PROYECTO

Bases de Datos NoSQL

ÍNDICE

1.	INT	RODUCCIÓN	3
2.	ANA	ÁLISIS DE DATOS	4
,	2.1	Insertar base de datos	4
,	2.1	Conocimiento sobre la base de datos	5
,	2.2	Ejercicios de filtrado	7
,	2.3	Ejercicios de proyección	. 11
,	2.4	Ejercicios de actualización	14
,	2.5	Ejercicios de inserción	15
,	2.6	Ejercicios de pipeline de agregación	17
3.	CON	NCLUSIONES	18

1. INTRODUCCIÓN

El objetivo de este proyecto consiste en elegir un dataset y realizar una serie de ejercicios sobre este.

Para la realización de este ejercicio, se ha obtenido un dataset sobre las emisiones de CO2 por vehículo. Esta base de datos se ha obtenido a través de la página web "Kaggle" del autor Batuhan Şahan.

Se adjunta el enlace correspondiente:

https://www.kaggle.com/datasets/brsahan/vehicle-co2-emissions-dataset

Previamente a la elección de este dataset, se ha realizado un análisis preliminar del archivo en Excel para comprobar la coherencia de los datos y si puede ser tratada correctamente con el objetivo de obtener conclusiones claras para la realización de esta práctica.

En concreto, esta base de datos contiene información sobre las especificaciones de los vehículos, consumo de combustible y emisiones de CO2 entre otros datos, recogidos en más de 7.000 filas y 12 columnas:

- Make: Fabricante o marca del vehículo.
- Model: Modelo específico del vehículo.
- Vehicle Class: Clasificación del modelo basado en el tamaño y uso.
- Engine Size (L): Tamaño del motor expresado en litros.
- Cylinders: Número de cilindros del motor.
- Transmission: Transmisión del vehículo
- Fuel type: Tipo de combustible
- Fuel Consumption City (L/100 km): Consumo del vehículo en ciudad, expresado en litros a los 100 kilómetros.
- Fuel Consumption Hwy (L/100 km): Consumo del vehículo en carretera, expresado en litros a los 100 kilómetros.

- Fuel Consumption Comb (L/100 km): Consumo combinado (mixto), expresado en litros a los 100 kilómetros.
- Fuel Consumption Comb (mpg): Consumo combinado de combustible, expresado en millas por galón (mpg).
- CO2 Emissions (g/km): Emisiones de CO2, expresadas en g/km.

NOTA: Todas las capturas que se detallan a lo largo del documento son recortes en los que se muestra la consulta realizada y un ejemplo de la consulta obtenida, ya que por el número de resultados que se obtienen, es inviable especificar todos.

2. ANÁLISIS DE DATOS

Tal y como se ha comentado anteriormente, el análisis de la base de datos se realizará mediante el programa MongoDB, mediante el cual se pretende obtener unas conclusiones claras sobre los datos objeto de estudio.

2.1 Insertar base de datos

El primer paso que se ha realizado es descargar la base de datos en el tipo de archivo disponible (csv), una vez que se ha verificado que los datos tienen sentido, se convierte el archivo "csv" a "json" mediante un convertidor tipo, para después descargar dicho archivo y subirlo a MongoDB, para posteriormente trabajarlo mediante la propia consola de la aplicación.

Se utiliza esta sentencia para introducir los datos y posteriormente se comprueba la correcta importación de estos mediante:

```
> db.proyectofinal.find()

< {
    _id: ObjectId('6745d28758b3738330e9321c'),
    Make: 'ACURA',
    Model: 'ILX',
    'Vehicle Class': 'COMPACT',
    'Engine Size(L)': 2,
    Cylinders: 4,
    Transmission: 'AS5',
    'Fuel Type': 'Z',
    'Fuel Consumption City (L/100 km)': 9.9,
    'Fuel Consumption Hwy (L/100 km)': 6.7,
    'Fuel Consumption Comb (L/100 km)': 8.5,
    'Fuel Consumption Comb (mpg)': 33,
    'CO2 Emissions(g/km)': 196
}</pre>
```

2.1 Conocimiento sobre la base de datos

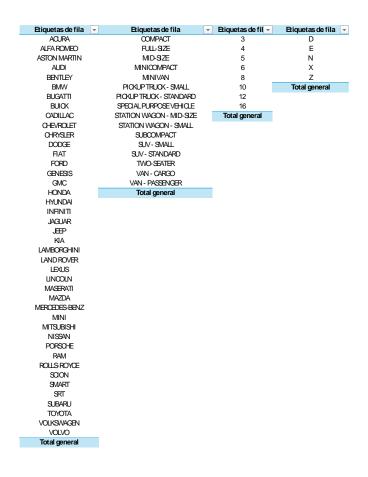
Una vez que se ha verificado que los datos se han cargado correctamente, se procede a realizar un análisis preliminar con el objetivo de conocer con mayor nivel de detalle mi base de datos.

