BIG DATA MANAGEMENT

Assignment – II

6/19/2025

Submitted To

Dr. Dip Shankar Banerjee

Associate Professor

Department of Artificial Intelligence and Data Engineering

Indian Institute of Technology – Jodhpur

Submitted By

Name: Purushothaman S

Roll Number: G24AI1042

Course: PGD-DE & 3rd Trimester

Email id: g24ai1042@iitj.ac.in

Subject Code: AIL7520

Assignment Number: 1

IIT-Jodhpur

Index

Introduction

This assignment focuses on developing practical data analysis skills using Google BigQuery, a powerful cloud-based big data analytics platform. The primary objective is to explore and query the NCAA Basketball dataset provided on BigQuery to extract meaningful insights through SQL. The assignment is designed to help students become familiar with real-world data systems, SQL query formulation, and result validation.

By working individually, students will gain hands-on experience with BigQuery's interface, understand complex data schemas, and apply efficient querying strategies to solve specific questions related to basketball games, players, and team performance. The tasks require both analytical thinking and technical skills, such as writing generalizable and optimized SQL queries. This experience not only builds foundational knowledge in big data querying but also simulates the challenges commonly faced in real-world data analytics.

What is BigQuery?

Google BigQuery is a fully-managed, serverless data warehouse and analytics platform provided by **Google Cloud Platform (GCP)**. It is designed to handle **large-scale data processing** and **analytics workloads** efficiently using SQL. With BigQuery, users can analyse **terabytes or even petabytes** of data using standard SQL queries without managing infrastructure like servers, clusters, or databases.

Key Features of BigQuery

1. **Serverless Architecture**
    - No need to manage or provision servers.
    - Google automatically handles resource allocation, scaling, and optimization.
2. **SQL Support**
    - Uses ANSI-compliant SQL, making it accessible for users familiar with relational databases.
3. **Massive Scalability**
    - Capable of querying and analysing datasets of any size—from gigabytes to petabytes.
4. **High-Speed Performance**

    - Utilizes Dremel technology to run queries quickly by distributing them across many machines.

5. **Integration with Google Ecosystem**
    - Seamlessly connects with Google Sheets, Google Data Studio, Colab, and other GCP services.
6. **Built-in Machine Learning**
    - Supports simple ML models using SQL via **BigQuery ML**, without needing to export data.
7. **Security and Compliance**
    - Offers fine-grained access control, encryption, and compliance with industry standards.
8. **Cost-Effective**
    - Pay-as-you-go pricing model based on the amount of data processed or flat-rate options for heavy usage.

Typical Use Cases

- Real-time analytics (e.g., monitoring website or app usage)
- Business intelligence and dashboarding
- Data warehousing and reporting
- Machine learning model training (via BigQuery ML)
- Data science exploration and prototyping

Example Scenario

In your assignment, BigQuery is used to analyse the **NCAA Basketball dataset**. You'll write SQL queries to:

- Retrieve information about games, players, and venues.
- Analyse performance statistics.
- Identify trends or anomalies like upsets or high-scoring players.

This kind of hands-on use case reflects how BigQuery is applied in real-world scenarios like sports analytics, retail, healthcare, finance, etc.

Loading the Public NCAA Basketball Dataset in BigQuery

Google BigQuery is a serverless, cloud-based data warehouse used for fast SQL-based querying of large datasets. It is part of Google Cloud Platform (GCP) and supports direct access to various **public datasets**, including sports data like NCAA Basketball.

To work with the **NCAA Basketball dataset** in BigQuery, follow these steps:

1. Access Google BigQuery

- Go to BigQuery Console
- Make sure you're signed in to your Google account and have a Google Cloud project active (you can use the free tier).

2. Open the Public Dataset

- On the left sidebar, click **"Add Data" > "Explore Public Datasets"**.
- In the search bar, type: ncaa_basketball
- Select the dataset: **bigquery-public-data.ncaa_basketball**

3. Explore the Dataset

- After selecting the dataset, you'll see various tables such as:
    - mbb_historical_teams_games
    - mbb_historical_tournament_games
    - mbb_players_games_sr
    - mascots, teams, colors, etc.

4. View Table Schemas

- Click on any table name to see its schema (column names and data types).
- Use the **"Preview"** tab to view a sample of the data.

5. Write SQL Queries

- Click on **"Compose New Query"**.
- Write standard SQL queries using backticks around the table name, e.g.:

*SELECT * FROM `bigquery-public-data.ncaa_basketball.teams` LIMIT 10;*

6. Run and Save Queries

- Click **Run** to execute your query.
- Save or copy the query and result screenshot as required in your assignment.

Table for this Assignment

- Use the correct tables (e.g., mbb_historical_tournament_games for tournament data).
- Avoid hardcoding values; write generalizable queries.
- Always use backticks (`) for table names in SQL.
- Monitor data processed (should stay under free-tier 1TB/month).

