# Data Preparation

**Course 2 end-of-course project scenarios**

**Cyclistic bike-share**

**Background:**

In this fictitious workplace scenario, the imaginary company Cyclistic has partnered with the city of New York to provide shared bikes. Currently, there are bike stations located throughout Manhattan and neighboring boroughs. Customers are able to rent bikes for easy travel among stations at these locations.

**Scenario:**

I am newly hired BI professional at Cyclistic. The company's Customer Growth Team is creating a business plan for next year. They want to understand how their customers are using their bikes; their top priority is identifying customer demand at different station locations. Previously, I gathered information from our meeting notes and completed important project planning documents. Now I am ready for the next part of project!

**Course 2 challenge:**

- Use project planning documents to identify key metrics and dashboard requirements
- Observe stakeholders in action to better understand how they use data
- Gather and combine necessary data
- Design reporting tables that can be uploaded to Tableau to create the final dashboard

**Course 2 workplace scenario overview: Cyclistic**

Previously, I started working with a fictional bike-share company, Cyclistic, to provide their team with key business intelligence insights. At the end of the last course, I consulted with stakeholders to develop project planning documents that establish their needs and expectations. The strategy and planning documents are key to helping me understand important details about this project.

Coming up, I am going to build on previous work to combine data from the tables I received for this project into one reporting table I will use to develop a dashboard that, I can share with stakeholders. The activities will guide me through uploading the data into my own project space, using SQL code in Dataflow or BigQuery, observing how stakeholders interact with data, and finalizing a reporting table to be used for the dashboard.

**Cyclistic datasets**

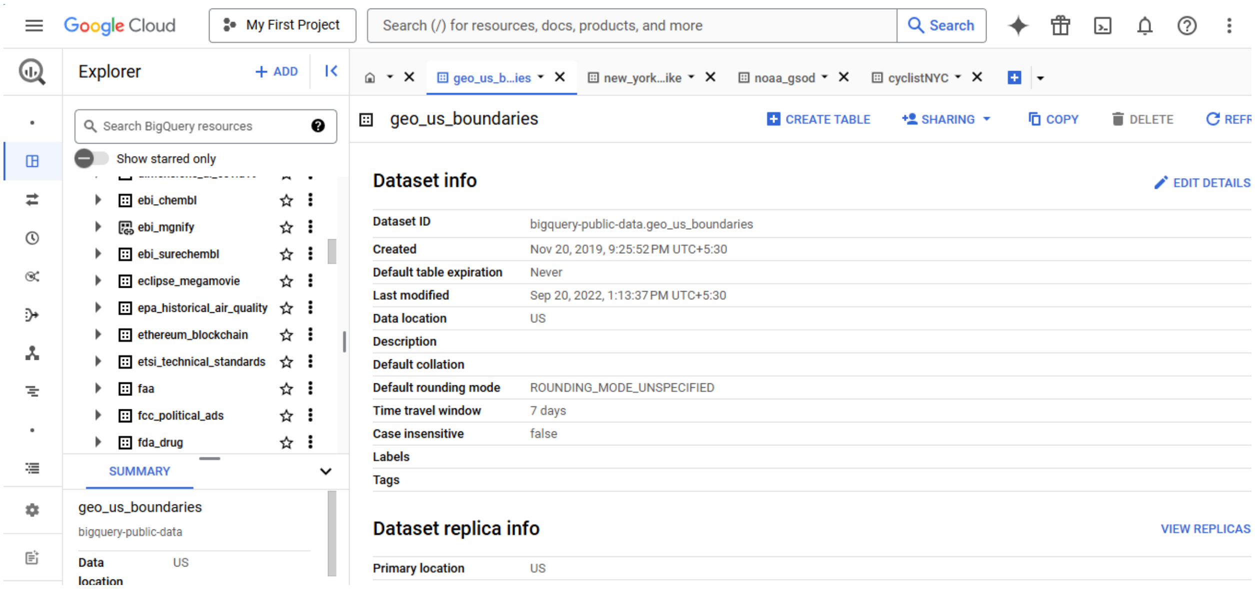
By now, I am getting ready to take the next steps with your Course 2 end-of-course project. To work with the Cyclistic project data, I will need to locate the appropriate public datasets and upload the zip code spreadsheet that my colleague shared into my BigQuery project space.

