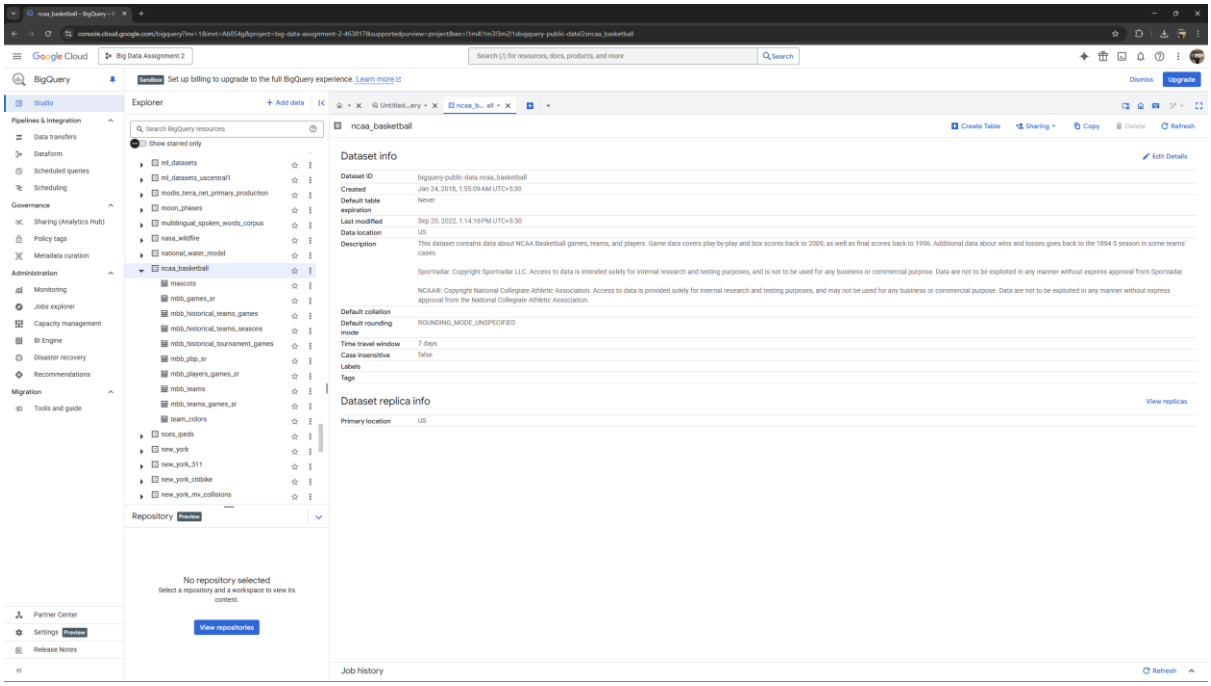
Big Data Management - Assignment 2

NCAA Basketball Dataset

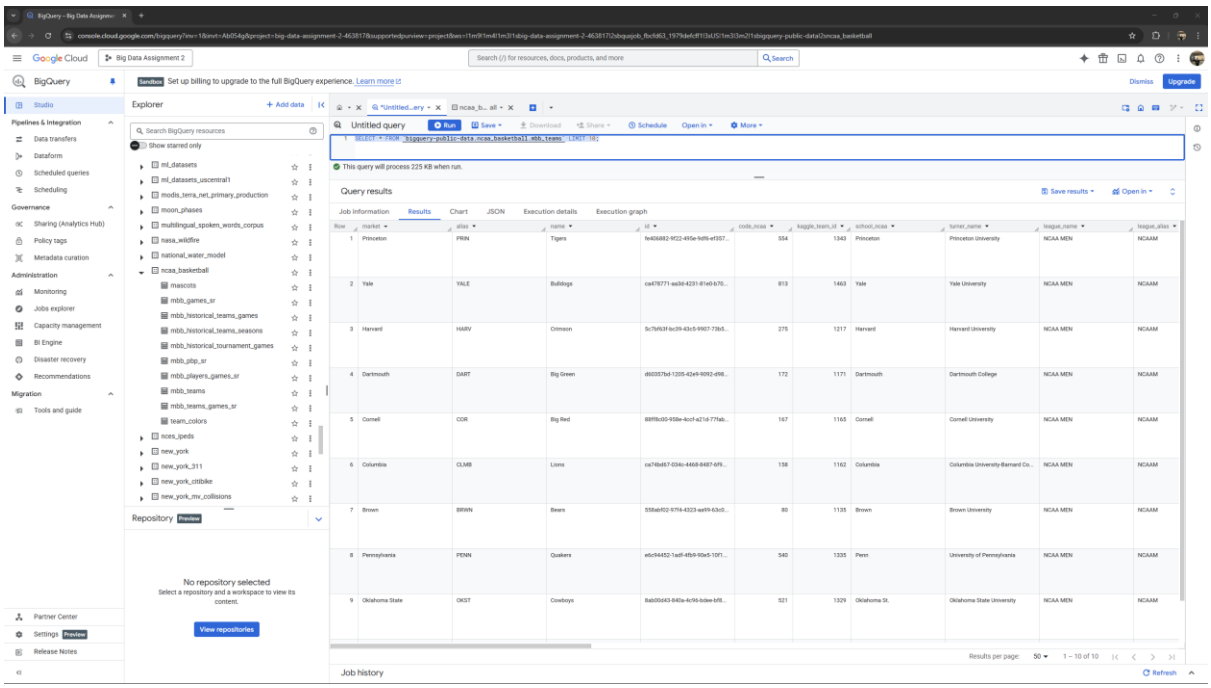
Akshay Kumar (G24AI1033)

