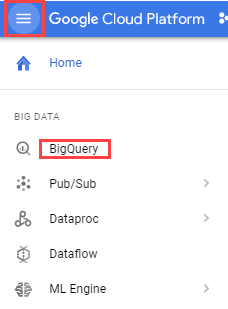
# Module-1 LAB - Using BigQuery to do Analysis

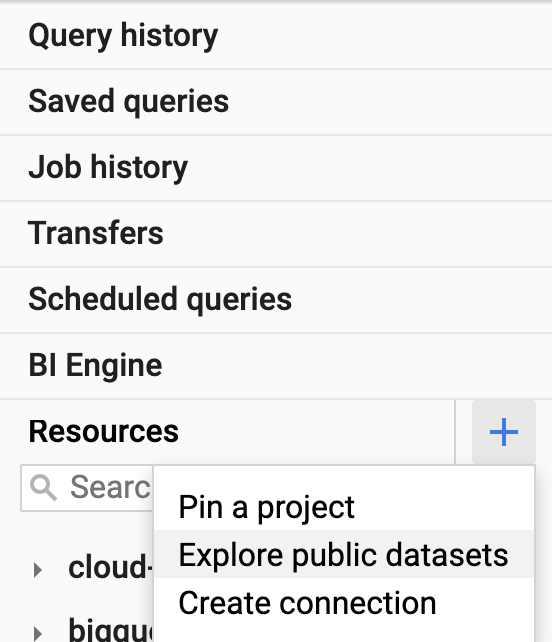
**Explore bicycle rental data**

**Open BigQuery Console**

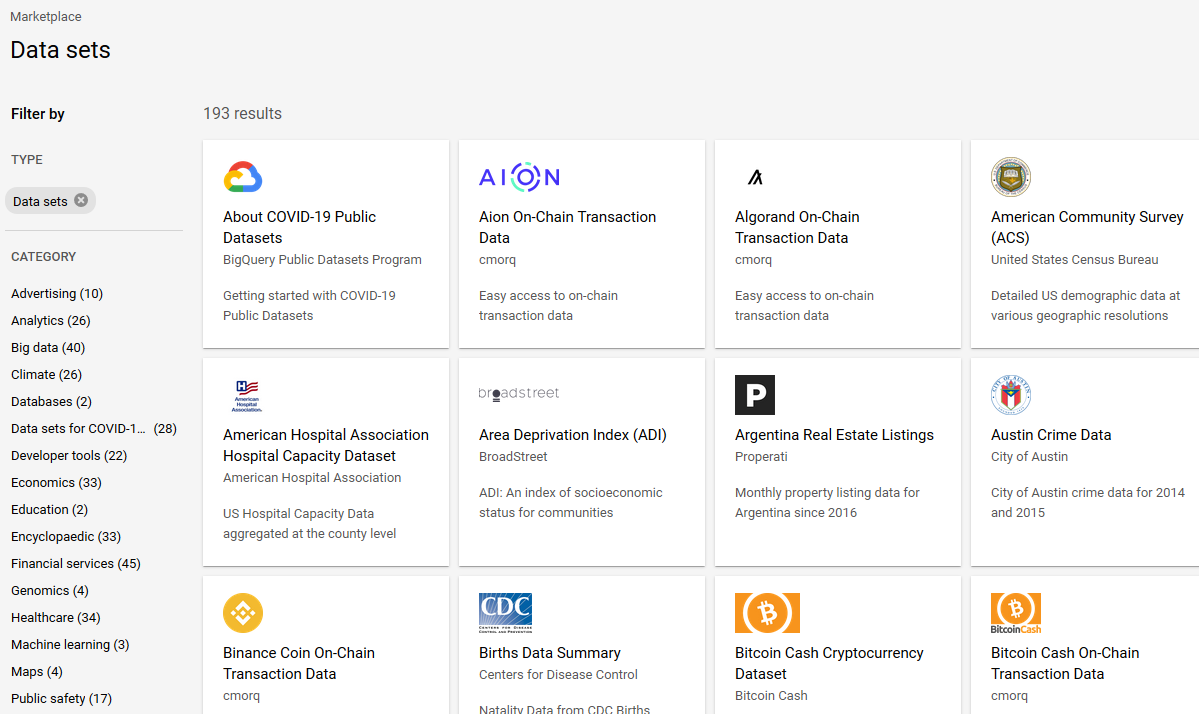
In the Google Cloud Console, select **Navigation menu** > **BigQuery**:



In the left pane, click **ADD DATA** > **Explore public datasets**.

