ncaa_basketball.mbb_players_games_sr`
JOIN `bigquery-public-data.ncaa_basketball.mbb_teams`
ON team_market = market AND birthplace_city = venue_city AND birthplace_state
= venue_state;
```



6. ***(2 points) What is the biggest margin of victory in the historical tournament data? Output the winning team name, losing team name, winning team points, losing team points, and the win margin of that game.***

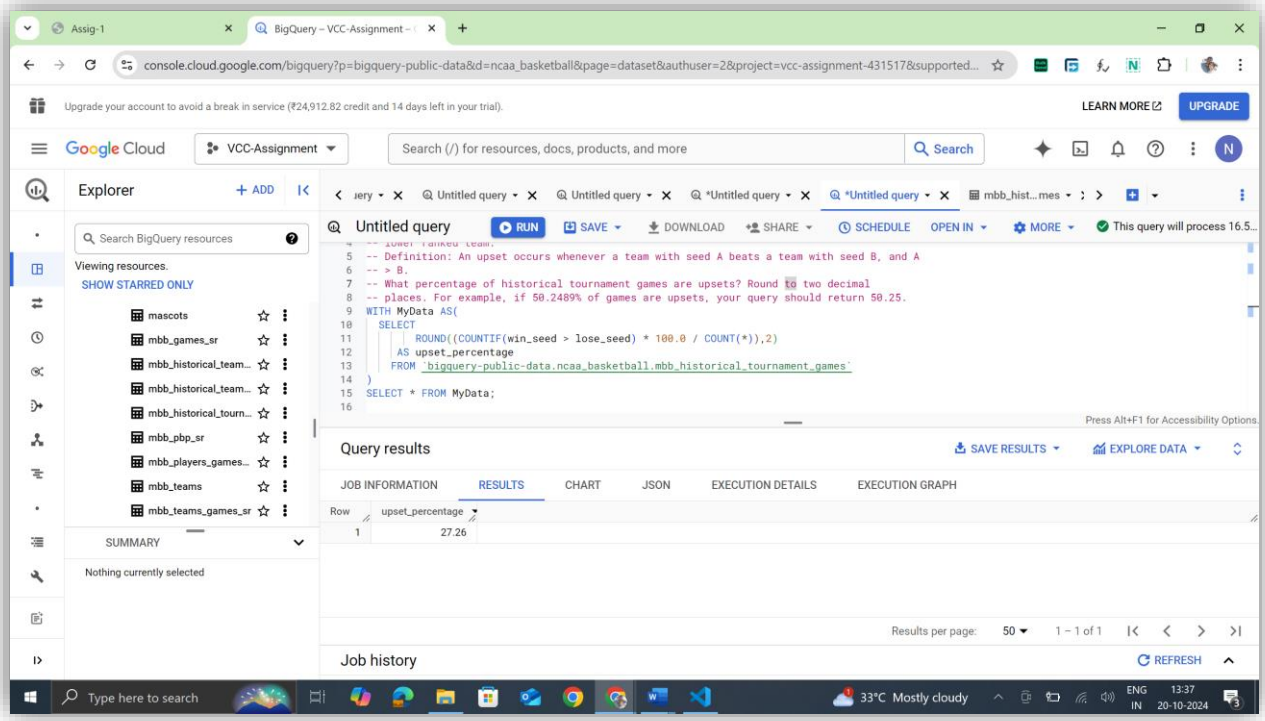
```
SELECT
  win_name, win_pts, lose_pts, (win_pts-lose_pts) AS margin
FROM
  `bigquery-public-data.ncaa_basketball.mbb_historical_tournament_games`
GROUP BY
  win_name, win_pts, lose_pts
ORDER BY
  Margin DESC
LIMIT 1;
```



7. ***(3 points) In a basketball tournament, teams are ranked from best to worst prior to starting the matches. This ranking is called the "seed" of the team (1 is the best team, and a higher number indicates a worse team). In general, a higher ranked team is expected to beat a lower ranked team. Definition: An