Matteo Botta

Car sales project

Showcasing skills: upload data to BigQuery, data cleaning and analysis **using SQL**

Dataset source: https://archive.ics.uci.edu/ml/datasets/Automobile

Task: The dataset contains historical car sales data from 1985, including details such as car features and prices. Clean the dataset and report the main characteristics of the top 5 most expensive cars.

Steps:

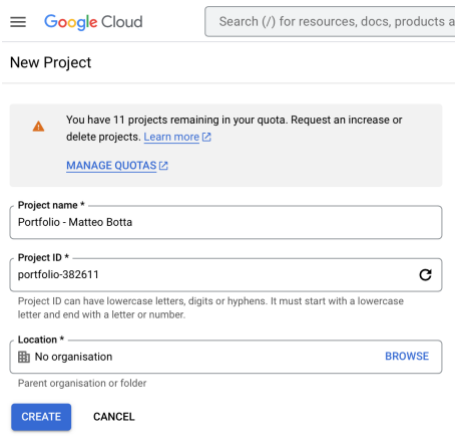
1) Data sourcing
2) Upload data to BigQuery
3) Data cleaning strategies using SQL language
4) Data analysis and reporting

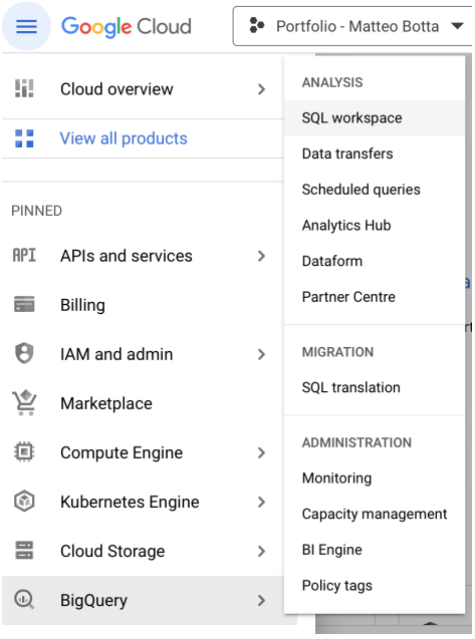
1) Data sourcing

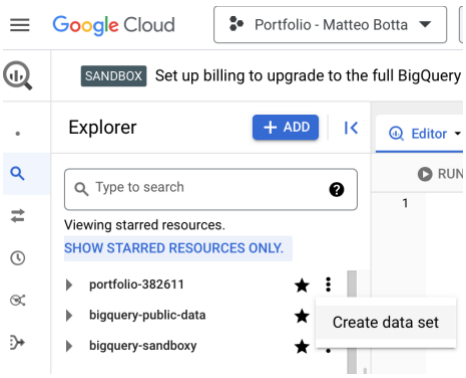
Relevant CSV datafile downloaded from data source

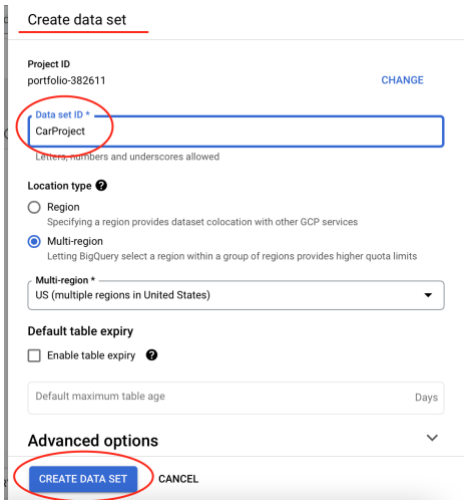
2) Upload data to BigQuery

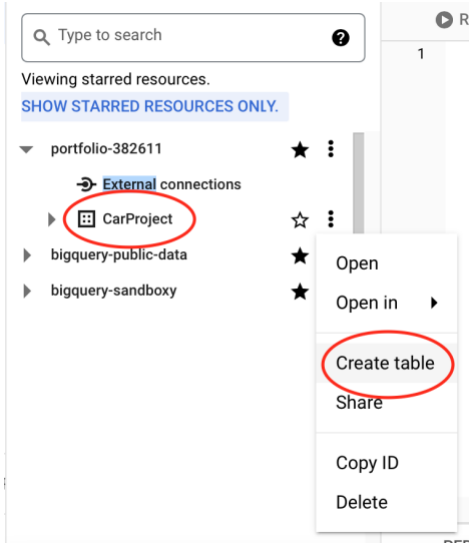
Project, dataset and table creation in BigQuery and upload of the CSV datafile



