

PROJECT: NBA DATASET CLEANING AND EXPORT TO CSV USING BIGQUERY SQL.

INTRODUCTION:

In this task, we performed data cleaning on the NBA dataset using BigQuery SQL. The cleaned dataset was then exported to a CSV file for further analysis or reporting.

DATASET OVERVIEW:

The original dataset contained player information such as Name, Team, College, Salary, Height, Weight, Age, and Position. Some rows had missing values and duplicate entries.

DATA CLEANING PROCESS:

STEP 1: Create a Dataset “nba_dataset”

STEP 2: Create a Table name as “Test_nba”

STEP 3: Give SQL Query to Cleaning the dataset

```
1  #head()
2  SELECT *
3  FROM `careful-synapse-461006-b5.test_dataset.test_nba`
4  LIMIT 5;
5  #tail()
6  SELECT *
7  FROM `careful-synapse-461006-b5.test_dataset.test_nba`
8  ORDER BY Name DESC
9  LIMIT 5;
10 #info()
11 SELECT
12     column_name,
13     data_type,
14     is_nullable
15 FROM `careful-synapse-461006-b5.test_dataset.INFORMATION_SCHEMA.COLUMNS`
16 WHERE table_name = 'test_nba';
17 #describe()
18 SELECT
19     COUNT(*) AS total_rows,
20     COUNT(Salary) AS salary_count,
21     MIN(SAFE_CAST(Salary AS FLOAT64)) AS min_salary,
22     MAX(SAFE_CAST(Salary AS FLOAT64)) AS max_salary,
23     AVG(SAFE_CAST(Salary AS FLOAT64)) AS avg_salary,
24     STDDEV(SAFE_CAST(Salary AS FLOAT64)) AS std_salary
25 FROM `careful-synapse-461006-b5.test_dataset.test_nba`;
26 #Null count per column
27 SELECT
```

```

28     COUNTIF(Name IS NULL OR TRIM(Name) = '') AS null_names,
29     COUNTIF(Team IS NULL OR TRIM(Team) = '') AS null_teams,
30     COUNTIF(College IS NULL OR TRIM(College) = '') AS null_colleges,
31     COUNTIF(SAFE_CAST(Salary AS FLOAT64) IS NULL) AS invalid_salaries
32 FROM `careful-synapse-461006-b5.test_dataset.test_nba`;
33 #Count of players by college()
34 SELECT College, COUNT(*) AS count
35 FROM `careful-synapse-461006-b5.test_dataset.test_nba`
36 GROUP BY College
37 ORDER BY count DESC;
38 #Duplicate()
39 SELECT Name, Team, College, Salary, COUNT(*) AS duplicate_count
40 FROM `careful-synapse-461006-b5.test_dataset.test_nba`
41 GROUP BY Name, Team, College, Salary
42 HAVING COUNT(*) > 1;
43 #Standardize Text Columns
44 SELECT
45     INITCAP(TRIM(Name)) AS Name,
46     INITCAP(TRIM(Team)) AS Team,
47     INITCAP(TRIM(College)) AS College,
48     SAFE_CAST(Salary AS FLOAT64) AS Salary
49 FROM `careful-synapse-461006-b5.test_dataset.test_nba`;
50 #Convert Salary to Numeric Format
51 SELECT
52     SAFE_CAST(Salary AS FLOAT64) AS salary_numeric
53 FROM `careful-synapse-461006-b5.test_dataset.test_nba`;
54
55
56
57
58

```

✓ Query completed

OUTPUT:

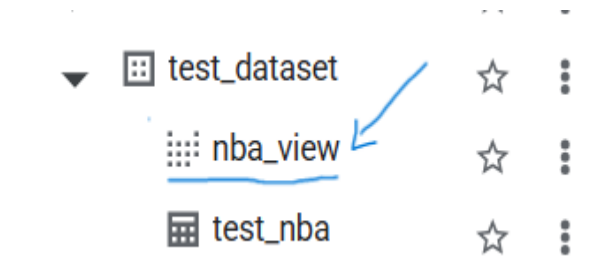
All results					
Elapsed time		Statements processed		Job status	
1 sec		4		✓ SUCCESS	
Status	End time	SQL	Stages completed	Bytes p	Action
✓	2:18PM [2:1]	SELECT *	0	0 B	View results
✓	2:18PM [6:1]	SELECT *	0	0 B	View results
✓	2:18PM [11:3]	SELECT	1	10 MB	View results
✓	2:18PM [18:1]	SELECT	2	3.48 KB	View results

STEP 4: Create View query to create view in table.

```
Untitled query [Run] [Save] [Download]

1 CREATE VIEW test_dataset.nba_view AS
2 SELECT *
3 FROM `careful-synapse-461006-b5.test_dataset.test_nba`
4 WHERE College IS NOT NULL
```

AFTER CREATED IT SHOULD BE:



STEP 5: VIEW THE “NBA_VIEW”:

The view should show the cleaned data, not the raw data, so we can easily write queries to explore the entire process.

