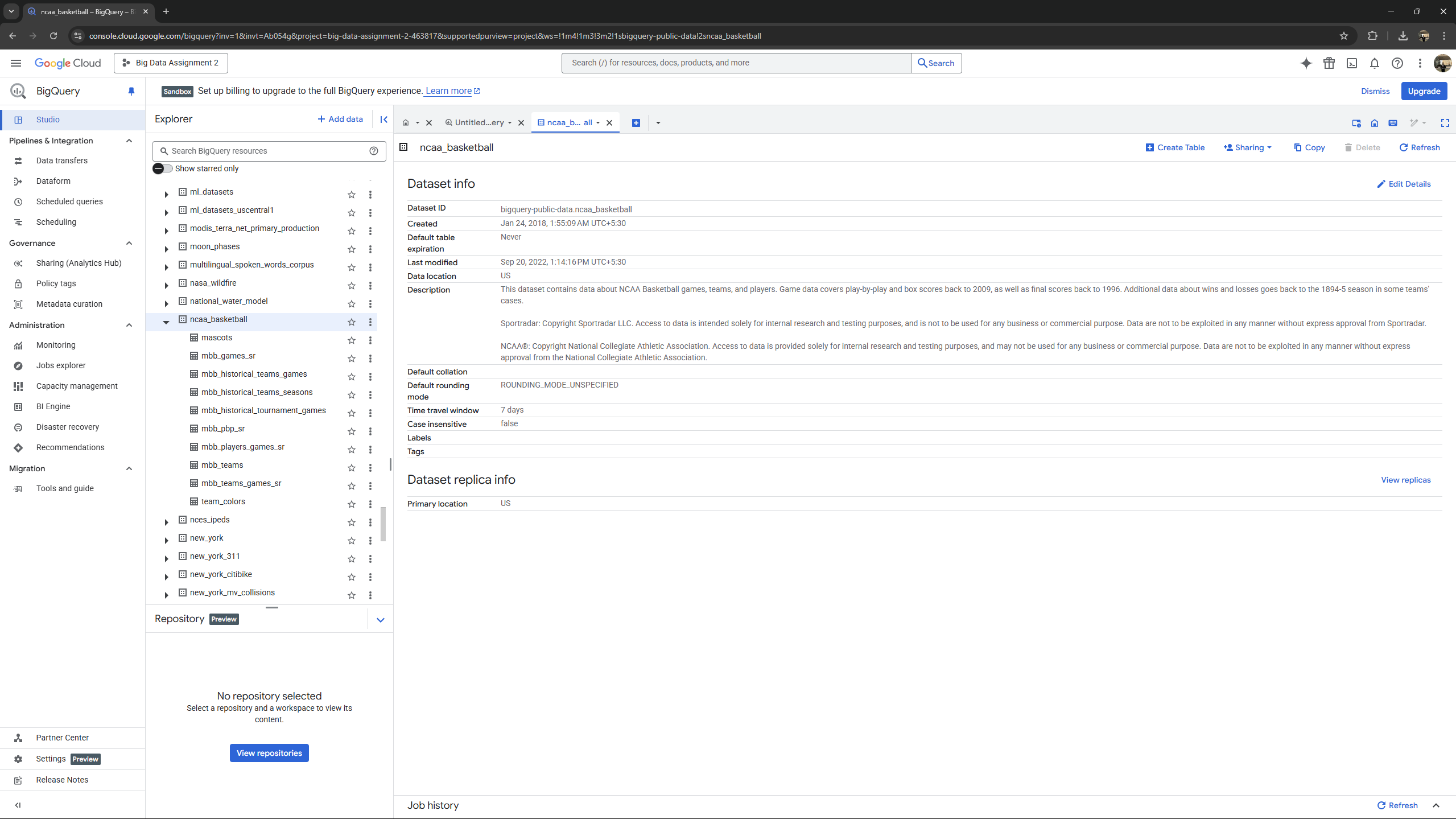
**Big Data Management - Assignment 2**

**NCAA Basketball Dataset**

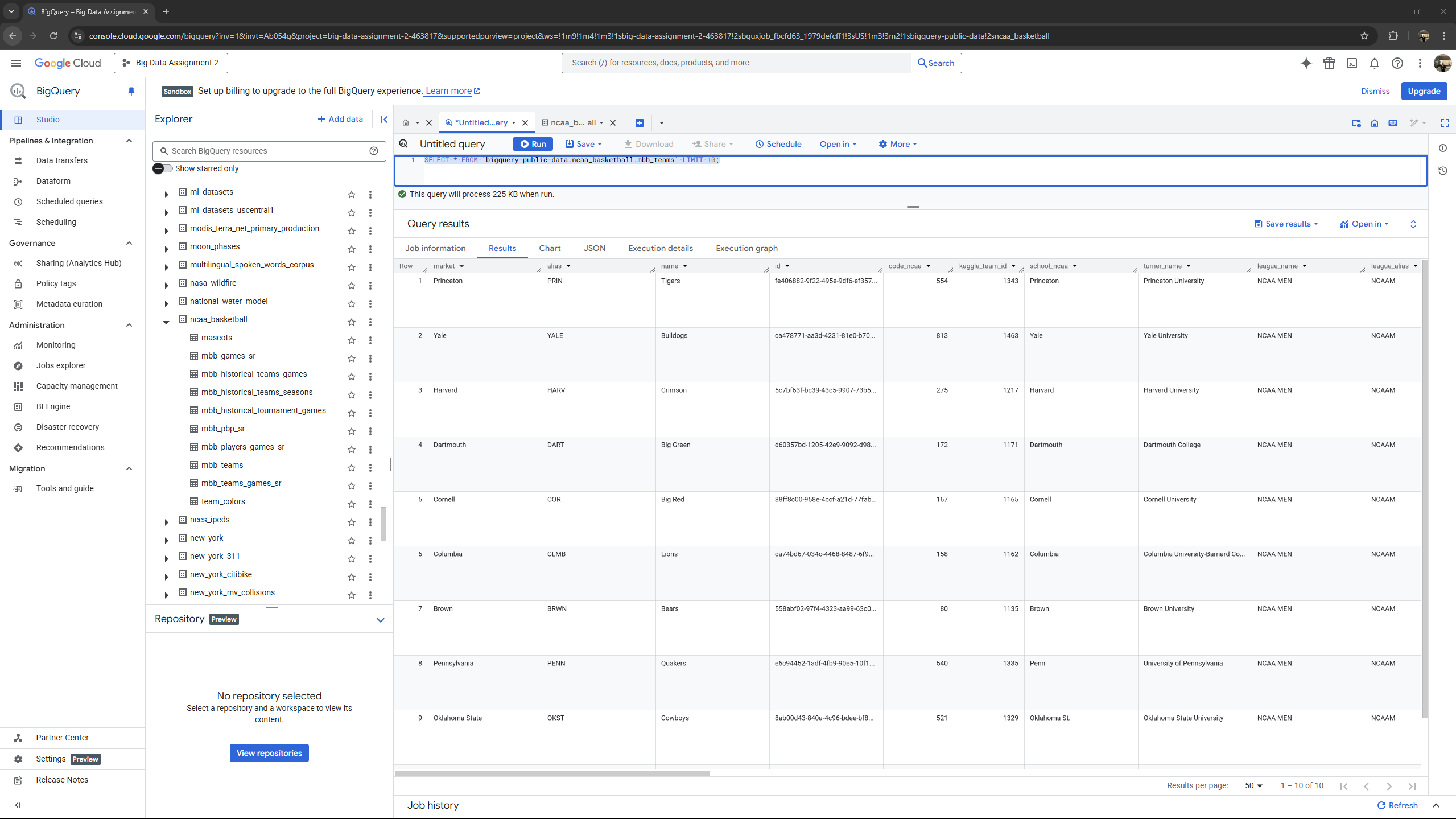
**Akshay Kumar (G24AI1033)**