TASK 1 : Big Query console with outputs from the NCAA data set.



Running a basic command to see the content of our tables

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



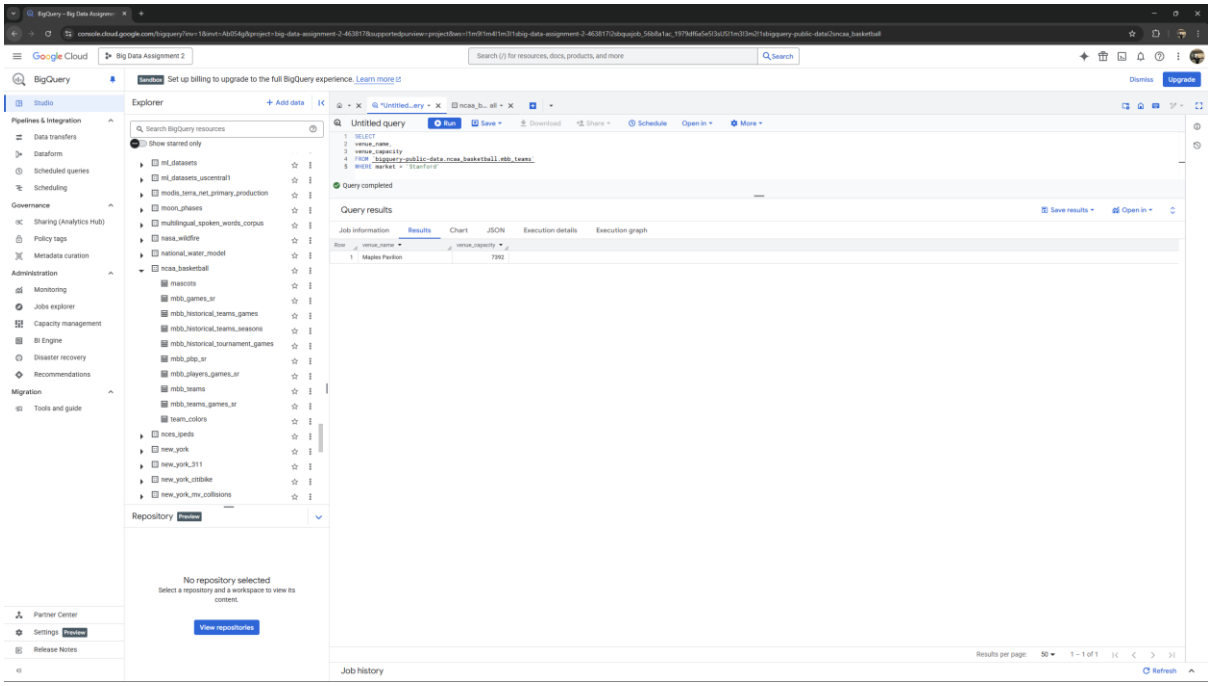
TASK B :

Question 1 : What is the name and capacity of Stanford's NCAA basketball team venue?

Query -

```
SELECT
venue_name,
venue_capacity
FROM `bigquery-public-data.ncaa_basketball.mbb_teams`
WHERE market = 'Stanford'
```