- **CLICK THE PREVIEW**

Schema

Details

Table Explorer

Preview

Insights

Lineage

Data Profile

Data Quality


Filter

Enter property name or value

<input type="checkbox"/>	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags ?	Description
<input type="checkbox"/>	Name	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Team	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Number	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Position	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Age	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Height	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Weight	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	College	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Salary	INTEGER	NULLABLE	-	-	-	-	-


STEP 6: SELECT THE NECESSARY COLUMN TO EXPLORE:

Select Fields

 This script will process 36.43 KB when run.

Filter

 Enter property name or value

	Field name	Type
<input checked="" type="checkbox"/>	Name	STRING
<input checked="" type="checkbox"/>	Team	STRING
<input checked="" type="checkbox"/>	Number	INTEGER
<input type="checkbox"/>	Position	STRING
<input type="checkbox"/>	Age	INTEGER
<input checked="" type="checkbox"/>	Height	STRING
<input type="checkbox"/>	Weight	INTEGER
<input type="checkbox"/>	College	STRING
<input type="checkbox"/>	Salary	INTEGER


Save

Cancel

STEP7: Choose related column and click apply

Distinct Values

Select Fields

 This script will process 48.65 KB when run.


Reset

Apply

Name

<input type="checkbox"/>	Value	Count
<input type="checkbox"/>	Al Horford	1
<input type="checkbox"/>	Jahlil Okafor	1
<input type="checkbox"/>	Festus Ezeli	1

Team

	Value	Count
<input checked="" type="checkbox"/>	Memphis Grizzl...	17
<input checked="" type="checkbox"/>	New Orleans P...	16
<input type="checkbox"/>	Detroit Pistons	15

Number

<input type="checkbox"/>	Value	Count
<input type="checkbox"/>	1	19
<input type="checkbox"/>	0	18
<input type="checkbox"/>	5	18

Generated Query

Copy To Query

```
1 SELECT
2   "Height",
3   "Name",
4   "Number",
5   "Team"
6 FROM
7   "careful-synapse-461006-b5.test_dataset.nba_view"
8 WHERE
9   ("Team" IN ('Memphis Grizzlies'))
```

STEP 8: Click Copy the query to paste the query in execute page.

The screenshot shows a query editor interface. On the left, there is a 'Table Explorer' tab with a 'Select Fields' button. Below it, a table is displayed with columns 'Name' and 'Count'. The table contains the following data:

Name	Count
Johnson	1
Conley	1
JaMychal Green	1
Jrue Holiday	1
Ennis	1
Farmar	1

Below the table, there is a 'Copy To Query' button. A pink arrow points to this button. On the right, the 'Untitled query' editor shows the following SQL query:

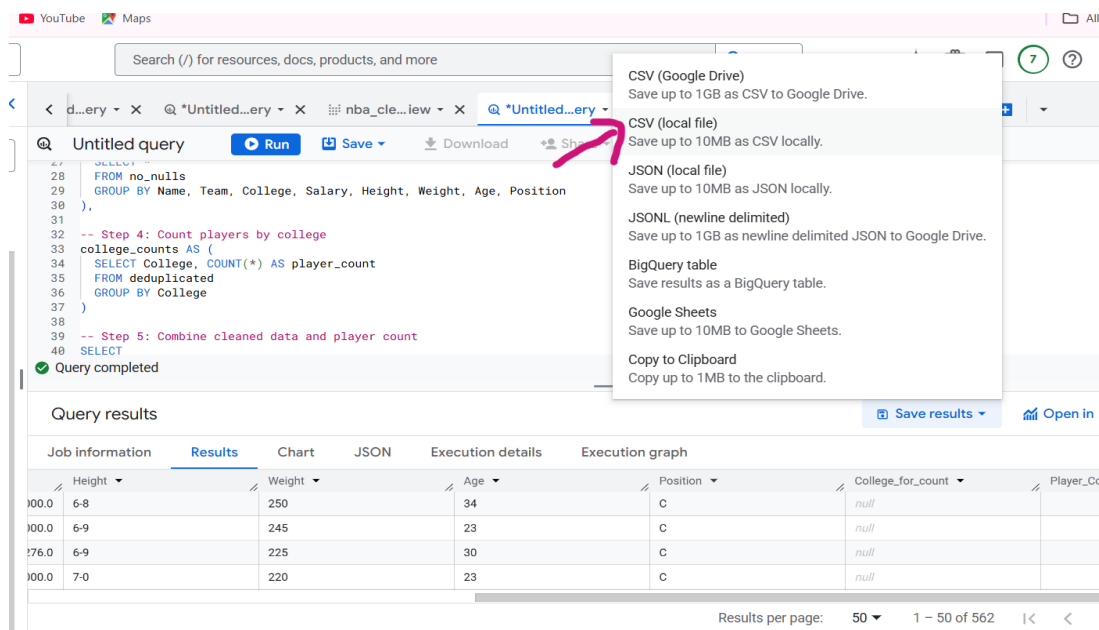
```
1 SELECT
2   `Height`,
3   `Name`,
4   `Number`,
5   `Team`
6 FROM
7   `careful-synapse-461006-b5.test_dataset.nba_view`
8 WHERE
9   (`Number` IN (0,
10    11,
11    4))
12 AND (`Team` IN ('Memphis Grizzlies',
13    'New Orleans Pelicans'));
```

