

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

The screenshot displays the Google Cloud BigQuery interface. At the top, there's a search bar and navigation tabs for various datasets: 'geo_us_b...ies', 'new_york...ike', 'noaa_gsod', and 'cyclistNYC'. The main panel shows the 'geo_us_boundaries' dataset with its details.

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
------------------	----

The interface also includes a left sidebar with a search bar and a list of datasets, and a bottom tab bar with the same dataset names.

Tables used to query the data

- [Census Bureau US Boundaries](#) (zip_codes)

zip_codesQUERYOPEN IN

SCHEMA	DETAILS	PREVIEW	TABL
<div><div>Filter</div><div>Enter property name or value</div></div>			
<input type="checkbox"/>	Field name	Type	
<input type="checkbox"/>	zip_code	STRING	
<input type="checkbox"/>	city	STRING	
<input type="checkbox"/>	county	STRING	
<input type="checkbox"/>	state_fips_code	STRING	
<input type="checkbox"/>	state_code	STRING	
<input type="checkbox"/>	state_name	STRING	
<input type="checkbox"/>	fips_class_code	STRING	
<input type="checkbox"/>	mtfcc_feature_class_code	STRING	
<input type="checkbox"/>	functional_status	STRING	

zip_codesQUERYOPEN IN

SCHEMA	DETAILS	PREVIEW	TABLE EX
<input type="checkbox"/>	state_code	STRING	
<input type="checkbox"/>	state_name	STRING	
<input type="checkbox"/>	fips_class_code	STRING	
<input type="checkbox"/>	mtfcc_feature_class_code	STRING	
<input type="checkbox"/>	functional_status	STRING	
<input type="checkbox"/>	area_land_meters	FLOAT	
<input type="checkbox"/>	area_water_meters	FLOAT	
<input type="checkbox"/>	internal_point_lat	FLOAT	
<input type="checkbox"/>	internal_point_lon	FLOAT	
<input type="checkbox"/>	internal_point_geom	GEOGRAPHY	
<input type="checkbox"/>	zip_code_geom	GEOGRAPHY	

- [NYC Citi Bike Trips](#), (citibike_trips)

citibike_tripsQUERYOPEN IN


SCHEMA	DETAILS	PREVIEW	TAE
<div><div>Filter</div><div>Enter property name or value</div></div>			
<input type="checkbox"/>	Field name	Type	
<input type="checkbox"/>	tripduration	INTEGER	
<input type="checkbox"/>	starttime	DATETIME	
<input type="checkbox"/>	stoptime	DATETIME	
<input type="checkbox"/>	start_station_id	INTEGER	
<input type="checkbox"/>	start_station_name	STRING	
<input type="checkbox"/>	start_station_latitude	FLOAT	
<input type="checkbox"/>	start_station_longitude	FLOAT	
<input type="checkbox"/>	end_station_id	INTEGER	
<input type="checkbox"/>	end_station_name	STRING	

citibike_tripsQUERYOPEN IN

SCHEMA	DETAILS	PREVIEW	TA
<input type="checkbox"/>	start_station_latitude	FLOAT	
<input type="checkbox"/>	start_station_longitude	FLOAT	
<input type="checkbox"/>	end_station_id	INTEGER	
<input type="checkbox"/>	end_station_name	STRING	
<input type="checkbox"/>	end_station_latitude	FLOAT	
<input type="checkbox"/>	end_station_longitude	FLOAT	
<input type="checkbox"/>	bikeid	INTEGER	
<input type="checkbox"/>	usertype	STRING	
<input type="checkbox"/>	birth_year	INTEGER	
<input type="checkbox"/>	gender	STRING	
<input type="checkbox"/>	customer_plan	STRING	


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

 gsod2024  QUERY [OPEN IN](#)

SCHEMA			DETAILS	PREVIEW	1
 Filter Enter property name or value					
<input type="checkbox"/>	Field name	Type			
<input type="checkbox"/>	stn	STRING			
<input type="checkbox"/>	wban	STRING			
<input type="checkbox"/>	date	DATE			
<input type="checkbox"/>	year	STRING			
<input type="checkbox"/>	mo	STRING			
<input type="checkbox"/>	da	STRING			
<input type="checkbox"/>	temp	FLOAT			
<input type="checkbox"/>	count_temp	INTEGER			
<input type="checkbox"/>	dewp	FLOAT			

 gsod2024  QUERY [OPEN IN](#)


