

INFORME. DATOS COCHES

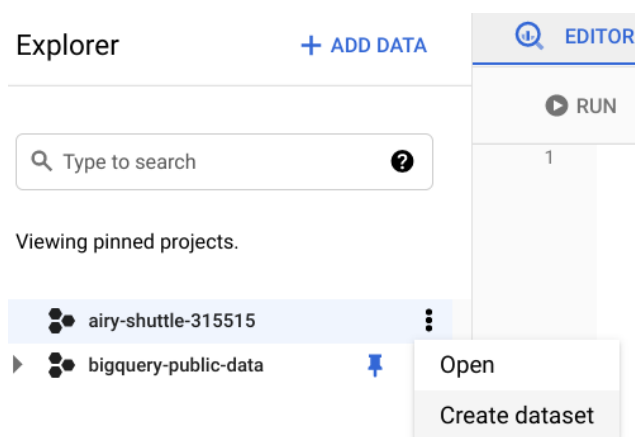
En este informe recojo los pasos realizados para explorar y limpiar los datos seleccionados.

Seleccionar conjunto de datos

A través de Big Query busco un conjunto de datos y creo una tabla personalizada para almacenarlos, en este caso automobile_data, con el fin de utilizar consultas SQL para explorarlos y limpiarlos.

Paso 1: Crear un conjunto de datos

Una vez que abrimos Big Query y descargamos el archivo de automobile_data, en el panel del explorador, seleccionamos Crear conjunto de datos.



Desde el menú Crear conjunto de datos, completamos la información sobre el conjunto de datos. Introduce el ID del conjunto de datos como coches y hacemos clic en CREAR CONJUNTO DE DATOS.

Create dataset

Dataset ID *
cars
Letters, numbers, and underscores allowed

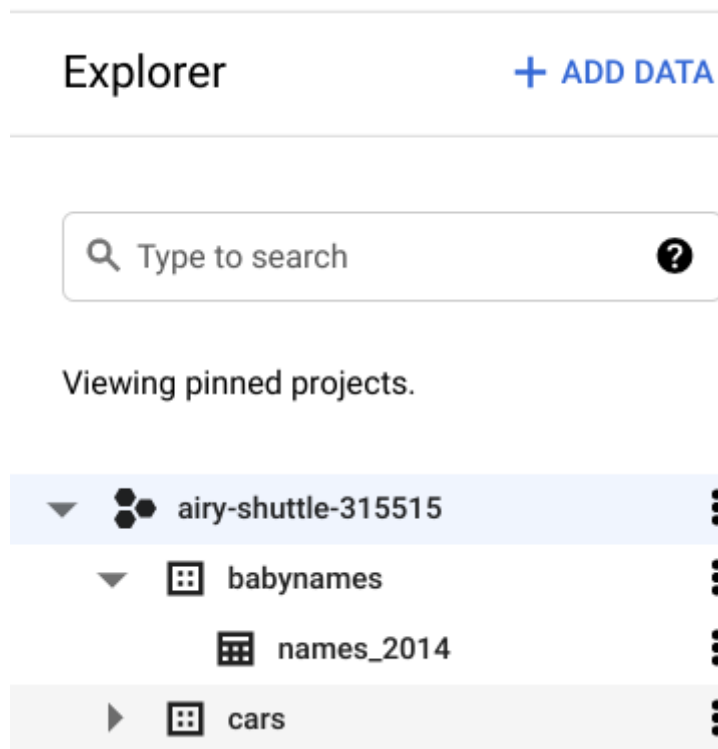
Data location
Default

Default table expiration
☐ Enable table expiration
Default maximum table age Days

Encryption
☒ Google-managed encryption key
No configuration required
☐ Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

CREATE DATASET CANCEL

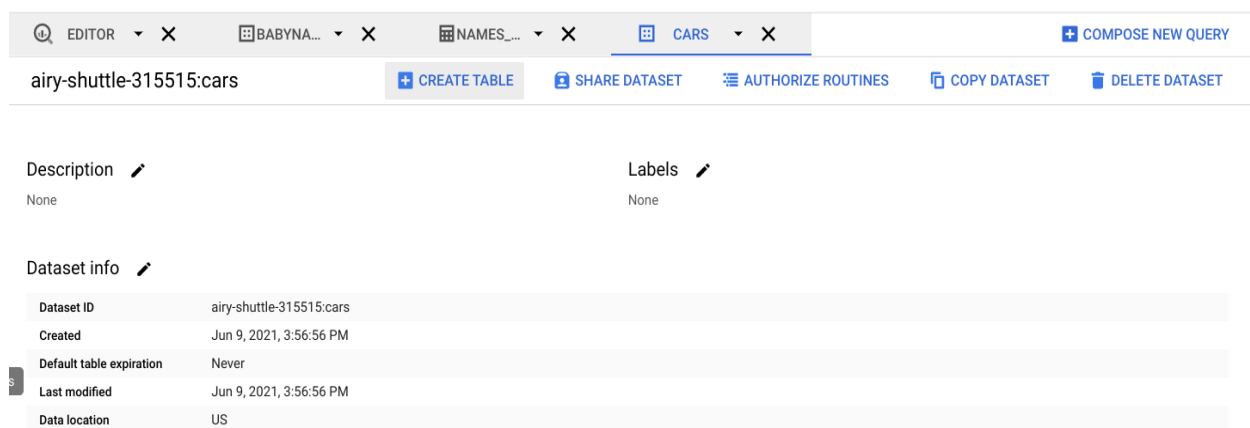
El conjunto de datos de coches debería aparecer bajo tu proyecto en el panel del Explorador como se muestra a continuación. Haz clic en los tres puntos junto al conjunto de datos de coches para abrirlo.