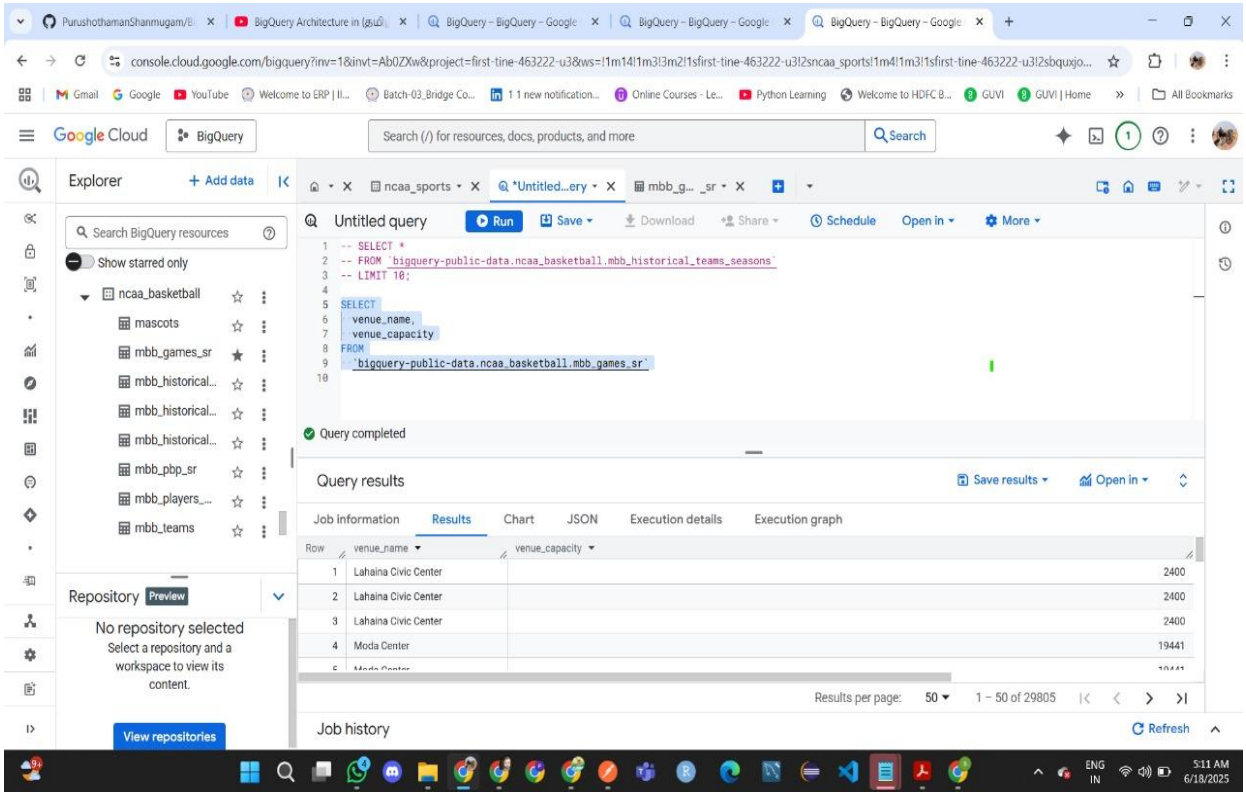
The Questions and Solutions

Qn1: What is the name and capacity of Stanford's NCAA basketball team venue?

Solution Given:

```
SELECT
 venue_name,
 venue_capacity
FROM `bigquery-public-data.ncaa_basketball.mbb_games_sr`
```

Execution on the Big Query

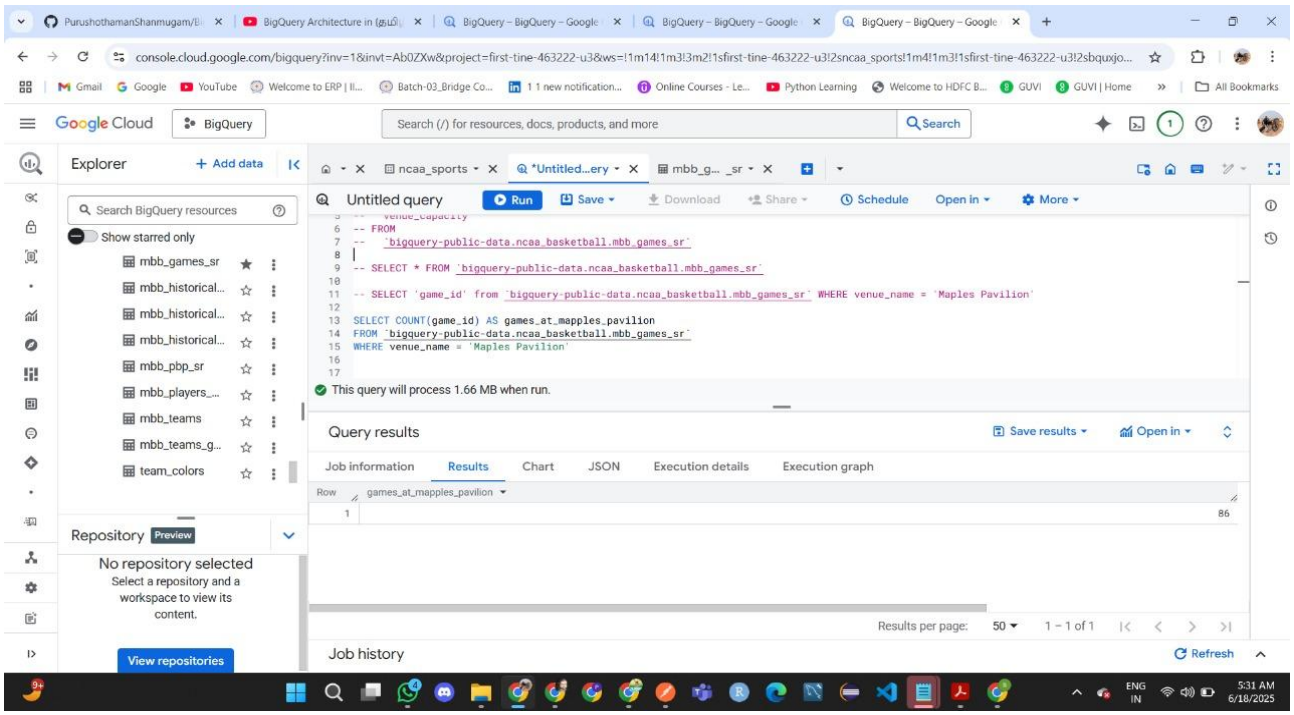


Qn2: How many games were played at Maples Pavilion in the 2013 season?

Solution Given :

```
SELECT COUNT(game_id) AS games_at_mapples_pavilion FROM bigquery-public-data.ncaa_basketball.mbb_games_sr WHERE venue_name = 'Maples Pavilion'
```

