Faculty of Engineering and Applied Science

Cloud Computing

Final Project Final Report

Group Q2	
Name	Student ID
Manreet Kaur	100766207
Nick Lang	100592096
Shreya Patel	100747036
Graem Sheppard	100700978

Introduction

As larger vehicles (SUVs and trucks) become more prevalent on our roads, it is an ever-growing concern as to the environmental damage that these vehicles cause. Bigger vehicles require more energy to move the same amount due to their increased mass and larger drag coefficients. As discussed in the proposal, there are various incentives car manufacturers have to encourage customers to buy these larger vehicles. The purpose of this project is to determine the impact that this social phenomenon has.

Dataset

For this project, we used the HighD dataset. The most important consideration for the data is that it is applicable to real-world situations and how these vehicles are actually used and driven. HighD provides high quality data for actual driving behaviour on highways. With it, we are able to precisely measure the information we need to conduct our analysis.

Tools

For this project, we leveraged Google BigQuery to manage and analyze our data. This cloud tool is capable of extremely rapid and efficient processing and querying of large datasets. In order to use this tool effectively we had to upload our dataset and then perform the queries as we needed. BigQuery proved to be an excellent tool given its scaling ability and ease of use.

Google Pub/Sub was used for data ingestion into BigQuery. We needed a powerful, flexible tool to allow us to transfer the relatively large dataset into our BigQuery database. Pub/sub allowed us to set up a messenger/receiver connection between local storage and BigQuery to transmit the data we needed.

Data Ingestion Code:

```
from google.cloud import pubsub
from google.oauth2 import service_account
import io
import pandas as pd
import os
from glob import glob

project_id = 'project-382903'

credentials = service_account.Credentials.from_service_account_file('creds.json')
publisher = pubsub.PublisherClient(credentials = credentials)

tracks_path = publisher.topic_path(project_id, 'highd-tracks')
tracks_meta_path = publisher.topic_path(project_id, 'highd-tracksMeta')
recording_meta_path = publisher.topic_path(project_id, 'highd-recordingMeta')

files = glob('highd-dataset-v1.0/data/*tracksMeta.csv')
for file in files:
   df = pd.read_csv(file)
   if 'tracksMeta' in file:
       df['recordingId'] = int(file.split('/')[2][0:2])
   tableName = file.split('/')[2].split('_')[1].split('.')[0]
   df['table'] = tableName
   for index, row in df.iterrows():
       topic_path = None
       if tableName == 'tracksMeta':
           topic_path = tracks_meta_path
       elif tableName == 'tracks':
           topic_path = tracks_path
       elif tableName == 'recordingMeta':
           topic_path = recording_meta_path

       json_string = row.to_json()
       future = publisher.publish(topic_path, json_string.encode('utf-8'))
       print(future.result())
```