**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*

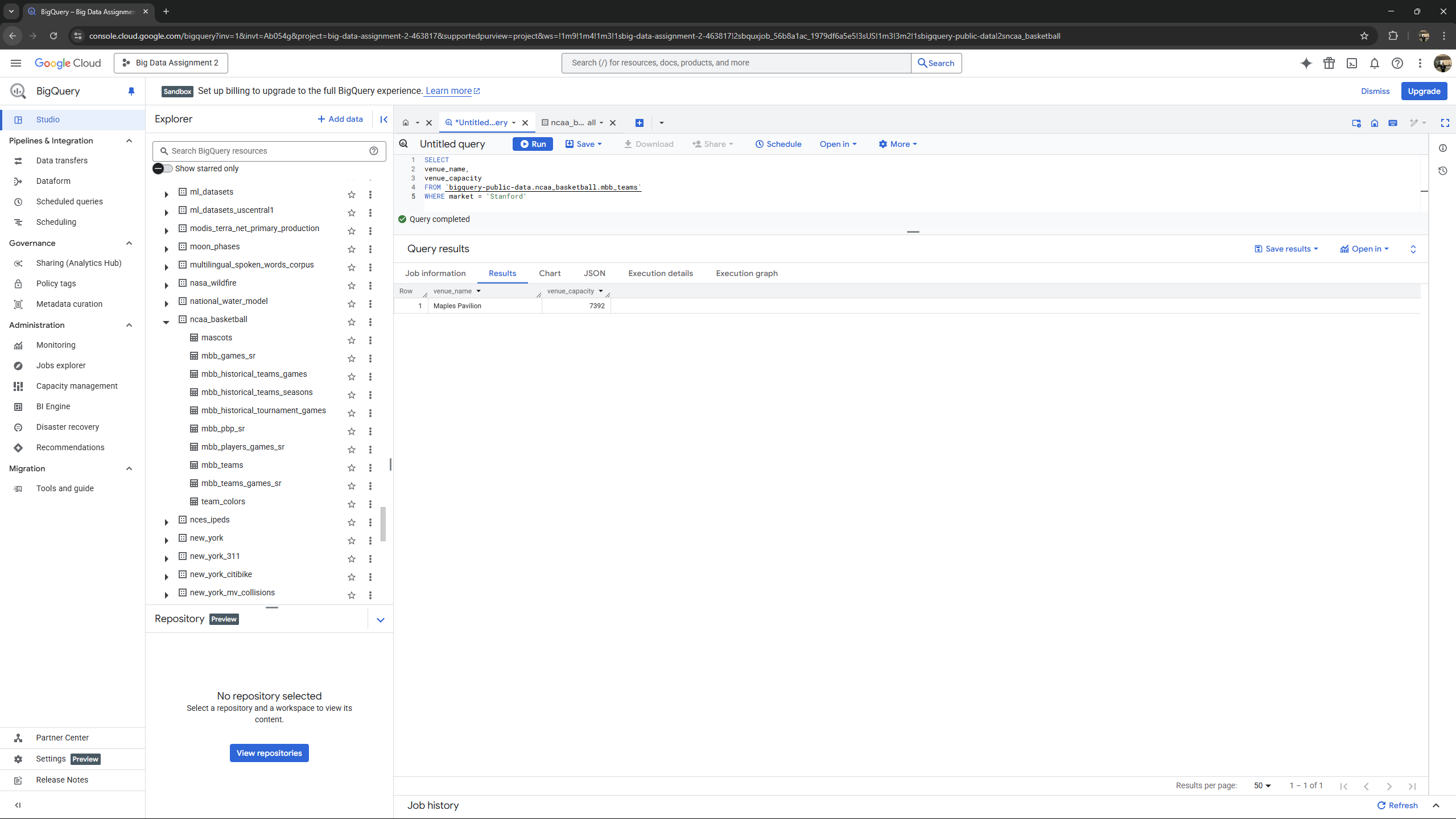
*venue\_name,*

*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*

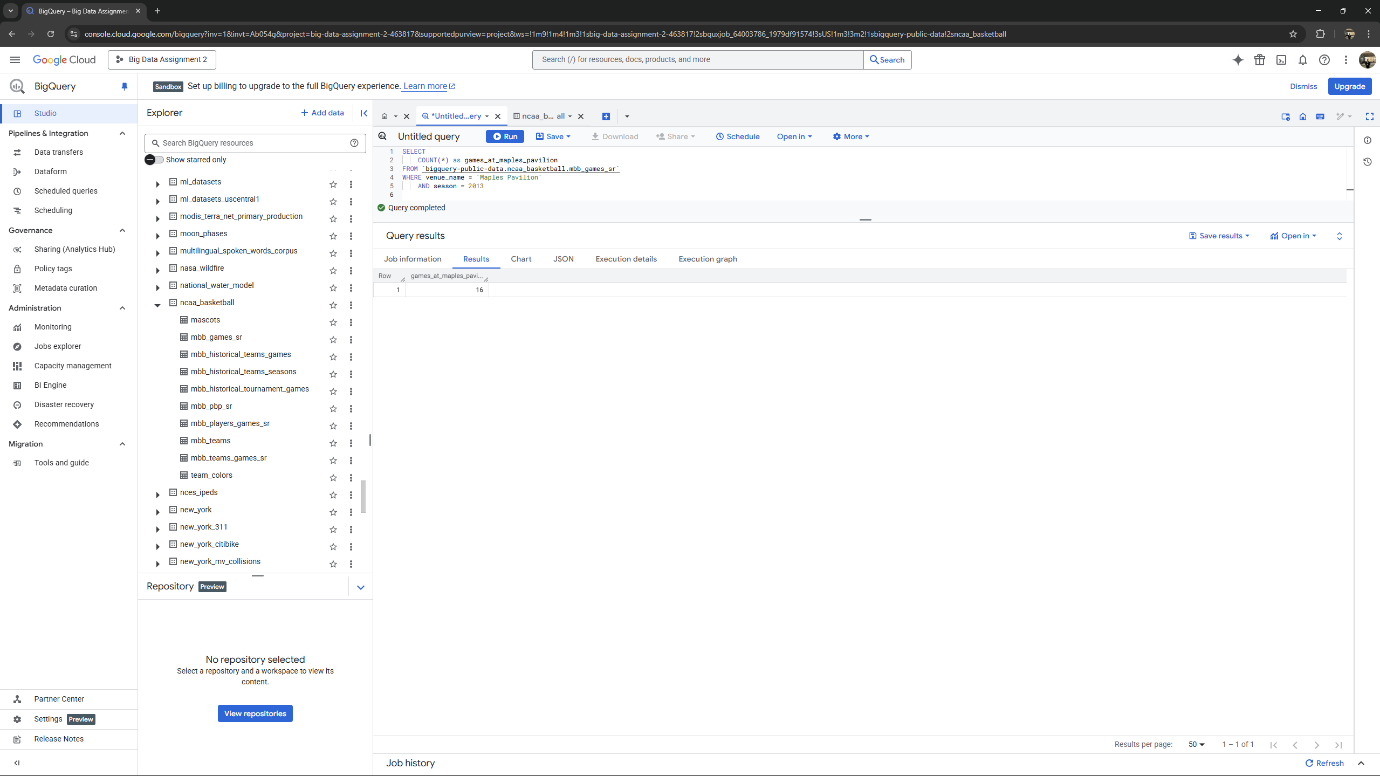