SCHEMA			DETAILS	PREVIEW
<input type="checkbox"/>	count_temp	INTEGER		
<input type="checkbox"/>	dewp	FLOAT		
<input type="checkbox"/>	count_dewp	INTEGER		
<input type="checkbox"/>	slp	FLOAT		
<input type="checkbox"/>	count_slp	INTEGER		
<input type="checkbox"/>	stp	FLOAT		
<input type="checkbox"/>	count_stp	INTEGER		
<input type="checkbox"/>	visib	FLOAT		
<input type="checkbox"/>	count_visib	INTEGER		
<input type="checkbox"/>	wdsp	STRING		
<input type="checkbox"/>	count_wdsp	STRING		
<input type="checkbox"/>	mxpsd	STRING		

 gsod2024  QUERY [OPEN IN](#)

SCHEMA			DETAILS	PREVIEW
<input type="checkbox"/>	mxpsd	STRING		
<input type="checkbox"/>	gust	FLOAT		
<input type="checkbox"/>	max	FLOAT		
<input type="checkbox"/>	flag_max	STRING		
<input type="checkbox"/>	min	FLOAT		
<input type="checkbox"/>	flag_min	STRING		
<input type="checkbox"/>	prcp	FLOAT		
<input type="checkbox"/>	flag_prcp	STRING		
<input type="checkbox"/>	sndp	FLOAT		
<input type="checkbox"/>	fog	STRING		
<input type="checkbox"/>	rain_drizzle	STRING		
<input type="checkbox"/>	snow_ice_pellets	STRING		
<input type="checkbox"/>	hail	STRING		

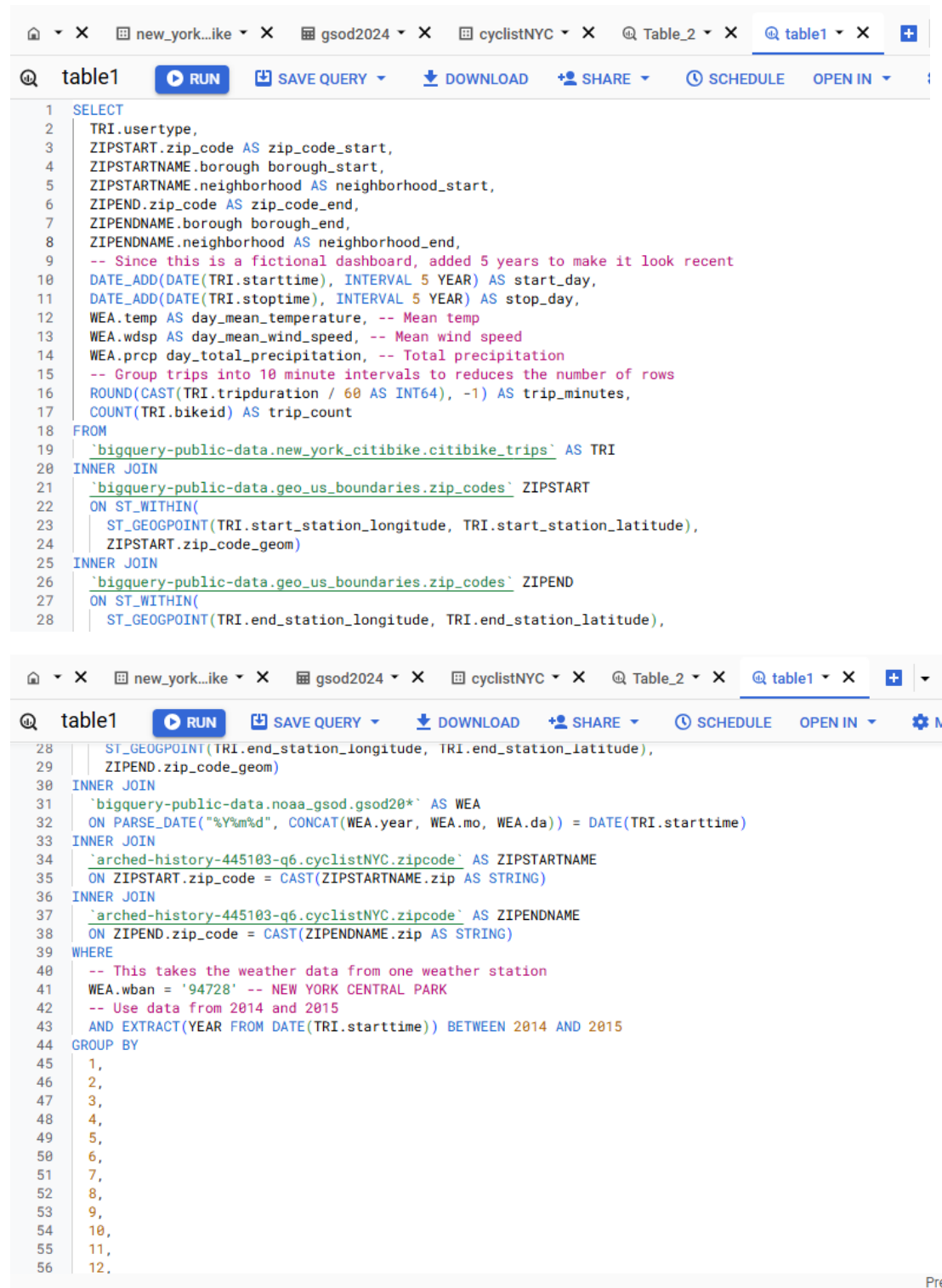
- [zip code spreadsheet](#) (cyclistNYC_zipcode)