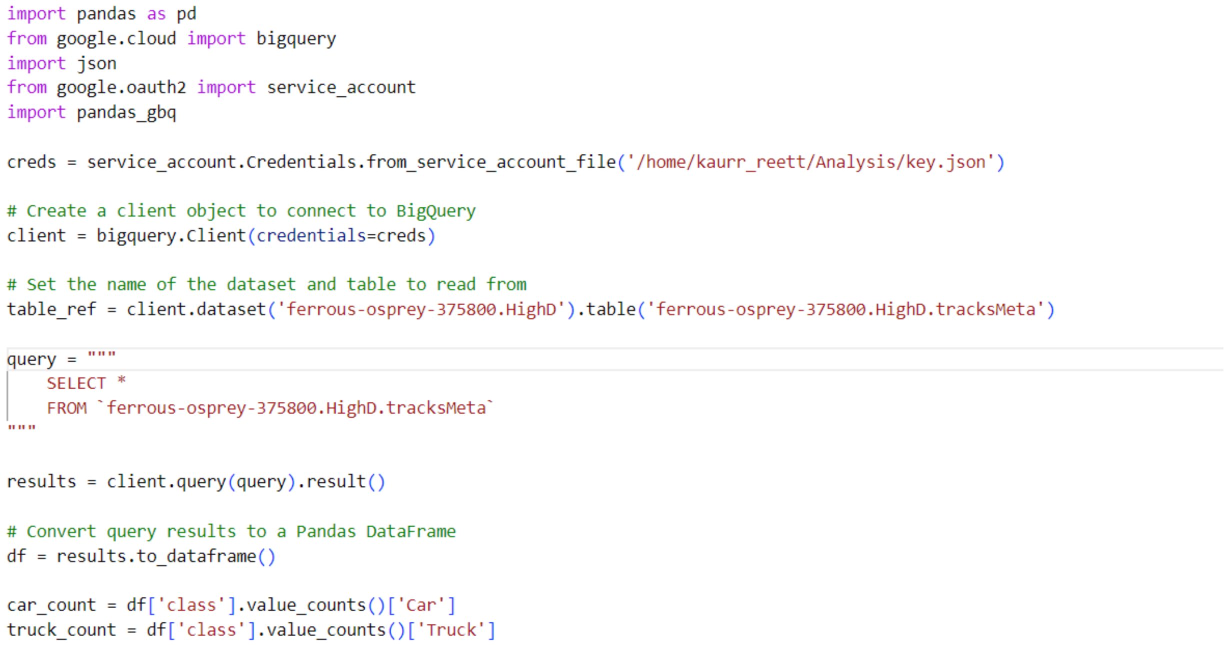
For data visualization, we used Google Looker Studio. This cloud tool is useful for converting raw data into attractive and readable visualizations. By using this tool we were able to take our results and create any graphs or charts that we needed to examine our solution and draw conclusions.

Finally, we used Github and the Google Drive suite for collaborating and sharing work.

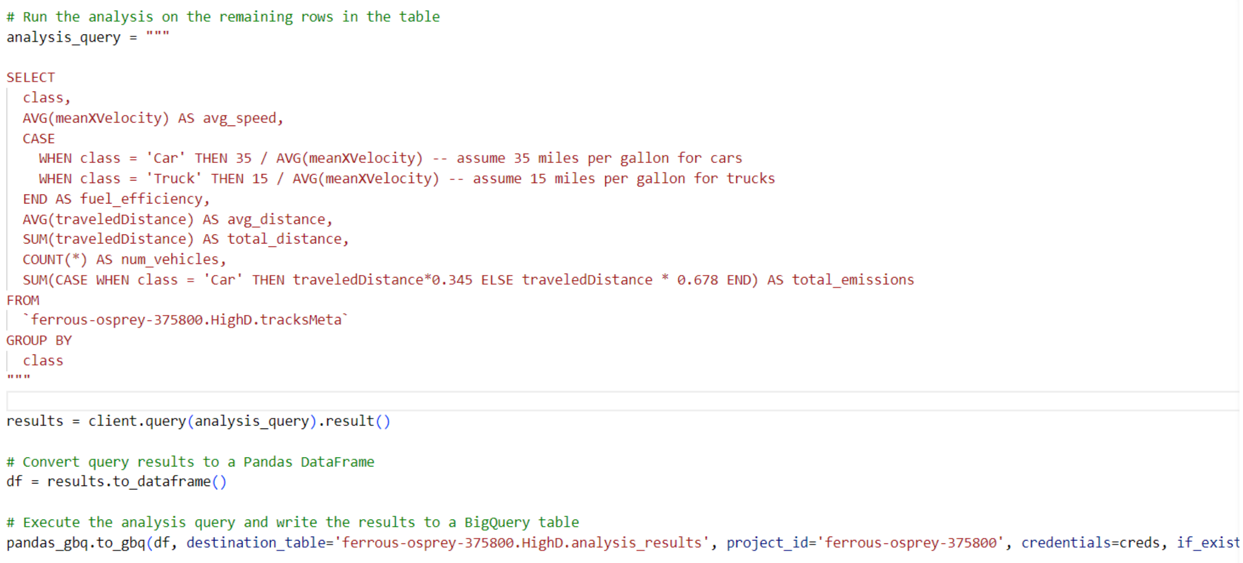
Data Analysis :

Analyzing our data involved taking the data stored in the database and using SQL queries to average or sum data as required, then returning it in its proper presentation from BigQuery. The two most important resulting pieces of data were the fuel efficiency metric and total emissions, which are generally inversely proportional metrics based on fuel consumption, distance travelled, and speed.