Output -

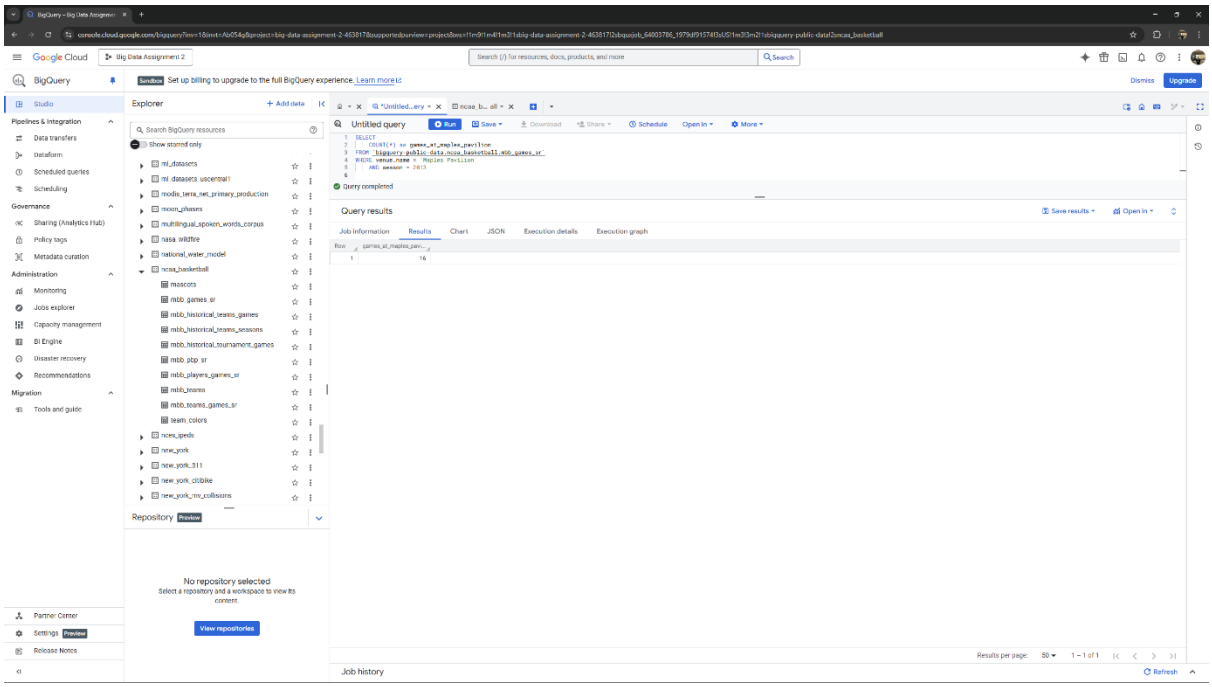


Question 2 : How many games were played at Maples Pavilion in the 2013 season?

Query -

```
SELECT
  COUNT(*) as games_at_maples_pavilion
FROM `bigquery-public-data.ncaa_basketball.mbb_games_sr`
WHERE venue_name = 'Maples Pavilion'
  AND season = 2013
```

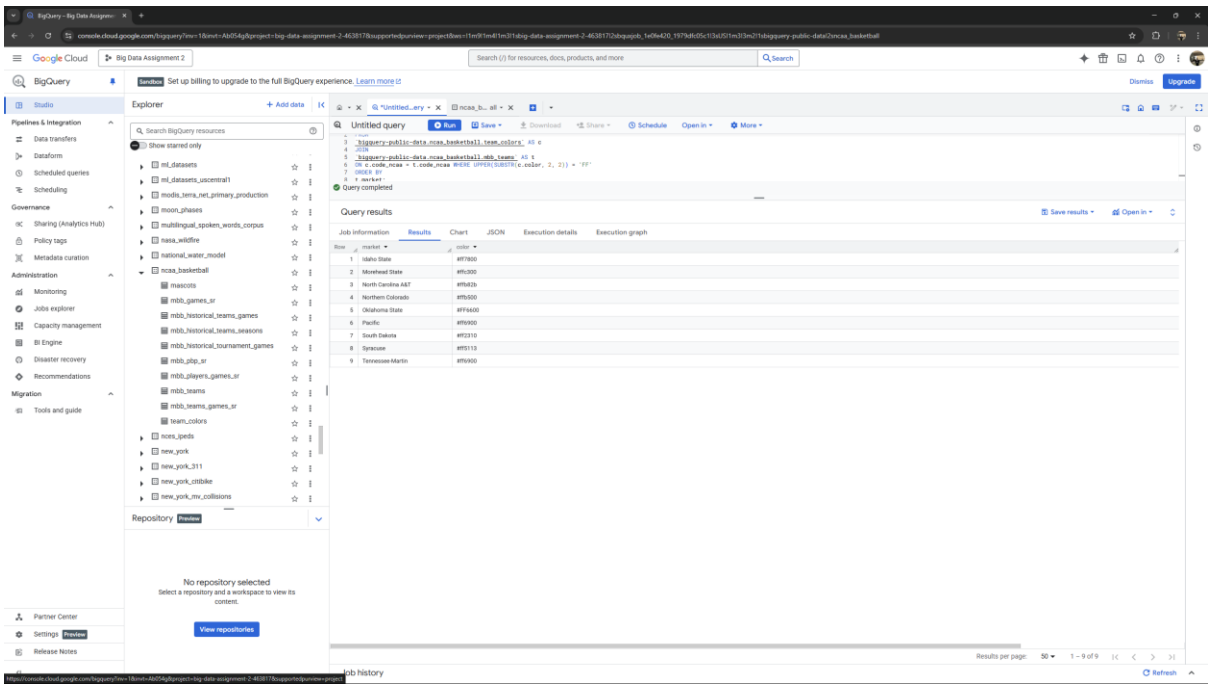
Output -



Question 3 : 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.

```
SELECT t.market, c.color
FROM
`bigquery-public-data.ncaa_basketball.team_colors` AS c
JOIN
`bigquery-public-data.ncaa_basketball.mbb_teams` AS t
ON c.code_ncaa = t.code_ncaa WHERE UPPER(SUBSTR(c.color, 2, 2)) = 'FF'
ORDER BY
t.market;
```