 zipcode  QUERY [OPEN IN](#)  SH/

SCHEMA			DETAILS	PREVIEW	TABLE EXPLC
 Filter Enter property name or value					
<input type="checkbox"/>	Field name	Type	Mode	K	
<input type="checkbox"/>	zip	INTEGER	NULLABLE	-	
<input type="checkbox"/>	borough	STRING	NULLABLE	-	
<input type="checkbox"/>	neighborhood	STRING	NULLABLE	-	

Querying the data

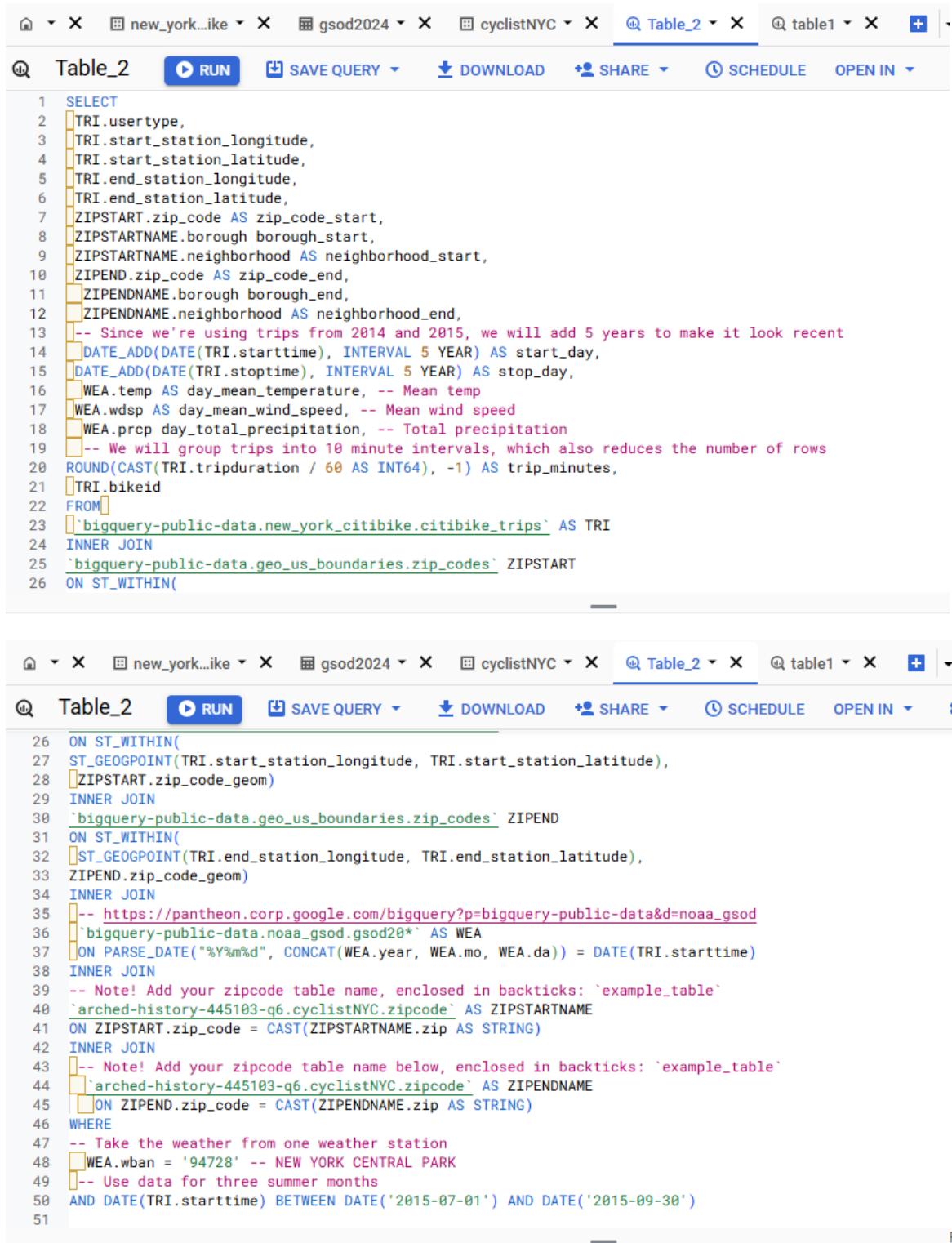
SQL query to create a summary table for the entire year:



The screenshot displays the Google BigQuery web interface. At the top, there are tabs for various datasets: 'new_york...ike', 'gsod2024', 'cyclistNYC', 'Table_2', and 'table1'. The 'table1' tab is active, showing a SQL query. The query is a SELECT statement that joins several tables: 'TRI' (from 'bigquery-public-data.new_york_citibike.citibike_trips'), 'ZIPSTART' (from 'bigquery-public-data.geo_us_boundaries.zip_codes'), 'ZIPEND' (from 'bigquery-public-data.geo_us_boundaries.zip_codes'), 'WEA' (from 'bigquery-public-data.noaa_gsod.gsod20*'), and 'ZIPSTARTNAME' and 'ZIPENDNAME' (from 'arched-history-445103-q6.cyclistNYC.zipcode'). The query includes various aliases and conditions, such as filtering for a specific weather station ('94728') and a time range (2014 to 2015). The query is grouped by a list of 12 columns. The results are shown in a table with 12 columns, corresponding to the grouped columns. The results are truncated, showing only the first few rows.

```
1 SELECT
2   TRI.usertype,
3   ZIPSTART.zip_code AS zip_code_start,
4   ZIPSTARTNAME.borough AS borough_start,
5   ZIPSTARTNAME.neighborhood AS neighborhood_start,
6   ZIPEND.zip_code AS zip_code_end,
7   ZIPENDNAME.borough AS borough_end,
8   ZIPENDNAME.neighborhood AS neighborhood_end,
9   -- Since this is a fictional dashboard, added 5 years to make it look recent