*COUNT(\*) as games\_at\_maples\_pavilion*

*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*

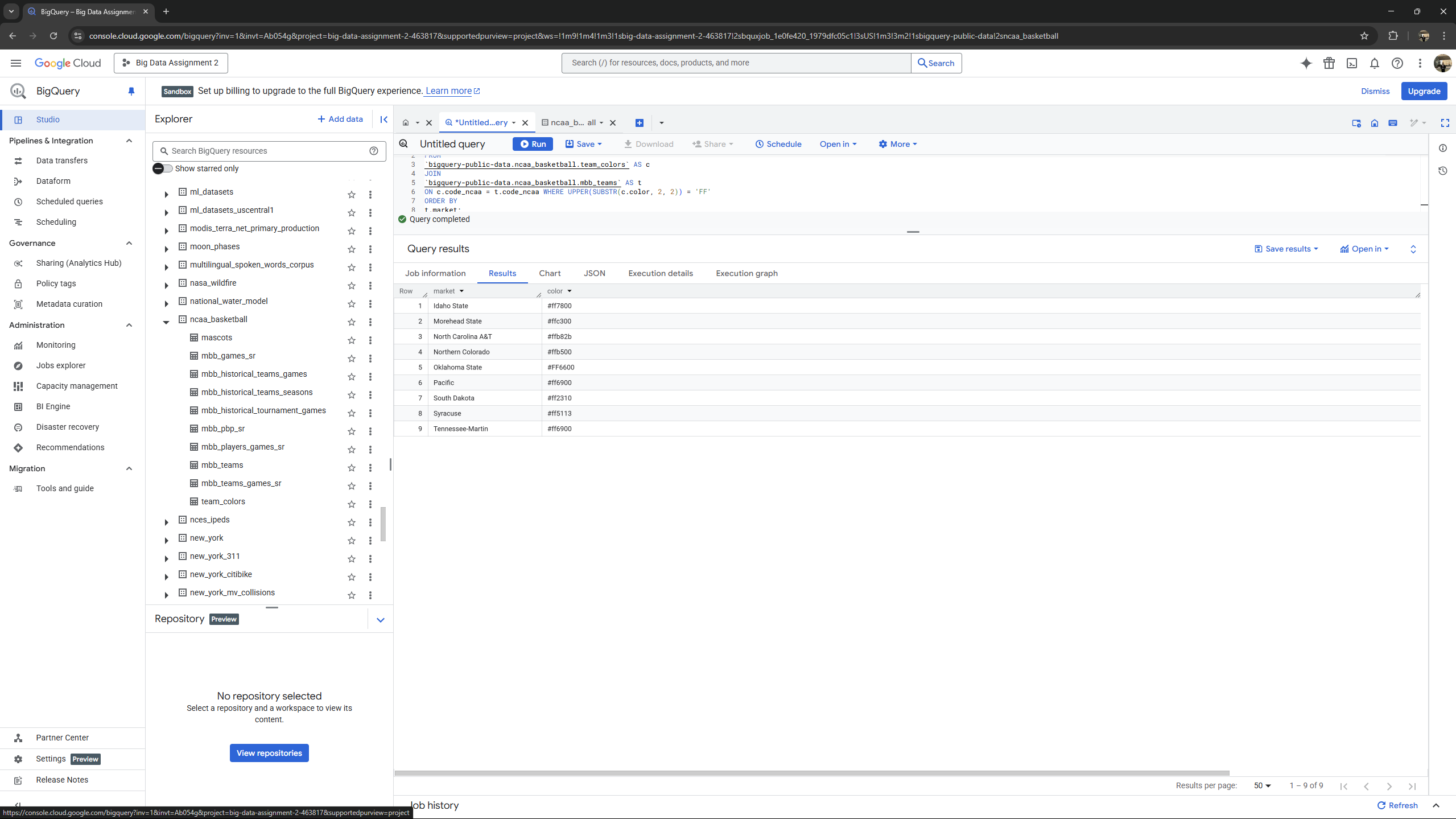
*`bigquery-public-data.ncaa\_basketball.mbb\_teams` AS t*