upset occurs whenever a team with seed A beats a team with seed B, and A > B. What percentage of historical tournament games are upsets? Round to two decimal places. For example, if 50.2489% of games are upsets, your query should return 50.25.***

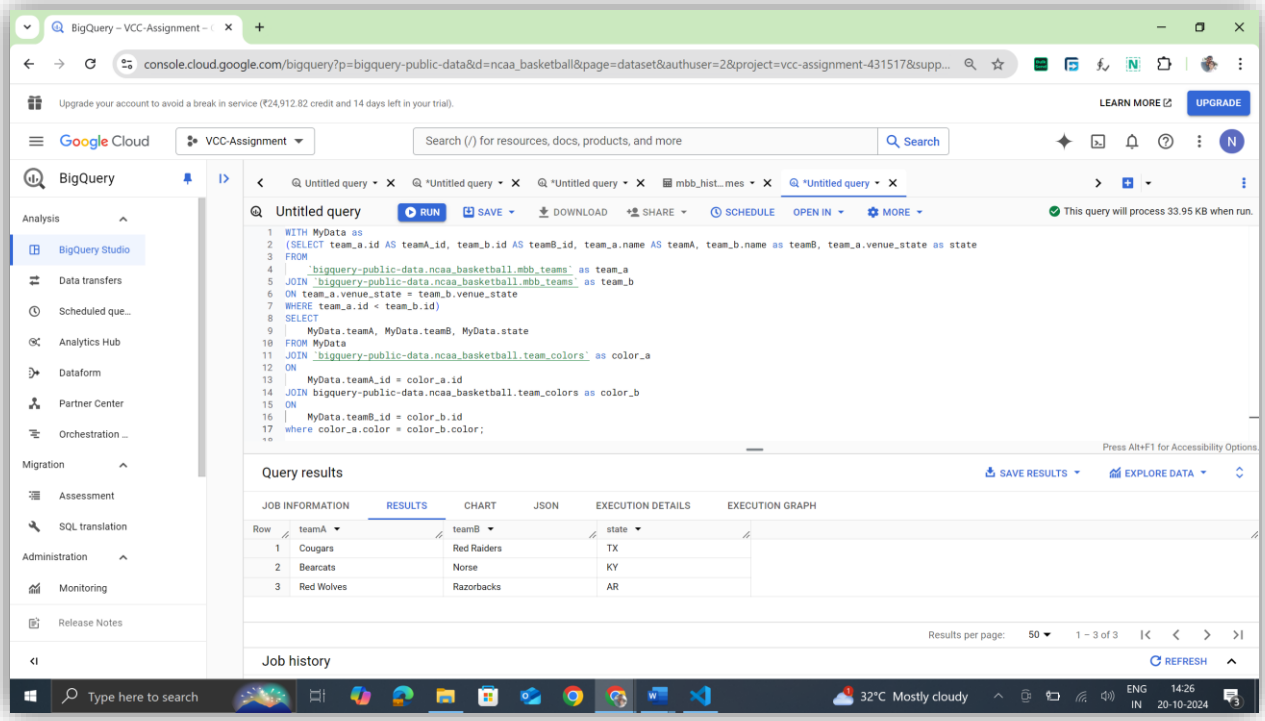
Answer:

```
WITH MyData AS(
  SELECT
        ROUND((COUNTIF(win_seed > lose_seed) * 100.0 / COUNT(*)),2)
     AS upset_percentage
    FROM `bigquery-public-
data.ncaa_basketball.mbb_historical_tournament_games`
)
SELECT * FROM MyData;
```



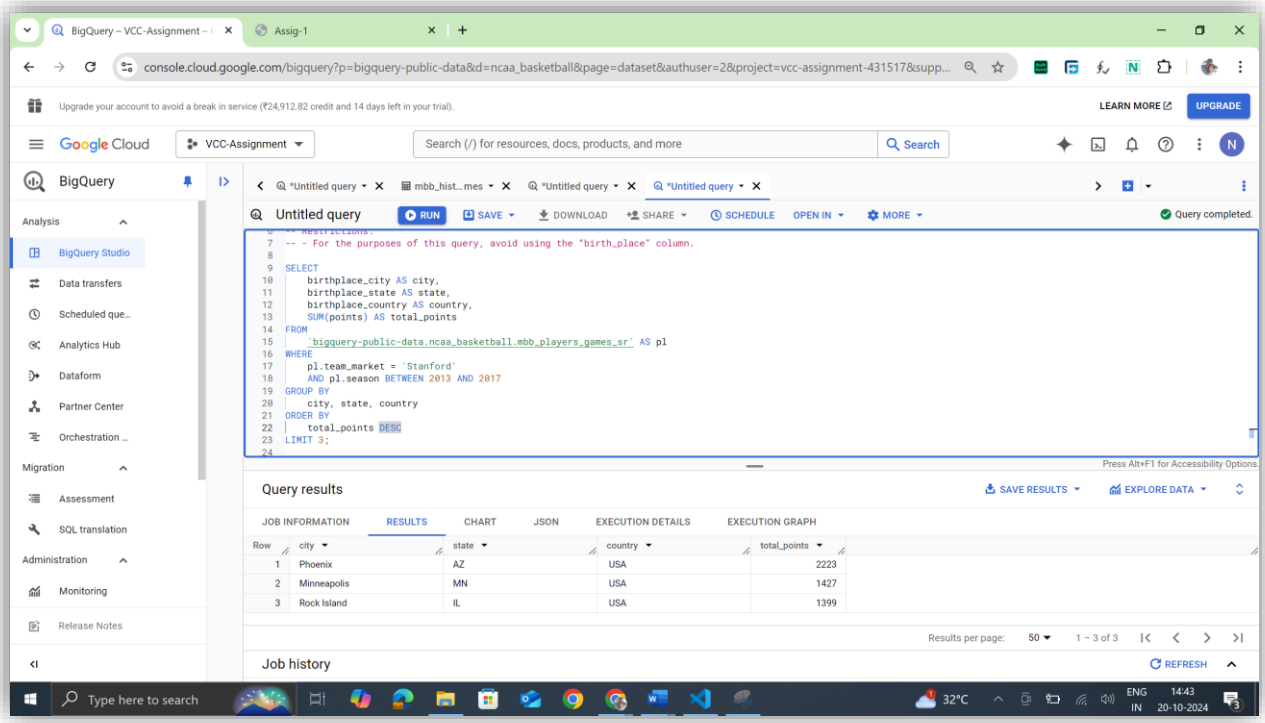
8. ***(3 points) Which pairs of NCAA basketball teams are 1) based in the same state and 2) have the same team color? Output the team names and the state. Put the team name that comes alphabetically first in each pair on the leftmost column, and order the rows alphabetically by the first column.***

```
WITH MyData as
(SELECT team_a.id AS teamA_id, team_b.id AS teamB_id, team_a.name AS teamA,
team_b.name as teamB, team_a.venue_state as state
FROM
    `bigquery-public-data.ncaa_basketball.mbb_teams` as team_a
JOIN `bigquery-public-data.ncaa_basketball.mbb_teams` as team_b
ON team_a.venue_state = team_b.venue_state
WHERE team_a.id < team_b.id)
SELECT
    MyData.teamA, MyData.teamB, MyData.state
FROM MyData
JOIN `bigquery-public-data.ncaa_basketball.team_colors` as color_a
ON
    MyData.teamA_id = color_a.id
JOIN bigquery-public-data.ncaa_basketball.team_colors as color_b
ON
    MyData.teamB_id = color_b.id
where color_a.color = color_b.color;
```



9. ***(3 points) What three geographical locations made the most points for Stanford's team in seasons 2013 through 2017, and how many points did they make?***
Definition: A geographical location L is a unique tuple (city, state, country). Definition: A geographical location L "makes" points for a team T whenever a player that was born in L scores points for T.
Restrictions: - For the purposes of this query, avoid using the "birth_place" column.

