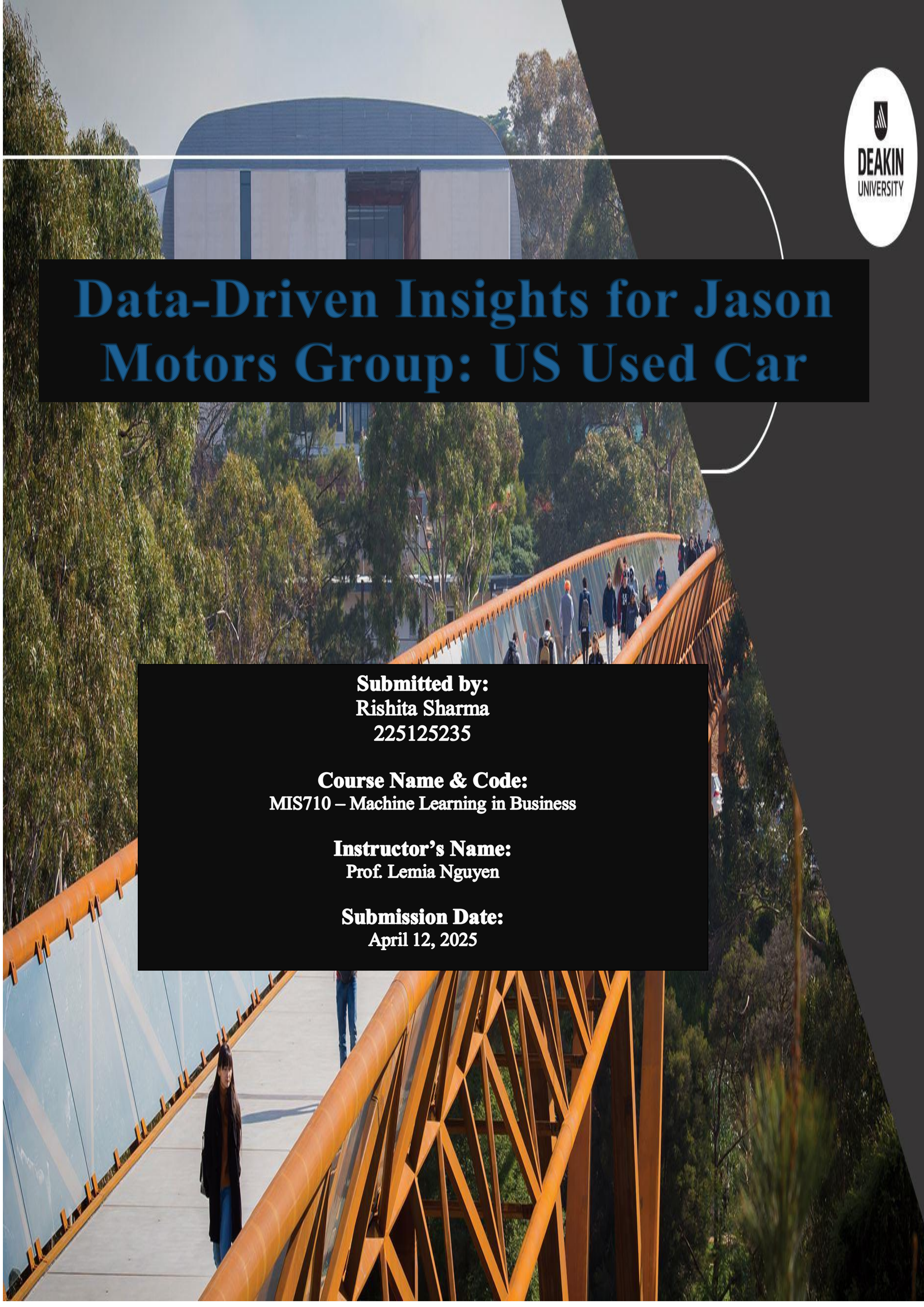


DEAKIN UNIVERSITY

Data-Driven Insights for Jason Motors Group: US Used Car

Submitted by:
Rishita Sharma
225125235

Course Name & Code:
MIS710 – Machine Learning in Business

Instructor's Name:
Prof. Lemia Nguyen

Submission Date:
April 12, 2025

Table of Contents

Summary

Due to increasing competition and market uncertainties, Jason Motors Group (JMG) wants to enter into the United States used car market. Multiple Linear Regression model was applied to examine the influence of various features on price. The key factors considered were cylinders, vehicle condition, drive type, vehicle age, fuel type, and odometer reading.

Condition and cylinders have a positive influence on price,according to model evaluation using R2, MAE, RMSE, and VIF. The drive-type, car-age, fuel-type, and odometer reading, had a negative effect.

Recommendations:

To increase resale value, JMG must target cars that are newer, well-maintained, have less mileage, and a greater number of cylinders to increase profit margins.