For this end-of-course project, I will be using two public datasets, which exist in the public data available from the Explorer pane.

- NYC Citi Bike Trips, Census Bureau US Boundaries,
- GSOD from the National Oceanic and Atmospheric Administration

Additionally, I will need to upload the zip code spreadsheet my colleague shared with me.

### Upload to BigQuery



geo_us_boundaries

bigquery-public-data

Data location US

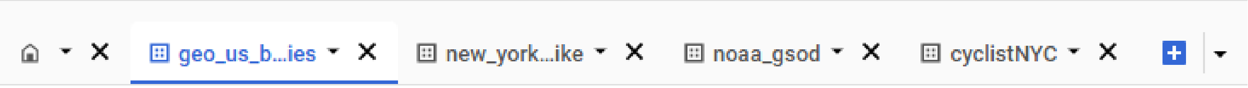
geo_us_boundaries

Dataset info

| | |
|---|---|
| Dataset ID | bigquery-public-data.geo_us_boundaries |
| Created | Nov 20, 2019, 9:25:52 PM UTC+5:30 |
| Default table expiration | Never |
| Last modified | Sep 20, 2022, 1:13:37 PM UTC+5:30 |
| Data location | US |
| Description | |
| Default collation | |
| Default rounding mode | ROUNDING_MODE_UNSPECIFIED |
| Time travel window | 7 days |
| Case insensitive | false |
| Labels | |
| Tags | |

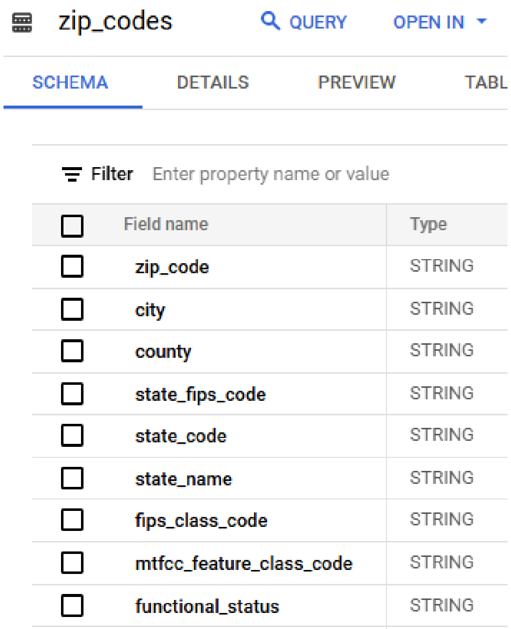
Dataset replica info

| | |
|---|---|
| Primary location | US |

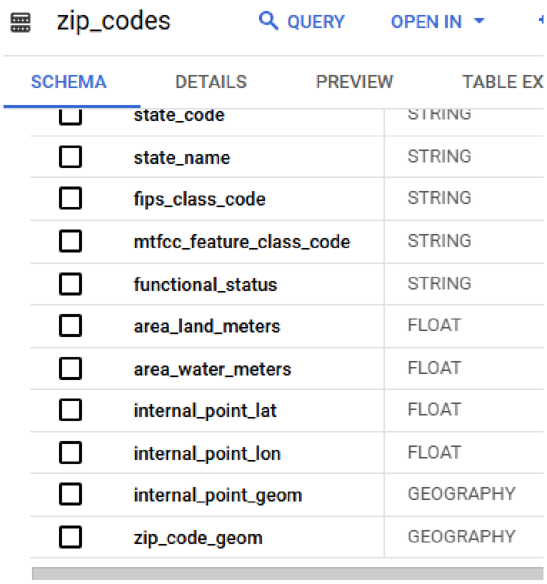


**Tables used to query the data**

- Census Bureau US Boundaries (zip_codes)