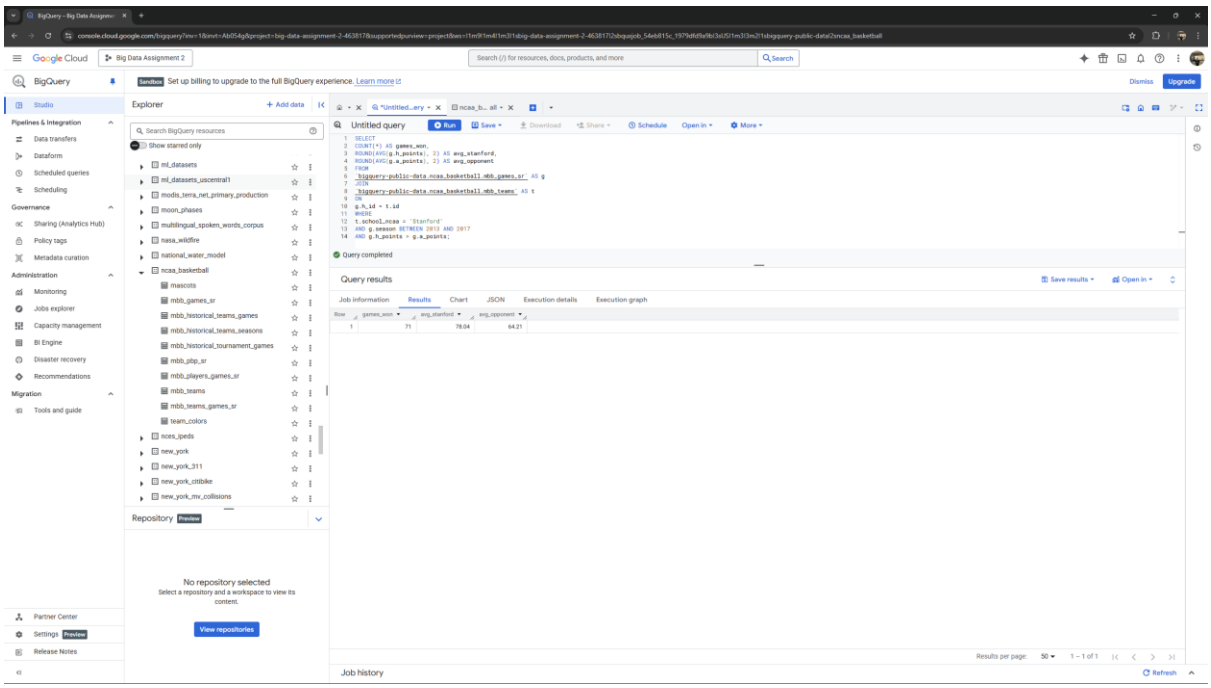
Output -



Question 4 : 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.

```
SELECT
COUNT(*) AS games_won,
ROUND(AVG(g.h_points), 2) AS avg_stanford,
ROUND(AVG(g.a_points), 2) AS avg_opponent
FROM
`bigquery-public-data.ncaa_basketball.mbb_games_sr` AS g
JOIN
`bigquery-public-data.ncaa_basketball.mbb_teams` AS t
ON
g.h_id = t.id
WHERE
t.school_ncaa = 'Stanford'
AND g.season BETWEEN 2013 AND 2017
AND g.h_points > g.a_points;
```