Future Works

The Support Vector Machine (SVM) represents a robust machine learning algorithm would be best fit. This model is particularly suitable for predicting the prices of used vehicles, as it identifies complex patterns within datasets and demonstrates effective performance in scenarios involving non-linear relationships **(Bishop, 2021)**.The XGBoost Model is renowned for being extremely fast and capable of dealing with big datasets. It is also extremely good at predicting the auto prices and achieving the highest prediction accuracy **(Chen & Guestrin, 2016).**

BAACM

Need:
Due to rising competition and uncertainty in Australia, JMG plans to enter the US market, where demand for used cars is high. They aim to understand consumer preferences and reduce inventory time to improve profit margins.

Solution:
EDA and Multiple Linear Regression were used to identify key price drivers—condition, age, odometer, fuel type, model, and cylinders. The focus is on newer, low-mileage, fuel-efficient cars with more cylinders to guide a data-driven inventory strategy.

Value:
The insights help JMG cut holding costs, align with market demand, and boost profitability, supporting a low-risk and successful US market entry.

Change:
JMG's pricing and inventory strategy will shift to prioritize high-demand vehicles—popular brands, newer models, low mileage, and efficient fuel use—based on regional preferences.

Stakeholders:
•Sales Manager: Uses popular car features to optimize inventory.
•Marketing Manager: Uses consumer patterns to target marketing efforts.
•CEO: Directs strategic decisions for growth and financial success.

Context:
JMG is using Craigslist data on US car listings to learn about market trends and price influences. So that, JMG can effectively modify its price and inventory for the competitive US market, for long-term growth and profitability.

Data Overview and Preprocessing

Dataset Summary:

The dataset includes **62,946** Craigslist-sourced used car listings with **18** attributes, providing crucial information about the US used automobile market.

Features Involved:

- **Categorical Features:**

⇒ *Listed_Date*
⇒ *Make*
⇒ *Model*
⇒ *Vehicle_Type*
⇒ *Size*
⇒ *Color*
⇒ *Transmission*
⇒ *Fuel_Type*
⇒ *Drive*
⇒ *Title_Status*
⇒ *State*
⇒ *Region*
⇒ *Condition*

- **Numerical Features:**

⇒ *Year*
⇒ *Cylinders*
⇒ *Odometer*
⇒ *Listed_Price (Target Variable*)
⇒ *CarID*

Data Quality Assessment:

- **Missing Values:**

⇒ *Cylinders:* There were 374 missing values in the Cylinders.
⇒ *Region:* 301 missing values were found in the Region.

- **Outliers Detected:**

⇒ *Odometer*: Outliers were found in certain car listings with odometer values higher than 270,000 miles.
⇒ *Year*: Outliers were found in vehicles that were made prior to 1990, and their prices differed greatly from those of more recent models.

⇒ *Listed_Price*: Outliers were found in prices ranging from $39,000 to $70,000.

⇒ *Age_Of_Car*: The majority of vehicles costing more than $28,000 and older than eight years were outliers.

⇒ *Vehicle Type*: The most popular car models, sedans and SUVs have outliers in the prices ranged from $20,000 to $75,000.

⇒ *Size:* Outliers were defined as full-size vehicles costing between $45,000 and $70,000.

⇒ Condition: A large number of expensive cars in excellent condition were found to be outliers.

⇒ Fuel Type: Outliers were found to be gas vehicles with prices ranging from $32,000 to $70,000.

Preprocessing Steps:

- **Handling Missing Values:**

⇒ *Cylinders***:** Replaced 374 missing values with the **median** due to the presence of skewness in the data. It preserved the true data distribution**(Osborne, 2013)**

⇒ *Region*: Replaced 301 missing values with the *mode* to ensure consistency**.**

- **Removing Inconsistencies:**

⇒ Where Price is 0 those records were not taken into consideration as they provided no meaningful insights.

- **Feature Encoding:**

⇒ Categorical Features were changed to Numerical features to compare them with target variable

Feature(Categorical Feature)	Encoded As
Size	*Size_N*
Fuel_Type	*Fuel_Type_Encoded*
Transmission	*Transmission_Encoded*
Drive	*Drive_Encoded*
Title_Status	*Title_Status_Encoded*

- **Feature Creation:**

⇒ **Age_of_Car**: Numerical feature derived from Year and Listed_Date to capture car age; converted to categorical for better visualization.

- **Dropping Unnecessary Columns:**

⇒ **Car_ID:** Removed as it was a unique identifier that didn’t contribute to the model.

