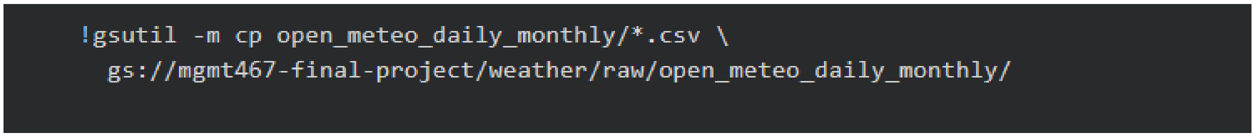
Infrastructure Setup Details

Gcloud Commands:

We didn’t use any direct gcloud commands during our project as almost everything in the infrastructure was created through the GCP web console. We did use a gsutil command that’s included below.



```
!gsutil -m cp open_meteo_daily_monthly/*.csv \
  gs://mgmt467-final-project/weather/raw/open_meteo_daily_monthly/
```

We used this command in our Colab notebook to copy the historical 2022, 2023, and 2024 forecast data files that we retrieved from the API to our GCS raw weather data bucket.

Enabling APIs:

BigQuery API: Enabling this API is important for our BigQuery SQL to create our external tables with the raw data, curated tables with the refined data & columns, and our BQML models that we use for trip predictions.

BigQuery Storage API: Dataflow utilizes this to read from/ write to BigQuery which is important for our Pub/Sub to write to our weather_forecast_stream raw weather data table.

Cloud Pub/Sub API: Used to manage messages through our topic (citibike-weather-topic) and our subscription (citibike-weather-sub) so that Cloud Run service can publish the daily forecast JSON messages to the topic. Dataflow subscribes to citibike-weather-sub and then allows us to push the data into our BigQuery table.

Dataflow API: Manages Dataflow jobs and allows us to read the JSON messages from Pub/Sub, then write them to our raw data table.

Cloud Run Admin API: Needed so we can deplay and update our open-meteo Cloud Run service.

Cloud Scheduler API: Used in our open-meteo-every-minute job to call the Cloud Run every minute, which in turn calls the Open-Meteo API and publishes to Pub/Sub

Cloud Storage JSON API: Used to read CSVs such as our CitiBike CSVs and the historical weather data we upload via our gsutil command.

Solution Architect Roles:

Lucas Gerbsch: Designed and built the full end-to-end data pipeline and Looker Studio using a combination of ChatGPT and Gemini to learn about Cloud Scheduling and Cloud Run services to automate the API retrieval process. Also designed the bucket structure to keep source data organized, the raw data database to combine all of the raw data into single tables, and the curated database that holds our curated tables for ML model training and Looker Studio visualizations. I had a lot of fun learning how to do the Cloud Scheduling and Run services, very cool seeing it all come together and update in real time.