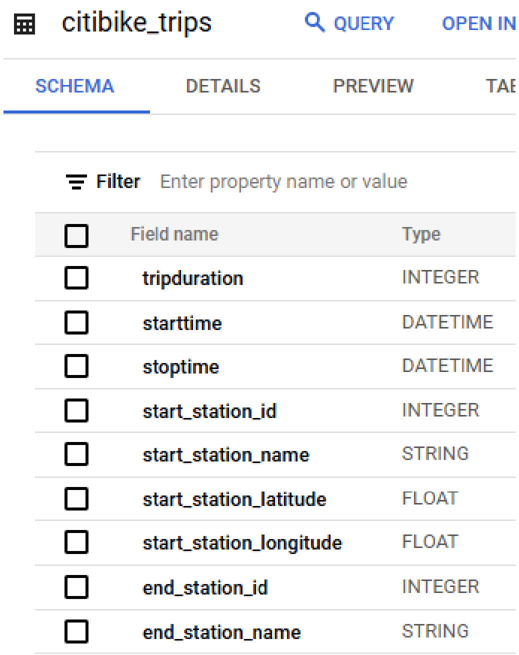


| □ | Field name | Type |
|---|---|---|
| □ | zip_code | STRING |
| □ | city | STRING |
| □ | county | STRING |
| □ | state_fips_code | STRING |
| □ | state_code | STRING |
| □ | state_name | STRING |
| □ | fips_class_code | STRING |
| □ | mtfcc_feature_class_code | STRING |
| □ | functional_status | STRING |

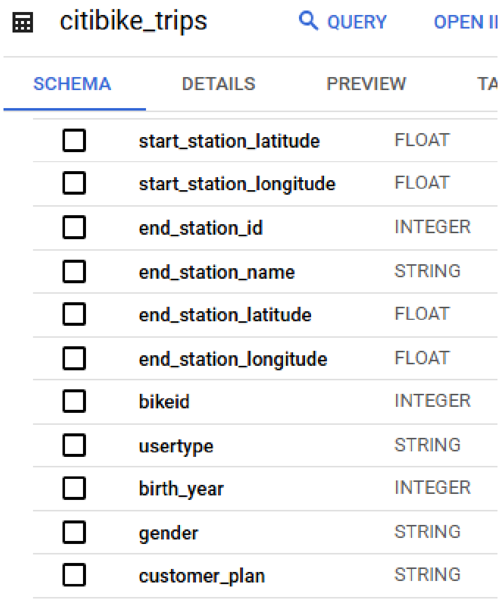


| | | |
|---|---|---|
| □ | state_code | STRING |
| □ | state_name | STRING |
| □ | fips_class_code | STRING |
| □ | mtfcc_feature_class_code | STRING |
| □ | functional_status | STRING |
| □ | area_land_meters | FLOAT |
| □ | area_water_meters | FLOAT |
| □ | internal_point_lat | FLOAT |
| □ | internal_point_lon | FLOAT |
| □ | internal_point_geom | GEOGRAPHY |
| □ | zip_code_geom | GEOGRAPHY |

- NYC Citi Bike Trips, (citibike_trips)

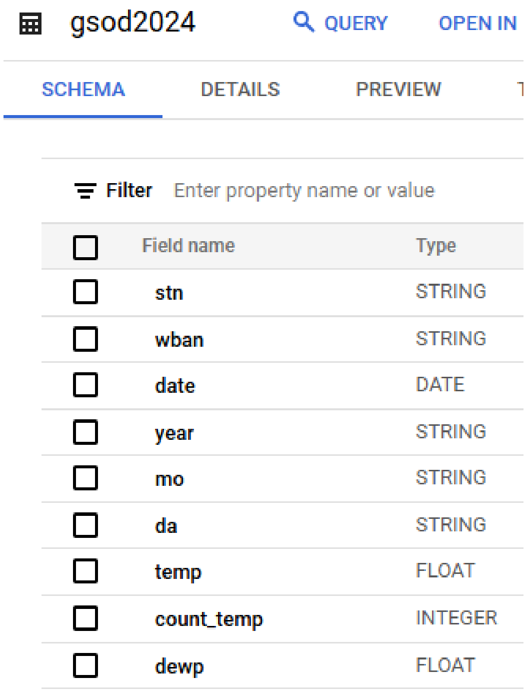


| □ | Field name | Type |
|---|---|---|
| □ | tripduration | INTEGER |
| □ | starttime | DATETIME |
| □ | stoptime | DATETIME |
| □ | start_station_id | INTEGER |
| □ | start_station_name | STRING |
| □ | start_station_latitude | FLOAT |
| □ | start_station_longitude | FLOAT |
| □ | end_station_id | INTEGER |
| □ | end_station_name | STRING |



| | | |
|---|---|---|
| □ | start_station_latitude | FLOAT |
| □ | start_station_longitude | FLOAT |
| □ | end_station_id | INTEGER |
| □ | end_station_name | STRING |
| □ | end_station_latitude | FLOAT |
| □ | end_station_longitude | FLOAT |
| □ | bikeid | INTEGER |
| □ | usertype | STRING |
| □ | birth_year | INTEGER |
| □ | gender | STRING |
| □ | customer_plan | STRING |