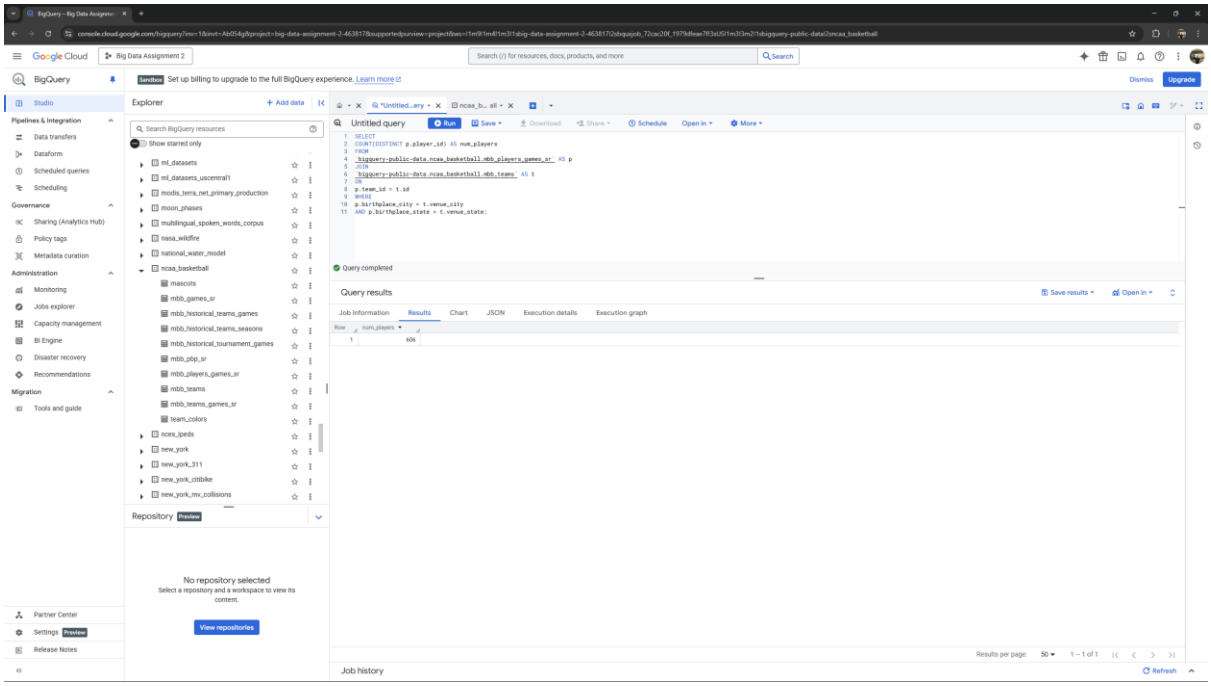
Output -



Question 5 : 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 p.player_id) AS num_players
FROM
`bigquery-public-data.ncaa_basketball.mbb_players_games_sr` AS p
JOIN
`bigquery-public-data.ncaa_basketball.mbb_teams` AS t
ON
p.team_id = t.id
WHERE
p.birthplace_city = t.venue_city
AND p.birthplace_state = t.venue_state;
```

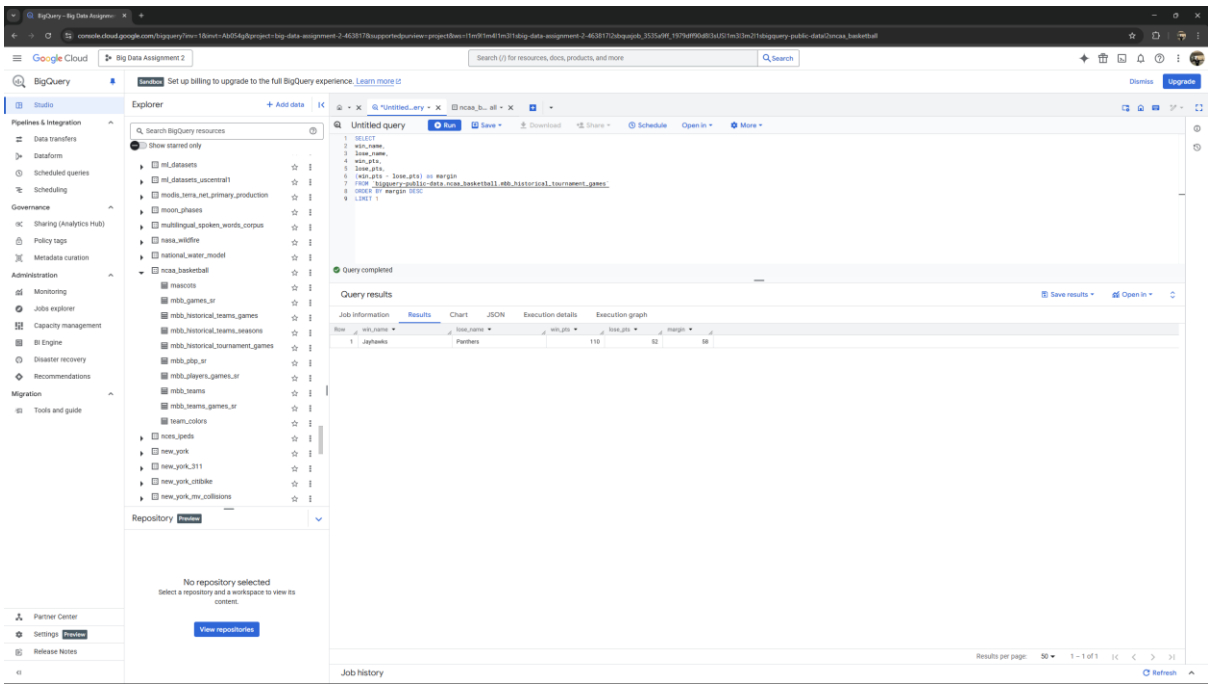
Output -



Question 6 : 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,
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 1
```