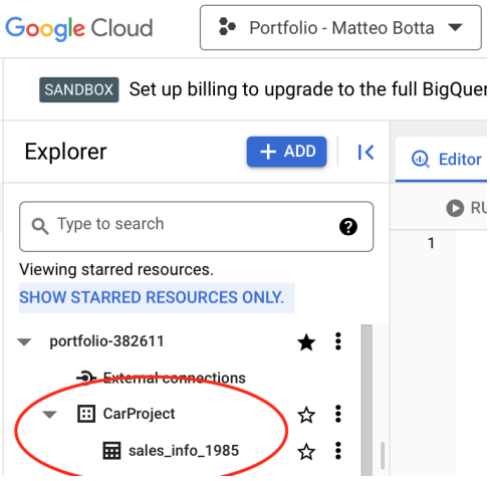

Google Cloud
Portfolio - Matteo Botta
Cloud overview
View all products
PINNED
APIs and services
Billing
IAM and admin
Marketplace
Compute Engine
Kubernetes Engine
Cloud Storage
BigQuery
ANALYSIS
SQL workspace
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Centre
MIGRATION
SQL translation
ADMINISTRATION
Monitoring
Capacity management
BI Engine
Policy tags


Google Cloud
Portfolio - Matteo Botta
SANDBOX
Set up billing to upgrade to the full BigQuery
Explorer
+ ADD
Editor
RUN
Type to search
Viewing starred resources.
SHOW STARRED RESOURCES ONLY.
portfolio-382611
bigquery-public-data
bigquery-sandboxy
Create data set


Create data set
Project ID
portfolio-382611
CHANGE
Data set ID *
CarProject
Letters, numbers and underscores allowed
Location type
Region
Specifying a region provides dataset colocation with other GCP services
Multi-region
Letting BigQuery select a region within a group of regions provides higher quota limits
Multi-region *
US (multiple regions in United States)
Default table expiry
Enable table expiry
Default maximum table age
Days
Advanced options
CREATE DATA SET
CANCEL


Type to search
Viewing starred resources.
SHOW STARRED RESOURCES ONLY.
portfolio-382611
External connections
CarProject
bigquery-public-data
bigquery-sandboxy
Open
Open in
Create table
Share
Copy ID
Delete


Create table
Source
Create table from
Upload
Select file *
automobile_data.csv
File format
CSV
Destination
Project *
portfolio-382611
Data set *
CarProject
Table *
sales_info_1985
Table type
Native table
CREATE TABLE
CANCEL


Google Cloud
Portfolio - Matteo Botta
SANDBOX
Set up billing to upgrade to the full BigQuer
Explorer
ADD
Editor
Type to search
Viewing starred resources.
SHOW STARRED RESOURCES ONLY.
portfolio-382611
External connections
CarProject
sales_info_1985

3) Data cleaning strategies using SQL language

- Check for duplicates:

```
SELECT
  DISTINCT *
FROM
  `portfolio-382611.CarProject.sales_info_1985`
```

We have 1 duplicate. Returned 202 distinct entries, but there are 203 rows.

- Inspect fuel_type column

```
SELECT
  DISTINCT fuel_type
FROM
  `portfolio-382611.CarProject.sales_info_1985`
```

Job information Results

Row	fuel_type
1	gas
2	diesel

As described in the data description table: “gas” and “diesel”.

- Inspect car length column

```
SELECT
  MIN(length) AS min_length,
  MAX(length) AS max_length
FROM
  `portfolio-382611.CarProject.sales_info_1985`
```

Row	min_length	max_length
1	141.1	208.1

Car length range as described.