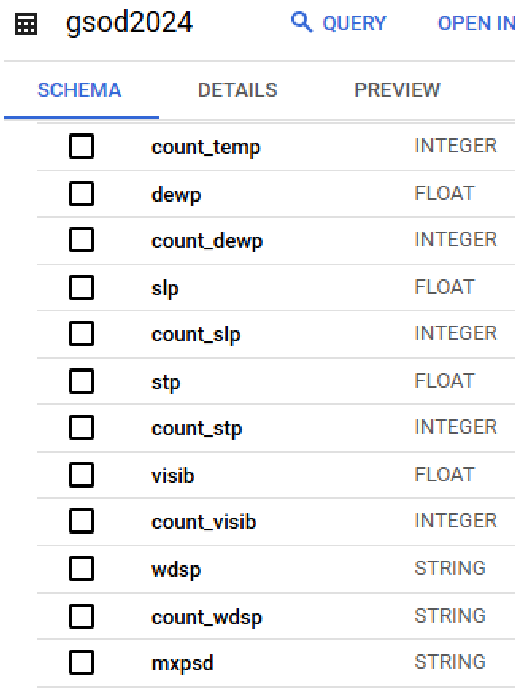
- [GSOD from the National Oceanic and Atmospheric Administration](#) (gsod20*)


gsod2024 QUERY OPEN IN

SCHEMA DETAILS PREVIEW T

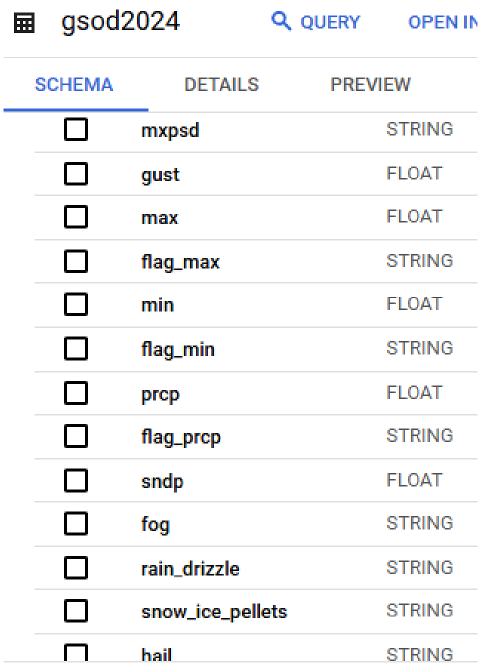
Filter Enter property name or value

| □ | Field name | Type |
|---|---|---|
| □ | stn | STRING |
| □ | wban | STRING |
| □ | date | DATE |
| □ | year | STRING |
| □ | mo | STRING |
| □ | da | STRING |
| □ | temp | FLOAT |
| □ | count_temp | INTEGER |
| □ | dewp | FLOAT |


gsod2024 QUERY OPEN IN

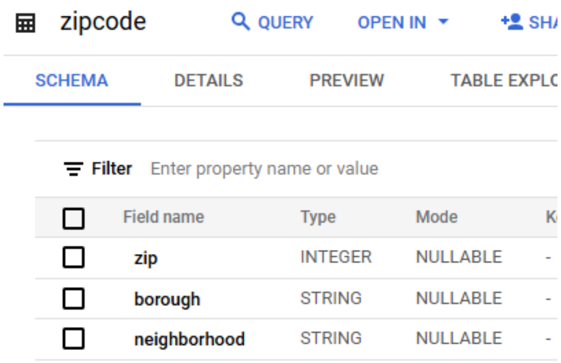
SCHEMA DETAILS PREVIEW

| | | |
|---|---|---|
| □ | count_temp | INTEGER |
| □ | dewp | FLOAT |
| □ | count_dewp | INTEGER |
| □ | slp | FLOAT |
| □ | count_slp | INTEGER |
| □ | stp | FLOAT |
| □ | count_stp | INTEGER |
| □ | visib | FLOAT |
| □ | count_visib | INTEGER |
| □ | wdsp | STRING |
| □ | count_wdsp | STRING |
| □ | mxpsd | STRING |