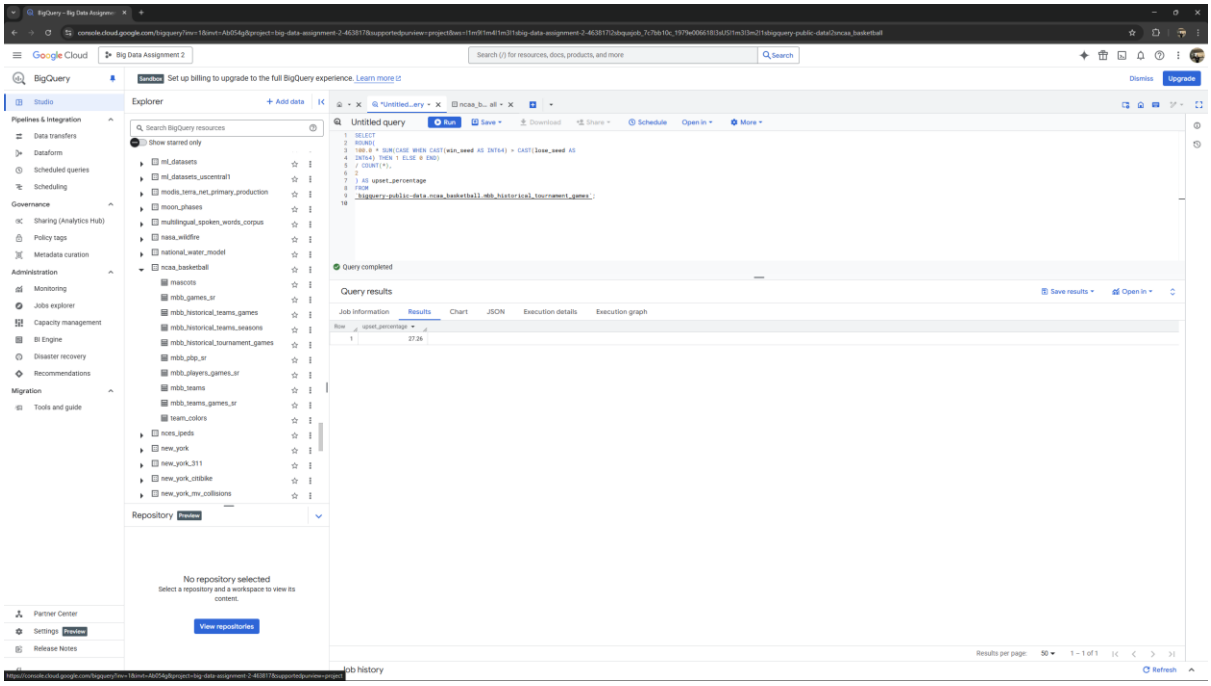
Output -



Question 7 : 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.

```
SELECT
ROUND(
100.0 * SUM(CASE WHEN CAST(win_seed AS INT64) > CAST(lose_seed AS
INT64) THEN 1 ELSE 0 END)
/ COUNT(*),
2
) AS upset_percentage
FROM
`bigquery-public-data.ncaa_basketball.mbb_historical_tournament_games`;
```

