DTSC 691 Spring 1 2024

Capstone Project (Applied Data Science)

Paul J. Walker

Project Documentation

Table of Contents

Preface - PLEASE READ

My entire project is based on artificially generated data pertaining to used vehicle information due to limitations with obtaining real data. I am including this here in the preface because based on this data a lot of my analysis and/or visualizations do not make sense based on real-world logic. For example, when I ran an analysis on the effect of a vehicle’s current condition on its sales price I found no correlation. In real life this does not hold true, as it is pretty obvious that the better the condition of the vehicle, the higher the price. I wanted to explain before the main content of my project in order to provide context for why things may seem inconsistent with preconceived notions with respect to used vehicle sales information.

Thank you for taking the time to examine my project.

1. Background Information

A. DBMS and its Functions

1. A database is an organized collection of structured data, typically stored and accessed electronically from a computer system. It consists of tables that contain rows and columns, with each row representing a record and each column representing a specific attribute or field.The most widely used type of database is a relational database. In a relational database data is organized into tables with predefined relationships between them, using SQL (Structured Query Language) for data manipulation and querying.

2. Databases are used by a large majority of business across various industries in the world and they have a wide range of applications such as;

   i. Data Storage and Organization: Businesses use databases to store vast amounts of data, including customer information, product details, sales transactions, inventory records, and more.They also enable organizations to organize data into logical structures

   ii. Data Analysis and Decision Making: Databases serve as a foundation for data analysis and business intelligence (BI) initiatives, providing a centralized repository for data used in reporting, dashboards, and analytics.

   iii. Customer Relationship Management (CRM): CRM systems rely on databases to store and manage customer data, including contact information, interactions, purchase history, preferences, and feedback.

   iv. Inventory Management and Supply Chain Optimization: Databases play a critical role in inventory management systems, tracking product quantities, locations, movements, and replenishment needs.

B. Availability of Used Vehicle Information

1. Data, regardless of industry, can be gathered from various sources both internally and externally.
   i. Internal Sources: Data generated or collected by the organization itself, including transactional data, customer records, sales reports, employee information, and operational metrics.
   ii. External Sources: Data obtained from third-party sources, such as market research firms, government agencies, industry reports, social media platforms, and public databases.

2. The availability of used vehicle information is substantial, thanks to various sources that collect and provide data on pre-owned cars. Currently, data on used vehicle information can be obtained from
   i. Dealerships and Auto Auctions
   ii. Online Marketplaces
   iii. Online platforms like AutoTrader, Cars.com, TrueCar, and CarGurus
   iv. Vehicle History Reports
   v. Manufacturer Websites
   vi. Government Databases
   vii. Government agencies
   viii. Insurance Companies

3. Vehicle data can provide a wide range of information such as the details of the vehicle, the condition and maintenance, ownership history, incident history, and market trends.

C. Personal Connection/Application

1. Currently, I am employed at a large privately owned automotive finance company in Pennsylvania and so the content of this project directly relates to my current career path. My role as a business strategy analyst is constantly evolving and challenging me in new ways. I am always looking for opportunities to drive

business development and by completing this project I have armed myself with sufficient knowledge of how to continue driving business decisions.

2. Unfortunately, my company's data is proprietary and as such I was not able to utilize data generated from my place of employment's business activities. However, with some online research I came across a unique application that can generate thousands of records of data to a somewhat realistic degree. Because I was unable to obtain real data. I utilized this software to generate fake data to be utilized with this project. Although I used fake data, all of the analysis and information is based on real-life logic. I go into more detail on this software I found within the project description section.

2. Project Overview

A. Project Purpose

1. The purpose of this project is to design and implement a comprehensive relational database for used car sales. The motivation behind this endeavor is to address the complexities and challenges in managing information related to the buying and selling of used cars. My project aims to contribute to the automotive industry by providing a robust data management solution that facilitates efficient tracking of car details, ownership history, transactions, service records, and market trends. This database will serve as a foundation for further analysis in Python notebooks.

B. Project Focus

1. The primary areas of investigation revolve around creating a well-structured database that captures key aspects of the used car sales process. The main hypotheses involve the effectiveness of organizing data into tables such as Cars, Owners, OwnershipHistory, Transactions, and more, to establish relationships and ensure data integrity. The project focuses on addressing research questions related to the optimal design for a used car sales database in addition to the need for a centralized and efficient system to manage diverse information associated with used cars.