*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*

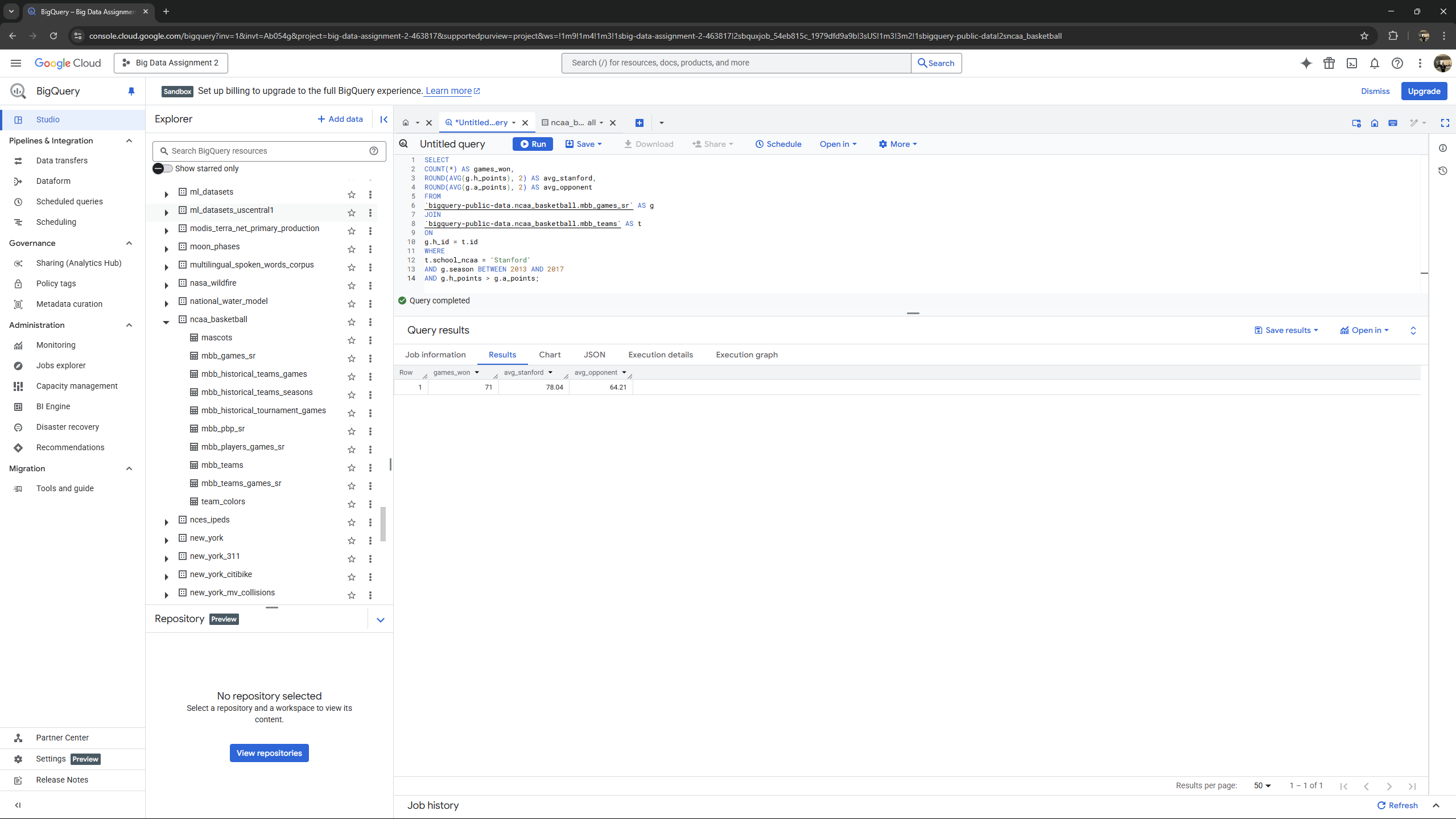
*WHERE*

*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*

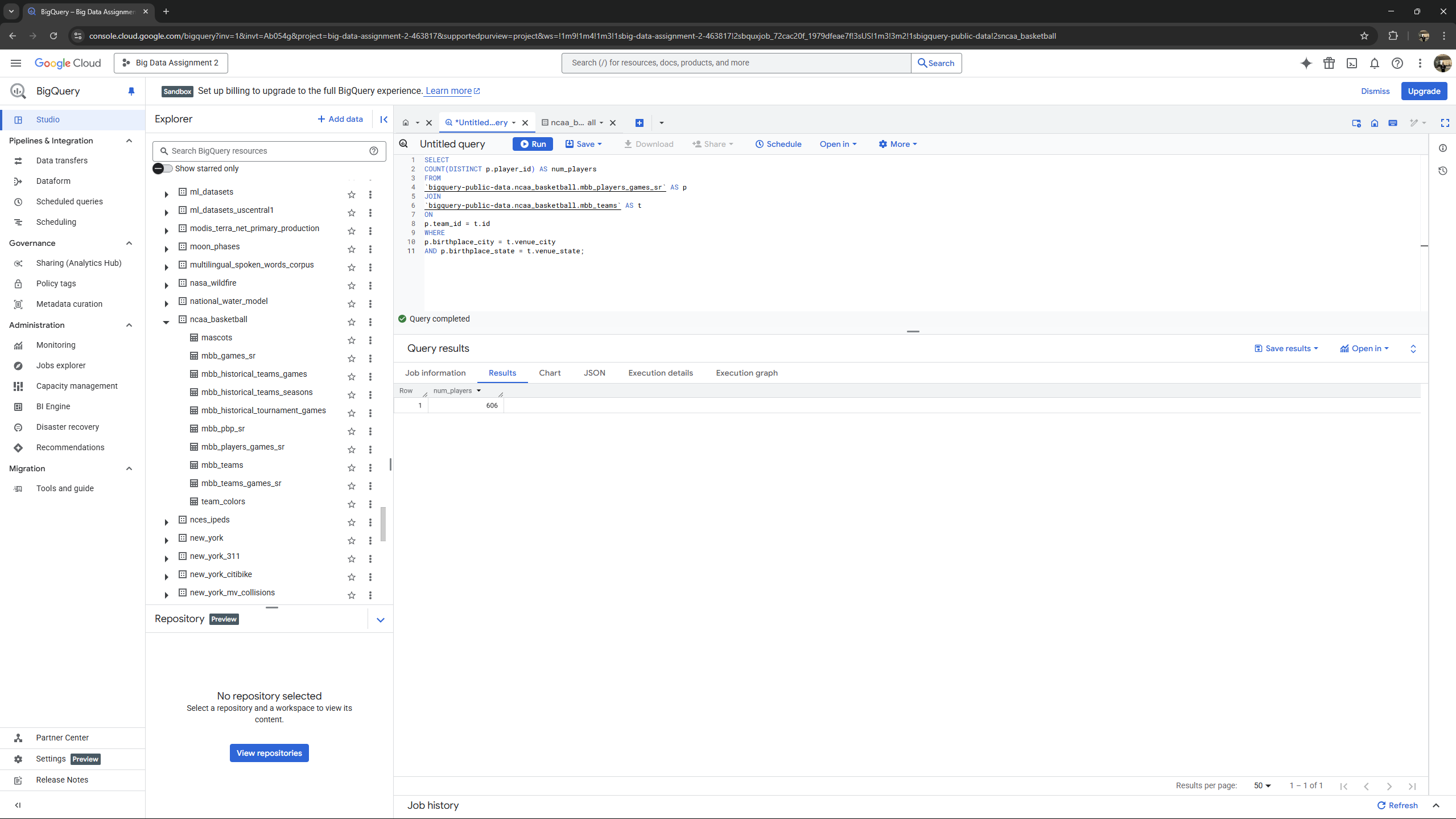