Para ello, se realiza un recuento de los registros disponibles:

```
> db.proyectofinal.countDocuments()
< 7385</pre>
```

Este resultado expresa que existen 7.385 registros en cuanto al número de emisiones por marca y modelo de este dataset.

A continuación, se realiza una query con el objetivo de ver cuántas marcas, tipos de vehículos, tipos de cilindrada y combustible distintos existen. Se verifica que los resultados obtenidos son correctos con una pivot en Excel (se adjunta resultado).



Como se puede observar, la base de datos no recoge todas las marcas disponibles en el mercado, ya que, la query realizada nos arroja un resultado de 42 marcas de automóviles distintas, cuando según estimaciones realizadas por periódicos automovilísticos, en 2024 existen alrededor de 250 marcas activas en todo el mundo.

El resultado obtenido no significa que esté mal, pues este total comentado anteriormente puede variar según los criterios utilizados para contar las marcas, como por ejemplo, marca principal y submarcas, marcas que fabrican automóviles para una parte en concreto del mundo, fabricantes exclusivos, emergentes, etc.

Sin embargo, el número de tipos de automóviles, cilindrada y tipo de combustible sí que refleja la realidad del mercado, teniendo siempre en cuenta que el dato puede diferir de otros encontrados por internet, por los motivos ya comentados.

2.2 Ejercicios de filtrado

2.2.1 Existencia de valores nulos en todo el dataset

Como se puede observar, esta query no arroja ningún resultado, sin embargo, esta información es buena, ya que verifica la no existencia de valores nulos.

2.2.2 Número de automóviles Bentley

```
> db.proyectofinal.find({"Make": "BENTLEY"})

< {
    _id: ObjectId('6745d28758b3738330e93261'),
    Make: 'BENTLEY',
    Model: 'CONTINENTAL GT',
    'Vehicle Class': 'SUBCOMPACT',
    'Engine Size(L)': 4,
    Cylinders: 8,
    Transmission: 'AS8',
    'Fuel Type': 'Z',
    'Fuel Consumption City (L/100 km)': 16.3,
    'Fuel Consumption Hwy (L/100 km)': 9.8,
    'Fuel Consumption Comb (L/100 km)': 13.4,
    'Fuel Consumption Comb (mpg)': 21,
    'C02 Emissions(g/km)': 314
}</pre>
```

Para conocer exactamente el número de elementos que cumplen esta condición, añadiremos a la query realizada anteriormente el ".count".

```
> db.proyectofinal.find({"Make": "BENTLEY"}).count()
< 46</pre>
```

Si dividimos este resultado por el número total de registros, obtendremos que de toda la base de datos, únicamente el 0,62% de los automóviles que existen en nuestra base de datos son Bentley.

A priori, el motivo de esto radicaría en el precio de estos vehículos, debido a que son vehículos de lujo con un precio muy elevado a la media del mercado.

<u>2.2.3 Automóviles Ford, Suv – Standard y con un motor de 5.4 litros.</u>

Realizamos el mismo proceso que el comentado anteriormente para obtener el número concreto de resultados que cumplan estos filtros:

A diferencia del resultado anterior, se observa como estos filtros tienen mayor coincidencia, con un 8,50% sobre el total del vehículo. Uno de los motivos que puede conllevar a esto es el precio de este vehículo, ya que, pese a ser un automóvil de alta gama, tiene un menor precio y es más frecuente en ciertas partes del mundo como Estados Unidos.

2.2.4 Automóviles de gama alta/lujo con emisiones de CO2 razonables.

Como se puede observar, al ejecutar esta query no se obtiene ningún resultado. Esta query, pese a no presentar ningún resultado es muy valiosa para nuestro análisis, ya que demuestra, que en nuestra base de datos, no existe ninguna marca de alta gama/lujo en la que sus vehículos tengan emisiones de CO2 razonablemente bajas.