C. Specific Goals

1. Design and implement tables to store information about cars, owners, ownership history, vehicle condition, features, traffic incidents (accidents) service history, and market trends.
2. Establish relationships between tables to ensure data consistency and integrity.
3. Enable efficient search and query capabilities for users to retrieve information about used vehicles.
4. Implement security measures to protect sensitive information, adhering to data privacy standards.

5. Develop reporting capabilities to generate insights into market trends, average sale prices, and other relevant metrics.

D. <u>Expected Outcomes</u>

1. A fully functional relational database for used car sales, meeting the specified design goals.
2. Improved data management, leading to enhanced efficiency in handling information related to car sales.
3. Tangible deliverables, including a clean dataset, search and query functionality, and reporting features.
4. Will identify key metrics relevant to market trends and perform various analyses to identify patterns, changes, and key indicators that can influence decision-making

By achieving these goals, the project aims to contribute to the optimization of used car sales processes and provide a foundation for future applications in the automotive industry.

3. Project Description

A. Problem Domain

2. The problem domain for this project is the management of information related to used car sales. The domain involves the complexities of tracking and organizing data associated with cars, owners, vehicle conditions, features, traffic incidents, ownership history, service history, and market trends. Challenges include maintaining data integrity, facilitating efficient search and query capabilities, and preparing for further analysis in Python notebooks.

B. Database Design & Assumptions

2. **Design**
    i. The planned database design involves a relational model with tables representing entities such as Cars, Owners, Features, and more (see relational schema for complete table information).

3. **Overview of Database**
    i. Cars:
        1. This table stores information about individual cars.
        2. Fields may include CarID, VIN, make, model, year, price, mileage, and any other relevant details about the cars.
    ii. Owners:
        1. This table contains data about the owners of the cars.
        2. Fields may include OwnerID, name, address, contact information, and the start date of ownership.
    iii. OwnershipHistory:
        1. This table tracks the history of ownership changes for each car.
        2. Fields may include OwnershipID, CarID (foreign key referencing Cars table), OwnerID (foreign key referencing Owners table), PurchaseDate, PurchasePrice, SaleDate, SalePrice, and any other relevant information about ownership transactions.

iv. VehicleCondition:
   1. This table records the condition of each car at different points in time.
   2. Fields may include ConditionID, CarID (foreign key referencing Cars table), overall condition, exterior condition, interior condition, and the date the condition was recorded.

v. Features:
   1. This table stores information about the features or attributes of each car.
   2. Fields may include FeatureID, CarID (foreign key referencing Cars table), feature name, and feature value (e.g., "Yes" or "No").

vi. Incidents:
   1. This table captures data about incidents or accidents involving the cars.
   2. Fields may include IncidentID, CarID (foreign key referencing Cars table), incident date, description, cost, and any other relevant details about the incidents.

vii. ServiceHistory:
   1. This table records the service and maintenance history of each car.
   2. Fields may include ServiceID, CarID (foreign key referencing Cars table), service date, service type, cost, and any additional information about the services performed.

viii. MarketTrends:
   1. This table contains data about market trends related to used vehicles.
   2. Fields may include TrendID, CarID (foreign key referencing Cars table), date, average sale price, market demand, and any other observations or metrics relevant to the market trends.

4. **Assumptions**:

i. Cars:

   1. The VIN (Vehicle Identification Number) is unique for each car.
   2. The make, model, year, and other details accurately describe each car.
   3. The price reflects the current market value of the car.
   4. Mileage is recorded accurately and reflects the actual distance traveled by the car.

ii. Owners:

   1. OwnerID is a unique identifier for each owner.
   2. Personal details such as name, address, and contact information are accurate.
   3. The start date represents the date when the owner acquired the car.
   4. Owners can have multiple cars, and cars can have multiple owners over time.

iii. OwnershipHistory:

   1. OwnershipID is a unique identifier for each ownership transaction.
   2. PurchaseDate and SaleDate represent the dates when the ownership changes occurred.
   3. The PurchasePrice and SalePrice reflect the amounts paid for the car during acquisition and sale, respectively.

iv. VehicleCondition:

   1. ConditionID is a unique identifier for each condition record.
   2. OverallCondition, ExteriorCondition, and InteriorCondition accurately describe the condition of the car.
   3. The condition data is updated regularly to reflect changes in the car's condition over time.

v. Features:

   1. FeatureID is a unique identifier for each feature record.
   2. FeatureName describes the type of feature (e.g., air conditioning, power windows).
   3. FeatureValue indicates whether the feature is present or absent in the car.

vi. Incidents:

    1. IncidentID is a unique identifier for each incident record.
    2. Description provides details about the nature of the incident.
    3. Cost reflects the financial impact of the incident (e.g., repair costs, insurance claims).