*p.team\_id = t.id*

*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,*

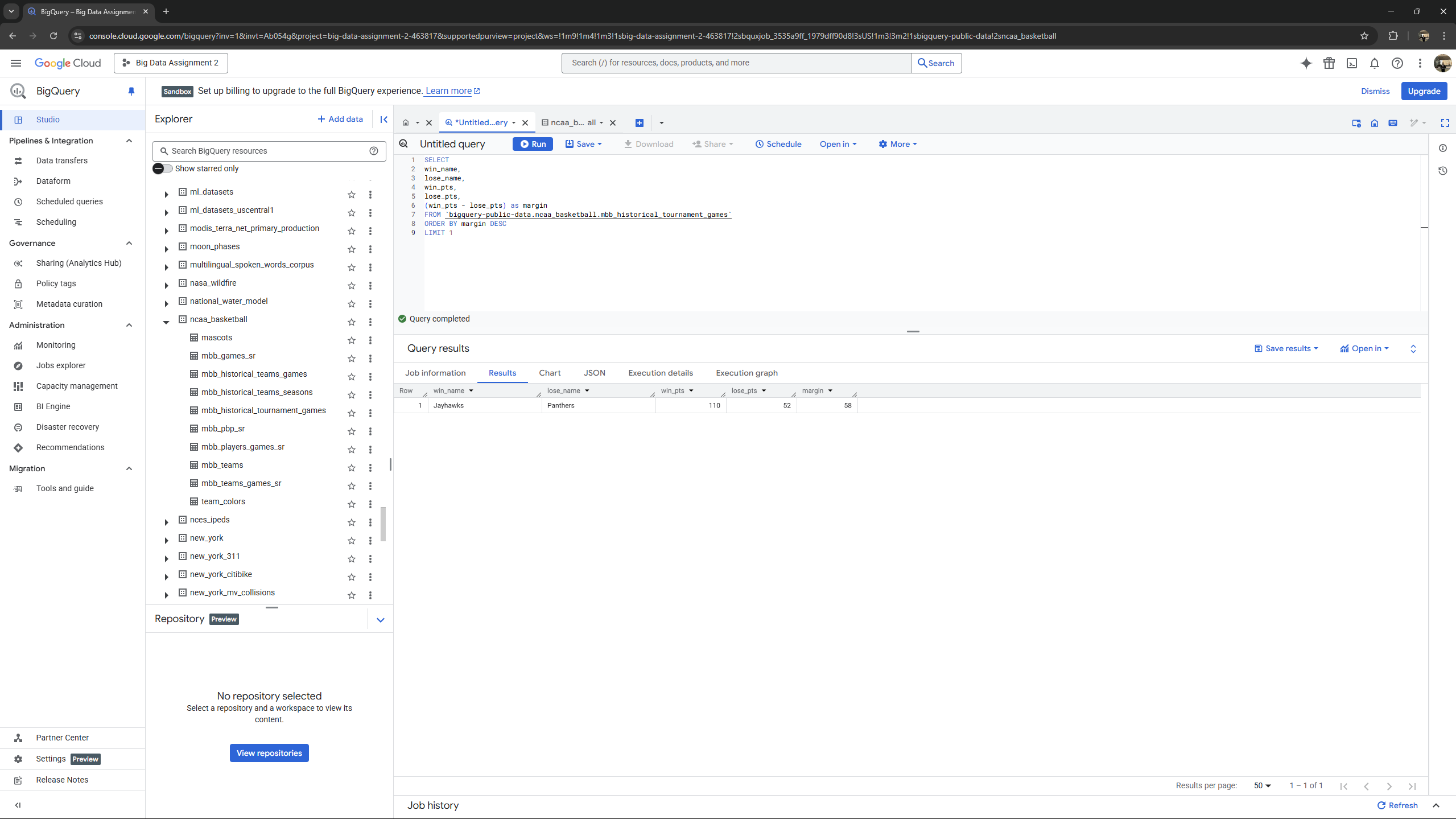
*(win\_pts - lose\_pts) as margin*

*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(\*),*