Exploratory Data Analysis & Key Findings

Univariate Analysis

1.***Make***

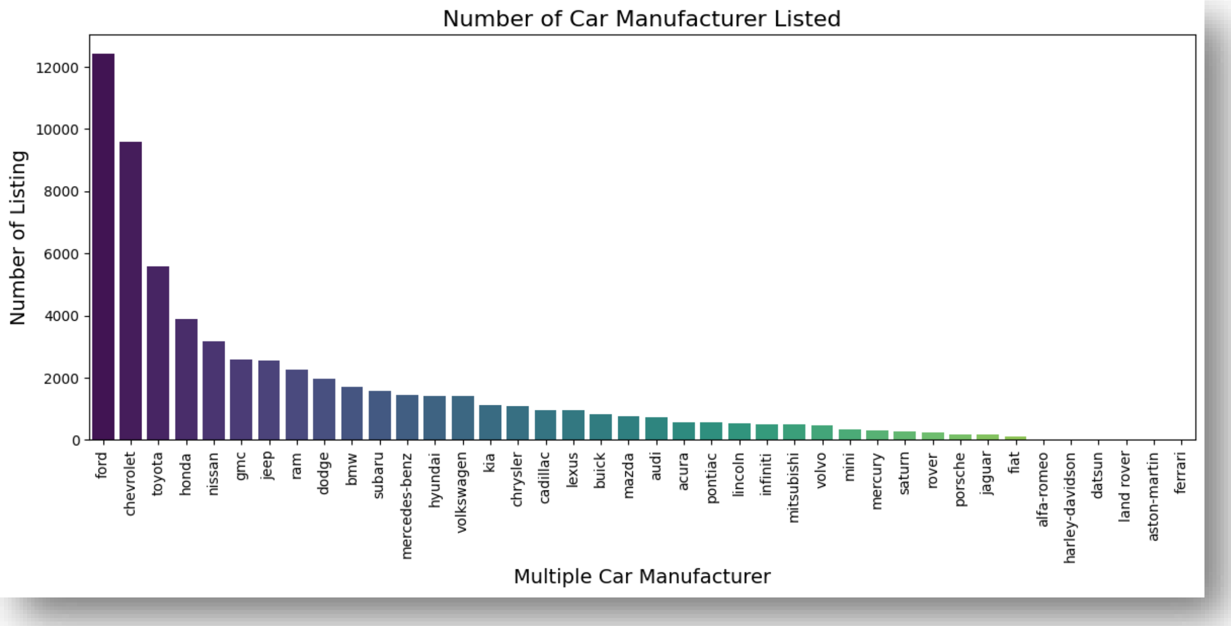


Figure 1:Countplot

EDA:

⇒ There are 40 different manufacturers.

⇒ Ford has over 12000 car listed followed by Chevrolet and Toyota.

Findings:

⇒ Ford, Chevrolet, and Toyota are the dominant Brands.

---

2. ***Vehicle Type***

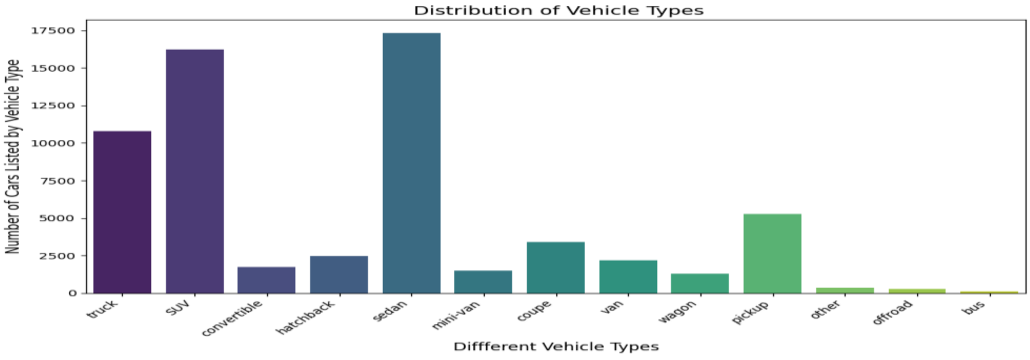


Figure 2:Countplot

EDA:

- ⇒ US Market offers 13 vehicle types.
- ⇒ Sedans has over 17000 car listed followed by SUV and Trucks.

Findings:

- ⇒ High demand for Sedans, SUV and Trucks

---

3. Odometer

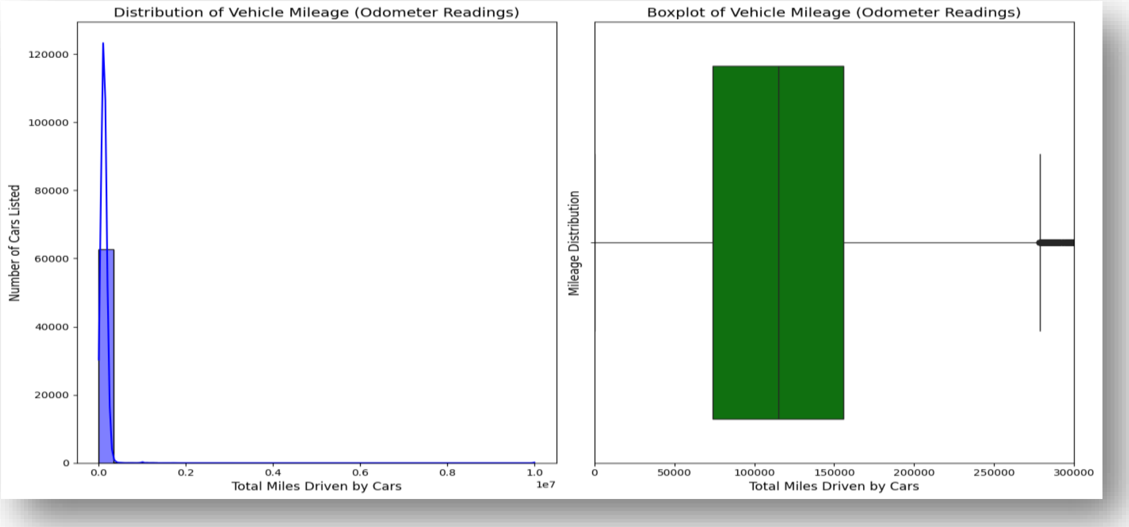


Figure 3:Histplot & Boxplot

EDA:

- ⇒ Most cars had mileage between 74,000 and around 150,000 miles.
- ⇒ average is recorded around 125,000 miles.
- ⇒ Cars with 200,000 miles were frequently listed.

Findings:

- ⇒ The US market is comfortable purchasing vehicles with higher mileage.

---

4. **Year**

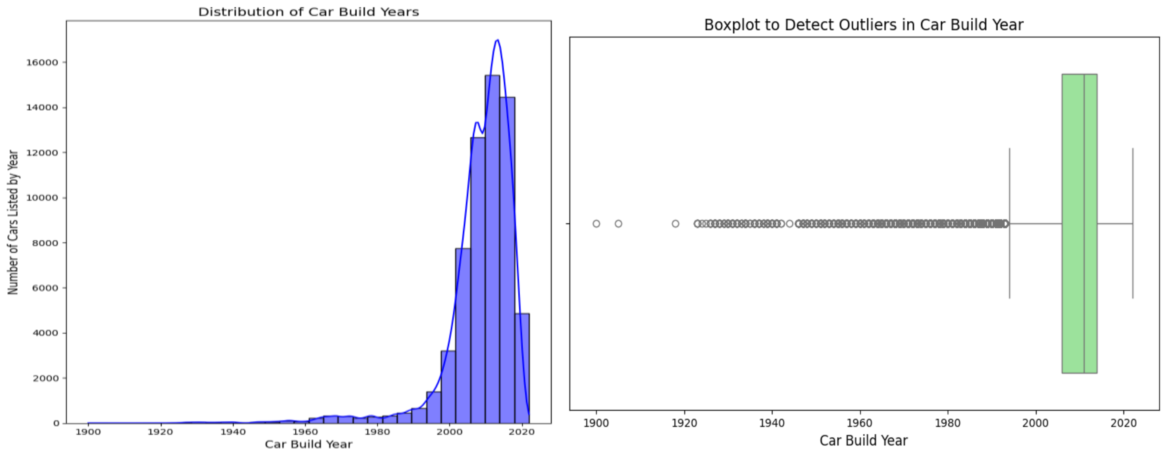


Figure 4:Histplot& Boxplot

EDA:

⇒ The majority of cars were manufactured between 2000 and 2020.
⇒ Vintage cars from 1900-1990 were limited in availability.

Findings:

⇒ High Demand for cars manufactured between 2000-2020, with less focus on vintage models.

5. ***Size***

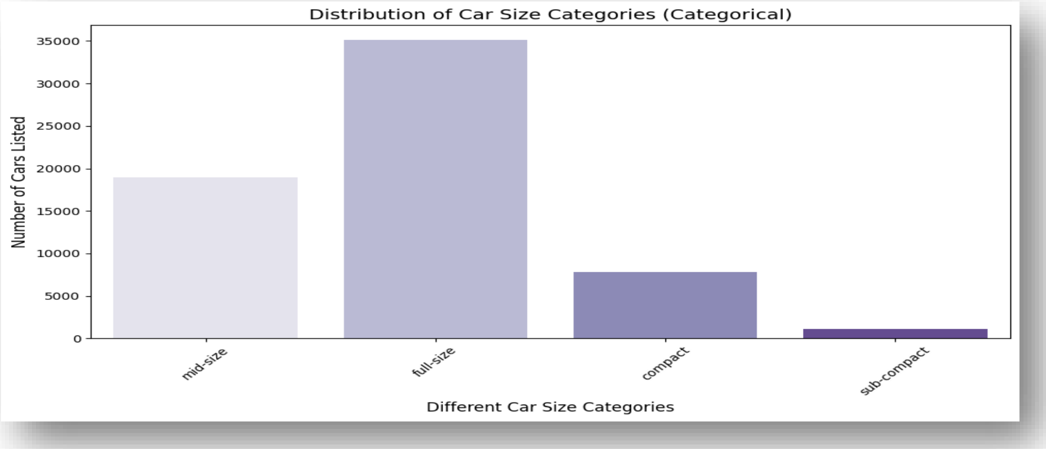


Figure 5:Countplot

EDA:

⇒ Around 35000 cars were listed as Full-Size followed by Mid-Size
⇒ Compact and Sub-Compact cars are less than around 7000 in total.