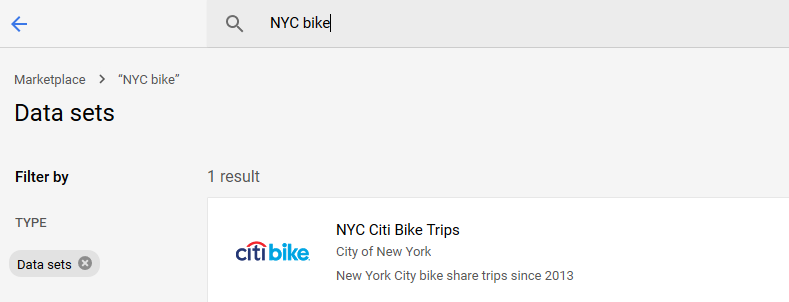


The Datasets window opens.



In the **Search** bar, type "NYC bike" then press **Enter**.

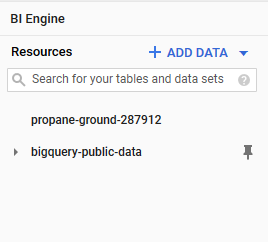
1 result, NYC Citi Bike Trips, displays.



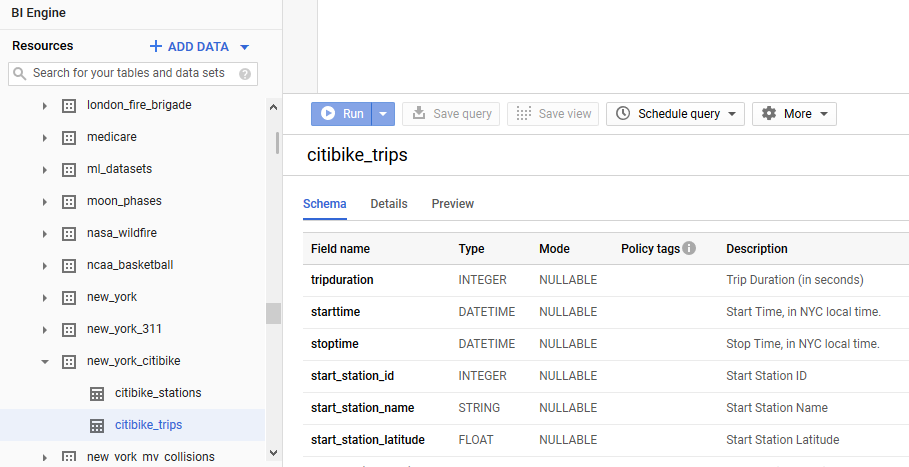
Click the NYC Citi Bike Trips dataset and then click **VIEW DATASET**.

Your The Google BigQuery console opens in a new browser tab.To keep your workspace organized, close this new browser tab and refresh the first tab.

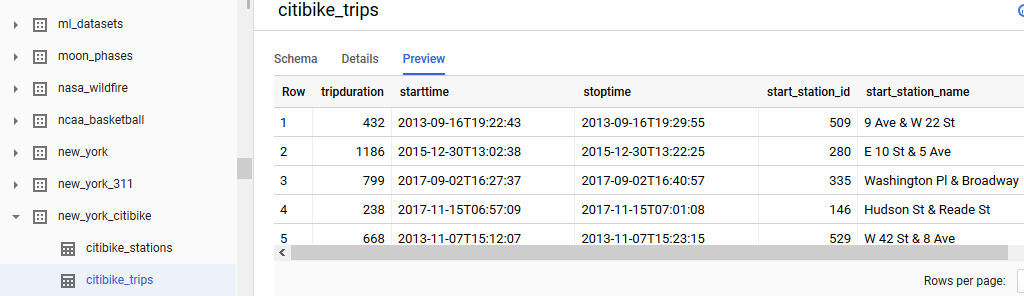
In the BigQuery console (in the first browser tab) you see two projects in the left pane, one named your Qwiklabs project ID, and one named **bigquery-public-data**.