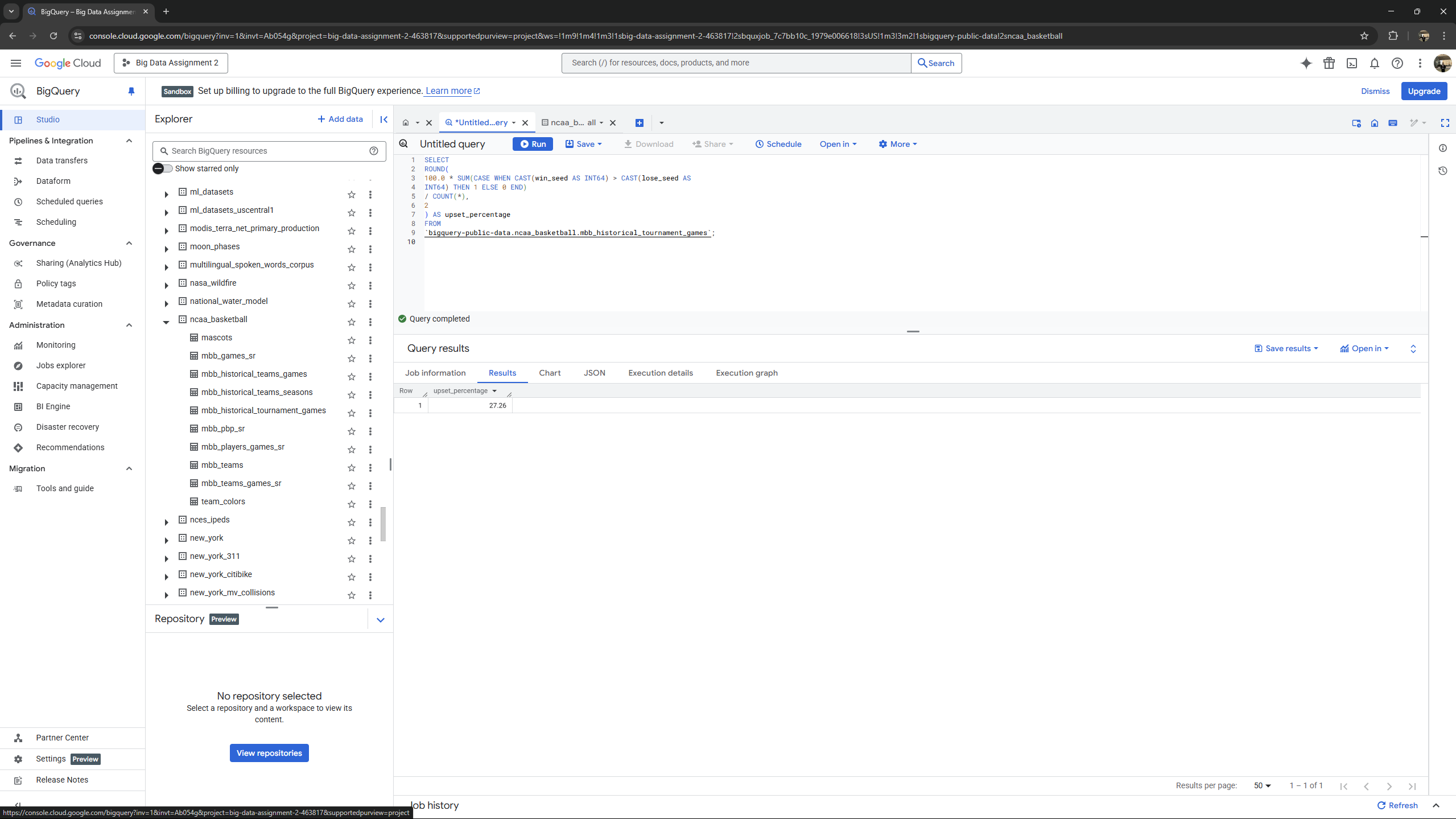
*2*

*) 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*

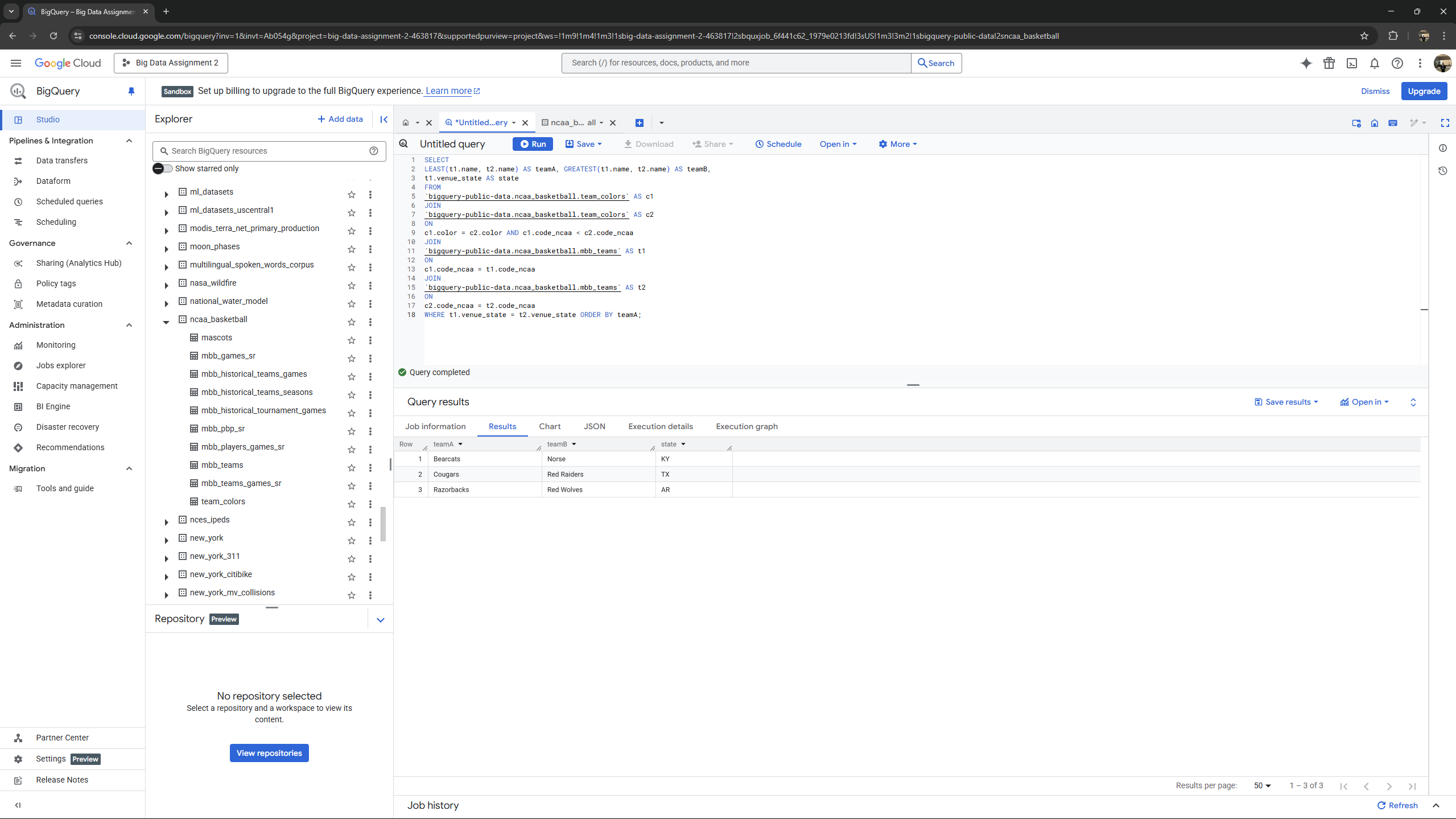
*`bigquery-public-data.ncaa\_basketball.mbb\_teams` AS t2*