Paso 2: Crear tabla

Después de abrir el conjunto de datos recién creado, podrás añadir una tabla personalizada para tus datos.

Desde el conjunto de datos de coches, haz clic en CREAR TABLA.



En Origen, carga el CSV de automobile_data. En Destino, asegúrate de que estás cargando en tu conjunto de datos de coches y nombra tu tabla *car_info*. Puedes configurar el esquema en Auto-detect. Luego haz clic en Crear tabla.

Create table

Source

Create table from: Upload Select file: automobile_data (1).csv Browse File format: CSV

Destination

☒ Search for a project ☐ Enter a project name

Project name: test Dataset name: cars Table type: Native table

Table name: car_info

Schema

Auto detect

☒ Schema and input parameters

Schema will be automatically generated.

Partition and cluster settings

Partitioning: No partitioning

Clustering order (optional): Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Comma-separated list of fields to define clustering order (up to 4)

Create table Cancel

Una vez creada la tabla, aparecerá en el panel del Explorador. Puedes hacer clic en la tabla para explorar el esquema y previsualizar los datos.

Limpieza de datos

El nuevo conjunto de datos contiene información histórica de las ventas, incluyendo detalles como las características y los precios de los coches.

Usaremos estos datos para encontrar los 10 coches y acabados más populares. Pero antes de realizar el análisis, tenemos que asegurarnos de que los datos están limpios, puesto que si analizamos datos sucios, podrías acabar presentando una lista de coches equivocada a los inversores. Eso podría hacerles perder dinero en su inversión para el inventario de coches.



Paso 1: Inspeccionar los datos

Lo primero que debemos hacer es inspeccionar los datos de la tabla para saber si hay que hacer alguna limpieza específica.

Según la descripción de los datos, la columna `fuel_type` solo debería tener dos valores de cadena, diésel o gasolina. Para comprobar y asegurar que es cierto, ejecutamos la siguiente consulta:

```
SELECT DISTINCT fuel_type FROM cars.car_info;
```

Esto devuelve los siguientes resultados:

Query results		 SAVE RESULTS	 EXPLORE DATA ▾
Query complete (0.6 sec elapsed, 1 KB processed)			
Job information		Results	JSON Execution details
Row	fuel_type		
1	gas		
2	diesel		

Esto confirma que la columna `fuel_type` no tiene valores inesperados.

A continuación, inspeccionamos la columna `length` que debe contener las longitudes mínimas y máximas de los coches, los cuales deben ajustarse a los datos recogidos en la descripción de los datos, es decir deben oscilar entre 141.1 y 208.1.

Ejecutamos esta consulta para confirmarlo:

```
SELECT MIN(length) AS min_length, MAX(length) AS max_length FROM cars.car_info;
```

Los resultados deberían confirmar que 141.1 y 208.1 son los valores mínimo y máximo, respectivamente, de esta columna.

Row	min_length	max_length
1	141.1	208.1

Paso 2: Completar los datos faltantes

La ausencia de valores puede dar lugar a errores o sesgar los resultados durante el análisis. Es necesario revisar los datos para ver si hay valores nulos o faltantes. Estos valores pueden aparecer como una celda en blanco o la palabra *null*.

Compruebo si la columna `num_of_doors` contiene valores nulos utilizando esta consulta:

```
SELECT * FROM cars.car_info  
WHERE num_of_doors IS NULL;
```

Esto seleccionará todas las filas con datos faltantes para la columna `num_of_doors` y las mostrará en la tabla de resultados.

Row	make	fuel_type	num_of_doors	body_style
1	dodge	gas	<i>null</i>	sedan
2	mazda	diesel	<i>null</i>	sedan

Obtenemos que tanto dodge como mazda contienen datos incompletos. Para rellenar estos valores, comprobamos que todos los sedanes de gasolina de Dodge y todos los sedanes diésel de Mazda vendidos tenían cuatro puertas.

Para actualizar la usamos la siguiente consulta:

```
UPDATE cars.car_info SET num_of_doors = "four" WHERE make = "dodge" AND fuel_type = "gas" AND body_style = "sedan";
```

Deberías recibir un mensaje indicando que se han modificado tres filas en esta tabla. Para asegurarte, puedes volver a ejecutar la consulta anterior:

```
SELECT * FROM cars.car_info
WHERE num_of_doors IS NULL;
```

Ahora, solo tienes una fila con un valor NULL para num_of_doors. Repetimos este proceso para sustituir el valor nulo del Mazda.