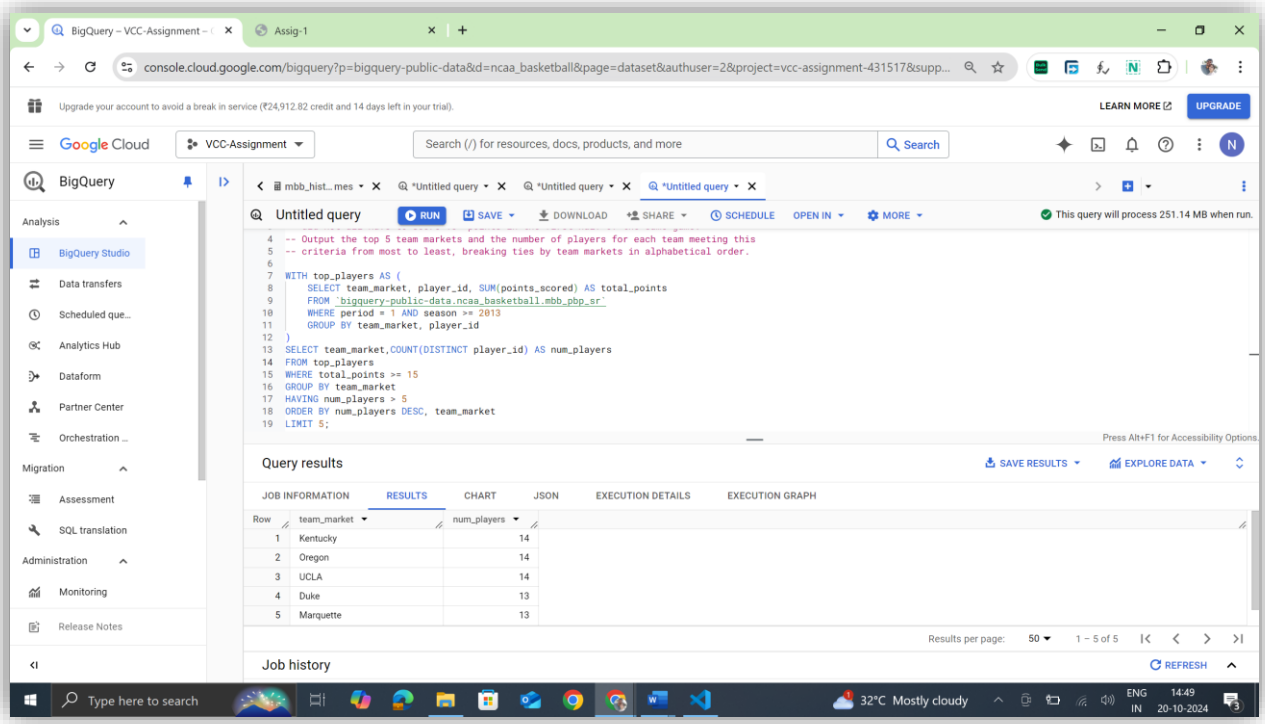
```sql
SELECT
    birthplace_city AS city,
    birthplace_state AS state,
    birthplace_country AS country,
    SUM(points) AS total_points
FROM
    `bigquery-public-data.ncaa_basketball.mbb_players_games_sr` AS pl
WHERE
    pl.team_market = 'Stanford'
    AND pl.season BETWEEN 2013 AND 2017
GROUP BY
    city, state, country
ORDER BY
    total_points DESC
LIMIT 3;
```



10. ***(4 points) Since the 2013 season (inclusive), which teams have had more than 5 players score 15 or more points in the first half (period) in a single game?***

Note: These players did not all have to score 15+ points in the first half of the same game. Output the top 5 team markets and the number of players for each team meeting this criteria from most to least, breaking ties by team markets in alphabetical order.

```sql
WITH top_players AS (
    SELECT team_market, player_id, SUM(points_scored) AS total_points
    FROM `bigquery-public-data.ncaa_basketball.mbb_pbp_sr`
    WHERE period = 1 AND season >= 2013
    GROUP BY team_market, player_id
)
SELECT team_market,COUNT(DISTINCT player_id) AS num_players
FROM top_players
WHERE total_points >= 15
GROUP BY team_market
HAVING num_players > 5
ORDER BY num_players DESC, team_market
LIMIT 5;
```



11. (4 points) What five teams (identify them here by their "markets") were top performers in the most seasons between 1900 and 2000 (inclusive), and how many times were they top performers?

Output the team markets and the number of times each team was a top performer. If there are ties in the final output, break them by giving a higher ranking to team markets that come first alphabetically. Ignore teams with NULL markets only in the final output.

Definition: Team X is a top performer on season Y if no other team had more wins than X in the same season. This includes teams with either null or non-null .

```
WITH max_wins AS (
    SELECT season, MAX(wins) AS highest_wins
    FROM `bigquery-public-
data.ncaa_basketball.mbb_historical_teams_seasons`
    WHERE season BETWEEN 1900 AND 2000
    GROUP BY season
),
top_wins AS (
    SELECT team.market, team.season
    FROM `bigquery-public-
data.ncaa_basketball.mbb_historical_teams_seasons` AS team
    JOIN max_wins
    ON  team.season = max_wins.season AND team.wins = max_wins.highest_wins
    WHERE team.market IS NOT NULL
)
SELECT market AS team_market, COUNT(season) AS top_performance_count
FROM top_wins GROUP BY market
ORDER BY top_performance_count DESC,market LIMIT 5;
```