10  DATE_ADD(DATE(TRI.starttime), INTERVAL 5 YEAR) AS start_day,
11  DATE_ADD(DATE(TRI.stoptime), INTERVAL 5 YEAR) AS stop_day,
12  WEA.temp AS day_mean_temperature, -- Mean temp
13  WEA.wdsp AS day_mean_wind_speed, -- Mean wind speed
14  WEA.prcp AS day_total_precipitation, -- Total precipitation
15  -- Group trips into 10 minute intervals to reduces the number of rows
16  ROUND(CAST(TRI.tripeduration / 60 AS INT64), -1) AS trip_minutes,
17  COUNT(TRI.bikeid) AS trip_count
18 FROM
19   `bigquery-public-data.new_york_citibike.citibike_trips` AS TRI
20 INNER JOIN
21   `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPSTART
22   ON ST_WITHIN(
23     ST_GEOGPOINT(TRI.start_station_longitude, TRI.start_station_latitude),
24     ZIPSTART.zip_code_geom)
25 INNER JOIN
26   `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPEND
27   ON ST_WITHIN(
28     ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
29     ZIPEND.zip_code_geom)
30 INNER JOIN
31   `bigquery-public-data.noaa_gsod.gsod20*` AS WEA
32   ON PARSE_DATE("%Y%m%d", CONCAT(WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
33 INNER JOIN
34   `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPSTARTNAME
35   ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
36 INNER JOIN
37   `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPENDNAME
38   ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
39 WHERE
40   -- This takes the weather data from one weather station
41   WEA.wban = '94728' -- NEW YORK CENTRAL PARK
42   -- Use data from 2014 and 2015
43   AND EXTRACT(YEAR FROM DATE(TRI.starttime)) BETWEEN 2014 AND 2015
44 GROUP BY
45   1,
46   2,
47   3,
48   4,
49   5,
50   6,
51   7,
52   8,
53   9,
54   10,
55   11,
56   12.
```