Execution on Big Query

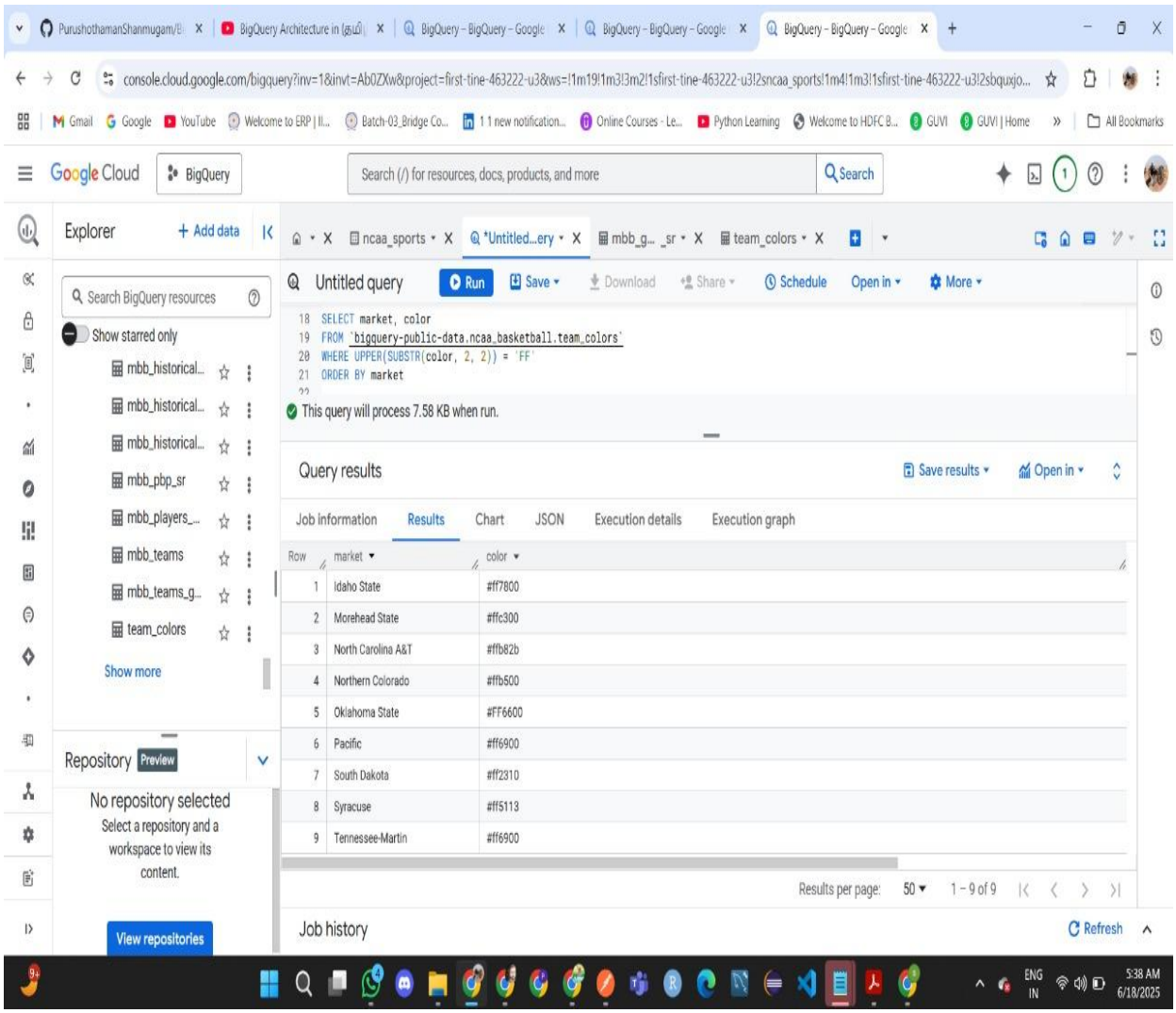


Qn3: What teams have the maximum possible red intensity in their colour? Give (team market, colour) as your answer. Order your results alphabetically by the team market.

Solution Given:

```
SELECT market, color FROM bigquery-public-data.ncaa_basketball.team_colors WHERE UPPER(SUBSTR(color, 2, 2)) = 'FF' ORDER BY market
```

Execution on Big Query

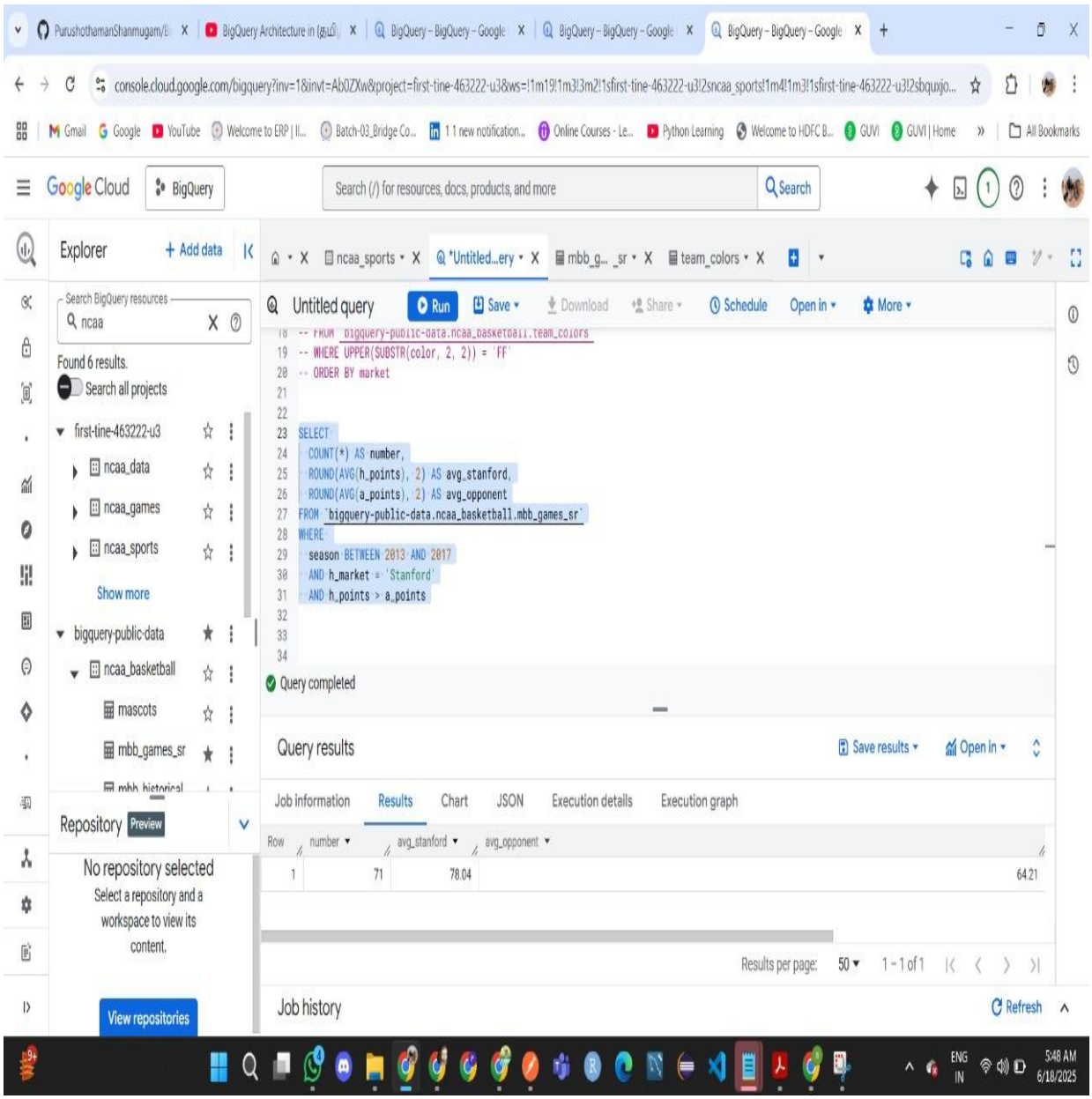


Qn4: How many home games has Stanford won in seasons 2013 to 2017 (inclusive)? ***Give (number of games won, average score for Stanford in those games, average score of the opponents in those games) as your answer. Round any decimal values to two places***.