Paso 3: Identificar posibles errores

Una vez que hayas terminado de asegurarte de que no faltan valores en tus datos, comprobamos si hay otros errores potenciales.

En este caso nos fijaremos en la columna num_of_cylinders, y usaremos esta consulta:

```
SELECT DISTINCT num_of_cylinders FROM cars.car_info;
```

Después de ejecutar esto, obtenemos que hay una fila de más. Hay dos entradas para dos cilindros: las filas 6 y 7. Pero el *dos* de la fila 7 está mal escrito.

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

Para corregir la falta de ortografía de todas las filas, usamos la consulta:

```
UPDATE cars.car_info SET num_of_cylinders = "two" WHERE num_of_cylinders = "tow";
```

Recibirás un mensaje alertando de que se ha modificado una fila después de ejecutar esta instrucción. Para comprobar que ha funcionado, volvemos a usar la consulta anterior:

```
SELECT DISTINCT num_of_cylinders FROM cars.car_info;
```

A continuación, comprobaremos la columna `compression_ratio`. Según la descripción de los datos, los valores de la columna deben oscilar entre 7 y 23. Al igual que cuando comprobaste los valores de longitud, usamos MIN y MAX para comprobar si es correcto.

```
SELECT MIN(compression_ratio) AS min_compression_ratio, MAX(compression_ratio) AS max_compression_ratio
FROM cars.car_info;
```

Obtenemos que el máximo es de 70, pero sabemos que esto es un error porque el valor máximo de esta columna debería ser 23. Así que lo más probable es que el 70 sea un 7.0. Vuelve a ejecutar la consulta anterior sin la fila con 70 para asegurarte de que el resto de los valores están dentro del rango esperado de 7 a 23.

```
SELECT MIN(compression_ratio) AS min_compression_ratio, MAX(compression_ratio) AS max_compression_ratio
FROM cars.car_info
WHERE compression_ratio <> 70;
```

Ahora el valor más alto es 23, que coincide con la descripción de los datos, así que corregimos el valor de 70. Consultamos con el director de ventas, que dice que esta fila se hizo por error y debe eliminarse. Antes de borrar nada, comprobamos cuántas filas contienen este valor erróneo como precaución para no acabar borrando el 50% de los datos. Si hay demasiados (por ejemplo, el 20% de las filas tienen el valor 70 incorrecto), entonces deberías volver a consultar con el director de ventas para preguntar si deben eliminarse o si el 70 debe actualizarse a otro valor. Utiliza la siguiente consulta para contar cuántas filas estarías borrando:

```
SELECT COUNT * AS num_of_rows_to_delete
FROM cars.car_info
WHERE compression_ratio = 70;
```

Resulta que solo hay una fila con el valor erróneo de 70. Así que podemos eliminar esa fila usando la consulta:

```
DELETE cars.car_info
WHERE compression_ratio = 70;
```

Paso 4: Asegurar la coherencia

Por último, debemos comprobar que los datos no presentan incoherencias que puedan provocar errores. Estas incoherencias pueden ser difíciles de detectar, a veces, incluso algo tan simple como un espacio extra puede causar un problema.

Comprobamos si la columna `drive_wheels` presenta incoherencias ejecutando una consulta con una instrucción `SELECT DISTINCT`:

```
SELECT DISTINCT drive_wheels FROM cars.car_info;
```

Parece que `4wd` aparece dos veces en los resultados, sin embargo como hemos usado una instrucción `SELECT DISTINCT` para obtener valores únicos, es probable que haya un espacio adicional en una de las entradas de `4wd` que la haga diferente de las otras.

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

Para comprobar si este es el caso, usamos la instrucción `LENGTH` para determinar la longitud de cada una de estas variables de cadena:

```
SELECT DISTINCT drive_wheels, LENGTH(drive_wheels) AS string_length  
FROM cars.car_info;
```

Según estos resultados, algunas instancias de la cadena `4wd` tienen cuatro caracteres en lugar de los tres esperados (`4wd` tiene 3 caracteres). En ese caso, usamos la función `TRIM` para eliminar todos los espacios extra en la columna `drive_wheels`:

```
UPDATE cars.car_info SET drive_wheels = TRIM(drive_wheels)  
WHERE TRUE;
```

A continuación, ejecutamos de nuevo la instrucción `SELECT DISTINCT` para asegurarnos de que solo hay tres valores distintos en la columna `drive_wheels`:

```
SELECT DISTINCT drive_wheels  
FROM cars.car_info;
```

En este caso los resultados nos devuelven tres valores únicos en esta columna.

Tras ejecutar estos pasos obtenemos unos datos que están limpios, son coherentes y están listos para el análisis.