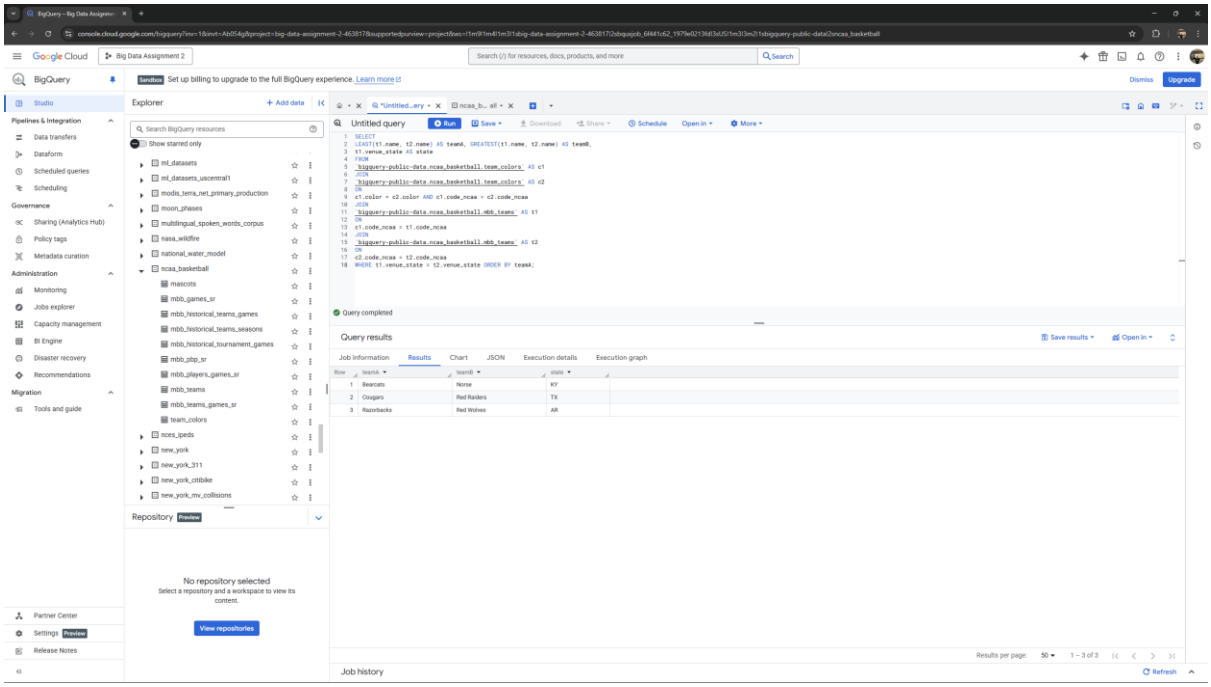
Output -



Question 8 : 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.

```
SELECT
LEAST(t1.name, t2.name) AS teamA, GREATEST(t1.name, t2.name) AS teamB,
t1.venue_state AS state
FROM
`bigquery-public-data.ncaa_basketball.team_colors` AS c1
JOIN
`bigquery-public-data.ncaa_basketball.team_colors` AS c2
ON
c1.color = c2.color AND c1.code_ncaa < c2.code_ncaa
JOIN
`bigquery-public-data.ncaa_basketball.mbb_teams` AS t1
ON
c1.code_ncaa = t1.code_ncaa
JOIN
`bigquery-public-data.ncaa_basketball.mbb_teams` AS t2
ON
c2.code_ncaa = t2.code_ncaa
WHERE t1.venue_state = t2.venue_state ORDER BY teamA;
```

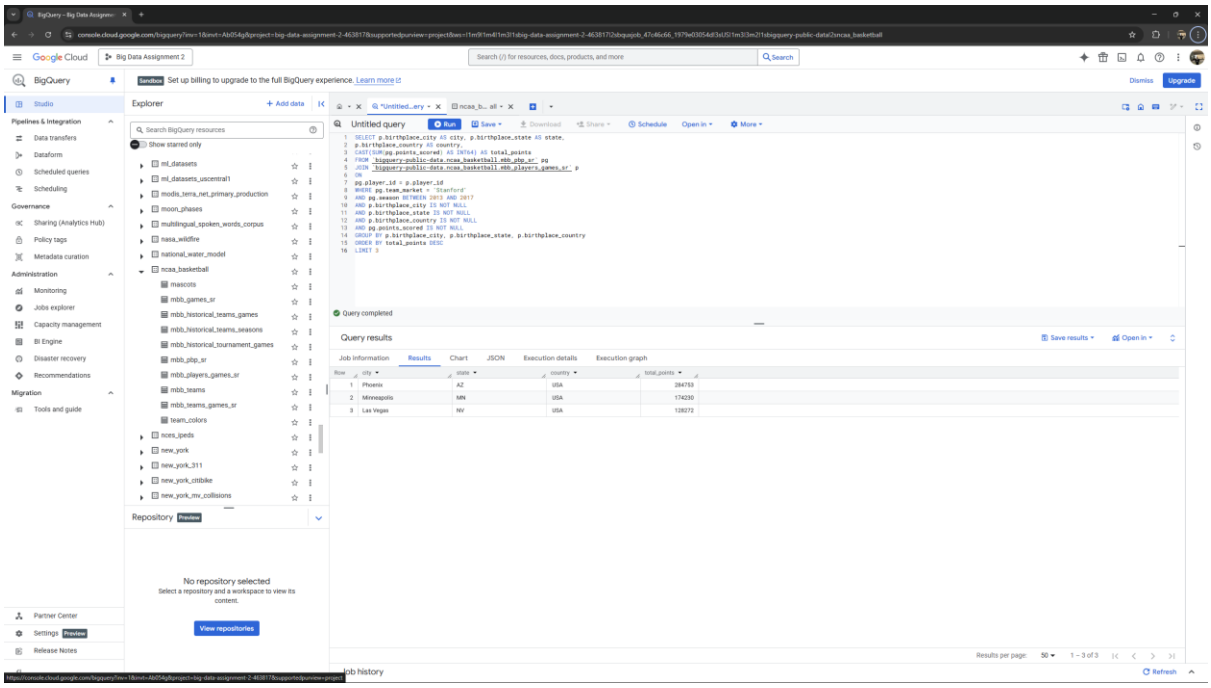
Output -



Question 9 : A geographical location L "makes" points for a team T whenever a player that was born in L scores points for T. (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?

```
SELECT p.birthplace_city AS city, p.birthplace_state AS state,
p.birthplace_country AS country,
CAST(SUM(pg.points_scored) AS INT64) AS total_points
FROM `bigquery-public-data.ncaa_basketball.mbb_pbp_sr` pg
JOIN `bigquery-public-data.ncaa_basketball.mbb_players_games_sr` p
ON
pg.player_id = p.player_id
WHERE pg.team_market = 'Stanford'
AND pg.season BETWEEN 2013 AND 2017
AND p.birthplace_city IS NOT NULL
AND p.birthplace_state IS NOT NULL
AND p.birthplace_country IS NOT NULL
AND pg.points_scored IS NOT NULL
GROUP BY p.birthplace_city, p.birthplace_state, p.birthplace_country
ORDER BY total_points DESC
LIMIT 3
```