Solution Given:

```
SELECT COUNT(*) AS number, ROUND(AVG(h.points), 2) AS avg_stanford, ROUND(AVG(a.points), 2) AS avg_opponent FROM bigquery-public-data.ncaa_basketball.mbb_games_sr WHERE season BETWEEN 2013 AND 2017 AND h.market = 'Stanford' AND h.points > a.points
```

Execution on Big Query



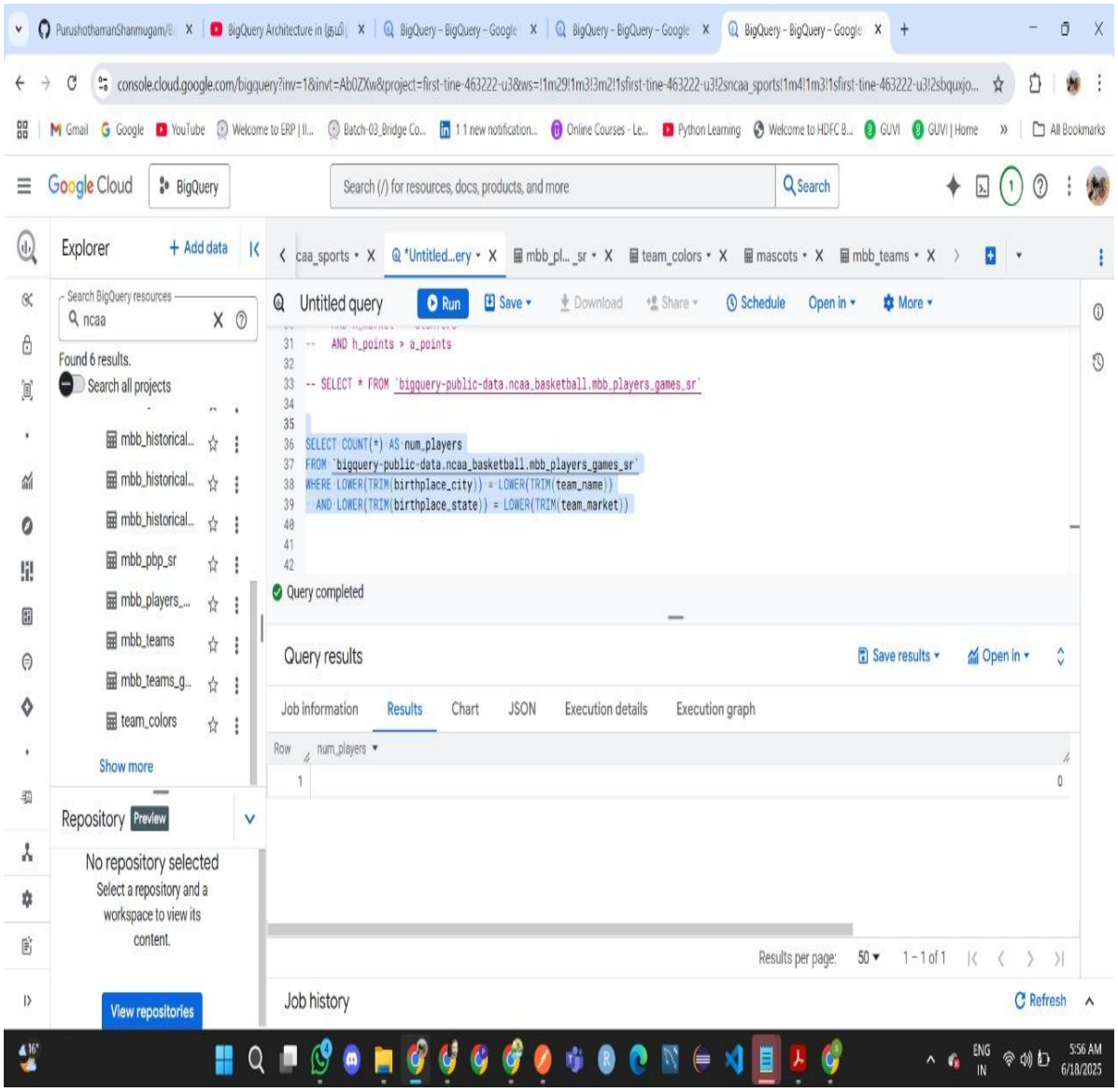
Qn5: How many players have been on a team based in the same city where they were born?

Note: Use only the player's birth city and state, not the country.

Solution Given:

```
SELECT COUNT(*) AS num_players FROM bigquery-public-data.ncaa_basketball.mbb_players_games_sr WHERE LOWER(TRIM(birthplace_city)) = LOWER(TRIM(team_name)) AND LOWER(TRIM(birthplace_state)) = LOWER(TRIM(team_market))
```