Findings:

- ⇒ Larger vehicles are more popular among US buyers.
- ⇒ smaller cars have less demand.

---

6. ***Condition***

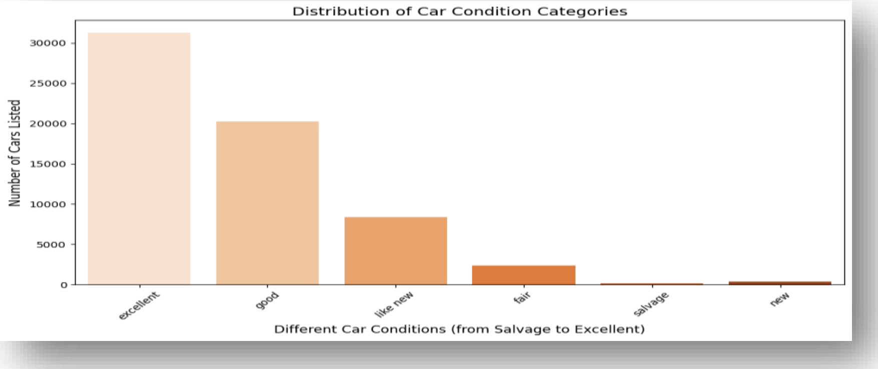


Figure 6:Countplot

EDA:

- ⇒ Around 35,000 cars were listed in Excellent condition, followed by Good condition and like new cars.
- ⇒ Less than 1000 cars were in Salvage condition.

Findings:

- ⇒ Customers prefer cars in Excellent or Good conditions.

---

7. ***Price***

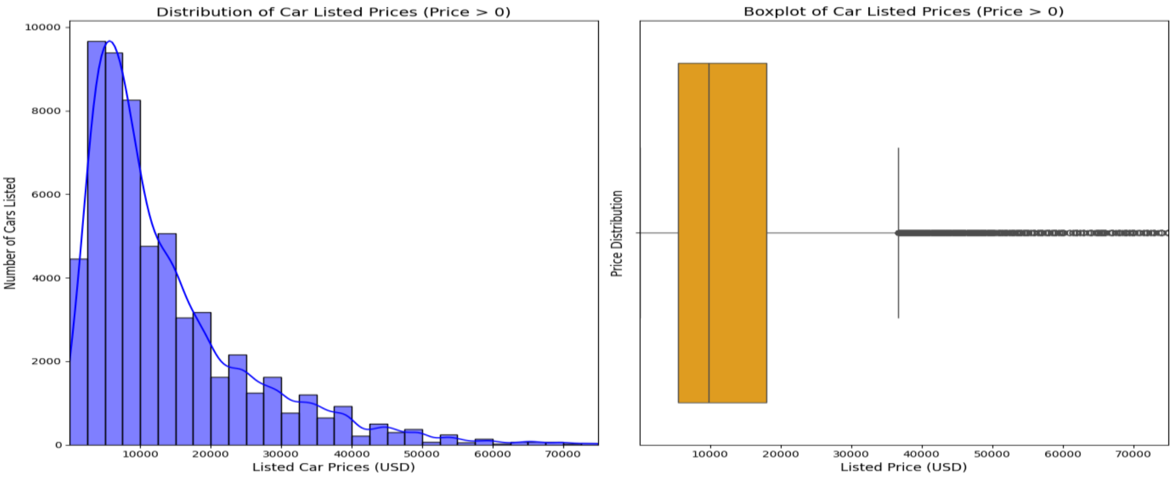


Figure 7:Histplot&Boxplot

EDA:

⇒ Most vehicles were listed around $10,000.
⇒ The highest prices went up to around $70,000.

Findings:

⇒ US has strong market demand for affordable cars, typically priced near $10,000.

8.*Region*

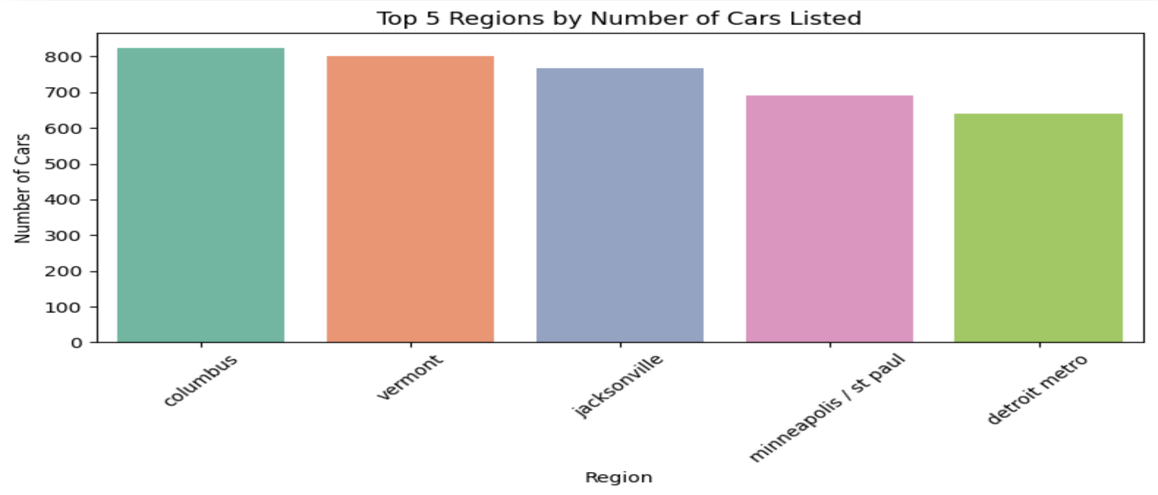


Figure 8:Countplot

EDA:

The top THREE regions with the highest number of car listings are:

1. **Columbus**
2. **Vermont**
3. **Jacksonville**

Findings:

- **Columbus** has the largest car markets in the dataset.
- **Vermont** and **Jacksonville** also has strong demand.

9.***Color***

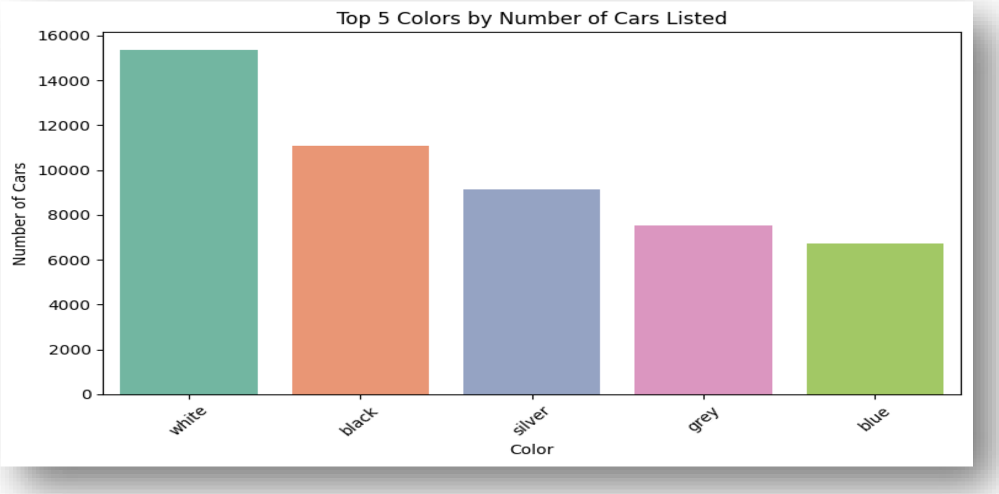


Figure 9:Countplot

EDA:

⇒ The top three car colors with the highest number of listings are:

- **White**
- **Black**
- **Silver**

⇒ White cars dominate the dataset with over 15,000 listings.

Findings:

⇒ Most customer prefer white cars.

⇒ Black and Silver are also popular choices.

Bivariate Analyses

1.***Make & Model***

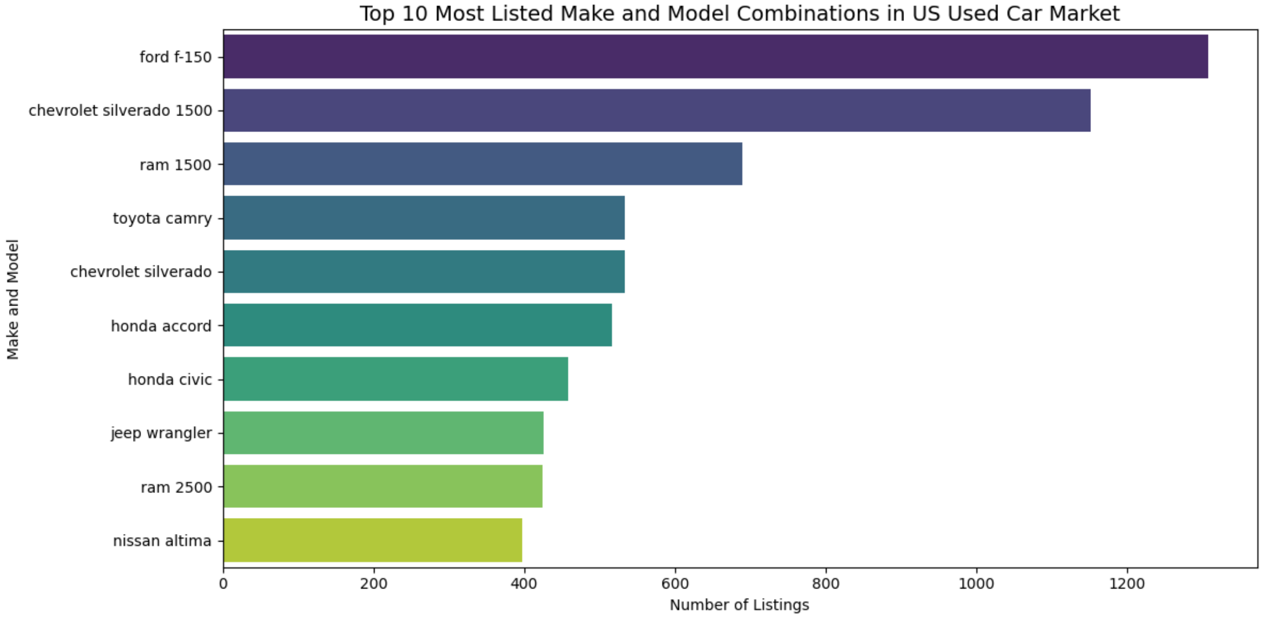
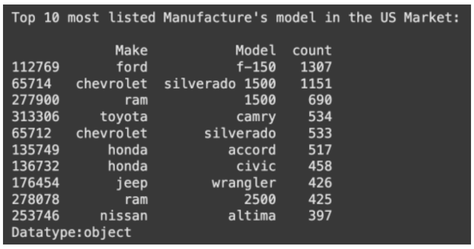