STEP 9: Run the Query

The screenshot shows the query editor interface after running the query. The 'Run' button is highlighted. Below the query, a message indicates 'Query completed'. At the bottom, the 'Query results' section shows a table with columns 'Row', 'Height', 'Name', and 'Number'. The table contains the following data:

Row	Height	Name	Number
1	6-9	JaMychal Green	0
2	6-1	Mike Conley	11
3	6-2	Jordan Farmar	4
4	6-4	Jrue Holiday	11
5	6-7	James Ennis	4

STEP 10: Download the Cleaned data into csv format for future analysis.



VISUALIZE THE CLEANED DATA IN LOOKER STUDIO:

I. POSITION BY SALARY (DONUT CHART)

QUESTION:

Which player position receives the highest overall salary percentage?

ANSWER:

The **Point Guard (PG)** position receives the highest overall salary percentage at **22.2%**.

II. SALARY BY NAME (BAR CHART)

QUESTION:

Who is the highest-paid NBA player among the listed names?

ANSWER:

Carmelo Anthony has the highest salary among the listed NBA players, earning over \$20 million.

III. STAFF REPORT (TABLE)

QUESTION:

Name any two players listed in the Staff Report along with their college.

ANSWER:

- Al Horford – College

- Lavoy Allen – College

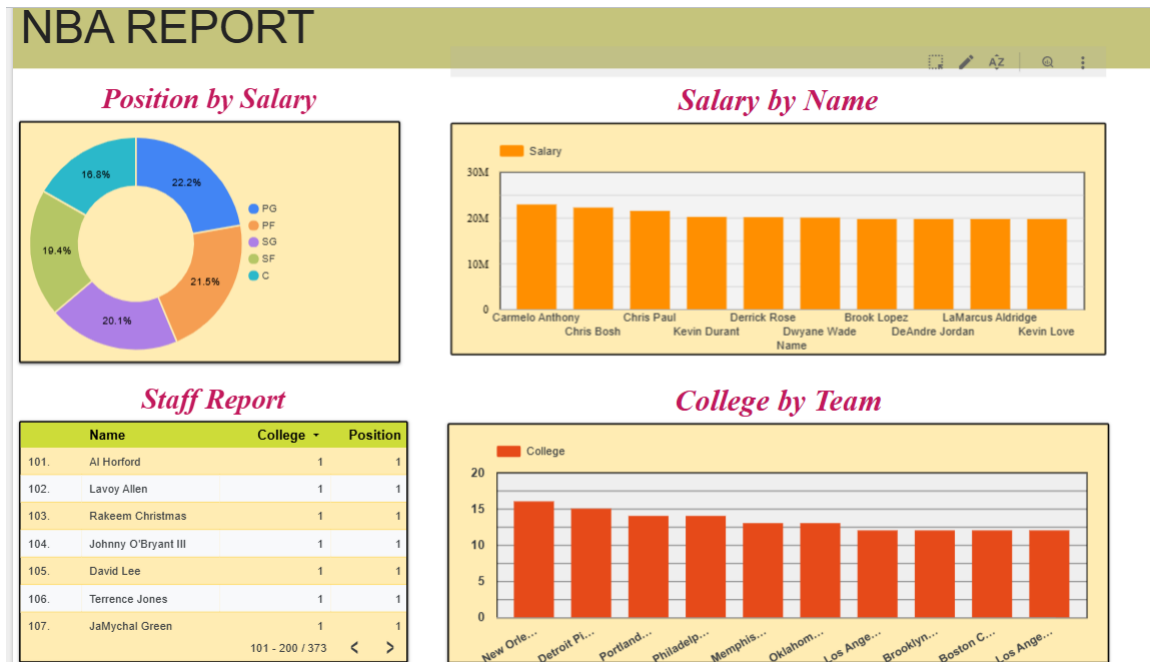
IV. COLLEGE BY TEAM (BAR CHART)

QUESTION:

Which NBA team has the highest number of players from colleges?

ANSWER:

New Orleans has the highest number of college players among the listed teams, with about **16–17 players**.



CONCLUSION:

In this project, we carried out a structured data analysis process on an NBA dataset to uncover meaningful insights. The entire workflow included data cleaning, transformation, and visualization, which helped simplify raw information into clear, interpretable patterns.