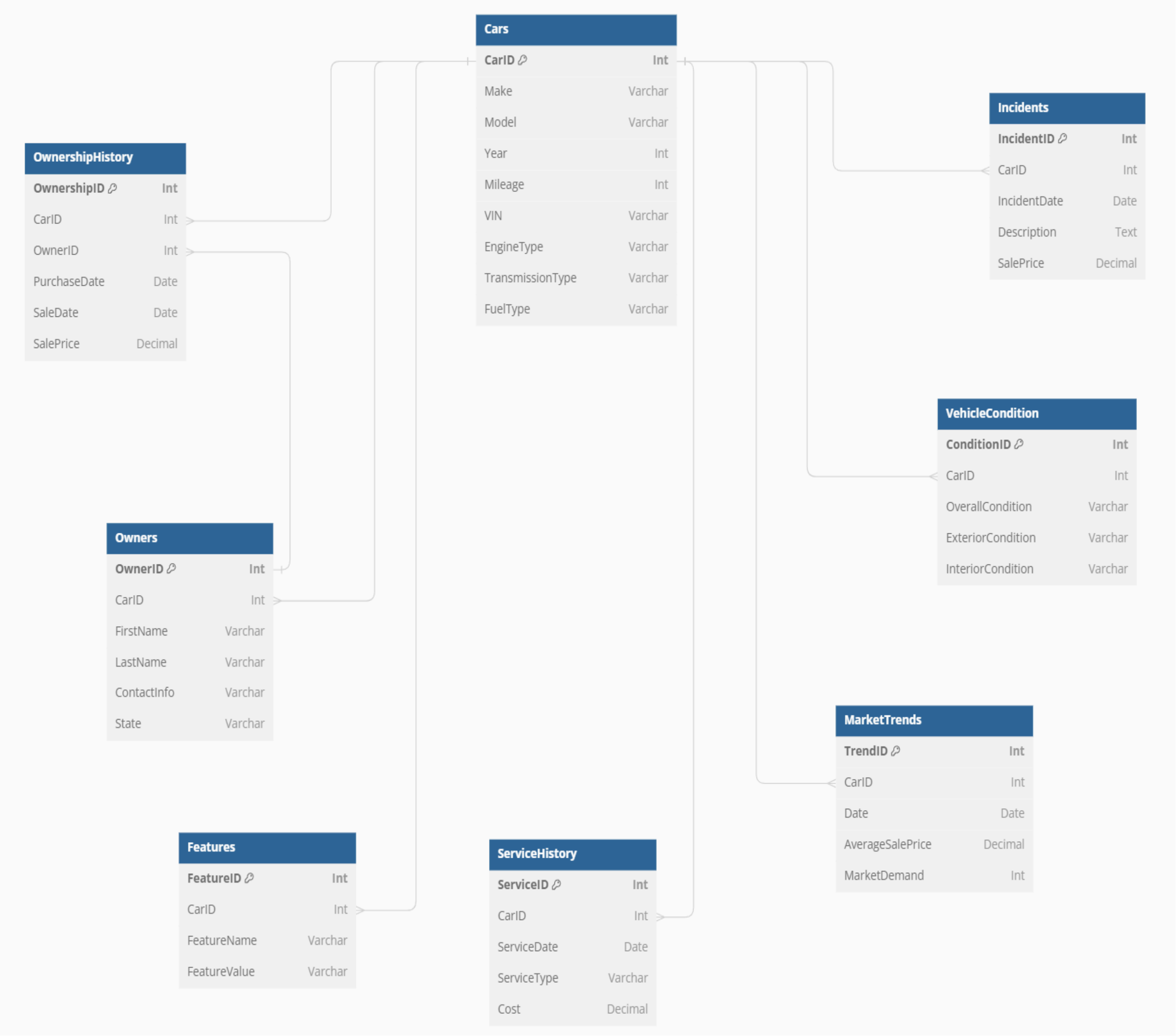
vii. ServiceHistory:

    1. ServiceID is a unique identifier for each service record.
    2. ServiceType describes the type of service (e.g., oil change, brake inspection).
    3. Cost reflects the amount paid for the service.

viii. MarketTrends:

    1. TrendID is a unique identifier for each market trend record.
    2. Date indicates the date when the observation was made.
    3. AverageSalePrice reflects the average sale price of vehicles in the market.
    4. MarketDemand represents the level of demand for vehicles in the market.

