Proyecto ETL

Santiago Gomez Castro

Juan Carlos Quintero

MIguel Angel Ruales

Docente: Javier Alejandro Vergara Zorrilla

Universidad Autónoma de Occidente

Facultad de Ingeniería

Santiago de Cali

2024

Metodología

Proyecto - Primera entrega:

Item	Repositorio de github	Readme	Gitignore	Migración de los datos a la bd (subset > 10k)	EDA	Análisis	Gráficos	Doc
Puntaje	0.3	0.4	0.3	0.5	1	0.5	1	1

Contexto

El siguiente trabajo busca darle solución a la actividad ETL Proyect, donde se nos pide un dataset para realizar la aplicación de EDA para la manipulación y transformación de los datos, almacenamiento en base de datos y visualización de los resultados, el dataset elegido fue un csv extraído de kaggle con la información de venta de carros de segunda y nuevos en Estados Unidos.

Introducción

Esta investigación se enfoca en el mercado de autos usados en Estados Unidos y analiza cómo se mueve el mercado automotor en un país caracterizado por su alta demanda de vehículos. Como una de las mayores economías consumidoras del mundo, Estados Unidos representa un mercado crucial para las marcas automotrices, que buscan capturar una porción significativa de esta demanda. El país se ha desarrollado en torno al automóvil, con ciudades y carreteras diseñadas para facilitar la movilidad, lo que ha convertido al automóvil en una pieza central de la vida estadounidense. Para muchos, el vehículo no solo es un medio de transporte esencial, sino también un símbolo de estatus y éxito, lo que explica la persistente y robusta demanda de autos nuevos y usados. Esta dinámica de compra de vehículos nuevos alimenta el mercado de reventa, que a su vez revela patrones de consumo específicos, como la preferencia por ciertas marcas y modelos. El análisis de este mercado secundario ofrece una ventana a las prioridades y valores de los consumidores estadounidenses cuando se trata de adquirir un automóvil.

Herramientas

Python: Se usó de un script de python para la conexión y subida del csv a la base de datos.

Jupyter: Usando Jupyter se creó el EDA donde se aplica la limpieza y transformación de los datos para ser consumidos por el Dashboard, creando visualizaciones más precisas sin datos atípicos o fuera de contexto, también fueron subidos los datos nuevos a la base de datos

Postgresql: Base de datos relacional para el almacenamiento, gestión y administración, siendo base de datos relacional significa que maneja la información

con filas y columnas permitiendo una mayor facilidad para almacenar nuestros datos.

Power BI: Herramienta para visualizar, analizar y compartir información de los datos, se utilizó Power BI para la creación de gráficas.

Librerías de Python: El uso de librerías de python como sqlalchemy para la conexión con la base de datos y obtener la información para su debido proceso de limpieza y transformación, posteriormente con la misma librería se subirán los nuevos datos.

VScode: Editor de código popular y fácil uso, también cuenta con amplia gama de extensiones para realizar un trabajo más eficiente y sencillo.

Neon: Almacenamiento en la nube para almacenar la información y de fácil acceso.

Poetry: Es una herramienta de gestión de dependencias y entornos virtuales en Python que facilita la instalación de las librerías necesarias sin causar conflictos con otras instaladas en el sistema.

Gut y GitHub: gestores de versiones donde guardamos el proyecto de forma que esté seguro y se puede acceder desde cualquier medio.

EDA

El análisis exploratorio de datos realizado comenzó con la identificación de las variables del dataset para comprender mejor que tipo de transformaciones y visualizaciones podrían ser necesarias para alcanzar el mejor analisis posible del dataset. Algunas de las variables fueron:

- VIN: Un identificador global único para los carros, gran potencial como llave primaria.
- Used/New: Objeto redundante, ya que especificaba si el carro era usado/nuevo y la marca de este. La marca ya se encontraba almacenada en la columna Make.
- Price: Referente al precio del auto. Variable tipo objeto con caracteres como ‘$’ o ‘,’ y con valores ‘No price’ que obstaculizaban la conversión a int, para el posterior análisis.