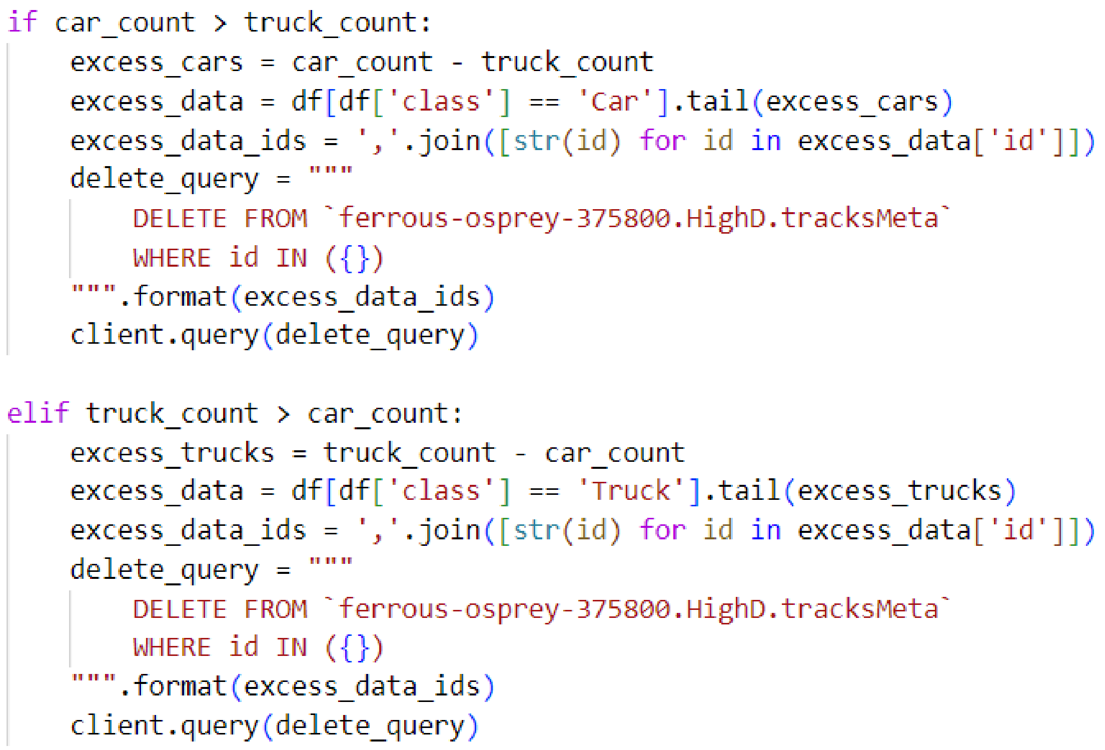
Code :



```python
import pandas as pd
from google.cloud import bigquery
import json
from google.oauth2 import service_account
import pandas_gbq

creds = service_account.Credentials.from_service_account_file('/home/kaurr_reett/Analysis/key.json')

# Create a client object to connect to BigQuery
client = bigquery.Client(credentials=creds)

# Set the name of the dataset and table to read from
table_ref = client.dataset('ferrous-osprey-375800.HighD').table('ferrous-osprey-375800.HighD.tracksMeta')

query = """
    SELECT *
    FROM `ferrous-osprey-375800.HighD.tracksMeta`
"""

results = client.query(query).result()

# Convert query results to a Pandas DataFrame
df = results.to_dataframe()

car_count = df['class'].value_counts()['Car']
truck_count = df['class'].value_counts()['Truck']
```