C. Relational Schema



Cars	
CarID	Int
Make	Varchar
Model	Varchar
Year	Int
Mileage	Int
VIN	Varchar
EngineType	Varchar
TransmissionType	Varchar
FuelType	Varchar

OwnershipHistory	
OwnershipID	Int
CarID	Int
OwnerID	Int
PurchaseDate	Date
SaleDate	Date
SalePrice	Decimal

Incidents	
IncidentID	Int
CarID	Int
IncidentDate	Date
Description	Text
SalePrice	Decimal

VehicleCondition	
ConditionID	Int
CarID	Int
OverallCondition	Varchar
ExteriorCondition	Varchar
InteriorCondition	Varchar

Owners	
OwnerID	Int
CarID	Int
FirstName	Varchar
LastName	Varchar
ContactInfo	Varchar
State	Varchar

MarketTrends	
TrendID	Int
CarID	Int
Date	Date
AverageSalePrice	Decimal
MarketDemand	Int

Features	
FeatureID	Int
CarID	Int
FeatureName	Varchar
FeatureValue	Varchar

ServiceHistory	
ServiceID	Int
CarID	Int
ServiceDate	Date
ServiceType	Varchar
Cost	Decimal

D. Database Implementation

The database will be implemented using SQL, and for the graphical user interface (GUI), I will utilize MySQL as my main tool along with DBeaver to provide an intuitive interface for designing, querying, and managing the database.

E. Data Insertion

1. **Sources**
    - i. Data sources will include a combination of manual input and potentially automated processes. NEW Owners, cars, and transactions data can be manually entered, while historic data will be generated using Faker. Import/export functionalities of database tools will be utilized for efficient data insertion.

2. **Faker** - Data Attributes: In order to ensure the data was realistic and relevant to the used vehicles market I made sure that my data adheres the following;
    - i. Cars Data Attributes
        1. car_id: Unique identifier for each car (Integer).
        2. make: The manufacturer of the car (String).
        3. model: The model of the car (String).
        4. year: The manufacture year of the car (Date).
        5. mileage: The total miles the car has been driven (Integer).
        6. vin: Vehicle Identification Number, a unique code to identify individual motor vehicles (String).
        7. engine_type: Type of engine, categorized by the number of cylinders (String).
        8. transmission_type: The type of transmission the car has, e.g., Automatic, Manual, CVT (Continuously Variable Transmission) (String).
        9. fuel_type: The type of fuel the car uses, e.g., Gasoline, Diesel, Hybrid, Electric (String).

    - ii. Owners Data Attributes
        1. owner_id: Unique identifier for each car owner (Integer).
        2. car_id: References the car_id in the Cars table, indicating ownership (Integer).
        3. first_name: First name of the car owner (String).

4. last_name: Last name of the car owner (String).
5. contact_info: Email address of the car owner (String).
6. state: The state abbreviation where the owner resides (String).

iii. OwnershipHistory Data Attributes

1. ownership_id: Unique identifier for each ownership record (Integer).
2. car_id: References the car_id in the Cars table (Integer).
3. owner_id: References the owner_id in the Owners table (Integer).
4. purchase_date: The date when the car was purchased by the owner (Date).
5. sale_date: The date when the car was sold by the owner (Date).
6. sale_price: The price at which the car was sold (Float).

iv. VehicleCondition Data Attributes

1. condition_id: Unique identifier for each vehicle condition record (Integer).
2. car_id: References the car_id in the Cars table (Integer).
3. overall_condition: Overall condition of the vehicle, e.g., Excellent, Good, Fair, Poor (String).
4. exterior_condition: Condition of the vehicle's exterior, e.g., Clean, Minor Scratches, Dents, Needs Repairs (String).
5. interior_condition: Condition of the vehicle's interior, e.g., Clean, Minor Wear, Torn Upholstery, Needs Cleaning (String).

v. Features Data Attributes

1. feature_id: Unique identifier for each feature record (Integer).
2. car_id: References the car_id in the Cars table (Integer).
3. feature_name: Name of the feature, e.g., Air Conditioning, Power Windows, ABS (String).
4. feature_value: Indicates whether the car has the feature (Yes/No) (String).