SQL query that captured data from just the summer season:



```
1 SELECT
2   TRI.usertype,
3   TRI.start_station_longitude,
4   TRI.start_station_latitude,
5   TRI.end_station_longitude,
6   TRI.end_station_latitude,
7   ZIPSTART.zip_code AS zip_code_start,
8   ZIPSTARTNAME.borough AS borough_start,
9   ZIPSTARTNAME.neighborhood AS neighborhood_start,
10  ZIPEND.zip_code AS zip_code_end,
11  ZIPENDNAME.borough AS borough_end,
12  ZIPENDNAME.neighborhood AS neighborhood_end,
13  -- Since we're using trips from 2014 and 2015, we will add 5 years to make it look recent
14  DATE_ADD(DATE(TRI.starttime), INTERVAL 5 YEAR) AS start_day,
15  DATE_ADD(DATE(TRI.stoptime), INTERVAL 5 YEAR) AS stop_day,
16  WEA.temp AS day_mean_temperature, -- Mean temp
17  WEA.wdsp AS day_mean_wind_speed, -- Mean wind speed
18  WEA.prpc AS day_total_precipitation, -- Total precipitation
19  -- We will group trips into 10 minute intervals, which also reduces the number of rows
20  ROUND(CAST(TRI.tripduration / 60 AS INT64), -1) AS trip_minutes,
21  TRI.bikeid
22 FROM
23   `bigquery-public-data.new_york_citibike.citibike_trips` AS TRI
24 INNER JOIN
25   `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPSTART
26   ON ST_WITHIN(
27     ST_GEOGPOINT(TRI.start_station_longitude, TRI.start_station_latitude),
28     ZIPSTART.zip_code_geom)
29 INNER JOIN
30   `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPEND
31   ON ST_WITHIN(
32     ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
33     ZIPEND.zip_code_geom)
34 INNER JOIN
35   -- https://pantheon.corp.google.com/bigquery?p=bigquery-public-data&d=noaa_gsod
36   `bigquery-public-data.noaa_gsod.gsod20*` AS WEA
37   ON PARSE_DATE("%Y%m%d", CONCAT(WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
38 INNER JOIN
39   -- Note! Add your zipcode table name, enclosed in backticks: `example_table`
40   `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPSTARTNAME
41   ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
42 INNER JOIN
43   -- Note! Add your zipcode table name below, enclosed in backticks: `example_table`
44   `arched-history-445103-q6.cyclistNYC.zipcode` AS ZIPENDNAME
45   ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
46 WHERE
47   -- Take the weather from one weather station
48   WEA.wban = '94728' -- NEW YORK CENTRAL PARK
49   -- Use data for three summer months
50   AND DATE(TRI.starttime) BETWEEN DATE('2015-07-01') AND DATE('2015-09-30')
```