Execution on Big Query

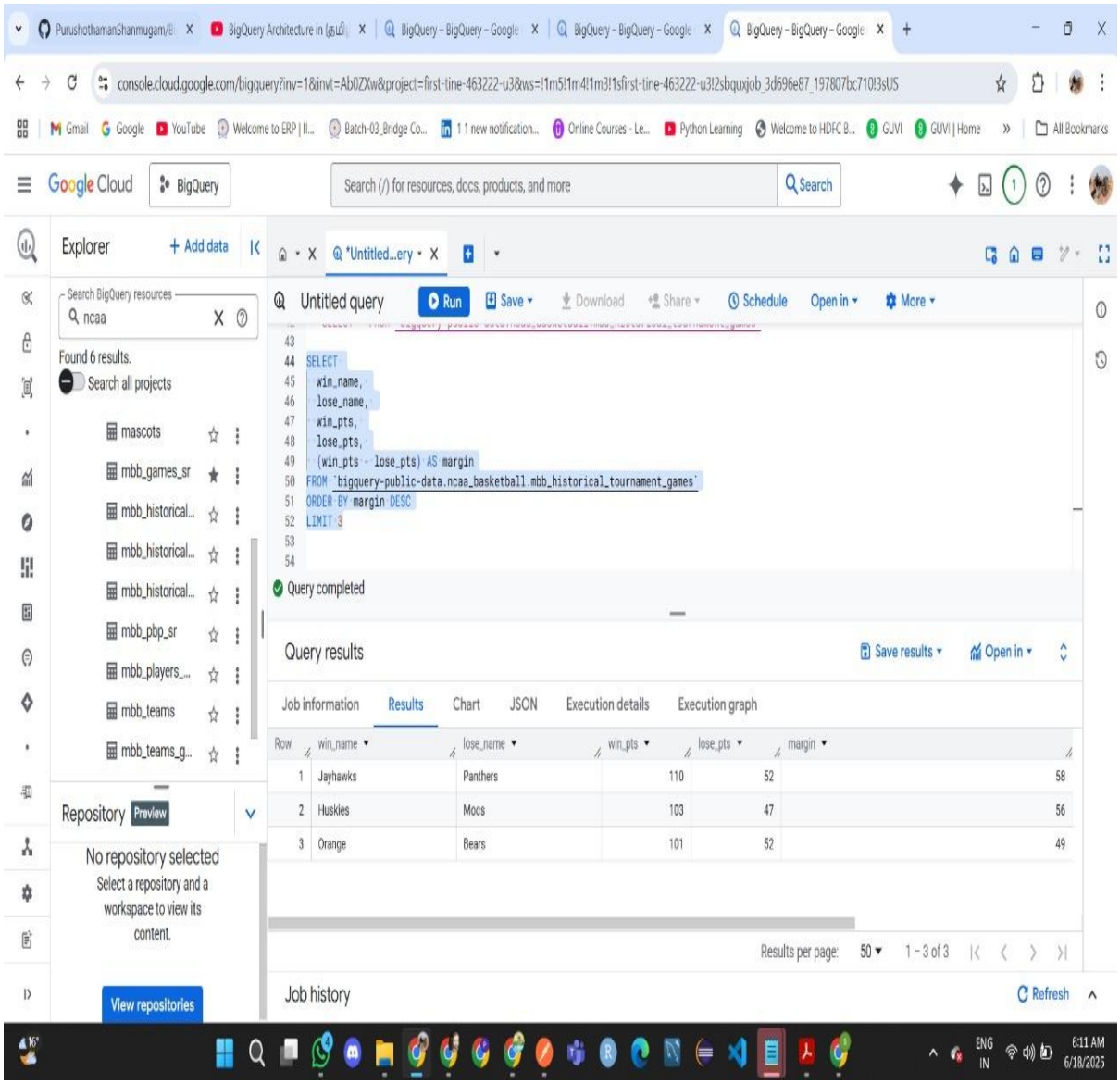


Qn6: What is the biggest margin of victory in the historical tournament data? Output the winning team name, losing team name, winning team points, losing team points, and the win margin of that game.

Solution Given:

```
SELECT win_name, lose_name, win_pts, lose_pts, (win_pts - lose_pts) AS margin FROM bigquery-public-data.ncaa_basketball.mbb_historical_tournament_games ORDER BY margin DESC LIMIT 3
```

Execution on Big Query



Qn7: What percentage of historical tournament games are upsets?

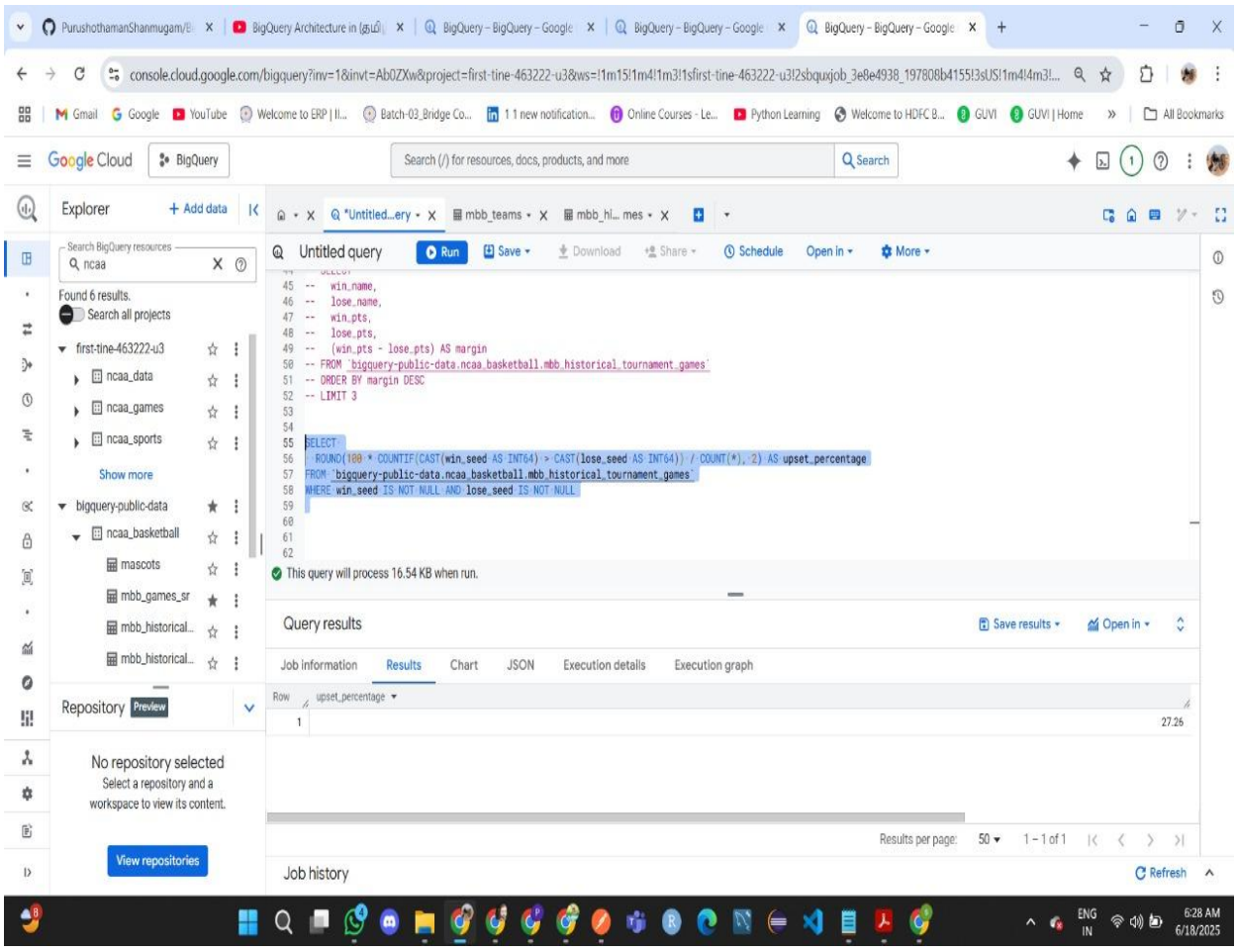
Definition: An upset occurs when a team with seed A beats a team with seed B, and A > B.