Se realiza la respectiva limpieza y transformación de los datos, eliminando duplicados, transformando columnas object a string (y así reducir un poco la carga que significaban). Se obtiene como resultado un archivo de datos limpios en CSV, listo para su migración a la base de datos.

Migración de Datos

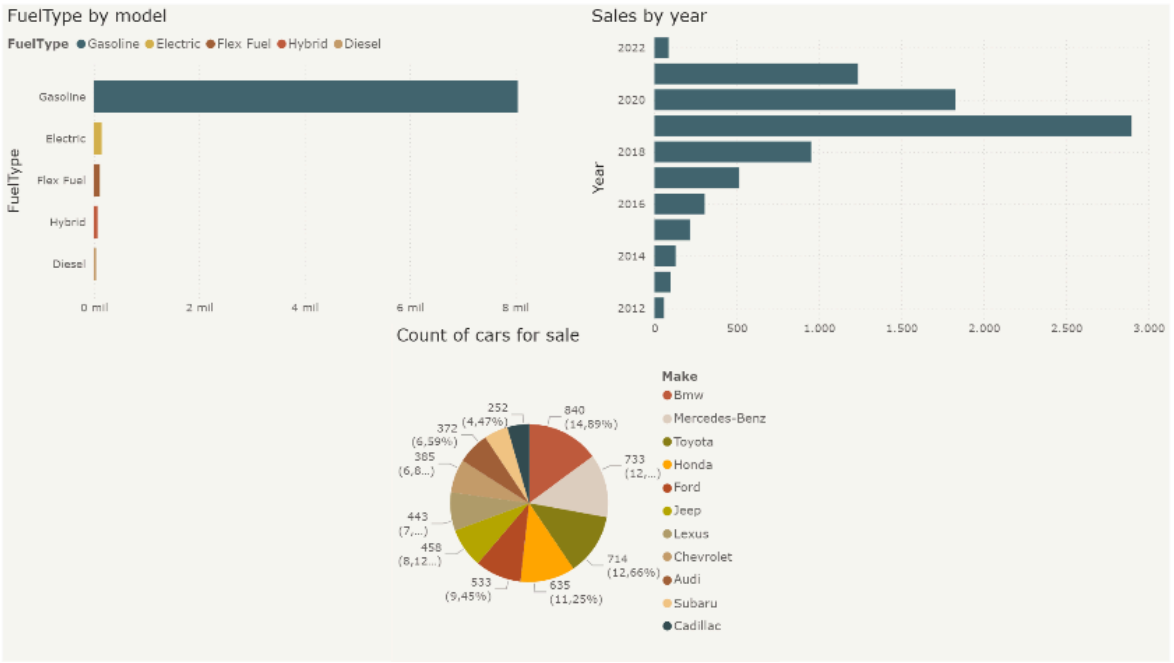
La migración de los datos del archivo CSV limpio a la base de datos en la nube se realizó mediante un script en Python que empleó pandas para la lectura de los datos, dotenv para las credenciales y SQLAlchemy para conectarse a la base de datos y realizar la inserción de los datos.