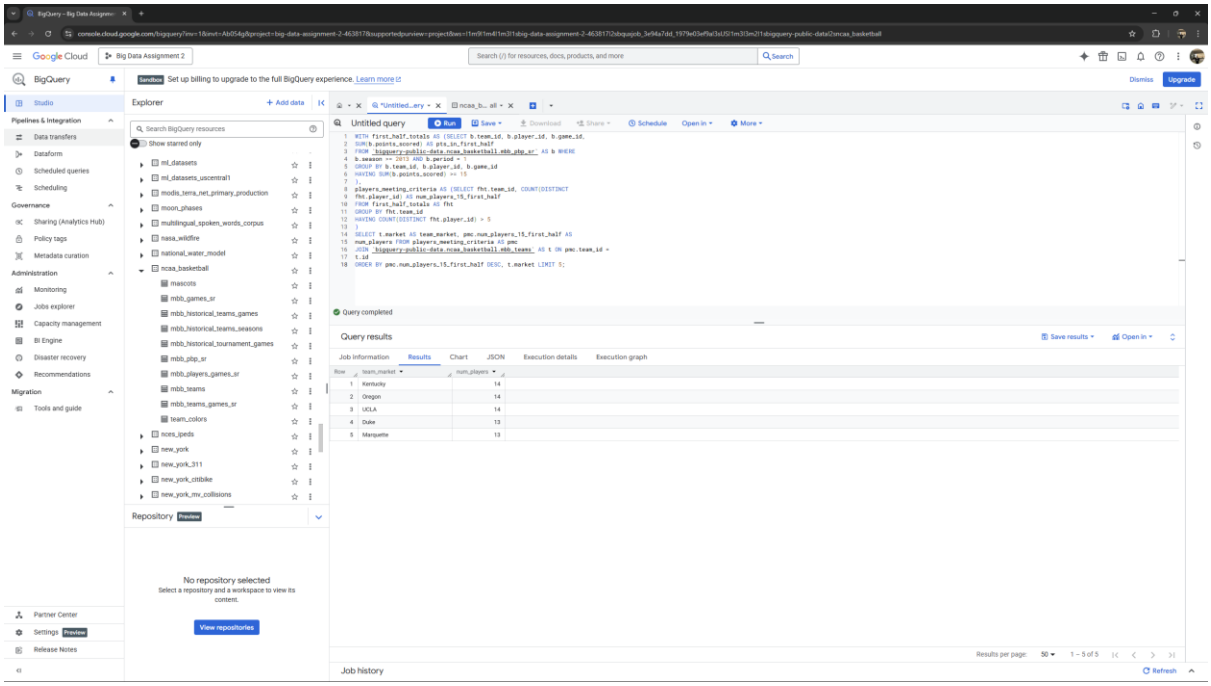
Output -



Question 10 : 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.

```
WITH first_half_totals AS (SELECT b.team_id, b.player_id, b.game_id,
SUM(b.points_scored) AS pts_in_first_half
FROM `bigquery-public-data.ncaa_basketball.mbb_pbp_sr` AS b WHERE
b.season >= 2013 AND b.period = 1
GROUP BY b.team_id, b.player_id, b.game_id
HAVING SUM(b.points_scored) >= 15
),
players_meeting_criteria AS (SELECT fht.team_id, COUNT(DISTINCT
fht.player_id) AS num_players_15_first_half
FROM first_half_totals AS fht
GROUP BY fht.team_id
HAVING COUNT(DISTINCT fht.player_id) > 5
)
SELECT t.market AS team_market, pmc.num_players_15_first_half AS
num_players FROM players_meeting_criteria AS pmc
JOIN `bigquery-public-data.ncaa_basketball.mbb_teams` AS t ON pmc.team_id =
t.id
ORDER BY pmc.num_players_15_first_half DESC, t.market LIMIT 5;
```

Output -



Question 11 : 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. (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.

```
WITH season_leaders AS ( SELECT market, season, wins, RANK() OVER
(PARTITION BY season ORDER BY wins DESC) as rank
FROM
`bigquery-public-data.ncaa_basketball.mbb_historical_teams_seasons`
WHERE season BETWEEN 1900 AND 2000 AND market IS NOT NULL AND wins IS
NOT NULL
)
SELECT market as team_market, COUNT(*) as top_performer_count
FROM season_leaders WHERE rank = 1
GROUP BY market
ORDER BY top_performer_count DESC, market ASC
LIMIT 5
```

Output -