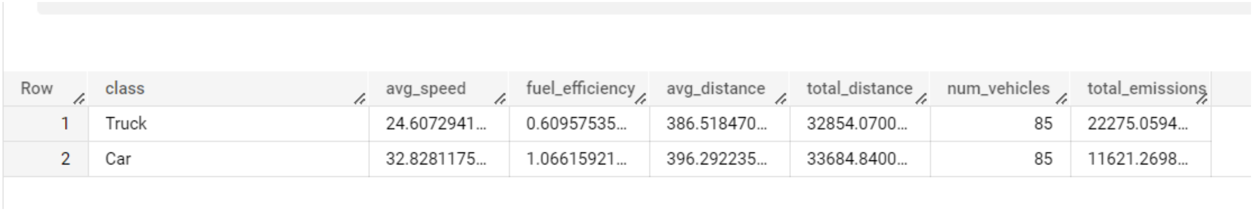
```python
# Run the analysis on the remaining rows in the table
analysis_query = """

SELECT
  class,
  AVG(meanXVelocity) AS avg_speed,
  CASE
    WHEN class = 'Car' THEN 35 / AVG(meanXVelocity) -- assume 35 miles per gallon for cars
    WHEN class = 'Truck' THEN 15 / AVG(meanXVelocity) -- assume 15 miles per gallon for trucks
  END AS fuel_efficiency,
  AVG(traveledDistance) AS avg_distance,
  SUM(traveledDistance) AS total_distance,
  COUNT(*) AS num_vehicles,
  SUM(CASE WHEN class = 'Car' THEN traveledDistance*0.345 ELSE traveledDistance * 0.678 END) AS total_emissions
FROM
  `ferrous-osprey-375800.HighD.tracksMeta`
GROUP BY
  class
"""

results = client.query(analysis_query).result()

# Convert query results to a Pandas DataFrame
df = results.to_dataframe()

# Execute the analysis query and write the results to a BigQuery table
pandas_gbq.to_gbq(df, destination_table='ferrous-osprey-375800.HighD.analysis_results', project_id='ferrous-osprey-375800', credentials=creds, if_exist
```



```
if car_count > truck_count:
    excess_cars = car_count - truck_count
    excess_data = df[df['class'] == 'Car'].tail(excess_cars)
    excess_data_ids = ','.join([str(id) for id in excess_data['id']])
    delete_query = """
        DELETE FROM `ferrous-osprey-375800.HighD.tracksMeta`
        WHERE id IN ({})
    """.format(excess_data_ids)
    client.query(delete_query)

elif truck_count > car_count:
    excess_trucks = truck_count - car_count
    excess_data = df[df['class'] == 'Truck'].tail(excess_trucks)
    excess_data_ids = ','.join([str(id) for id in excess_data['id']])
    delete_query = """
        DELETE FROM `ferrous-osprey-375800.HighD.tracksMeta`
        WHERE id IN ({})
    """.format(excess_data_ids)
    client.query(delete_query)
```

Results

Our results show that trucks are causing massively more damage to the environment than cars are. BigQuery allows us to easily determine this data given our dataset, and Looker Studio allows us to visualize the data in a clear way.



Row	class	avg_speed	fuel_efficiency	avg_distance	total_distance	num_vehicles	total_emissions
1	Truck	24.6072941...	0.60957535...	386.518470...	32854.0700...	85	22275.0594...
2	Car	32.8281175...	1.06615921...	396.292235...	33684.8400...	85	11621.2698...