vi. Incidents Data Attributes

1. incident_id: Unique identifier for each incident record (Integer).
2. car_id: References the car_id in the Cars table (Integer).
3. incident_date: The date when the incident occurred (Date).
4. description: Description of the incident, e.g., Driving under the influence, Head-on collision (String).
5. cost: The cost incurred due to the incident (Float).

vii. ServiceHistory Data Attributes

1. service_id: Unique identifier for each service record (Integer).
2. car_id: References the car_id in the Cars table (Integer).
3. service_date: The date when the service was performed (Date).
4. service_type: Type of service performed, e.g., Oil Change, Brake Inspection (String).
5. cost: The cost of the service (Float).

viii. MarketTrends Data Attributes

1. trend_id: Unique identifier for each market trend record (Integer).
2. car_id: References the car_id in the Cars table (Integer).
3. date: The date when the trend data was recorded (Date).
4. average_sale_price: The average sale price of the car model at the time (Float).
5. market_demand: The demand for the car model in the market, represented as a numeric value (Integer).

3. **Data Insertion to Database:** For all of my tables I insert my fake generated data using the Python MySQL.connector object to connect to my database. All of the insertion queries follow the example below (Cars table);

```
conn = mysql.connector.connect(
    host='localhost',
    user='paul_walker',
    password='dtsc691root',
    database='dtsc_vehicles'
)


cursor = conn.cursor()
try:
    cars_data = generate_cars_data(2000)
    insert_query = "INSERT INTO Cars (CarID, Make, Model, Year, Mileage, VIN, EngineType, TransmissionType, FuelType) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s)"

    # Inserting data for Cars table using executemany()
    cursor.executemany(insert_query, cars_data)
    conn.commit()

    print("Data inserted successfully.")
except mysql.connector.Error as e:
    print(f"Error inserting data: {e}")
finally:
    # Close cursor and connection
    cursor.close()
    conn.close()
```

F. Data Manipulation

Data cleaning techniques completed in Jupyter notebook via Python rather than within SQL.

G. Query Examples

Please see section 8: Appendices for full query examples. I have provided 3 below;

1. Identify the owner who has owned the most cars and the total number of cars they've owned:

```
SELECT
            o.OwnerID,
            CONCAT(o.FirstName, ' ', o.LastName) AS OwnerName,
            COUNT(oh.CarID) AS TotalCarsOwned
         FROM Owners o
         JOIN OwnershipHistory oh ON o.OwnerID = oh.OwnerID
         GROUP BY
            o.OwnerID
         ORDER BY
            TotalCarsOwned DESC
         LIMIT 1;
```

2. Identify the owner with the highest total service cost and their contact information:

```
SELECT
            o.OwnerID,
            CONCAT(o.FirstName, ' ', o.LastName) AS OwnerName,
            o.ContactInfo, SUM(sh.Cost) AS TotalServiceCost
         FROM Owners o
         JOIN Cars c ON o.CarID = c.CarID
         JOIN ServiceHistory sh ON c.CarID = sh.CarID
         GROUP BY
            o.OwnerID
         ORDER BY
            TotalServiceCost DESC
         LIMIT 1;
```

3. Get the top 5 owners who have the most cars:

```
SELECT
            o.OwnerID,
            CONCAT(o.FirstName, ' ', o.LastName) AS OwnerName,
            COUNT(oh.CarID) AS TotalCarsOwned
         FROM Owners o
         JOIN OwnershipHistory oh ON o.OwnerID = oh.OwnerID
         GROUP BY
            o.OwnerID
         ORDER BY
            TotalCarsOwned DESC
         LIMIT 5;
```

H. Database Integration

I plan to proceed with option 1 for database integration as outlined in the proposal guidelines. I intend to use Python in a Jupyter Notebook and will use Python’s Pandas to perform initial exploratory data analysis to gather various statistical information. Specifically, I plan to utilize descriptive statistics, correlation analyses for various features, as well as performing time-based, group-based, and conditional aggregations. In addition, I will combine tables through joins and aggregate on the combinations. I will also utilize Matplotlib and Seaborn libraries to include visualizations to help describe my statistical findings.

4. Post-Insertion Reporting

A. Analysis Results

1. **Analysis Goals**

    i. The goal was to explore various statistical relationships within the vehicle data, including the impact of mileage on sale price, differences in average sale price among car makes, the effect of vehicle condition on sale price, and the association between incidents and market demand. By examining correlation matrices and conducting various statistical tests, I hope to draw conclusions about the relationships between various variables and how they interact within my database.