*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*

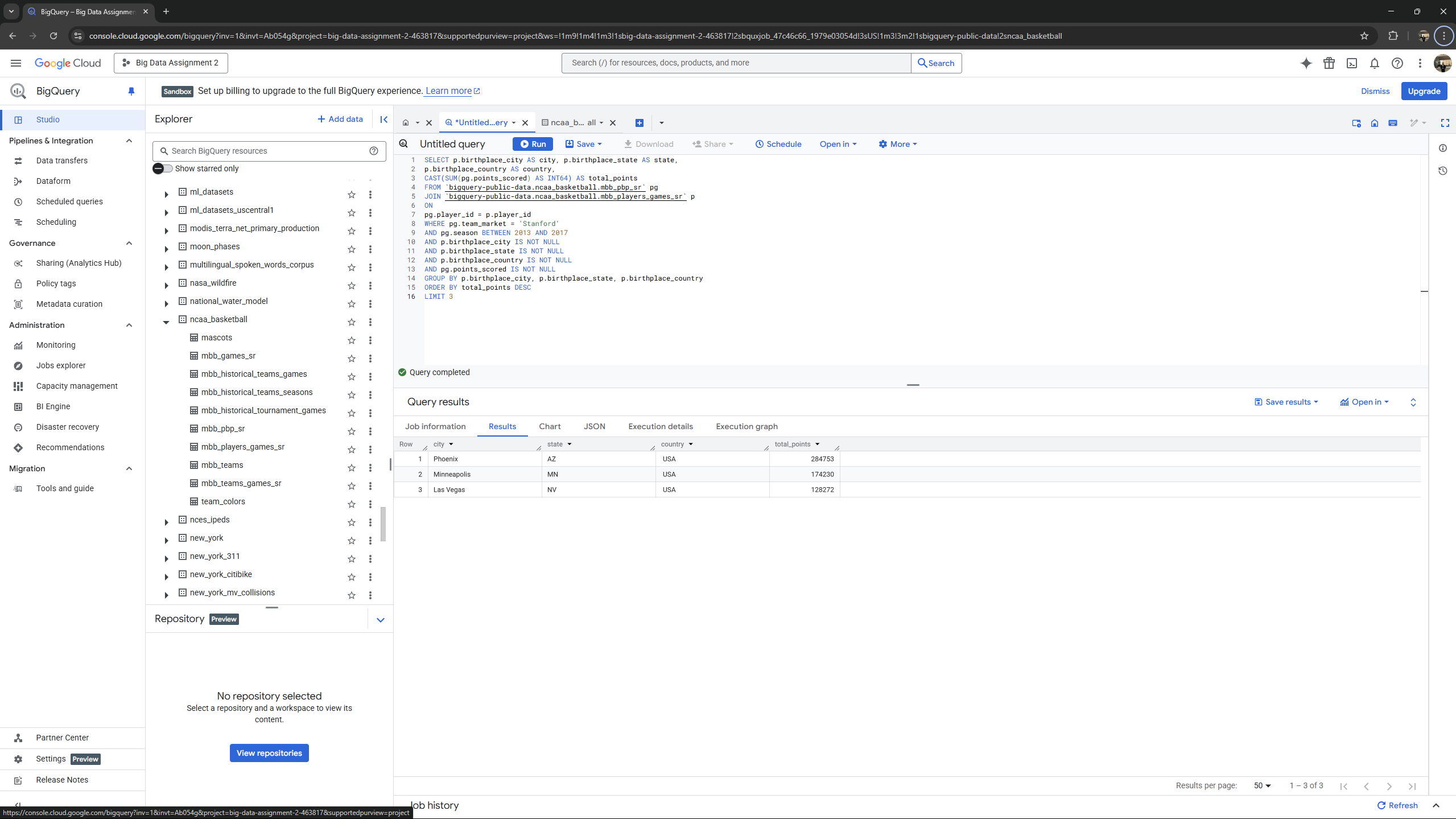
*AND pg.points\_scored IS NOT NULL*

*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*

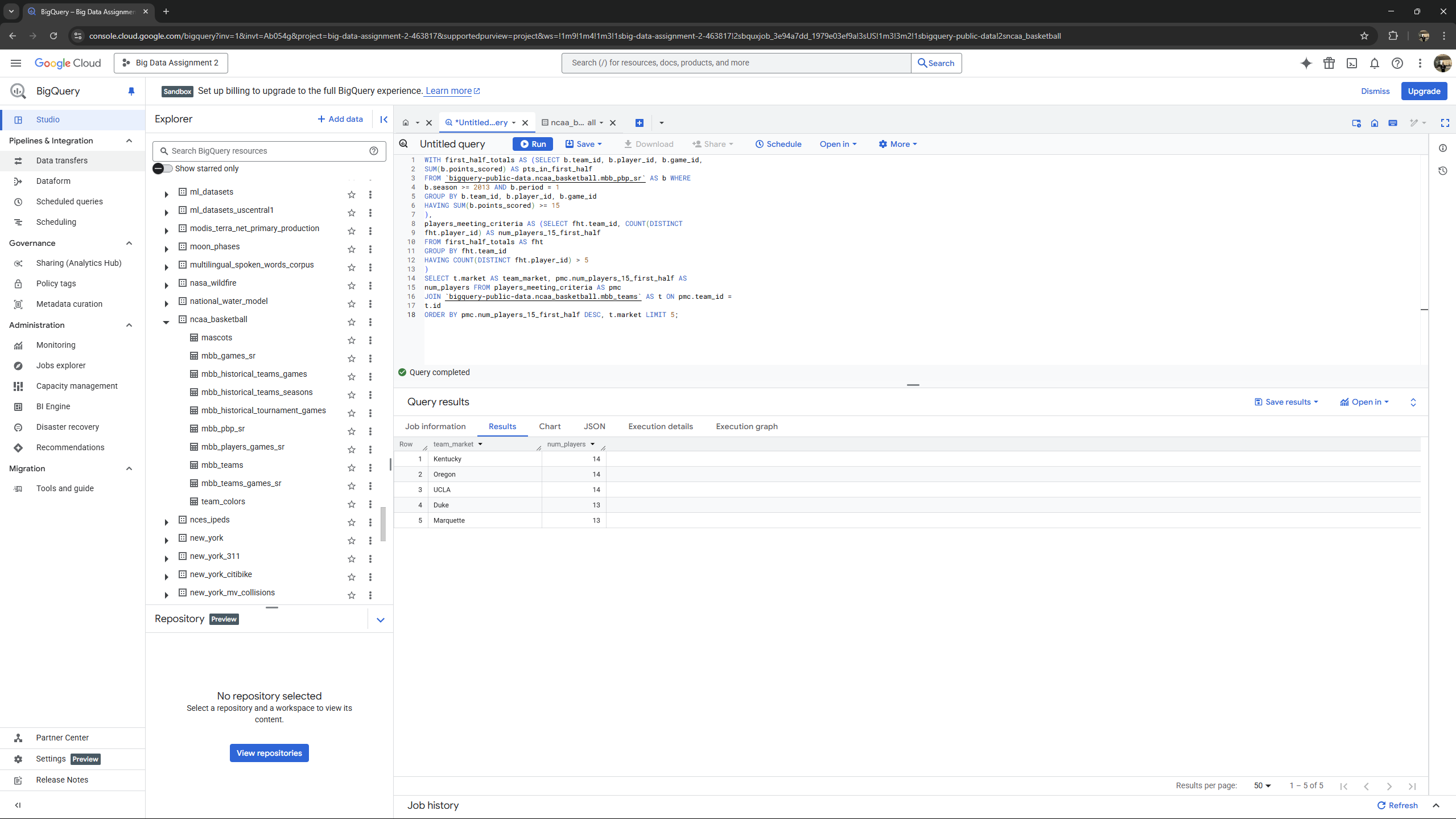
*num\_players FROM players\_meeting\_criteria AS pmc*

