Activity Overview

In this activity, you will use your knowledge of SQL and potentially Google Dataflow to combine and move the key datasets you identified for the Cyclistic project into a target table. This represents the extraction phase of an ETL pipeline, when data is pulled from different sources and moved to its destination. You will use the table you create in this activity to develop the final dashboard for stakeholders. As you complete this activity, remember to refer to the previous work you did when completing the business intelligence project documents for Cyclistic for details about the Cyclistic project, as well as the activity to create a target table in BigQuery for a refresher on target tables.

Be sure to complete this activity before moving on. The next course item will provide you with a completed exemplar to compare to your own work. You will not be able to access the exemplar until you have completed this activity.

Scenario

Review the following scenario. Then, complete the activity. As a reminder, the end-of-course project activities are more open to your personal interpretation than other activities in this program. This is to give you an opportunity to practice the skills you have been learning in your own way. If you need help or feel stuck, you can always discuss your work with other learners in the discussion forums or review the exemplar to help guide your process.

The product development team at Cyclistic has begun developing their business plan for next year. In order to build a better Cyclistic, the team needs to understand how customers are currently using the bikes, how location and other factors impact demand, and what stations get the most traffic. The Cyclistic team has a few goals:

- Understand current customers needs, what makes a successful product, and how new stations might alleviate demand in different geographical areas
- Understand current usage of bikes at different locations
- Apply customer usage insights to inform new station growth
- Understand how different users (subscribers and non-subscribers) use the bikes

You met with stakeholders to complete project planning documents and uploaded the necessary tables into your BigQuery project space.

Instructions

Follow the instructions and answer the following question to complete the activity. Then, go to the next course item to compare your work to a completed exemplar.

Step 1: Log into your GCP tool

To begin this activity, log into your Google Cloud account and navigate to the BigQuery console. You can complete this activity using the BigQuery Sandbox, which does not require a Google Cloud billing account. You can learn more about enabling the Sandbox from the BigQuery help guide. You can also use Dataflow to execute SQL code as a Job by navigating to the Dataflow console instead; this will require you to have a Google Cloud account. Both tools are useful for this project, so choose the tool you are more interested in working with for this project.

Step 2: Querying your data

For this step, keep in mind the key metrics you and your stakeholders have identified, their business questions, and what data you'll need to develop the final dashboard. Previously, you explored the different public datasets your stakeholders provided and uploaded the zip code table your colleague shared with you. For the final dashboard, you will need to create two target tables: a table to capture the entire year and a table that focuses on summer trends. Here is an example of a query to capture a table with data from the entire year:

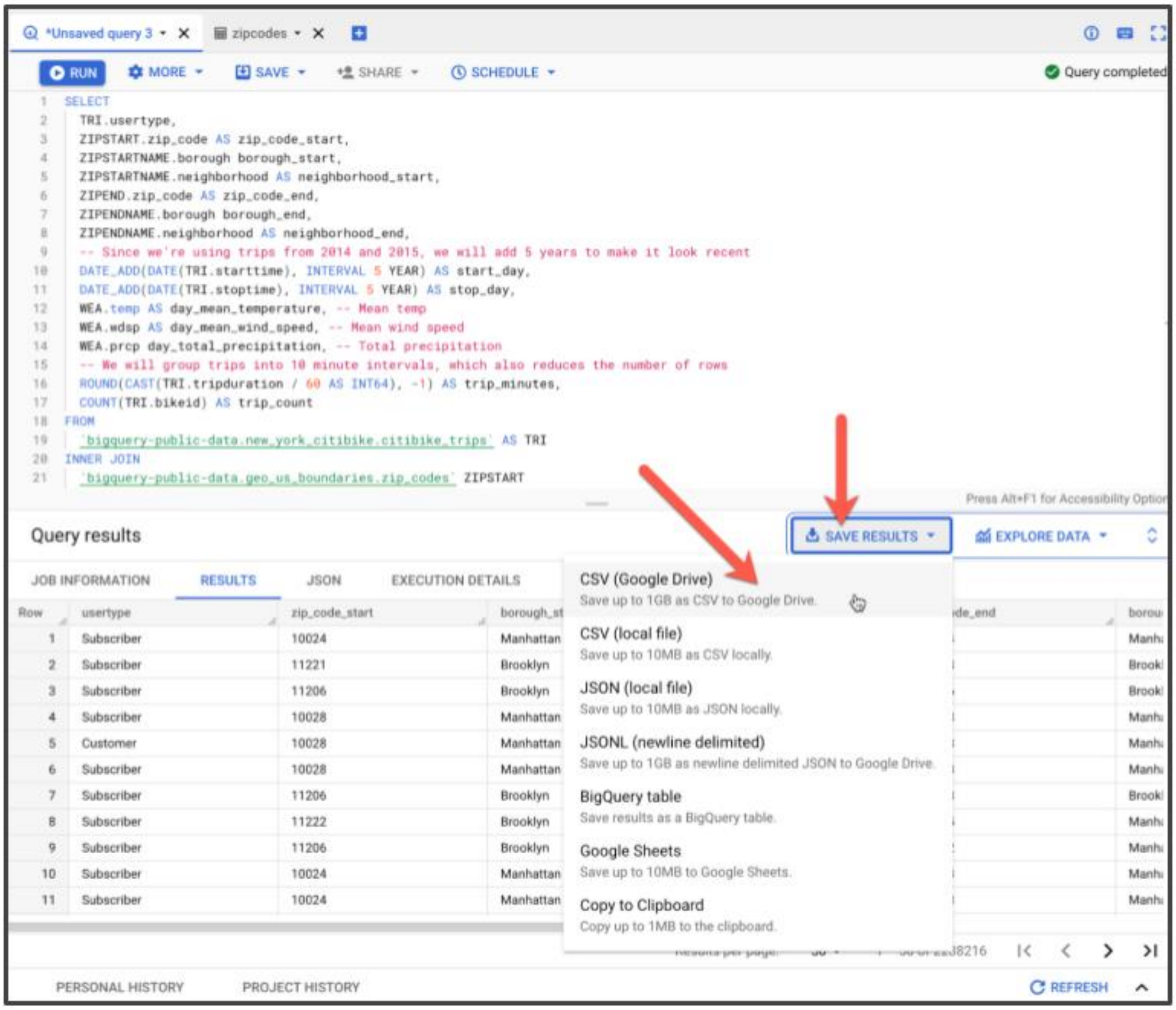
```
SELECT
  TRI.usertype,
  ZIPSTART.zip_code AS zip_code_start,
  ZIPSTARTNAME.borough borough_start,
  ZIPSTARTNAME.neighborhood AS neighborhood_start,
  ZIPEND.zip_code AS zip_code_end,
  ZIPENDNAME.borough borough_end,
  ZIPENDNAME.neighborhood AS neighborhood_end,
  -
- Since this is a fictional dashboard, you can 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
  -- 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),
    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
```

```
  -- Note! Add your zip code table name, enclosed in backticks: `example_table`
  `(insert your table name) zipcodes` AS ZIPSTARTNAME
  ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
INNER JOIN
  -- Note! Add your zipcode table name, enclosed in backticks: `example_table`
  `(insert your table name) zipcodes` 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,
  13
```

Note that this query includes a DATE_ADD function to add five years to the data. The public data you are using to create this dashboard is from 2014 and 2015, so this is a way to make your dashboard appear more recent. Normally, you would not change the dates in a dataset, but because this is a fictional project, you can include this in your own query. This part of the query is optional; however, your exemplar will appear differently if you don't include it. You will need to develop a similar query to capture a table with data from July through September to explore summer trends specifically.

Step 3: Finish the job

Once you execute the code, it will take a few moments to process. After the query has finished running, you will be able to download the tables as CSV files by using the Save Results dropdown and selecting the appropriate file type.



*Unsaved query 3
zipcodes
RUN
MORE
SAVE
SHARE
SCHEDULE
Query completed
Press Alt+F1 for Accessibility Option
Query results
SAVE RESULTS
EXPLORE DATA
JOB INFORMATION
RESULTS
JSON
EXECUTION DETAILS
CSV (Google Drive)
Save up to 1GB as CSV to Google Drive.
CSV (local file)
Save up to 10MB as CSV locally.
JSON (local file)
Save up to 10MB as JSON locally.
JSONL (newline delimited)
Save up to 1GB as newline delimited JSON to Google Drive.
BigQuery table
Save results as a BigQuery table.
Google Sheets
Save up to 10MB to Google Sheets.
Copy to Clipboard
Copy up to 1MB to the clipboard.
38216
PERSONAL HISTORY
PROJECT HISTORY
REFRESH

```
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 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,
  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
```

Row	usertype	zip_code_start	borough_st	de_end	borou
1	Subscriber	10024	Manhattan		Manh
2	Subscriber	11221	Brooklyn		Brook
3	Subscriber	11206	Brooklyn		Brook
4	Subscriber	10028	Manhattan		Manh
5	Customer	10028	Manhattan		Manh
6	Subscriber	10028	Manhattan		Manh
7	Subscriber	11206	Brooklyn		Brook
8	Subscriber	11222	Brooklyn		Manh
9	Subscriber	11206	Brooklyn		Manh
10	Subscriber	10024	Manhattan		Manh
11	Subscriber	10024	Manhattan		Manh

This might take a few minutes. Once you have downloaded the table, you will be ready to upload it to Tableau to create your dashboard!

What to Include in Your Response

Be sure to address the following criteria:

- Necessary tables are successfully combined into summary tables
- Appropriate tables are downloaded and ready to upload to Tableau