**Car Resale Data Analysis Project**

**Aim:** The aim of this project is to conduct a comprehensive analysis of car resale trends and factors influencing resale prices in the Indian market. Leveraging a dataset sourced from Kaggle, encompassing resale car information from the year 2023, this project demonstrates proficiency in utilizing AWS S3, Databricks, and PySpark for data ingestion, cleaning, transformation, and visualization. By meticulously examining the dataset, the project seeks to unravel insights into the dynamics of the Indian car resale market, shedding light on key determinants shaping resale prices.

**Tools Used:** AWS s3 (As external storage system)

: Databricks (For data ingestion,cleaning,transformation and visualization)

: Pyspark (Programming Language)

**Dataset:**

Link: <https://www.kaggle.com/datasets/rahulmenon1758/car-resale-prices>

The dataset contains 17446 records and 15 columns in it.

**Procedure**: The following processes are utilized to draw conclusions in a chronological order.

1. Dataset Upload to AWS S3: The raw dataset is uploaded from the local machine to an AWS S3 bucket, ensuring secure and scalable storage.
2. Data Ingestion with Databricks: Utilizing Databricks, the raw dataset stored in AWS S3 is ingested for further processing and analysis.
3. Data Cleaning and Transformation: This crucial stage involves several key steps:

* Dropping unnecessary columns to streamline the dataset.
* Removing duplicate entries to ensure data integrity.
* Transforming various columns, including renaming, changing data formats, altering column values, modifying data types, and creating new columns as required.
* Handling null values through imputation or removal, ensuring data completeness.
* Exporting the cleaned dataset back to AWS S3 for storage and creating a table for facilitating advanced SQL queries.
* Generating a statistical summary of the cleaned dataset to gain insights into its distribution and characteristics.

1. Data Analysis and Visualization:

* Leveraging the cleaned dataset, conduct thorough data analysis to uncover trends and patterns.
* Utilize visualization techniques to present insights effectively, aiding in the interpretation of findings and decision-making processes.

**Key Questions to Address:**

1. Identifying the market share of various car manufacturers in the resale market, providing insights into the competitive landscape.
2. Unveiling the most popular car models from the leading car manufacturers, highlighting consumer preferences and market trends.
3. Analyzing how the resale market varies across different registration years, indicating trends in consumer behavior and preferences over time.
4. Investigating how ownership type influences car resale trends, shedding light on differences in resale patterns between individual owners, dealerships, and other entities.
5. Understanding consumer preferences regarding insurance types for used car purchases, providing insights into factors influencing decision-making in the secondary car market.
6. Examining the geographical distribution of car resale activities across cities in India, identifying regional preferences and market dynamics.

**Conclusion:**

By addressing these key questions, the project provides valuable insights into the dynamics of the Indian car resale market, facilitating informed decision-making for stakeholders in the automotive industry.