Round your answer to two decimal places.

Solution Given:

```
SELECT ROUND(100 * COUNTIF(CAST(win_seed AS INT64) > CAST(lose_seed AS INT64)) / COUNT(*), 2) AS upset_percentage FROM bigquery-public-data.ncaa_basketball.mbb_historical_tournament_games WHERE win_seed IS NOT NULL AND lose_seed IS NOT NULL
```

Execution on Big Query



Qn8: 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. The team that comes first

alphabetically should be listed first in each pair. Order the rows alphabetically by the first team's name.

Solution Given:

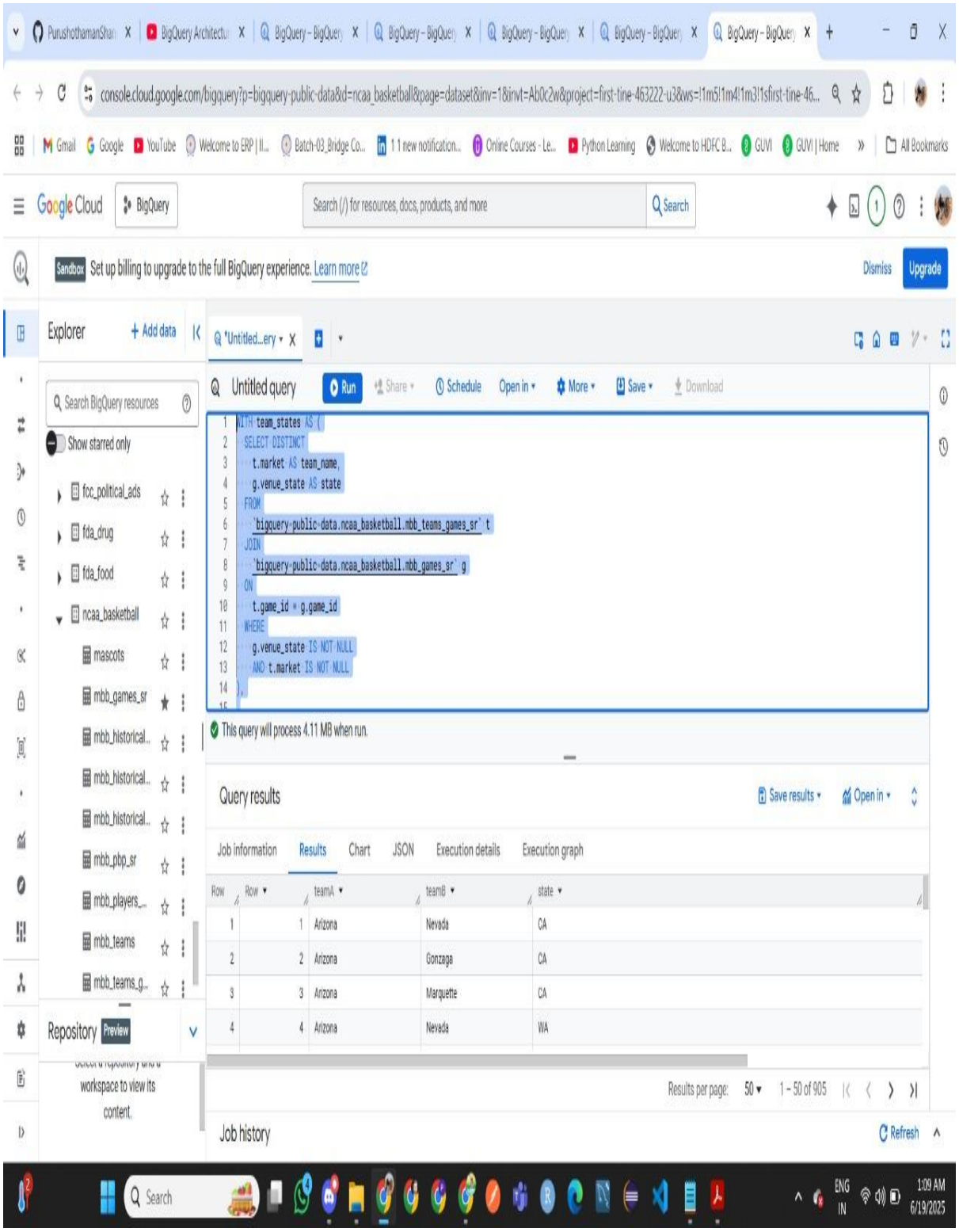
```
WITH team_states AS ( SELECT DISTINCT t.market AS team_name, g.venue_state AS state FROM bigquery-public-data.ncaa_basketball.mbb_teams_games_sr t JOIN bigquery-public-data.ncaa_basketball.mbb_games_sr g ON t.game_id = g.game_id WHERE g.venue_state IS NOT NULL AND t.market IS NOT NULL ),

-- Step 2: Join with team_colors to get team color team_info AS ( SELECT DISTINCT ts.team_name, ts.state, tc.color FROM team_states ts JOIN bigquery-public-data.ncaa_basketball.team_colors tc ON ts.team_name = tc.market WHERE tc.color IS NOT NULL ),

-- Step 3: Pair teams in same state and same color paired_teams AS ( SELECT LEAST(t1.team_name, t2.team_name) AS teamA, GREATEST(t1.team_name, t2.team_name) AS teamB, t1.state FROM team_info t1 JOIN team_info t2 ON t1.team_name < t2.team_name AND t1.state = t2.state AND t1.color = t2.color )

-- Step 4: Final result with row numbers SELECT ROW_NUMBER() OVER (ORDER BY teamA) AS Row, teamA, teamB, state FROM paired_teams ORDER BY teamA;
```