gsod2024 QUERY OPEN IN

SCHEMA DETAILS PREVIEW

| | | |
|---|---|---|
| □ | mxpsd | STRING |
| □ | gust | FLOAT |
| □ | max | FLOAT |
| □ | flag_max | STRING |
| □ | min | FLOAT |
| □ | flag_min | STRING |
| □ | prcp | FLOAT |
| □ | flag_prcp | STRING |
| □ | sndp | FLOAT |
| □ | fog | STRING |
| □ | rain_drizzle | STRING |
| □ | snow_ice_pellets | STRING |
| □ | hail | STRING |


zipcode QUERY OPEN IN SHA

SCHEMA DETAILS PREVIEW TABLE EXPLC

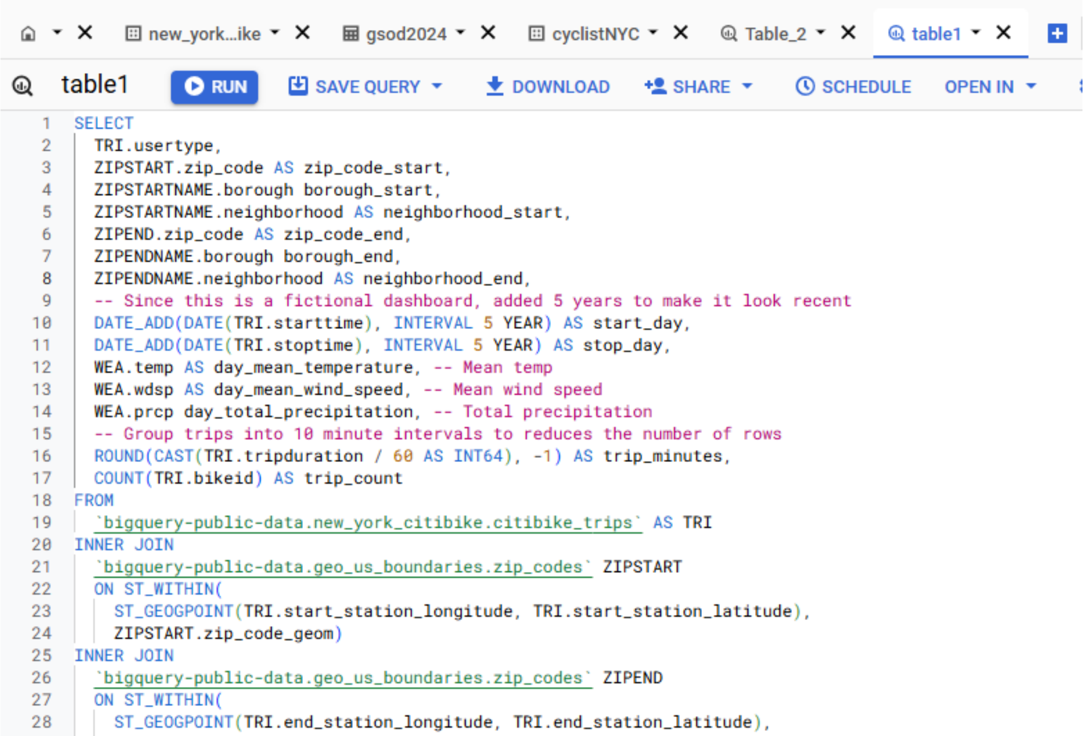
Filter Enter property name or value

| □ | Field name | Type | Mode | K |
|---|---|---|---|---|
| □ | zip | INTEGER | NULLABLE | - |
| □ | borough | STRING | NULLABLE | - |
| □ | neighborhood | STRING | NULLABLE | - |

