Homework #1

In this assignment, you will practice Data Modeling and writing analytical SQL statements:

- Part 1: Get started with Google BigQuery and write analytical SQL queries on Google Public datasets.

- Part 2: Create an Entity Relationship Diagram (ERD) for a set of 4 tables already defined on Google BigQuery.

To begin this assignment

- Please download the HW #1 Assignment Template from the HW #1 folder in Blackboard. You must use the template to submit this homework assignment.

Homework Questions

Part 1

1. If you have not already done so, follow the tutorial on Blackboard “Getting Started with Google BigQuery” to create a new project in Google BigQuery.

   Once complete, take a screenshot of the results of any query you write. For example:

```
1  SELECT *
2  FROM nba_data.nba_mvps
3  WHERE age > 30
4  ORDER BY age DESC;
```

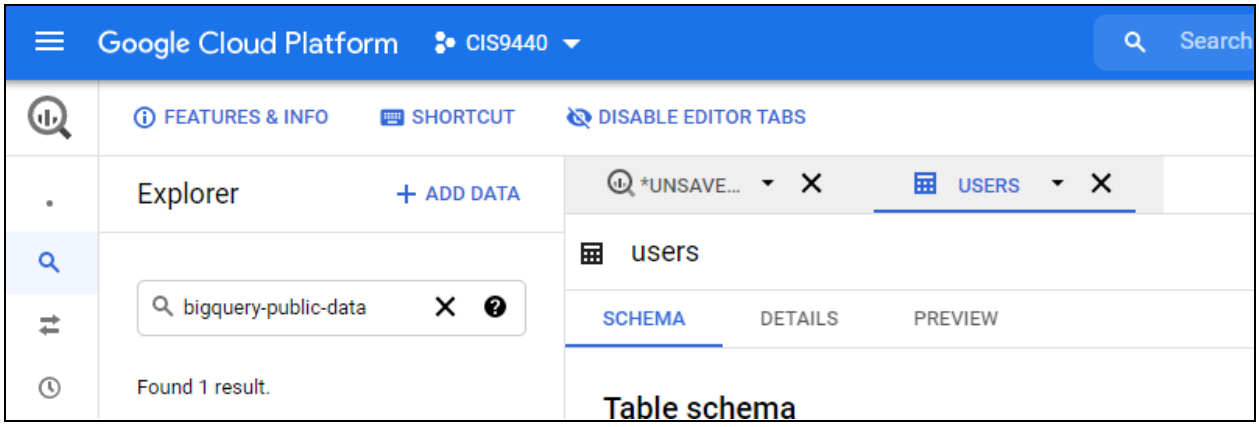
Query results SAVE RESULTS EXPLORE DATA

Query complete (0.3 sec elapsed, 3.4 KB processed)

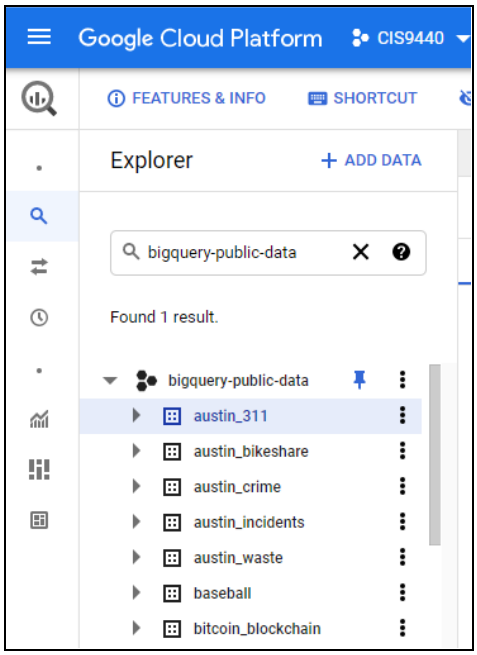
Job information | Results | JSON | Execution details

