Used Vehicle Database Capstone Project (Applied Data Science)

DTSC 691

Spring 1 2024

Paul J. Walker

CONTENTS OF THIS PRESENTATION

General Outline:

1. Background information for context
2. Introduction to project and its relevance
3. Project overview
4. Software and technology utilized
5. Post database creation analyses & visualizations
6. Project Outcomes
7. Conclusions and Future Improvements

“Data is a precious thing and will last longer than the systems themselves.”

— Tim Berners-Lee
Inventor of the World Wide Web

Background Information

1. **DBMS Fundamentals**
   a. **Core to business operations across industries, enabling efficient data storage, organization, and access using relational databases and SQL**
   b. **Critical for data analysis, CRM, and inventory management, facilitating informed decision-making and optimization**

2. **Used Vehicle Data Sources**
   a. **Quite diverse - including dealerships, online marketplaces, vehicle history reports, and government databases, offering comprehensive insights into the pre-owned car market**

3. **Impact on the Automotive Industry**
   a. **Data analysis supports strategic decision-making, market trend analysis, and enhances customer relationship management within the used vehicle sector.**

Intro to Project

1. **Objective and Motivation**
   a. **Design a relational database to manage used car sales, addressing industry challenges by tracking car details, ownership, transactions, and service records**

2. **Key Components**
   a. **Implement tables for Cars, Owners, Ownership History, Transactions, etc., with relationships to ensure data consistency and integrity**

3. **Impact and Contributions**
   a. **Provide a comprehensive data management solution to improve the automotive industry's efficiency and decision-making regarding used car sales**

An Overview

1. **Comprehensive Database Design**
   a. **Structured relational model ensuring detailed and organized data management**

2. **Data Integrity and Accessibility**
   a. **Emphasis on maintaining data accuracy and facilitating efficient search and query operations to support dynamic data retrieval needs**

3. **Strategic Data Insertion and Analysis**
   a. **Utilization of manual inputs and automated processes via Faker for data generation, integrated with advanced Python analysis**

4. **Visualization and Insight Generation**
   a. **Application of Pandas for data manipulation, along with Matplotlib and Seaborn for creating compelling visualizations, enabling insightful trends**

My Technologies of Choice

- Integrated Database System - Utilizing MySQL for robust data storage and management, catering to the complex needs of the used car sales market
- Dynamic Data Analysis and Modeling - Leveraging Python within Jupyter Notebooks for interactive data exploration, manipulation, and documentation, facilitating an innovative data analysis approach
- Realistic Data Generation - Employing the Faker API to produce realistic and comprehensive datasets for system development, testing, and analysis
- Advanced Data Manipulation and Visualization: Utilizing Pandas and NumPy for sophisticated data manipulation, and Matplotlib and Seaborn for creating compelling data visualizations, enabling deep insights into the used car sales data

Post Database Creation: Analysis - Background

Goal of Analysis

- 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, etc

Python Tools Used

- Pandas for data manipulation
- Scipy.stats for hypothesis testing (e.g., Pearson correlation, ANOVA)
- Statsmodels for regression analysis and time-series decomposition

Post Database Creation: Analysis - Key Findings

- Mileage vs. Sale Price: The relationship between vehicle mileage and sale price is weaker than expected, suggesting other factors are more influential in determining price
  - Using OLS Regression

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	1.232
Date:	Fri, 23 Feb 2024	Prob (F-statistic):	0.267
Time:	19:15:49	Log-Likelihood:	-30166.
No. Observations:	2805	AIC:	6.034e+04
Df Residuals:	2803	BIC:	6.035e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.679e+04	425.410	62.968	0.000	2.6e+04	2.76e+04
Mileage	0.0042	0.004	1.110	0.267	-0.003	0.012

Omnibus:	110.166	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	48.260
Skew:	0.019	Prob(JB):	3.31e-11
Kurtosis:	2.359	Cond. No.	2.25e+05

Post Database Creation: Analysis - Key Findings

- Car Makes and Sale Price: Variability in average sale prices among car makes exists but is less pronounced than hypothesized, indicating brand perception's limited effect moderated by external factors
  - Using One-Way ANOVA

```
F-statistic: 1.0713232445354648
P-value: 0.3332891648335326
Fail to reject null hypothesis: There is no significant difference in average sale price between different car makes.
```

Post Database Creation: Analysis - Key Findings

- Vehicle Condition Impact: Vehicle condition has a surprisingly minimal impact on sale price, hinting at buyers valuing brand, model, or features more
  - ➢ Using One-Way ANOVA

```
F-statistic: 0.17672068641108232
P-value: 0.9122067271383474
Fail to reject null hypothesis: There is no significant impact of vehicle condition on sale price.
```