- Fill in missing data

```
SELECT
  *
FROM
  `portfolio-382611.CarProject.sales_info_1985`
WHERE
  num_of_doors IS NULL
```

Row	make	fuel_type	num_of_doors	body_style
1	dodge	gas	*null*	sedan
2	mazda	diesel	*null*	sedan

Two entries have NULL values. Was told that Dodge gas sedans and Mazda diesel sedans where all 4-doors, so carried on with replacing the NULL value with “four” doors, for both entries.

```
UPDATE
  `portfolio-382611.CarProject.sales_info_1985`
SET
  num_of_doors = "four"
WHERE
  make = "dodge"
  AND fuel_type = "gas"
  AND body_style = "sedan";
```

- Inspect num_of_cylinders column

```
SELECT
  DISTINCT num_of_cylinders
FROM
  `portfolio-382611.CarProject.sales_info_1985`
```

JOB INFORMATION RESULTS

Row	num_of_cylinders
1	four
2	six
3	five
4	three
5	twelve
6	two
7	tow
8	eight

Found a typo in row 7. Proceeded to fix it.

```
UPDATE
  `portfolio-382611.CarProject.sales_info_1985`
SET
  num_of_cylinders = "two"
WHERE
  num_of_cylinders = "tow"
```

- Inspect compression_ratio column

```
SELECT
  MIN (compression_ratio) AS min_compr_ratio,
  MAX (compression_ratio) AS max_compr_ratio
FROM
```

```
  `portfolio-382611.CarProject.sales_info_1985`
```

JOB INFORMATION RESULTS

Row	min_compr_ratio	max_compr_rati
1	7.0	70.0

Found an error. Data description gives 7-23 as the range for compression ratio. 70 was most likely meant to be a 7. Running a new query excluding the value 70 to make sure that the rest of the values fall within the expected range of 7 to 23.

```
SELECT
  MIN (compression_ratio) AS min_compr_ratio,
  MAX (compression_ratio) AS max_compr_ratio
FROM
  `portfolio-382611.CarProject.sales_info_1985`
WHERE
  compression_ratio <> 70
```

JOB INFORMATION RESULTS

Row	min_compr_ratio	max_compr_rati
1	7.0	23.0

This confirms that 70 was a typo and that the rest of the values fall within the expected range. Before deleting it, I checked how many rows contain wrong value.

```
SELECT
 COUNT(*) AS rows_wrong_value
FROM
  `portfolio-382611.CarProject.sales_info_1985`
WHERE
  compression_ratio = 70
```

JOB INFORMATION RESULTS

Row	rows_wrong_value
1	1

Only one row contains the wrong value in the compression_ratio column. It can be deleted

```
DELETE
  `portfolio-382611.CarProject.sales_info_1985`
WHERE
  compression_ratio = 70
```

- Inspect drive_wheels column

```
SELECT
  DISTINCT drive_wheels
FROM
  `portfolio-382611.CarProject.sales_info_1985`
```

JOB INFORMATION RESULTS

Row	drive_wheels
1	rwd
2	fwd
3	4wd
4	4wd

4wd is reported twice. I checked for extra white spaces.

```
SELECT
  DISTINCT drive_wheels,
  LENGTH (drive_wheels) AS string_length
FROM
  `portfolio-382611.CarProject.sales_info_1985`
```

JOB INFORMATION RESULTS JSON

Row	drive_wheels	string_length
1	rwd	3
2	fwd	3
3	4wd	4
4	4wd	3

Error found. In row three the length of the string is 4 when it should be 3. Preceding with trimming the extra white spaces.

```
UPDATE
  `portfolio-382611.CarProject.sales_info_1985`
SET
  drive_wheels = TRIM(drive_wheels)
WHERE TRUE
```

4) Data analysis and reporting

Objective: Find out characteristics of the top 5 most expensive cars

```
SELECT
  *
FROM
  `portfolio-382611.CarProject.sales_info_1985`
ORDER BY price DESC
LIMIT 5
```

I then exported the results in Google Sheets and made the following table that summarises the features of the top 5 most expensive cars.

Make	Price ($)	Body style	Num. of doors	Engine location	Num. of cylinders	Horsepower	City mpg
Mercedes-Benz	45400	hardtop	two	front	eight	184	14
BMW	41315	sedan	two	front	six	182	16
Mercedes-Benz	40960	sedan	four	front	eight	184	14
Porsche	37028	convertible	two	rear	six	207	17
BMW	36880	sedan	four	front	six	182	15