Figure 10:Barchart



```
Top 10 most listed Manufacture's model in the US Market:

               Make           Model  count
112769         ford           f-150   1307
65714     chevrolet  silverado 1500   1151
277900          ram            1500    690
313306       toyota           camry    534
65712     chevrolet       silverado    533
135749        honda          accord    517
136732        honda           civic    458
176454         jeep        wrangler    426
278078          ram            2500    425
253746       nissan          altima    397
Datatype:object
```

Figure 11:CrossTable

EDA:

Ford F-150 is the most listed model with over1,300 cars, followed by Chevrolet Silverado 1500 and Ram 1500 Other popular models include Toyota Camry, Honda Accord, Honda Civic, Jeep wrangler, Ram 2500 and Nissan Altima.

Findings:

The US used car market have a strong demand for Ford F-150 and Chevrolet Silverado 1500. Toyota Camry and Honda Accord also have high market presence.

2. *Odometer & Price*

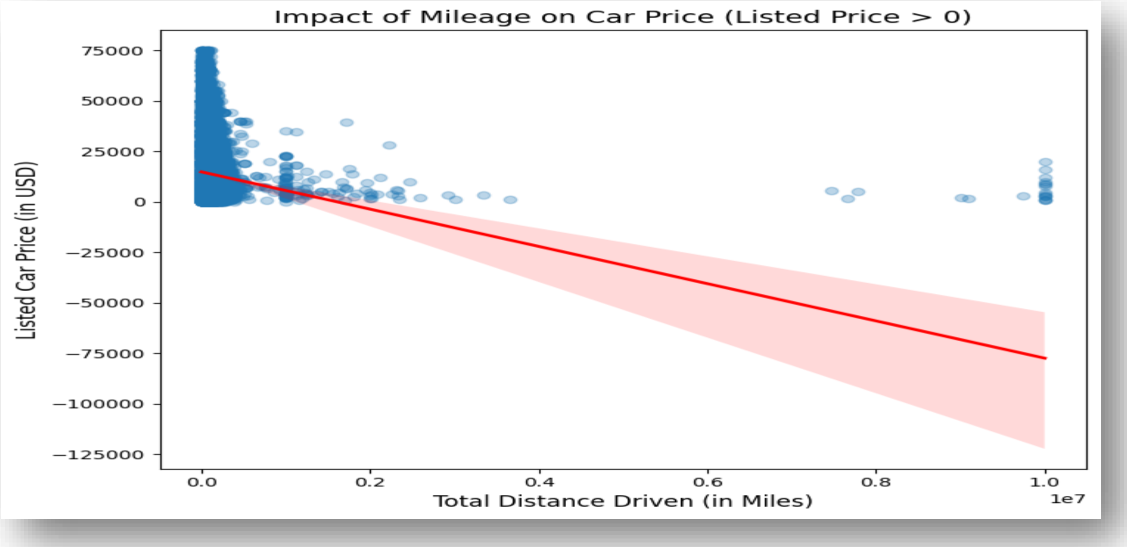


Figure 12:ScatterPlotWithRegressionLine

EDA:

- ⇒ The average odometer reading is around 125,000 miles.
- ⇒ Cars with lower mileage have average price around $22,000, while cars with higher mileage dropped to an average around $8,000.

Findings:

- ⇒ Used cars with lower mileage command higher prices, while cars with mileage beyond 150,000 miles see a sharp price drop.