Post Database Creation: Analysis - Key Findings

- Incidents and Market Demand: There's no direct correlation between incidents and market demand, pointing to economic factors or vehicle specifics as more significant influences
  - Using Chi-Square and Logistic Regression

```
Chi-square statistic: 434.0845975882778
P-value: 0.752402325698046
Fail to reject null hypothesis: There is no significant association between incidents and market demand.
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.000000
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:     MarketDemand_binary   No. Observations:                 2671
Model:                           Logit   Df Residuals:                     2669
Method:                            MLE   Df Model:                            1
Date:                 Fri, 23 Feb 2024   Pseudo R-squ.:                     inf
Time:                         19:15:52   Log-Likelihood:            -6.5998e-08
converged:                       False   LL-Null:                        0.0000
Covariance Type:             nonrobust   LLR p-value:                     1.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             23.6915   3892.801      0.006      0.995   -7606.058    7653.441
Incidents_binary   9.0203   3.41e+05   2.65e-05      1.000   -6.67e+05    6.67e+05
==============================================================================
```

Post Database Creation: Visualizations - Background

Visualization Techniques

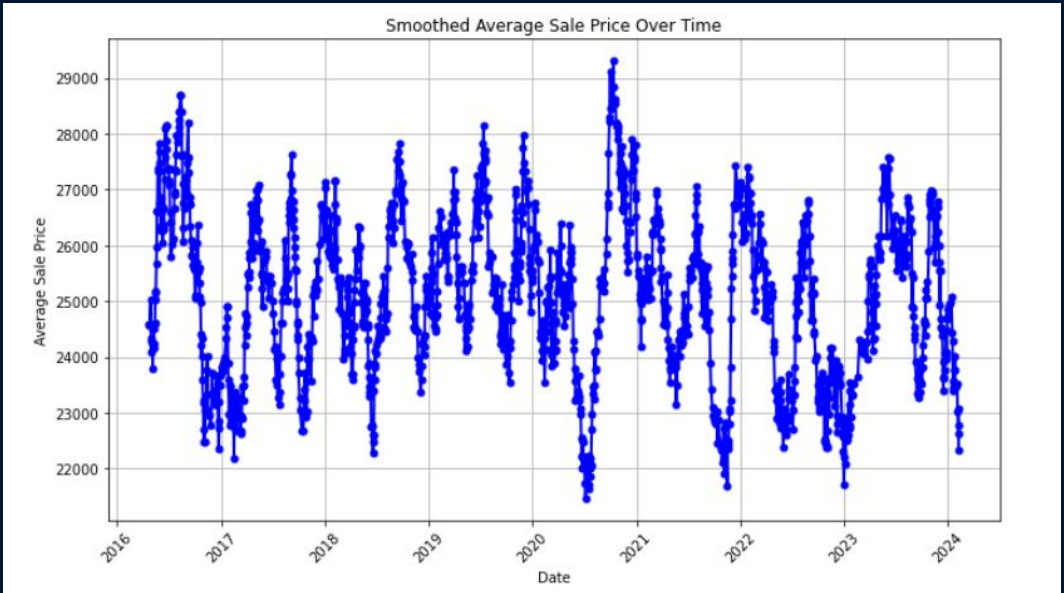
- Utilized histograms, bar charts, box plots, scatter plots, line charts, pie charts, heatmaps, pair plots, violin plots, and word clouds

Python Tools Used

- Employed matplotlib.pyplot, seaborn, and wordcloud libraries in Python for a comprehensive graphical analysis of the used vehicle data

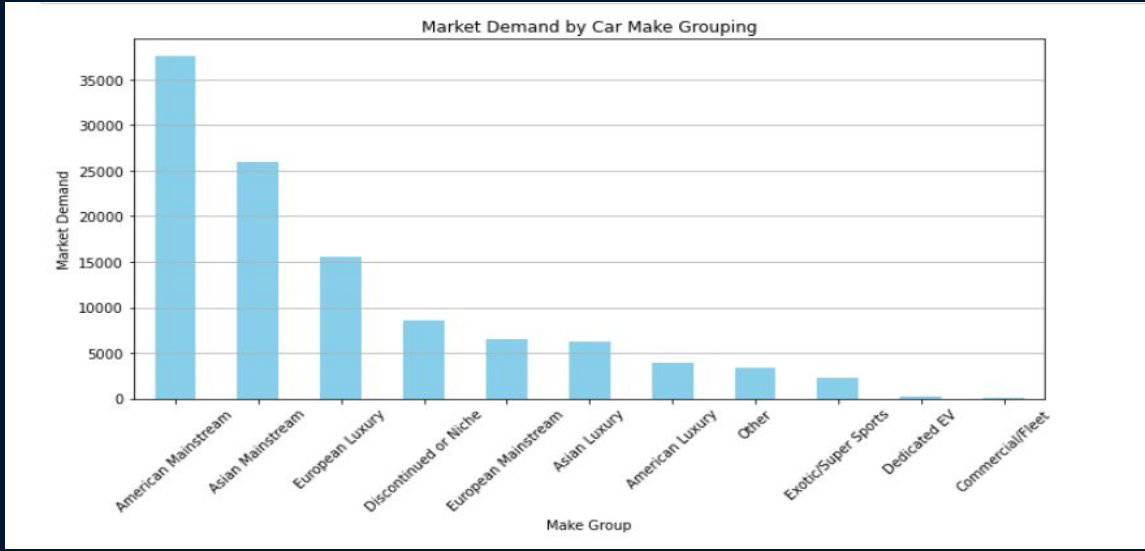
Post Database Creation: Visualizations - Key Findings