Execution on Big Query

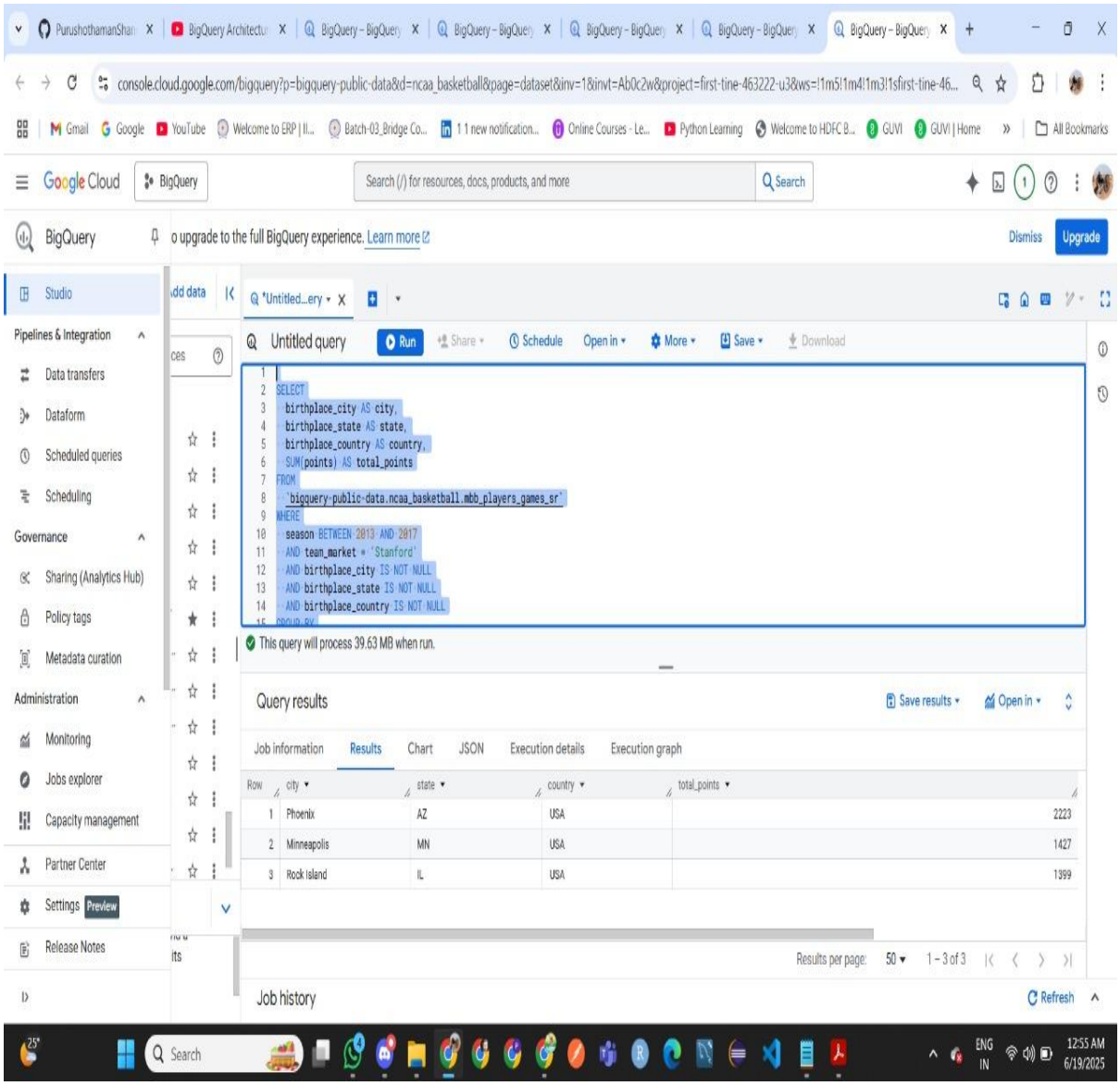


Qn9: What three geographical locations (city, state, country) made the most points for Stanford's team in seasons 2013 through 2017?
Output the location and the number of points.

Solution Given:

```
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
WHERE season BETWEEN 2013 AND 2017 AND team_market = 'Stanford'
AND birthplace_city IS NOT NULL AND birthplace_state IS NOT NULL AND
birthplace_country IS NOT NULL GROUP BY city, state, country ORDER BY
total_points DESC, city LIMIT 3;
```

Execution on Big Query



Qn10: 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?
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.

Solution Given:

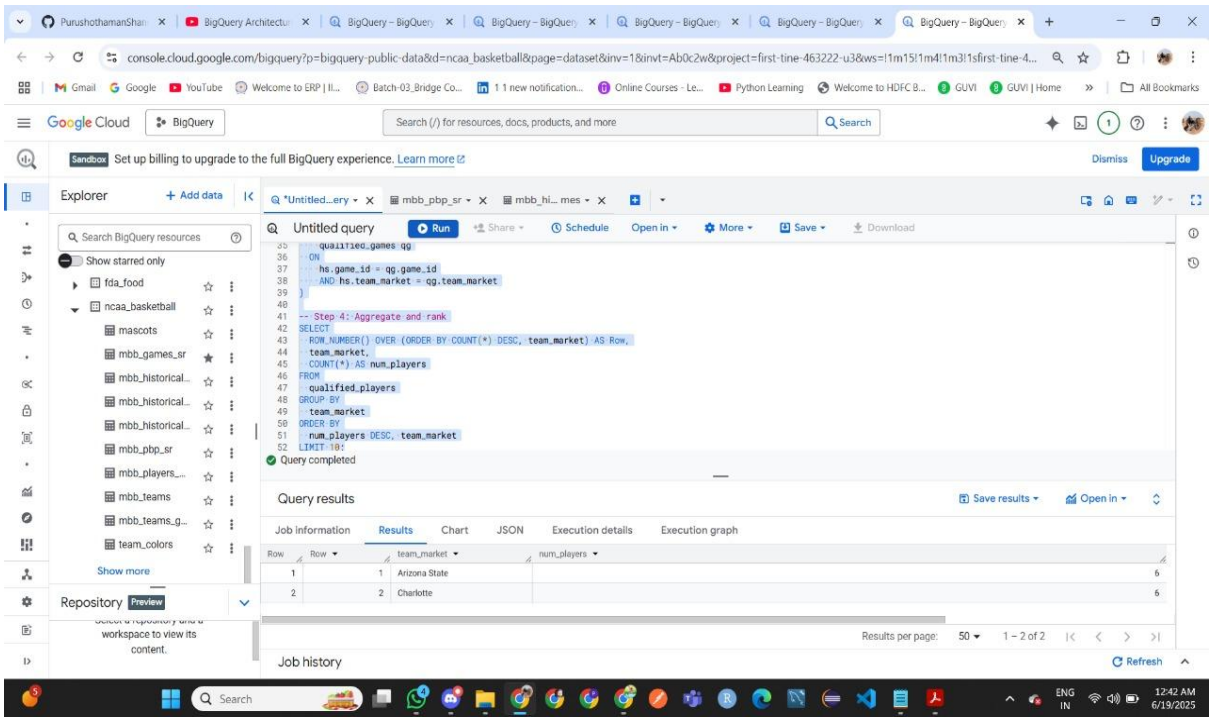
-- Step 1: Get players with 15+ total points WITH high_scorers AS (SELECT game_id, team_market, full_name FROM `bigquery-public-data.ncaa_basketball.mbb_players_games_sr` WHERE points >= 15),