2.2.5 Automóviles con emisiones de CO2 menores a 140gm/km.

Esta query se ha realizado para ver el porcentaje de automóviles que generan unas emisiones de CO2 menores a 140 gr/km, y se ha obtenido el siguiente resultado:

```
db.proyectofinal.find(
    {"CO2 Emissions(g/km)": {$lt: 140}}
)

<{
    _id: ObjectId('6745d28758b3738330e9321e'),
    Make: 'ACURA',
    Model: 'ILX HYBRID',
    'Vehicle Class': 'COMPACT',
    'Engine Size(L)': 1.5,
    Cylinders: 4,
    Transmission: 'AV7',
    'Fuel Type': 'Z',
    'Fuel Consumption City (L/100 km)': 6,
    'Fuel Consumption Hwy (L/100 km)': 5.8,
    'Fuel Consumption Comb (L/100 km)': 5.9,
    'Fuel Consumption Comb (mpg)': 48,
    'CO2 Emissions(g/km)': 136
}</pre>
```

```
> db.proyectofinal.find(
          {"CO2 Emissions(g/km)": {$lt: 140}}
).count()
< 121</pre>
```

Con estos resultados se puede concluir que el parque automovilístico de nuestra base de datos es poco respetuoso con el medio ambiente, ya que, únicamente un 1,63% de los vehículos tiene unas emisiones de CO2 razonables. Por último, si vemos el número de coches que no generan emisiones de CO2, constararemos que es de 0.

```
> db.proyectofinal.find(
          {"CO2 Emissions(g/km)": 0}
).count()
< 0</pre>
```

2.3 Ejercicios de proyección

2.3.1 Automóviles con 3 cilindros y tipo de combustible X

```
> db.proyectofinal.find({"Cylinders": 3}, {"Fuel Type": "X"})
< {
    __id: ObjectId('6745d28758b3738330e93558'),
    'Fuel Type': 'X'
}
{
    __id: ObjectId('6745d28758b3738330e93559'),
    'Fuel Type': 'X'
}
{
    __id: ObjectId('6745d28758b3738330e935cb'),
    'Fuel Type': 'X'
}
{
    __id: ObjectId('6745d28758b3738330e935cb'),
    'Fuel Type': 'X'
}
</pre>
```

Tal y como se ha realizado anteriormente, vemos cuántos vehículos cumplen exactamente con la función de proyección aplicada:

```
> db.proyectofinal.find({"Cylinders": 3}, {"Fuel Type": "X"}).count()
< 95</pre>
```

Determinando que únicamente el 1,28% de los automóviles del dataset cumplen con estas condiciones.

2.3.2 Ilustración únicamente marcas automóviles ordenadas alfabéticamente

Comprobamos que el dato obtenido es correcto comparando con el apartado 2.1 de esta práctica.

Validamos este dato, pues coincide con el apartado mencionado anteriormente, tras haber hecho las comprobaciones oportunas.

2.3.3 Automóviles y la media de sus emisiones ordenadas de mayor a menor

Como se puede observar, aquellas marcas de automóvil que mayores emisiones de CO2 emiten, no son aquellas marcas de gama baja, media o alta, si no que son aquellas marcas de lujo.

2.3.4 Obtener los resultados del modelo Acura Rdx Awd y mostrar sólo las emisiones

2.4 Ejercicios de actualización

Para los ejercicios de filtrado se crea un nuevo proyecto con el mismo dataset, dado que los datos de los ejercicios a realizar pueden diferir de la realidad.

2.4.1 Modificar el número de cilindros de los motores de la marca AUDI a 1

```
> db.proyectofinal2.updateMany(
          {Make: "AUDI"},
          {$set: {"Cylinders": 1}}
)
< {
          acknowledged: true,
          insertedId: null,
          matchedCount: 286,
          modifiedCount: 286,
          upsertedCount: 0
}</pre>
```

2.4.2 Modificar el número de cilindros de los motores de la marca AUDI a 1 y sus emisiones de CO2 a 0

2.4.3 Modificar todas las emisiones de CO2 de las marcas de los vehículos a 0.

2.5 Ejercicios de inserción

2.5.1 Inserción de una nueva marca de automóviles

Se realiza una inserción de una marca de automóviles española que fabrica superdeportivos:

```
> db.proyectofinal.insertOne([{
 Make: "GTA",
 Model: "Spano",
 'Vehicle Class': "SUPER-DEPORTIVO",
 'Engine Size(L)': 8,
 Cylinders: 12,
 Transmission: "Automático",
 'Fuel Type': 'Z',
 'Fuel Consumption City (L/100 km)': 30,
 'Fuel Consumption Hwy (L/100 km)': 20,
 'Fuel Consumption Comb (L/100 km)': 25,
 'Fuel Consumption Comb (mpg)': 200,
 'CO2 Emissions(g/km)': 1500
 }])
< {
   insertedId: ObjectId('674963fb0bae12ff6f1280f0')
```

