

Big Data Management - Assignment 2

NCAA Basketball Dataset

Akshay Kumar (G24AI1033)

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

The screenshot shows the Google Cloud BigQuery console interface. On the left, the 'Explorer' pane displays a tree view of datasets, with 'ncaa_basketball' selected. The main pane shows the 'Dataset info' for 'ncaa_basketball', including details like Dataset ID, Created date, Default table, Last modified, Data location, and Description. The description states that the dataset contains data about NCAA Basketball games, teams, and players, covering play-by-play and box scores back to 1996, as well as final scores back to 1996. Additional data about wins and losses goes back to the 1984-5 season in some teams' cases. The 'Dataset replica info' section shows the primary location as 'US'. The bottom pane shows 'Job history'.

Running a basic command to see the content of our tables

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

The screenshot shows the Google Cloud BigQuery console interface with a query executed. The query is `SELECT * FROM `bigquery-public-data.ncaa_basketball.mbb_teams` LIMIT 10;`. The 'Query results' pane displays a table with 10 rows and 12 columns. The columns are: Row, market, alias, name, US, mbb_teams, ncaa_basketball, school_name, league_name, and league_abbr. The results show the first 10 teams in the dataset.

Row	market	alias	name	US	mbb_teams	ncaa_basketball	school_name	league_name	league_abbr
1	Pittston	PBR	Tigers	6458822-922-450e-9b8e-ef057...	554	1342	Pittston University	NCAA MEN	NCAAM
2	Yale	YALE	Bulldogs	ea78771-5a3b-423f-81e5-8175...	813	1463	Yale University	NCAA MEN	NCAAM
3	Harvard	HARV	Crimson	5c79d32-6c39-4b5-9187-738d...	275	1217	Harvard University	NCAA MEN	NCAAM
4	Dartmouth	DART	Big Green	4b03576-1335-42e5-9b82-096...	172	1171	Dartmouth College	NCAA MEN	NCAAM
5	Cornell	CCR	Big Red	8876d03-938e-4ecf-af16-77ab...	147	1145	Cornell University	NCAA MEN	NCAAM
6	Columbia	CLUB	Lions	ca78a77-034b-4468-9487-07f...	158	1162	Columbia University Barnard Co...	NCAA MEN	NCAAM
7	Brown	BROWN	Bears	538a8f02-6716-4323-a678-63d5...	80	1135	Brown University	NCAA MEN	NCAAM
8	Pennsylvania	PENN	Quakers	4c03462d-1a7f-4839-9b45-10f1...	540	1235	University of Pennsylvania	NCAA MEN	NCAAM
9	Oklahoma State	OKST	Cowboys	84d05483-84b5-4c95-b08e-981...	521	1229	Oklahoma State University	NCAA MEN	NCAAM

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 -

The screenshot displays the Google Cloud BigQuery interface. On the left, the 'Explorer' pane shows a tree of datasets, with 'ncaa_basketball' expanded. The main area shows the query editor with the following SQL:

```
1 SELECT
2 venue_name,
3 venue_capacity
4 FROM `bigquery-public-data.ncaa_basketball.mbb_teams`
5 WHERE market = 'Stanford'
```

Below the query editor, a message indicates 'Query completed'. The 'Query results' section shows a table with the following data:

venue_name	venue_capacity
Maple Pavilion	7992

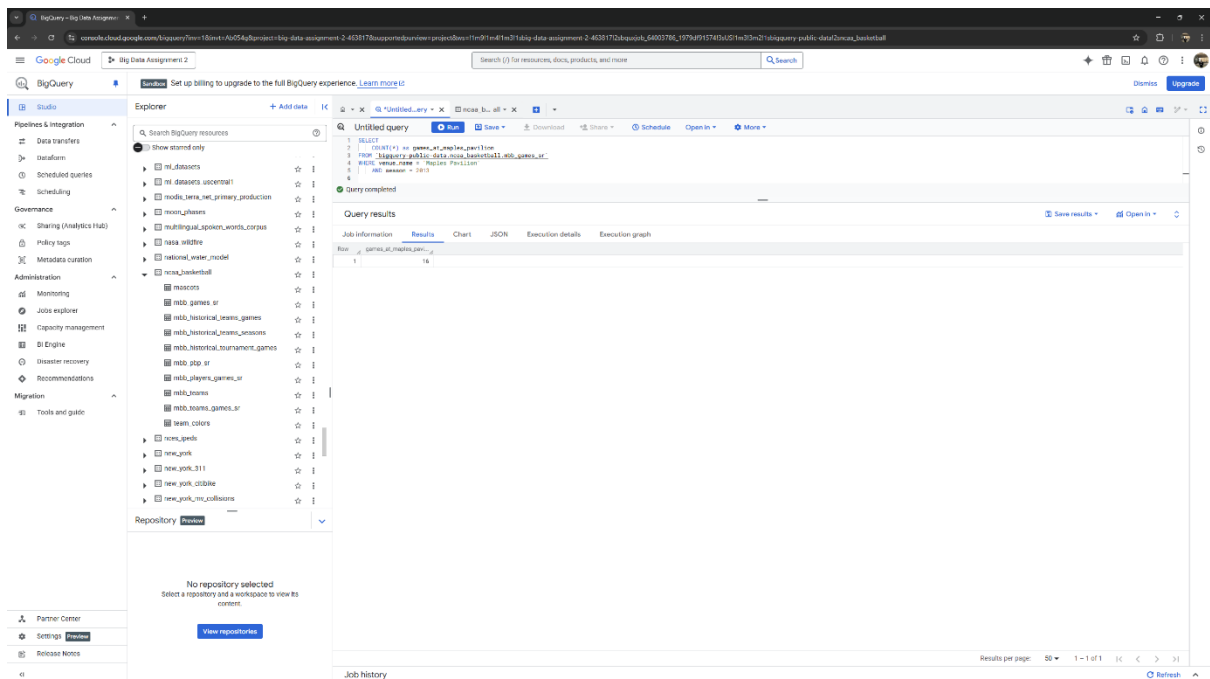
The bottom of the interface shows a 'Job history' section with a 'Refresh' button.

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 -



The screenshot shows the BigQuery console interface. On the left is a navigation menu with categories like Pipelines & Integration, Governance, Administration, and Migration. The main area is divided into an Explorer on the left and a query editor on the right. The Explorer shows a tree of datasets, with 'ncaa_basketball' expanded. The query editor contains the following SQL 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
```

Below the query editor, the 'Query results' section is visible, showing a table with one row and one column:

Row	games_at_maples_pavilion
1	16

At the bottom of the console, there is a 'Job history' section with a 'Refresh' button.

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 -

The screenshot shows the Google Cloud BigQuery Studio interface. On the left is a sidebar with navigation options like Pipelines & Integration, Data transfers, and Governance. The main area is divided into an Explorer on the left and a query editor on the right. The query editor contains a SQL query that filters for teams with a red color code (hex codes starting with 'FF'). Below the query editor, the 'Query results' section displays a table with 9 rows of data. The table has two columns: 'market' and 'color'. The results list various NCAA basketball teams and their corresponding market names and color codes.

market	color
1 Idaho State	#FF7040
2 Mississippi State	#FFC300
3 North Carolina A&T	#FF6628
4 Northern Colorado	#FF5020
5 Oklahoma State	#FF6600
6 Pacific	#FF6600
7 South Dakota	#FFC310
8 Spacow	#FF5110
9 Tennessee Martin	#FF6600

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 -

The screenshot shows the Google Cloud BigQuery console interface. On the left is a navigation menu with categories like Pipelines & Integration, Governance, Administration, and Migration. The main area is divided into three panes. The top pane shows the 'Explorer' with a search bar and a list of datasets, including 'ncaa_basketball'. The middle pane displays the SQL query that was executed, which is the same query as provided in the previous block. The bottom pane shows the 'Query results' table, which contains one row of data.

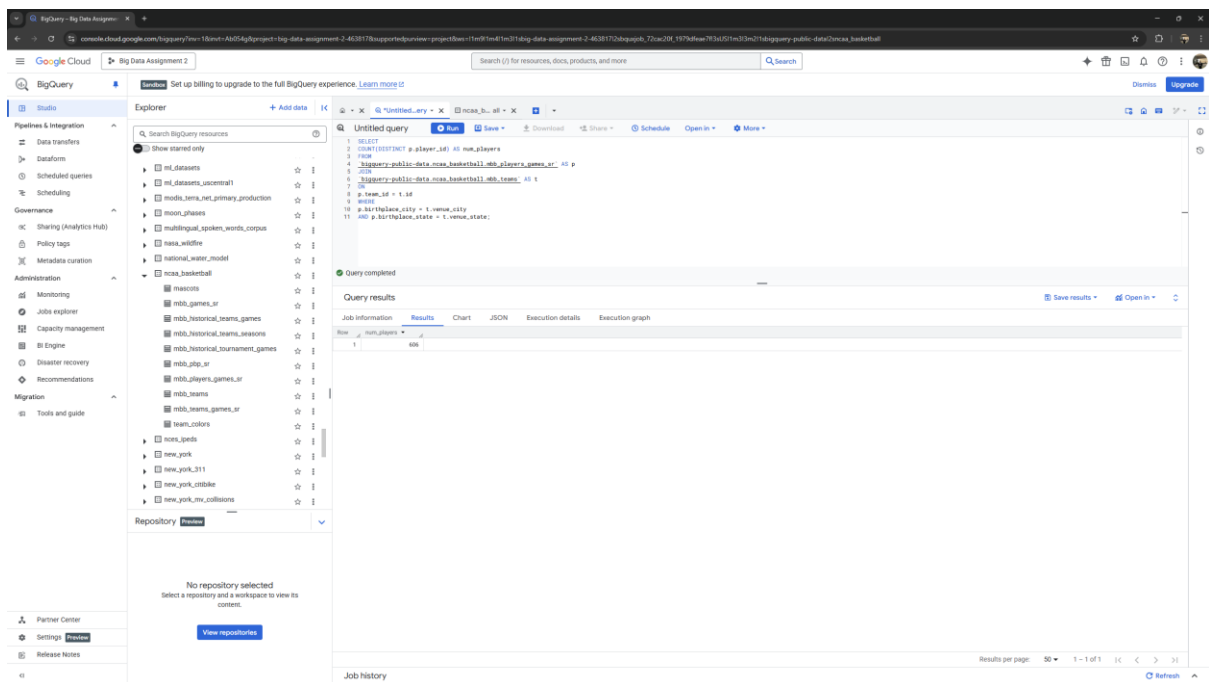
Job information	Results	Chart	JSON	Execution details	Execution graph
Row	games_won	avg_stanford	avg_opponent		
1	71	78.04	64.21		

At the bottom of the console, there is a 'Job history' section and a 'Results per page' dropdown set to 50. The status 'Query completed' is visible above the results table.

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 -



The screenshot shows the BigQuery console interface. On the left is a sidebar with navigation options like Pipelines & Integration, Governance, Administration, and Migration. The main area is divided into three sections: Explorer, Query Editor, and Query Results.

Explorer: Displays a tree view of BigQuery resources. The 'ncaa_basketball' dataset is expanded, showing tables like 'mbb_players_games_sr', 'mbb_historical_teams_seasons', and 'mbb_teams'.

Query Editor: Contains the SQL query from the previous block. The query is: `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;`

Query Results: Shows the execution results. The 'Results' tab is active, displaying a table with one row and one column, 'num_players', with a value of 406. The table has a header row and a data row.

num_players
406

At the bottom of the console, there is a 'Job history' section and a 'Results per page' dropdown set to 50.

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 -

The screenshot displays the Google Cloud BigQuery Studio interface. On the left, the 'Explorer' pane shows a tree of datasets, with 'ncaa_basketball' expanded. The main area shows an 'Untitled query' with the following SQL code:

```
1 SELECT
2 win_name,
3 lose_name,
4 win_pts,
5 lose_pts,
6 (win_pts - lose_pts) as margin
7 FROM `bigquery-public-data.ncaa_basketball.mbb_historical_tournament_games`
8 ORDER BY margin DESC
9 LIMIT 1
```

Below the query, a message states 'Query completed'. The 'Query results' section shows a table with the following data:

Row	win_name	lose_name	win_pts	lose_pts	margin
1	Japhania	Parthers	110	92	18

At the bottom right, the 'Job history' section is visible, and the 'Results per page' is set to 50.

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 -

The screenshot displays the Google Cloud BigQuery Studio interface. On the left, the 'Explorer' pane shows a tree view of available datasets, with 'ncaa_basketball' selected. The main 'Untitled query' editor contains the SQL query provided in the previous block. Below the editor, a message indicates 'Query completed'. The 'Query results' section shows a table with one row and one column, 'upset_percentage', with a value of 27.26. The bottom of the interface shows the 'Repository' section with a message 'No repository selected' and a 'View repositories' button.

Row	upset_percentage
1	27.26

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 -

The screenshot shows the Google Cloud BigQuery console. On the left is the 'Explorer' sidebar with a tree view of datasets. The main area displays an 'Untitled query' with the following SQL code:

```
1 SELECT
2 LEAST(t1.name, t2.name) AS teamA, GREATEST(t1.name, t2.name) AS teamB,
3 t1.venue_state AS state
4 FROM
5 `bigquery-public-data.ncaa_basketball.team_colors` AS c1
6 JOIN
7 `bigquery-public-data.ncaa_basketball.team_colors` AS c2
8 ON
9 c1.color = c2.color AND c1.code_ncaa < c2.code_ncaa
10 JOIN
11 `bigquery-public-data.ncaa_basketball.mbb_teams` AS t1
12 ON
13 c1.code_ncaa = t1.code_ncaa
14 JOIN
15 `bigquery-public-data.ncaa_basketball.mbb_teams` AS t2
16 ON
17 c2.code_ncaa = t2.code_ncaa
18 WHERE t1.venue_state = t2.venue_state ORDER BY teamA;
```

Below the query, the 'Query results' section shows a table with 3 columns: teamA, teamB, and state. The results are as follows:

Row	teamA	teamB	state
1	Swartz	None	KY
2	Chapman	Red Raiders	TX
3	Paetzels	Red Raiders	AR

The bottom of the console shows the 'Job history' section, which is currently empty.

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 -

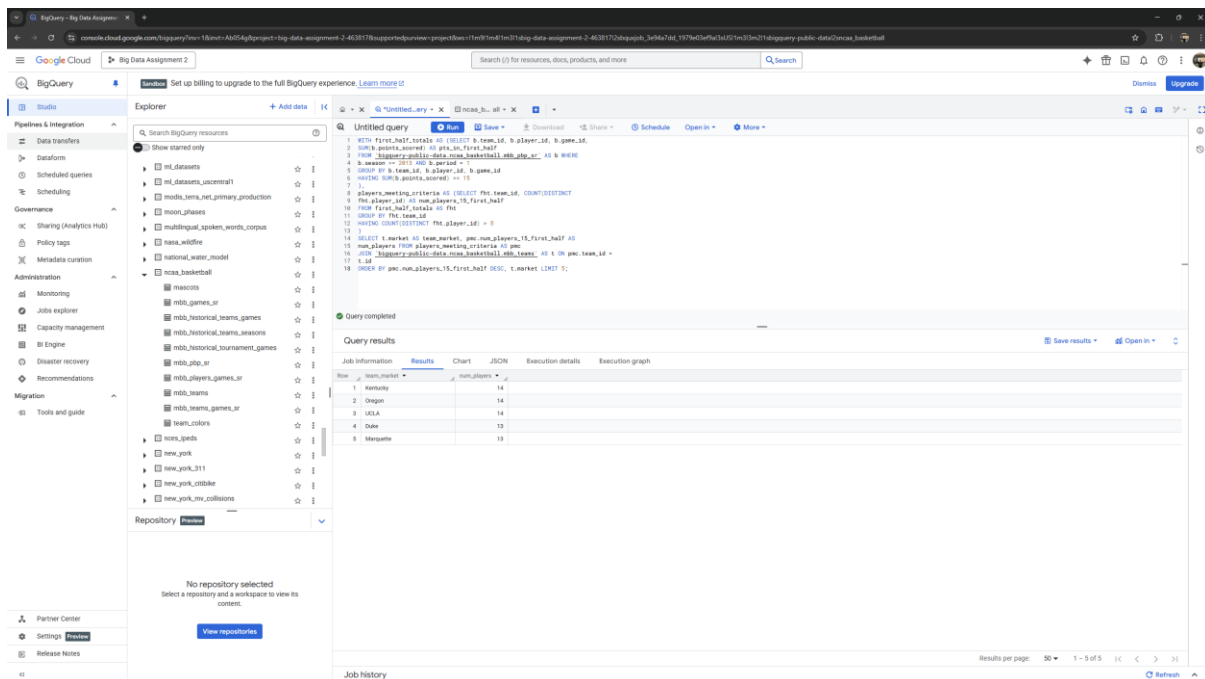
The screenshot shows the Google Cloud BigQuery console interface. The query editor on the right contains the SQL query from the previous block. Below the query, the 'Query results' section displays a table with 3 rows and 5 columns: city, state, country, and total_points. The results are sorted by total_points in descending order.

Row	city	state	country	total_points
1	Pasadena	CA	USA	28479
2	Minneapolis	MN	USA	17430
3	Las Vegas	NV	USA	13272

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 -



The screenshot shows the BigQuery web interface. On the left is a sidebar with navigation options like 'Pipelines & Integration', 'Data transfers', 'Governance', 'Administration', and 'Migration'. The main area is divided into three panes. The top pane shows the 'Untitled query' with the SQL code from the previous block. The bottom-left pane shows a 'Repository' section with a message 'No repository selected'. The bottom-right pane shows the 'Query results' table.

team_market	num_players
Tennessee	14
Oregon	14
UCLA	14
Duke	13
Marquette	13

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 -

The screenshot shows the Google Cloud BigQuery console. The query editor on the right contains the SQL query from the previous block. The query results are displayed in a table below the query editor.

Row	team_market	top_performer_count
1	University of California, Los Ang.	6
2	University of Kentucky	6
3	University of Pennsylvania	6
4	Fordham University	5
5	Texas Southern University	5

The left sidebar shows the BigQuery Explorer with a tree view of datasets. The 'ncaa_basketball' dataset is expanded, showing various tables like 'mbb_games_sr', 'mbb_historical_teams_seasons', etc. The bottom of the console shows the 'Repository' section with a message 'No repository selected' and a 'View repositories' button.