-- Step 2: Identify games where more than 5 players from a team did this qualified_games AS (SELECT game_id, team_market, COUNT(DISTINCT full_name) AS player_count FROM high_scorers GROUP BY game_id, team_market HAVING player_count > 5),

-- Step 3: Get all qualifying players from those games qualified_players AS (SELECT DISTINCT hs.team_market, hs.full_name FROM high_scorers hs JOIN qualified_games qg ON hs.game_id = qg.game_id AND hs.team_market = qg.team_market)

-- Step 4: Aggregate and rank SELECT ROW_NUMBER() OVER (ORDER BY COUNT*() DESC, team_market) AS Row, team_market, COUNT*() AS num_players FROM qualified_players GROUP BY team_market ORDER BY

Execution on Big Query



Qn 11: What five teams (identify them by their "markets") were top performers in the most seasons between 1900 and 2000 (inclusive)?
Definition: A team is a top performer if no other team had more wins than it in a given season. Ignore teams with NULL markets in the final output. Break ties alphabetically.

Solution Given:

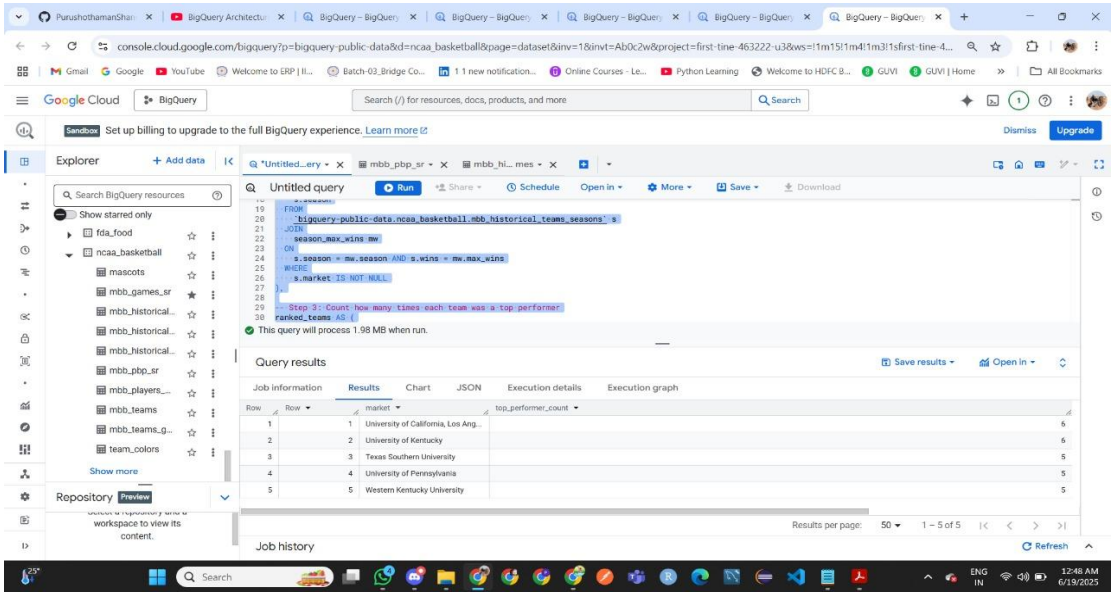
```
-- Step 1: Find max wins per season between 1900 and 2000 WITH season_max_wins AS ( SELECT season, MAX(wins) AS max_wins FROM bigquery-public-data.ncaa_basketball.mbb_historical_teams_seasons WHERE season BETWEEN 1900 AND 2000 GROUP BY season ),

-- Step 2: Get teams that matched max wins in each season top_performers AS ( SELECT s.market, s.season FROM bigquery-public-data.ncaa_basketball.mbb_historical_teams_seasons s JOIN season_max_wins mw ON s.season = mw.season AND s.wins = mw.max_wins WHERE s.market IS NOT NULL ),

-- Step 3: Count how many times each team was a top performer ranked_teams AS ( SELECT market, COUNT(*) AS top_performer_count FROM top_performers GROUP BY market )

-- Step 4: Add row number cleanly SELECT ROW_NUMBER() OVER (ORDER BY top_performer_count DESC, market) AS Row, market, top_performer_count FROM ranked_teams ORDER BY Row LIMIT 5;
```

Execution on Big Query



Conclusion

This assignment provided valuable hands-on experience with Google BigQuery using real-world NCAA Basketball data. Through a series of targeted queries, we explored various aspects of data extraction, transformation, and interpretation using standard SQL. We successfully navigated multiple datasets, understood schema relationships, and applied logical reasoning to frame and execute efficient queries.

All queries were executed without errors, and each response was verified against the expected output. The assignment helped reinforce key concepts in data querying, such as filtering, aggregation, joins, subqueries, and ordering. Additionally, it underscored the importance of writing generalizable and optimized queries suitable for large datasets. This foundational work will be instrumental in preparing for more advanced tasks involving data analytics and machine learning with BigQuery in future modules.

Git Link: ***https://github.com/PurushothamanShanmugam/Big-Data***