Se realiza una comprobación de que se ha insertado correctamente:

```
> db.proyectofinal.find({"Make": "GTA"})

< {
    _id: ObjectId('674963e50bae12ff6f1280ef'),
    Make: 'GTA',
    Model: 'Spano',
    'Vehicle Class': 'SUPER-DEPORTIVO',
    'Engine Size(L)': 8,
    Cylinders: 12,
    Transmission: 'Automático',
    'Fuel Type': 'Z',
    'Fuel Consumption City (L/100 km)': 30,
    'Fuel Consumption Hwy (L/100 km)': 20,
    'Fuel Consumption Comb (L/100 km)': 25,
    'Fuel Consumption Comb (mpg)': 200,
    'CO2 Emissions(g/km)': 1500
}</pre>
```

2.5.2 Inserción una fila a cada registro en relación con las emisiones de CO2

Habiendo analizado con anterioridad las emisiones de CO2 de los automóviles objeto de estudio, y habiendo constatado que no existe ningún modelo que no emita emisiones de CO2, es decir, no existe ningún coche eléctrico, si extrapolamos esto a nuestra ciudad, en un contexto de aumento restrictivo sobre ciertas zonas de esta, se decide crear un una fila por cada registro, especificando que con ese nivel de emisiones no podría entrar a Madrid Central.

```
> db.proyectofinal.updateMany(
     {},
     {$set: {"¿Puede entrar a Madrid Central?": false}}
)

< {
    acknowledged: true,
    insertedId: null,
    matchedCount: 7387,
    modifiedCount: 7387,
    upsertedCount: 0
}</pre>
```

Comprobamos la correcta inserción a través de la búsqueda de una marca:

```
> db.proyectofinal.find({Make: "AUDI"})
< {
    _id: ObjectId('6745d28758b3738330e93230'),
    Make: 'AUDI',
    Model: 'A4',
    'Vehicle Class': 'COMPACT',
    'Engine Size(L)': 2,
    Cylinders: 4,
    Transmission: 'AV8',
    'Fuel Type': 'Z',
    'Fuel Consumption City (L/100 km)': 9.9,
    'Fuel Consumption Hwy (L/100 km)': 7.4,
    'Fuel Consumption Comb (L/100 km)': 8.8,
    'Fuel Consumption Comb (mpg)': 32,
    'CO2 Emissions(g/km)': 202,
    '¿Puede entrar a Madrid Central?': false
}</pre>
```

2.6 Ejercicios de pipeline de agregación

2.6.1 Cálculo promedio de todas las emisiones de CO2

Mediante un pipeline de agregación se calcula el número de emisiones de CO2 del conjunto de vehículos y se calcula en concreto para las marcas Toyota y Bentley.

3. CONCLUSIONES

Podemos obtener varias conclusiones claras de este análisis, teniendo en cuenta que es un dataframe extraído de una fuente no oficial como es Kabble.

- No se recogen todas las marcas y modelos existentes en el mercado mundial, únicamente se tienen en cuenta un cierto número de automóviles a iniciativa propia del creador del dataframe. Como se ha detallado anteriormente, es normal que no se recojan todas las marcas, puesto que hay empresas que únicamente comercializan sus vehículos en cierta parte del mundo y el creador del df desconozca que existen más.
 - El dataframe está bien organizado puesto que los valores objeto de estudio reflejan la realidad y no se ha detectado ningún valor nulo
- Existe un mayor número de automóviles en aquellas marcas que tienen una categoría baja, media, en comparación con las marcas de alta gama o de lujo. Uno de los motivos por lo que sucede esto es debido a que no todas las personas pueden acceder a determinados vehículos con un precio de compra muy superior a la media del mercado.
- Con los resultados obtenidos, la gran mayoría de marcas no están concienciadas con el medio ambiente, pues no tienen ningún modelo con emisiones de CO2 razonables.
 - Existe una gran desviación respecto de la media de aquellas marcas de alta gama en cuanto a las emisiones de CO2.

Como conclusión final, podemos afirmar que aquellas marcas que inviertan en combustibles no provenientes de energías fósiles como electricidad e hidrógeno tendrán una posición aventajada respecto de sus competidores y podrán disfrutar de un cierto monopolio en el sector, siempre y cuando las variables exógenas que rodean a las empresas y a este sector en concreto sean favorables.