In the left pane of the BigQuery console, select **bigquery-public-data** > **new\_york\_citibike** > **citibike\_trips** table.



In the Table (citibike\_trips) window, click the Preview tab.



Examine the columns and some of the data values.

Paste the following in the Query editor textbox:

SELECT

MIN(start\_station\_name) AS start\_station\_name,

MIN(end\_station\_name) AS end\_station\_name,

APPROX\_QUANTILES(tripduration, 10)[OFFSET (5)] AS typical\_duration,

COUNT(tripduration) AS num\_trips

FROM

`bigquery-public-data.new\_york\_citibike.citibike\_trips`

WHERE

start\_station\_id != end\_station\_id

GROUP BY

start\_station\_id,

end\_station\_id

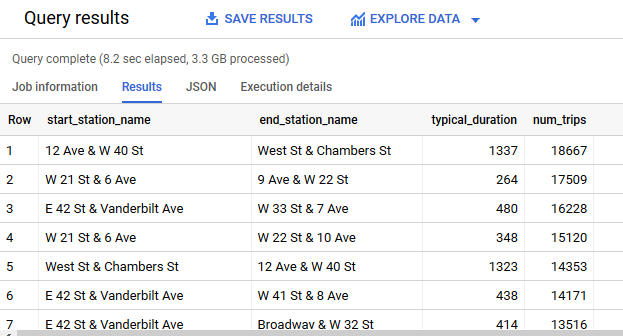
ORDER BY

num\_trips DESC

LIMIT

10

Click **Run**. Look at the result and try to determine what this query does ? (Hint: typical duration for the 10 most common one-way rentals)



Next, run the below to find another interesting fact: total distance travelled by each bicycle in the dataset. Note that the query limits the results to only top 5.

WITH

trip\_distance AS (

SELECT

bikeid,

ST\_Distance(ST\_GeogPoint(s.longitude,

s.latitude),

ST\_GeogPoint(e.longitude,

e.latitude)) AS distance

FROM

`bigquery-public-data.new\_york\_citibike.citibike\_trips`,

`bigquery-public-data.new\_york\_citibike.citibike\_stations` as s,

`bigquery-public-data.new\_york\_citibike.citibike\_stations` as e

WHERE

start\_station\_id = s.station\_id

AND end\_station\_id = e.station\_id )