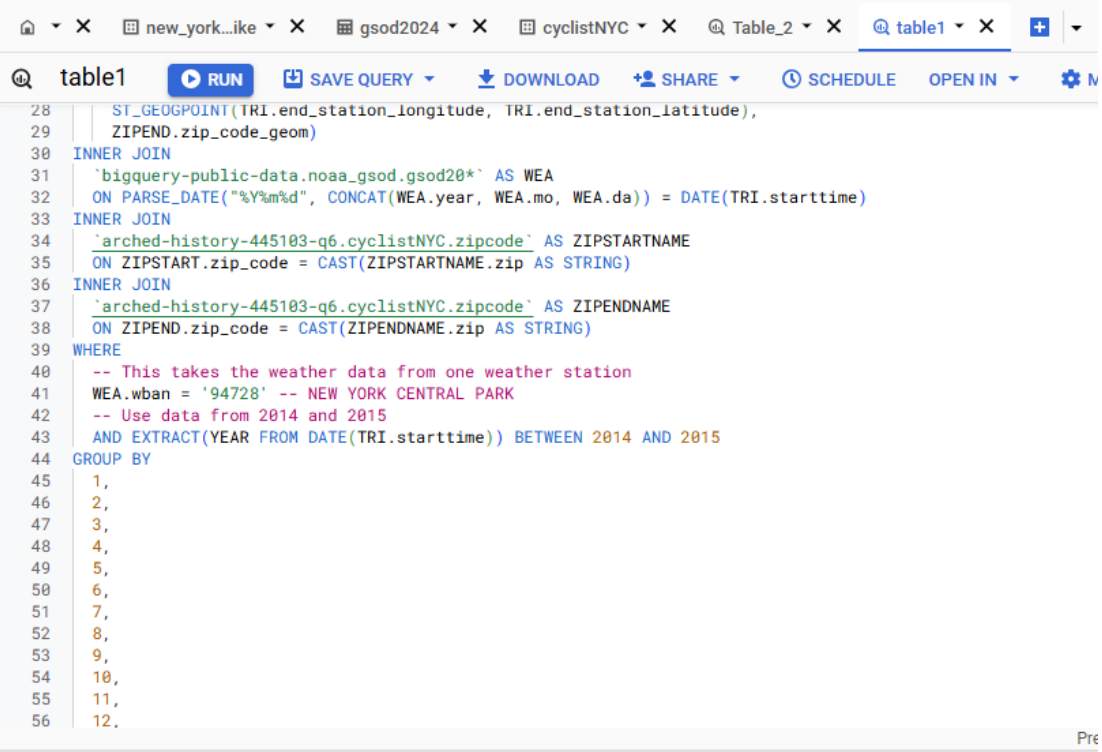
- [zip code spreadsheet](#) (cyclistNYC_zipcode)

## Querying the data

SQL query to create a summary table for the entire year:

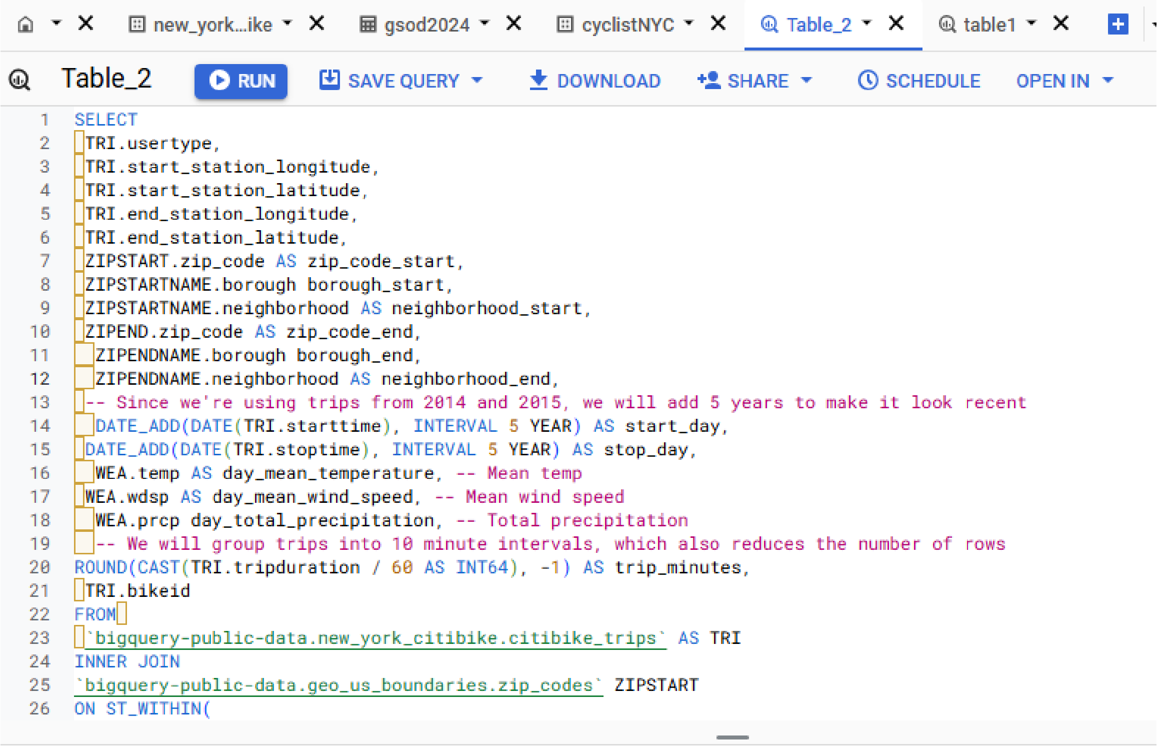


```sql
SELECT
  TRI.usertype,
  ZIPSTART.zip_code AS zip_code_start,
  ZIPSTARTNAME.borough borough_start,
  ZIPSTARTNAME.neighborhood AS neighborhood_start,
  ZIPEND.zip_code AS zip_code_end,
  ZIPENDNAME.borough borough_end,
  ZIPENDNAME.neighborhood AS neighborhood_end,
  -- Since this is a fictional dashboard, added 5 years to make it look recent
  DATE_ADD(DATE(TRI.starttime), INTERVAL 5 YEAR) AS start_day,
  DATE_ADD(DATE(TRI.stoptime), INTERVAL 5 YEAR) AS stop_day,
  WEA.temp AS day_mean_temperature, -- Mean temp
  WEA.wdsp AS day_mean_wind_speed, -- Mean wind speed
  WEA.prcp day_total_precipitation, -- Total precipitation
  -- Group trips into 10 minute intervals to reduces the number of rows
  ROUND(CAST(TRI.tripduration / 60 AS INT64), -1) AS trip_minutes,
  COUNT(TRI.bikeid) AS trip_count
FROM
  `bigquery-public-data.new_york_citibike.citibike_trips` AS TRI
INNER JOIN
  `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPSTART
  ON ST_WITHIN(
    ST_GEOGPOINT(TRI.start_station_longitude, TRI.start_station_latitude),
    ZIPSTART.zip_code_geom)
INNER JOIN
  `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPEND
  ON ST_WITHIN(
    ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
```

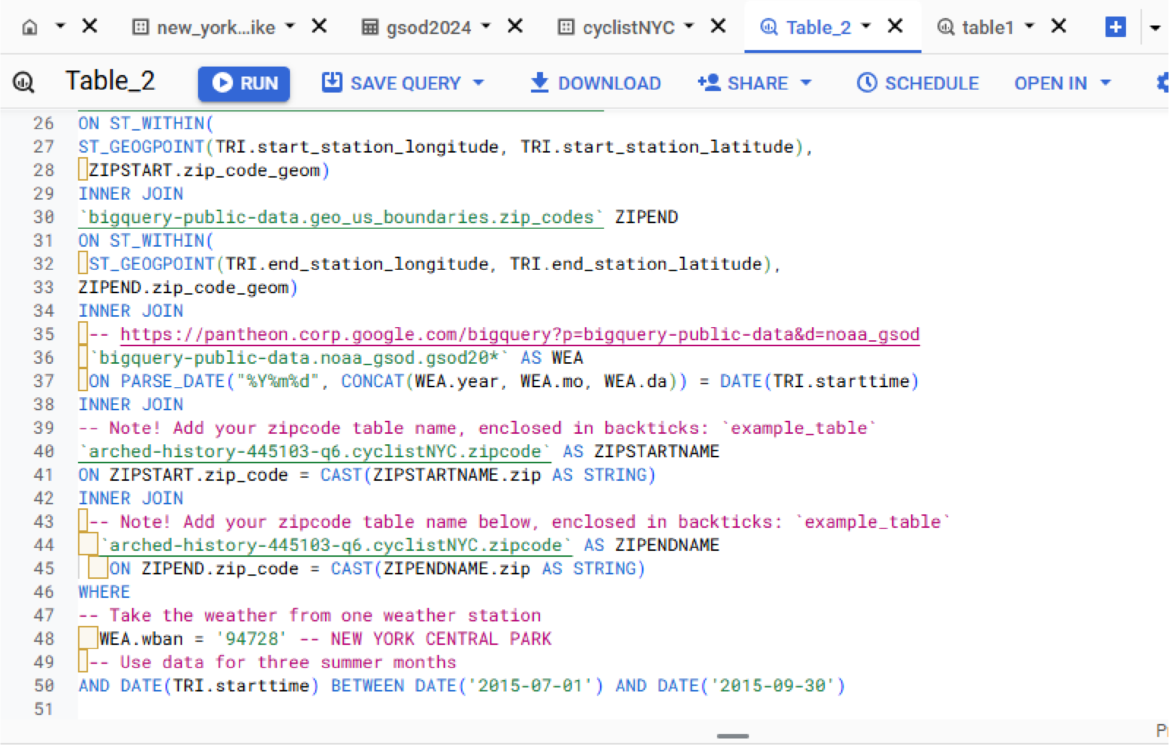


```sql
    ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
    ZIPEND.zip_code_geom)
INNER JOIN
  `bigquery-public-data.noaa_gsod.gsod20*` AS WEA
  ON PARSE_DATE("%Y%m%d", CONCAT(WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
INNER JOIN
  `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPSTARTNAME
  ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
INNER JOIN
  `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPENDNAME
  ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
WHERE
  -- This takes the weather data from one weather station
  WEA.wban = '94728' -- NEW YORK CENTRAL PARK
  -- Use data from 2014 and 2015
  AND EXTRACT(YEAR FROM DATE(TRI.starttime)) BETWEEN 2014 AND 2015
GROUP BY
  1,
  2,
  3,
  4,
  5,
  6,
  7,
  8,
  9,
  10,
  11,
  12,
```