*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*

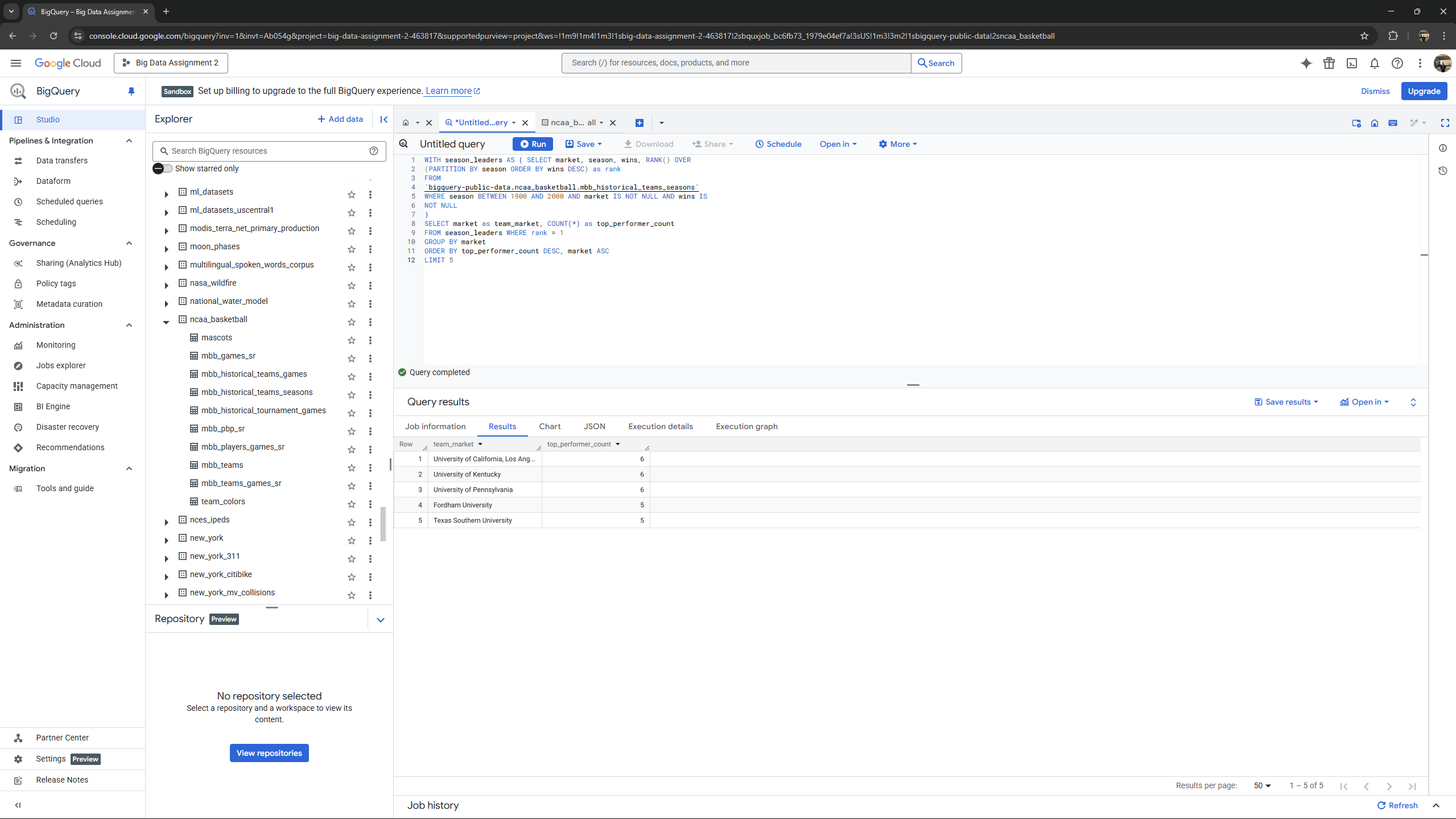
*FROM season\_leaders WHERE rank = 1*

*GROUP BY market*

*ORDER BY top\_performer\_count DESC, market ASC*

*LIMIT 5*

**Output -**

**