- Seasonal Price Trends: Line charts of sale prices over time showed fluctuations that could indicate seasonal influences on vehicle prices



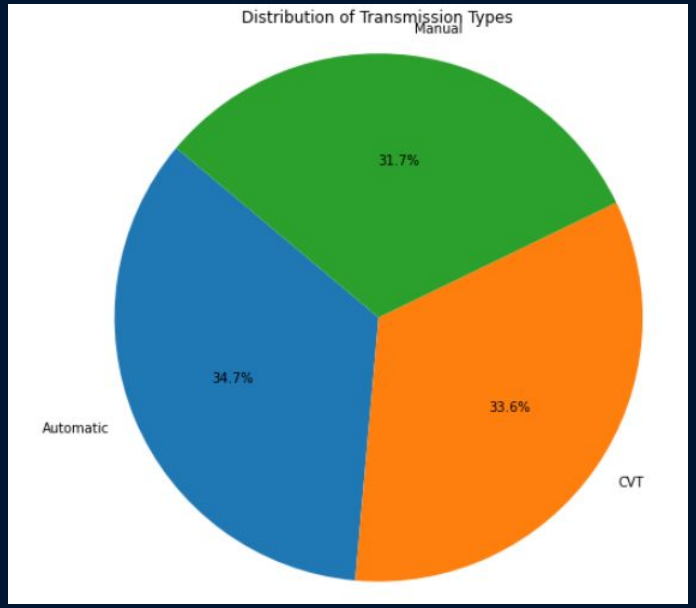
Post Database Creation: Visualizations - Key Findings

- Market Demand Variations: Bar graphs revealed that American Mainstream vehicle manufacturer's are the most sought after by consumers



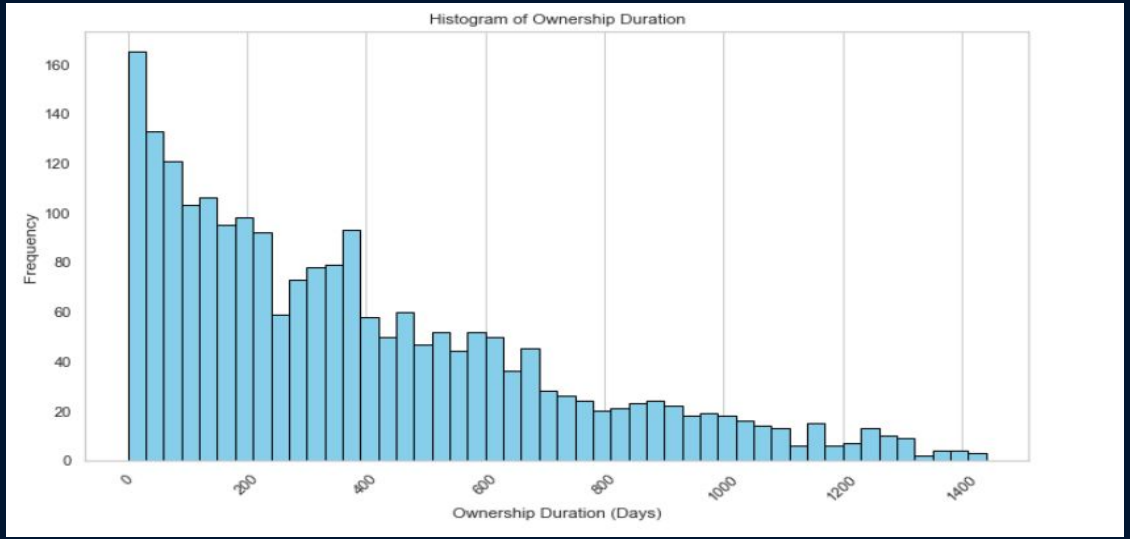
Post Database Creation: Visualizations - Key Findings

- Transmission Type Distribution: Analysis of transmission types through pie charts highlighted a diverse set of preferences or availabilities in the market



Post Database Creation: Visualizations - Key Findings

- 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



Project Outcomes

Insightful Market Analysis

- Revealed intricate dynamics and consumer preferences in the used vehicle market through advanced data generation, analysis, and visualization techniques

Factors Influencing Market Performance

- Uncovered key insights into the impact of vehicle attributes, features, and conditions on sale prices, demand, and overall market value

Closing Remarks

To Summarize;

- My project successfully designed and implemented a relational database for managing used car sales
- I was able to generate fake data to be used for subsequent analysis and reporting
- Advanced data analysis and visualization techniques provided deep insights into market dynamics, consumer preferences, and factors influencing vehicle sale prices and demand

Future Enhancements;

- Enhancing the project with real-world data validation, advanced predictive analytics, machine learning for refined price predictions, and incorporating geographic and diverse vehicle data will provide deeper market insights and improve forecasting accuracy.
- Extending analysis to include regional trends and the impact of external economic factors will offer a comprehensive global view of the used vehicle market, revealing new opportunities and trends