SQL query that captured data from just the summer season:



```sql
SELECT
 TRI.usertype,
 TRI.start_station_longitude,
 TRI.start_station_latitude,
 TRI.end_station_longitude,
 TRI.end_station_latitude,
 ZIPSTART.zip_code AS zip_code_start,
 ZIPSTARTNAME.borough borough_start,
 ZIPSTARTNAME.neighborhood AS neighborhood_start,
 ZIPEND.zip_code AS zip_code_end,
  ZIPENDNAME.borough borough_end,
  ZIPENDNAME.neighborhood AS neighborhood_end,
 -- Since we're using trips from 2014 and 2015, we will add 5 years to make it look recent
  DATE_ADD(DATE(TRI.starttime), INTERVAL 5 YEAR) AS start_day,
 DATE_ADD(DATE(TRI.stoptime), INTERVAL 5 YEAR) AS stop_day,
  WEA.temp AS day_mean_temperature, -- Mean temp
 WEA.wdsp AS day_mean_wind_speed, -- Mean wind speed
  WEA.prcp day_total_precipitation, -- Total precipitation
  -- We will group trips into 10 minute intervals, which also reduces the number of rows
ROUND(CAST(TRI.tripduration / 60 AS INT64), -1) AS trip_minutes,
 TRI.bikeid
FROM
 `bigquery-public-data.new_york_citibike.citibike_trips` AS TRI
INNER JOIN
`bigquery-public-data.geo_us_boundaries.zip_codes` ZIPSTART
ON ST_WITHIN(
```



```sql
ON ST_WITHIN(
ST_GEOGPOINT(TRI.start_station_longitude, TRI.start_station_latitude),
 ZIPSTART.zip_code_geom)
INNER JOIN
`bigquery-public-data.geo_us_boundaries.zip_codes` ZIPEND
ON ST_WITHIN(
 ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
ZIPEND.zip_code_geom)
INNER JOIN
 -- https://pantheon.corp.google.com/bigquery?p=bigquery-public-data&d=noaa_gsod
 `bigquery-public-data.noaa_gsod.gsod20*` AS WEA
 ON PARSE_DATE("%Y%m%d", CONCAT(WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
INNER JOIN
-- Note! Add your zipcode table name, enclosed in backticks: `example_table`
`arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPSTARTNAME
ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
INNER JOIN
 -- Note! Add your zipcode table name below, enclosed in backticks: `example_table`
  `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPENDNAME
  ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
WHERE
-- Take the weather from one weather station
 WEA.wban = '94728' -- NEW YORK CENTRAL PARK
 -- Use data for three summer months
AND DATE(TRI.starttime) BETWEEN DATE('2015-07-01') AND DATE('2015-09-30')

```