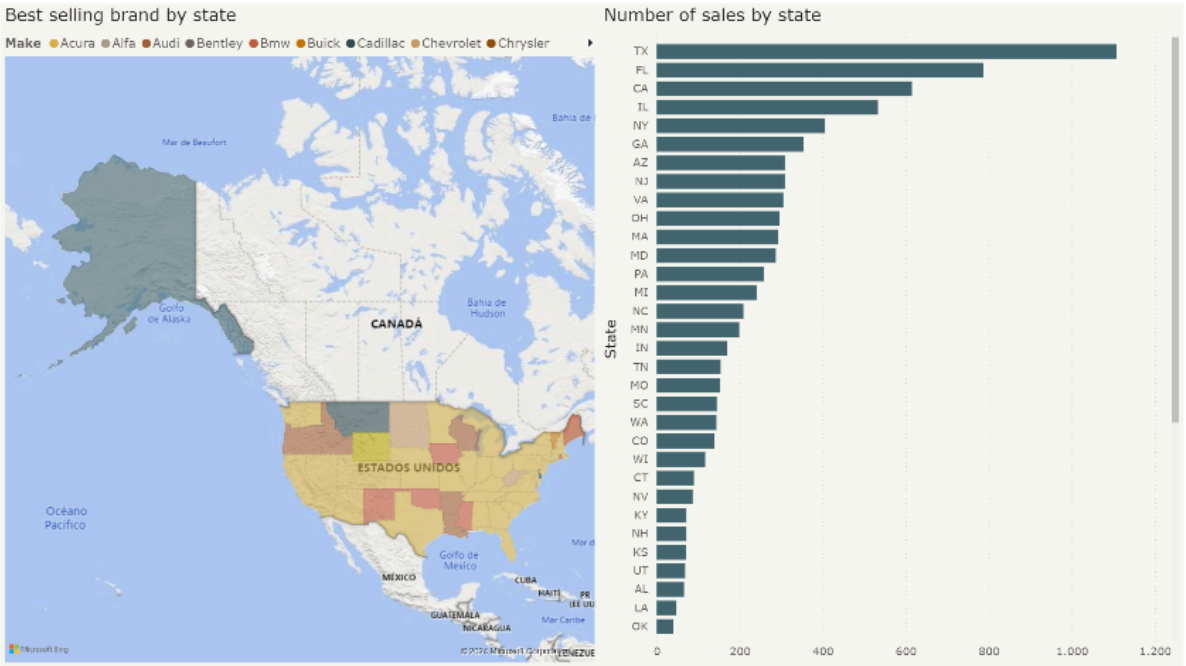
Análisis y Visualización de Datos

Con los datos ya en la base de datos PostgreSQL, se procedió a realizar el análisis y crear las visualizaciones utilizando Power BI. Las visualizaciones generadas incluyen:

- Tipo de gasolina que más se utiliza
- Ventas por año
- Conteo de autos vendidos por cada marca
- Marcas mejor vendidas por estado
- Número de ventas por estado

Aquí se puede ver el dashboard en detalle:





Conclusión del análisis

Las ventas de autos en Estados Unidos siguen siendo predominantemente dominadas por vehículos que utilizan gasolina, con una adopción aún limitada de alternativas como los vehículos eléctricos, híbridos o diésel. Esto sugiere que, a pesar de los avances y la disponibilidad de nuevas tecnologías, la gasolina continúa siendo la opción preferida por la mayoría de los consumidores.

El análisis de las ventas anuales revela fluctuaciones significativas, con picos notables en años específicos, como en 2020. Estos cambios podrían estar influenciados por factores económicos, eventos externos o el lanzamiento de nuevos modelos en el mercado, lo que impacta directamente en el comportamiento de compra de los consumidores.

En cuanto a las preferencias regionales, se observa una variabilidad considerable en las marcas más vendidas en diferentes estados. Esto refleja cómo las preferencias de los consumidores pueden variar según la región, posiblemente debido a factores como el clima, la economía local y la cultura automovilística de cada área.

Finalmente, la concentración de ventas es más alta en estados como Texas, Florida y California, indicando que estos mercados son particularmente activos. La gran población y la robusta economía de estos estados contribuyen a su posición como líderes en ventas automotrices a nivel nacional.

En resumen, el mercado automotriz de Estados Unidos muestra una clara preferencia por vehículos a gasolina, a pesar del creciente interés en opciones más

sostenibles. Las ventas anuales están sujetas a fluctuaciones que parecen responder a factores económicos y eventos específicos, mientras que las preferencias de marca varían significativamente entre las diferentes regiones del país. Además, los estados con economías grandes y poblaciones densas, como Texas, Florida y California, lideran en número de ventas, subrayando su importancia en el mercado nacional. En conjunto, estos hallazgos destacan la diversidad y complejidad del mercado automotriz estadounidense, así como las diferencias regionales que influyen en las tendencias de consumo.