SELECT

bikeid,

SUM(distance)/1000 AS total\_distance

FROM

trip\_distance

GROUP BY

bikeid

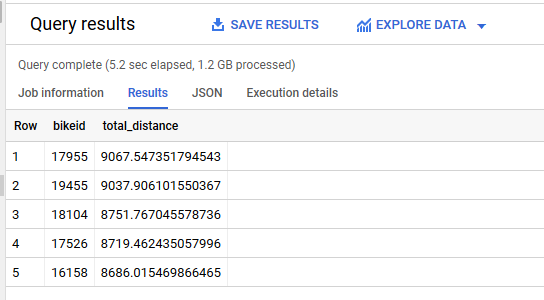
ORDER BY

total\_distance DESC

LIMIT

5

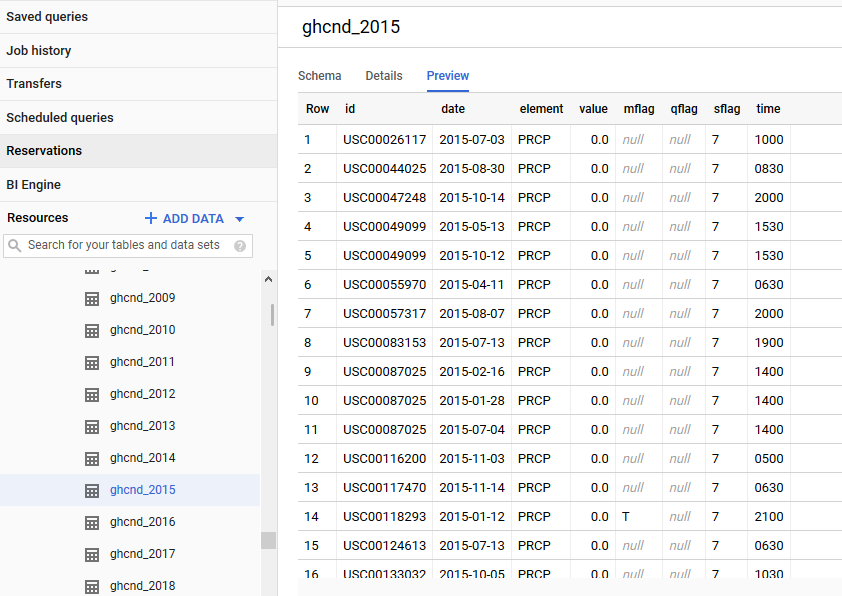
**Note:** For this query, we also used the other table in the dataset called **citibike\_stations** to get bicycle station information.



Click **HIDE EDITOR** on top right to close the Query editor.

## Explore the weather dataset

In the left pane of the BigQuery Console, select the newly added bigquery-public-data project and select **ghcn\_d** > **ghcnd\_2015**. Then click on the **Preview** tab. Your console should resemble the following:



Examine the columns and some of the data values.

Click **COMPOSE NEW QUERY** and enter the following:

SELECT

wx.date,

wx.value/10.0 AS prcp

FROM

`bigquery-public-data.ghcn\_d.ghcnd\_2015` AS wx

WHERE

id = 'USW00094728'

AND qflag IS NULL

AND element = 'PRCP'

ORDER BY

wx.date

Click **Run**.

This query will return rainfall (in mm) for all days in 2015 from a weather station in New York whose id is provided in the query (the station corresponds to NEW YORK CNTRL PK TWR)

## Find correlation between rain and bicycle rentals

How about joining the bicycle rentals data against weather data to learn whether there are fewer bicycle rentals on rainy days?

Click **COMPOSE NEW QUERY** and run the following query:

WITH bicycle\_rentals AS (

SELECT

COUNT(starttime) as num\_trips,

EXTRACT(DATE from starttime) as trip\_date

FROM `bigquery-public-data.new\_york\_citibike.citibike\_trips`

GROUP BY trip\_date

),

rainy\_days AS

(

SELECT

date,

(MAX(prcp) > 5) AS rainy

FROM (

SELECT

wx.date AS date,

IF (wx.element = 'PRCP', wx.value/10, NULL) AS prcp

FROM

`bigquery-public-data.ghcn\_d.ghcnd\_2015` AS wx

WHERE

wx.id = 'USW00094728'

)

GROUP BY

date

)

SELECT

ROUND(AVG(bk.num\_trips)) AS num\_trips,

wx.rainy

FROM bicycle\_rentals AS bk

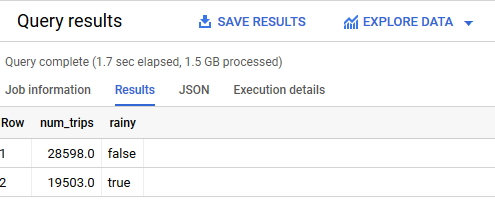
JOIN rainy\_days AS wx

ON wx.date = bk.trip\_date

GROUP BY wx.rainy

Click **Run**.

Now you can see the results of joining the bicycle rental dataset with a weather dataset that comes from a completely different source.



Running the query yields that, yes, New Yorkers ride the bicycle 47% fewer times when it rains.