Fig.-1: Raw average data from BigQuery

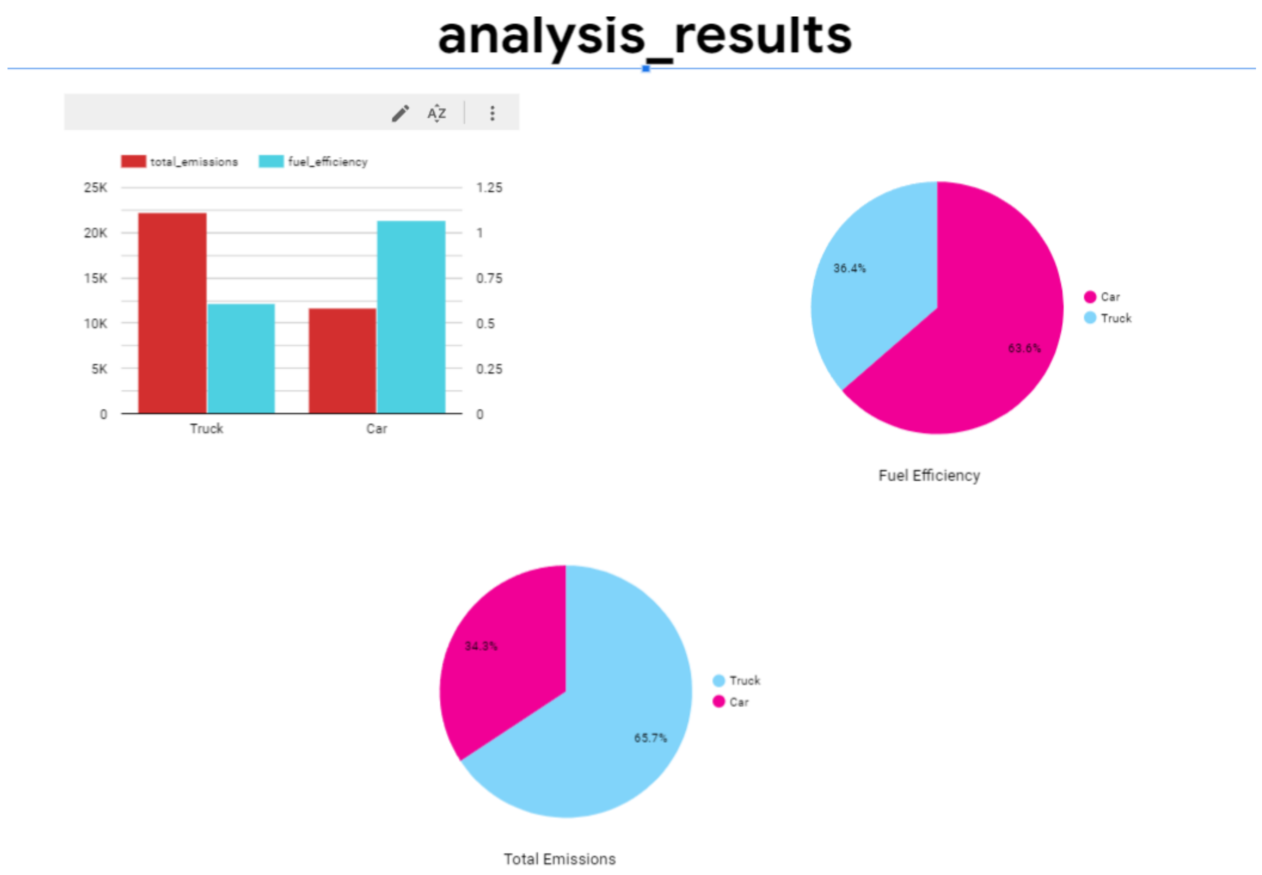


Fig.-2: Visualized data using Looker Studio

Demo Video(Data Ingestion with Pub/Sub)
https://youtu.be/L9dyIdWMvPU
Demo Video(Data Analysis and Visualization):
https://drive.google.com/file/d/1AUqB6R3ABId5RS-9DIM_t4CEmkiKEgRG/view?usp=sharing

Conclusions

Using these cloud tools, we were able to show the difference in environmental impact between cars and trucks from the perspective of fuel consumption and emissions. These tools allowed us to perform this study in an efficient and performant way. We believe that overall, we have conclusively shown that there is a large difference in impact between the two classes of vehicles and that more effort should be put in to curb the growing trend of larger vehicle purchases.