Row	Season	Player	Age	Tm	G	MP	PTS	TRB	AST	STL	BLK	WS
1	1999	Karl Malone	35	Utah Jazz	49	37.4	23.8	9.4	4.1	1.3	0.6	9.6
2	1998	Michael Jordan	34	Chicago Bulls	82	38.8	28.7	5.8	3.5	1.7	0.5	15.8
3	1997	Karl Malone	33	Utah Jazz	82	36.6	27.4	9.9	4.5	1.4	0.6	16.7
4	1996	Michael Jordan	32	Chicago Bulls	82	37.7	30.4	6.6	4.3	2.2	0.5	20.4
5	2006	Steve Nash	31	Phoenix Suns	79	35.4	18.8	4.2	10.5	0.8	0.2	12.4
6	1994	Hakeem Olajuwon	31	Houston Rockets	80	41.0	27.3	11.9	3.6	1.6	3.7	14.3

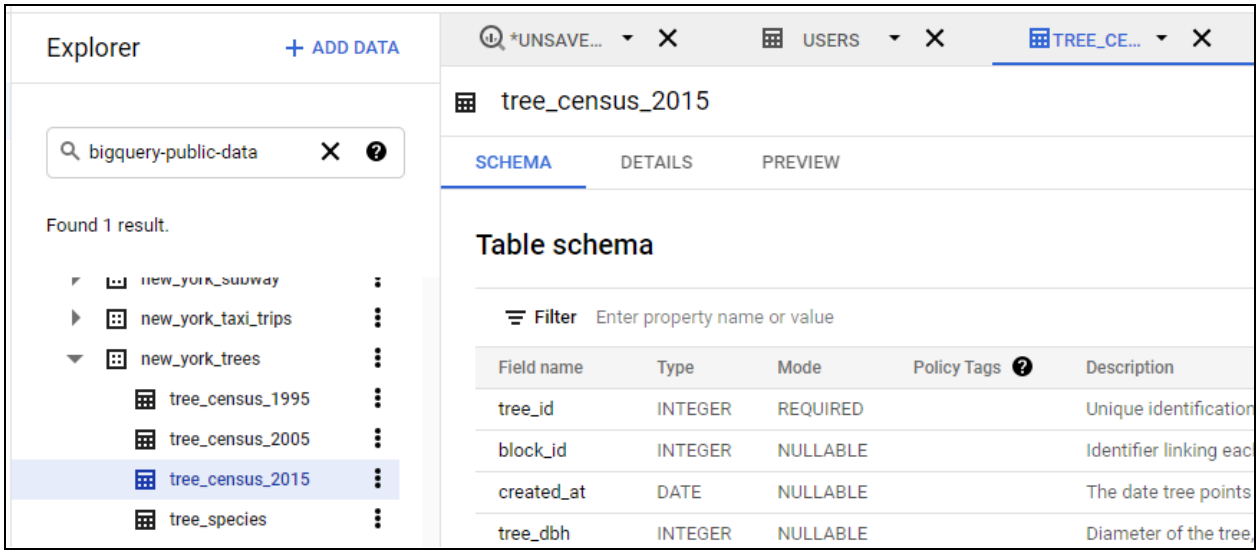
2. Now, we will move onto Google's Public Datasets in BigQuery. To do so, type "bigquery-public-data" into the "Explorer" search bar on the left-hand side of your BigQuery window and click enter. You may have to click on the blue highlighted text below the search bar "expand search to all projects" to see results. Example below:



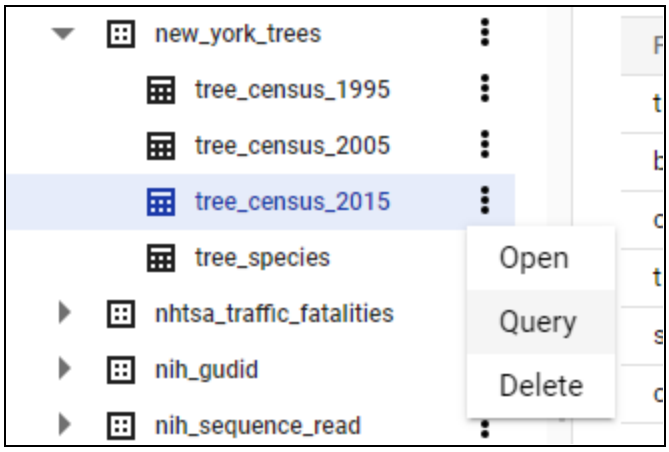
Once you can see the "bigquery-public-data" projects, click on the grey triangle to expand the project and view all datasets. Example below:

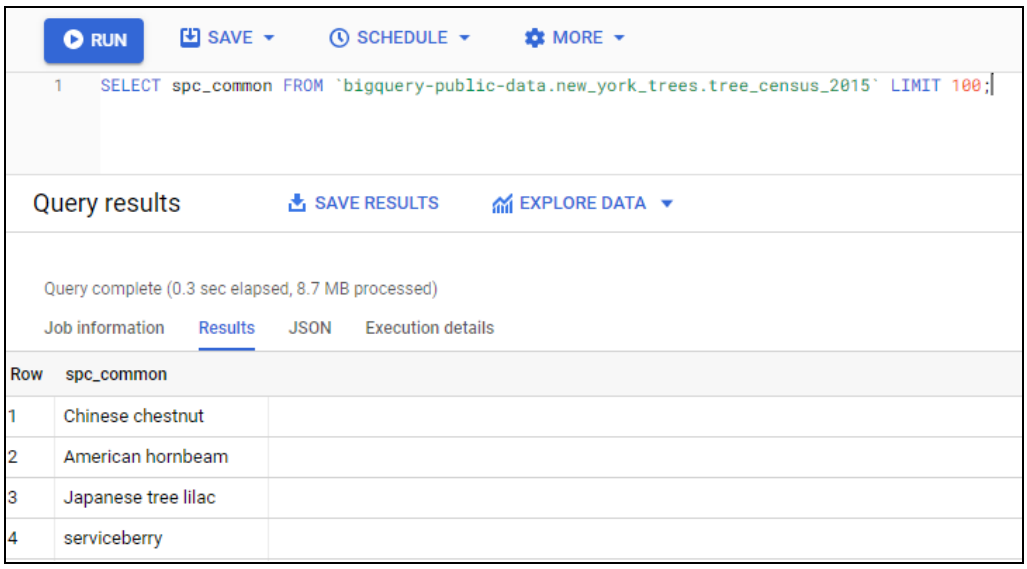


Scroll down and expand the dataset labeled “new_york_trees”. Then, click on the table “tree_census_2015”. Example below:



Finally, click on the 3 dots to the right of the “tree_census_2015” table and click on “Query”. Then, complete the query with a column name (such as “spc_common”) and click on the blue “Run” button. Examples below:





Take a screenshot of your results and post them as the answer to this question.

3. Continuing to use the “new_york_trees” dataset, write a query to find the top 5 most common trees in the “tree_census_2015” table. More specifically, you are looking for the top 5 most common “spc_common” in the table.

   Paste a screenshot of your **query** and **query results** as your answer.

4. Continuing to use the “new_york_trees” dataset, write a query to find the average tree diameter of trees in “Good” health by Borough in the “tree_census_2015” table.

For more details, in the “tree_census_2015” table the tree diameter is in column “tree_dbh”, tree health is in column “health”, and Boroughs are in column “boroname”.

Paste a screenshot of your **query** and **query results** as your answer.

5. Continuing to use the “new_york_trees” dataset and the “tree_census_2015” table, write a query to find the common name of the tree with the largest tree diameter in the Borough of “Brooklyn”.

   For more details, in the “tree_census_2015” table the tree diameter is in column “tree_dbh”, tree common name is in column “spc_common”, and Boroughs are in column “boroname”.

   Paste a screenshot of your **query** and **query results** as your answer.

6. Continuing to use the “new_york_trees” dataset and the “tree_census_2015” table, write a query to determine which “curb_loc” has the largest average “tree_dbh”.

   For more details, in the “tree_census_2015” table the tree diameter is in column “tree_dbh”, tree common name is in column “spc_common”, and Boroughs are in column “boroname”.

   Paste a screenshot of your **query** and **query results** as your answer.

7. Continuing to use the “new_york_trees” dataset and the “tree_census_2015” table, write a query to determine the zip code with the most trees in “Good” health.