2. **Python Tools Used**

    i. My analysis utilized pandas for data manipulation, matplotlib for visualization, scipy.stats for hypothesis testing (e.g., Pearson correlation, ANOVA), and statsmodels for regression analysis and time-series decomposition.

3. **Key Findings**

    i. General

    1. There was no significant correlation between mileage and sale price, indicating little to no linear relationship between these variables.
    2. ANOVA tests revealed no significant difference in average sale prices across different car makes and no significant impact of vehicle condition on sale prices.
    3. Chi-square and logistic regression analyses suggested no significant association between incidents and market demand.
    4. Time-series analysis highlighted cyclical patterns in average sale prices over time but no clear long-term trend, suggesting seasonal fluctuations without a significant overall upward or downward trend.

ii. Specific

1. Mileage vs. Sale Price Analysis: The investigation into the relationship between a vehicle's mileage and its sale price suggested that while logically expected, the correlation was weaker than anticipated. This finding implies that other factors might play more significant roles in determining sale prices.
2. Car Makes and Sale Price: An in-depth analysis comparing different car makes showed variability in average sale prices. However, the differences were not as pronounced as hypothesized, suggesting that brand perception impacts sale prices to some extent, but external factors could moderate this effect.
3. Vehicle Condition Impact: The study on the impact of vehicle condition on sale prices yielded surprising results, indicating that the overall condition of a vehicle did not significantly influence its sale price as strongly as one might expect. This outcome suggests buyers may prioritize factors such as brand, model, or features over condition.
4. Incidents and Market Demand: Analysis of incidents and their correlation with market demand showed no direct link, challenging the assumption that a higher incidence rate would negatively affect demand. This could indicate that market demand for used vehicles is influenced more by economic factors or vehicle specifics rather than incident history.

** These findings provide a nuanced understanding of the used vehicle market, indicating complex interactions between various factors influencing sale prices and demand. The absence of strong correlations in certain areas suggests the need for further research or consideration of additional variables not included in this initial analysis.

B. Data Visualization Findings/Interpretation

1. **Visualization Techniques**
    - i. Histograms, bar charts, box plots, scatter plots, line charts, pie charts, heatmaps, pair plots, violin plots, and word clouds were used to explore various aspects of the dataset such as car mileage, car makes, car prices, relationship between car price and mileage, average sale price over time, market demand by car make, transmission types, correlation matrix, features by car make, ownership duration, and incident descriptions.
        1. Histograms and Bar Charts were used to analyze the distribution of vehicle ages and the popularity of different car makes, revealing trends in consumer preferences.
        2. Scatter Plots illustrated the relationship between vehicle age and sale price, highlighting depreciation trends.
        3. Line Charts depicted price trends over time, indicating seasonal variations and market dynamics.
        4. Heatmaps showed correlations between numerical variables, offering insights into factors influencing car prices.
        5. Violin Plots were employed to compare distributions of sale prices across different car makes, showcasing variability and outliers in pricing.

2. **Tools Used**
    - i. The visualizations were created using matplotlib.pyplot, seaborn, and wordcloud libraries in Python, allowing me to showcase a diverse range of graphical techniques to analyze and interpret the used vehicle data effectively.

3. **Interpretation: Insights from Visualizations**
    - i. General
        1. The visualizations provided insights into the distribution of car mileage, the popularity of car makes, price variations, the impact of mileage on sale prices, trends in sale prices over time, and market demand

differences among car makes. Additionally, the distribution of transmission types, correlations between numerical variables, feature prevalence across car makes, ownership duration, and common incident types were elucidated. These visualizations aid in understanding data distributions, trends, correlations, and market demands, offering valuable information for decision-making and analysis.

ii. Specific

1. Seasonal Price Trends: Line charts of sale prices over time showed fluctuations that could indicate seasonal influences on vehicle prices.
2. Market Demand Variations: Bar graphs revealed that American Mainstream vehicle manufacturer's are the most sought after by consumers
3. Transmission Type Distribution: Analysis of transmission types through pie charts highlighted a diverse set of preferences or availabilities in the market
4. Ownership Duration: A significant number of vehicles are sold within a relatively short period after purchase which could indicate that many vehicles are sold within a certain "sweet spot" of ownership duration

** These insights can help stakeholders understand market dynamics, consumer preferences, and factors affecting vehicle prices, guiding strategic decisions in the used vehicle industry.