   For more details, in the “tree_census_2015” table the tree diameter is in column “tree_dbh”, tree common name is in column “spc_common”, and Boroughs are in column “boroname”.

   Paste a screenshot of your **query** and **query results** as your answer.

8. Continuing to use the “new_york_trees” dataset and the “tree_census_2015” table, write a query to help determine if “guards” improve the trees’ “health”. Be creative with this query, and explain your answer.

For more details, in the “tree_census_2015” table the tree diameter is in column “tree_dbh”, tree common name is in column “spc_common”, and Boroughs are in column “boroname”.

Paste a screenshot of your **query** and **query results** as your answer.

9. Continuing to use the “new_york_trees” dataset and the “tree_census_2015” table, write a query to determine the most common “user_type” for trees that are “London planetree”.

   For more details, in the “tree_census_2015” table the tree diameter is in column “tree_dbh”, tree common name is in column “spc_common”, and Boroughs are in column “boroname”.

   Paste a screenshot of your **query** and **query results** as your answer.

10. Now, we will move onto a different Google Public Dataset. Find the “chicago_taxi_trips” dataset in the “bigquery-public-data” project.

    In the table “taxi_trips”, write a query to find the average tip left by taxi riders that paid with a “Credit Card” for rides that were longer than 15 minutes.

    For more details, the tip is in the “tips” column, the payment type is in the “payment_type” column, and the trip duration is in the “trip_seconds” column.

    Paste a screenshot of your **query** and **query results** as your answer.

11. Still using the “taxi_trips” table in the “chicago_taxi_trips” dataset,

    Write a query to find the payment type that resulted in the largest average tip for rides that were longer than 10 minutes and between 5 and 10 miles.

    For more details, the tip is in the “tips” column, the payment type is in the “payment_type” column, the trip duration is in the “trip_seconds” column, and the trip distance is in the “trip_miles” column.

    Paste a screenshot of your **query** and **query results** as your answer.

12. Still using the “taxi_trips” table in the “chicago_taxi_trips” dataset,

Use a SQL query to find the most expensive taxi “company”. You choose how to define “expensive”. Please paste a screenshot of your **query**, your **query results**, and **your explanation of “expensive”** as your answer.

13. Now, we will move onto a different Google Public Dataset. Find the “stackoverflow” dataset in the “bigquery-public-data” project.

    Write a SQL query to find the “title”, “tags”, “view_count”, and “score” of the 5 posts with the highest “favorite_count” in the table “stackoverflow_posts”.

    Paste a screenshot of your **query** and **query results** as your answer.

14. Continue in the “stackoverflow” dataset and “stackoverflow_posts” table in the “bigquery-public-data” project.

    Write a SQL query to find the titles of the 10 most viewed posts about BigQuery.

    Hint: leverage the LIKE operator.

    Paste a screenshot of your **query** and **query results** as your answer.

15. Continue in the “stackoverflow” dataset in the “bigquery-public-data” project.

    Write a SQL query that joins the “stackoverflow_posts” and “users” tables to return the “title”, “view_count”, “owner_display_name”, and “reputation” of the 10 titles with the most comments (“comment_count”). Please add a WHERE clause to filter out NULLs in the “title” column.

    Hint: “reputation” is from the “users” table.

    Paste a screenshot of your **query** and **query results** as your answer.

Part 2

16. Look at the following 3 tables in the “stackoverflow” datasets:

    - stackoverflow_posts

- users
- badges

Use Lucidchart to create an Entity Relationship Diagram (ERD) of the above 3 tables. Please include Primary and Foreign Key information for each Entity, and Data Types of each attribute

Paste a picture of your ERD as your answer.

Additional Notes:

- This is an individual assignment.
- Assemble all the responses to the questions in one MS Word or Google Docs document.
- Do not use MS Word to write your queries as MS Word will make “smart quotes” out of your text strings and SQL will not understand these.
- If the query results would take up more than 1 page, just copy the first dozen rows or enough to fill one page. Make a note of the total number of records in the full result
- Use tools like *Snipping Tool* on Windows for example to snip exactly what is needed to answer the question. Please do not clip the entire window, only what is needed to answer the question.