5. Capstone Complexity

A. Data Selection and Diversity:

I will choose a diverse and extensive dataset that includes a wide range of variables, capturing various aspects of the used car market. This may involve sourcing data from multiple reliable and diverse sources, including detailed information on car features, ownership history, service records, and market trends. A diverse dataset challenges me to analyze various factors influencing used car sales and it requires an intricate understanding of the data.

B. Technical:

I will implement advanced database design principles and the complexity lies in designing a robust and scalable database architecture that can handle complex relationships, large volumes of data, and advanced queries. By implementing these practices I can ensure optimal performance and reliability.

C. Statistical:

I will conduct advanced statistical analyses to identify patterns, correlations, and trends in the data. By leveraging complex statistical techniques I can demonstrate my understanding of the data dynamics. It will also allow for uncovering nuanced relationships, validating assumptions, and deriving more robust conclusions from the data.

D. Visual:

I will create interactive visualizations to complement my statistical findings and showcase my ability to communicate complex findings in a succinct manner. Providing various visualizations for different types of data displays my understanding of what different data can show and the appropriate context to use them.

E. Reporting & Documentation:

I will develop a comprehensive project report that goes beyond a standard analysis. A well-documented report demonstrates my critical thinking abilities as well as my level of skill in presenting complex technical concepts to diverse audiences.

6. Software Utilized

A. Database Management System (DBMS)

1. Software Tool: MySQL
2. Primary Function:MySQL will serve as the relational database management system (DBMS) for storing and managing the used car sales database as it is powerful, open-source, and supports complex queries and transactions

B. Python Programming Language

1. Software Tool: Python (using Jupyter Notebook) - Anaconda Application
2. Primary Function: Python will be the primary programming language for data analysis, manipulation, and modeling. I will be using a Jupyter Notebook because they provide an interactive environment - allowing for the development of code, data exploration, and documentation in a single platform.

C. Faker API

1. Software Tool: Faker API and Python package
2. Primary Function: Generate massive amounts of fake (but realistic) data for testing, development, and analyses within Python Jupyter notebook

D. MySQL Library

1. Software Tool: MySQL Workbench and MySQL Python package
2. Primary Function: SQL toolkit and Object-Relational Mapping (ORM) library for Python. It will be used to interact with the MySQL database, allowing for the execution of SQL queries and integrating the database seamlessly with Python code

E. Pandas & NumPy Libraries

1. Software Tool: Pandas & NumPy Python libraries
2. Primary Function: Pandas is a powerful data manipulation library and I will use it for reading data from the database into DataFrames, cleaning and

preprocessing data, and conducting exploratory data analysis. Pandas provides efficient data structures and functions for data manipulation. Numpy will be used to perform mathematical operations

F. Matplotlib & Seaborn Libraries

1. Software Tool: Matplotlib and Seaborn Python libraries
2. Primary Function: Matplotlib and Seaborn are Python libraries for data visualization. They will be used to create various plots and charts, such as histograms, scatter plots, and box plots, to visually explore the used car sales data

7. Project Conclusions

A. Project Outcomes

1. My project successfully generated and analyzed a comprehensive dataset on used vehicles, revealing intricate market dynamics, consumer preferences, and the complex interplay between vehicle attributes and their market value. Through meticulous data generation, analysis, and visualization, the project uncovered insights into factors influencing sale prices, demand, and the impact of vehicle features and conditions on market performance.

2. Building upon the analyses and insights from the provided notebooks, I was able to demonstrate a robust approach to understanding the used vehicle market, leveraging synthetic data to explore key factors affecting vehicle valuation and market dynamics. The analysis identified nuanced relationships between vehicle attributes and market behavior, offering a foundation for deeper exploration.

B. Future Work

1. For future enhancements, the project could integrate real-world data to validate the findings from the fake dataset and enhance the accuracy of market insights.

2. Integrating advanced predictive analytics and machine learning models could refine sale price predictions and demand forecasting.

3. Additionally, incorporating geographic data to analyze market trends on a regional basis and extending the dataset to include more diverse vehicle categories would offer a more granular view of the used vehicle market.

4. Finally, exploring regional market trends and the impact of external economic factors would provide a more comprehensive view of the global used vehicle market, potentially revealing untapped opportunities and trends.