---

3. *Odometer & Age*

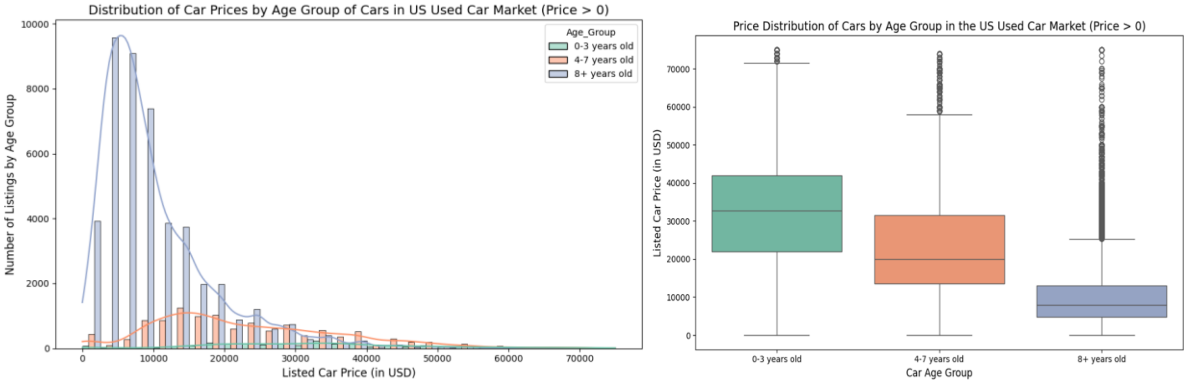


Figure 13:HistPlot&Boxplot

EDA:

- ⇒ Most cars priced up to $20,000 are 8+ years old.
- ⇒ 0-3 year-old cars are priced higher.

Findings

- ⇒ 8+ years old cars dominate the affordable segment, whereas 0-3 years old cars are fewer and priced higher.

---

4. ***Vehicle Type vs Price***

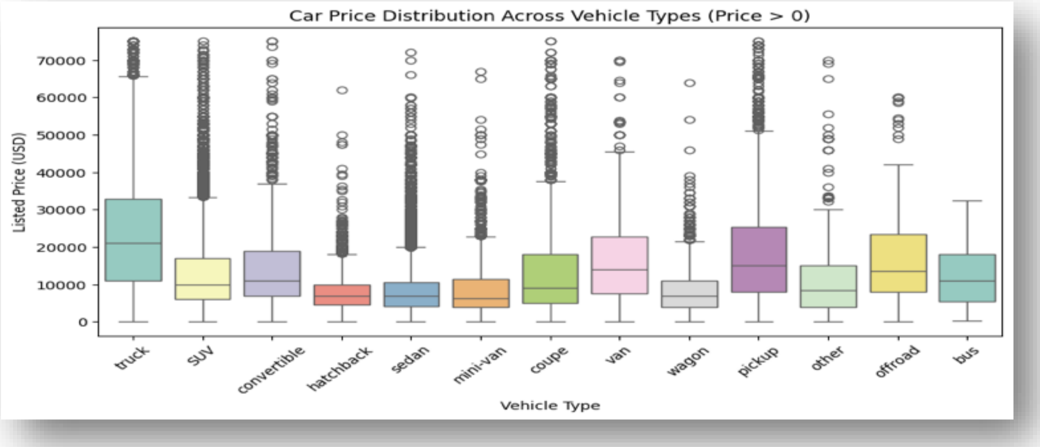


Figure 14:Boxplot

EDA:

- ⇒ Convertible, Coupe, Pickup, and Truck type vehicles have the highest listed price around $75,000.
- ⇒ The average price for these vehicle types is around $15,000.

Findings:

- ⇒ Higher vehicle prices in the US used car market are commonly associated with specific vehicle types like Convertibles, Coupes, Pickups, and Trucks.

---

5. ***Size vs Price***

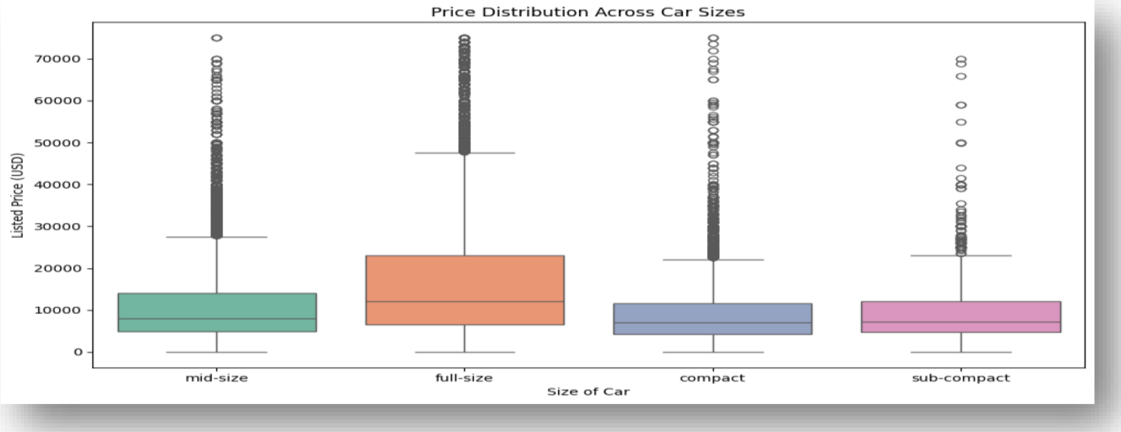


Figure15:Boxplot

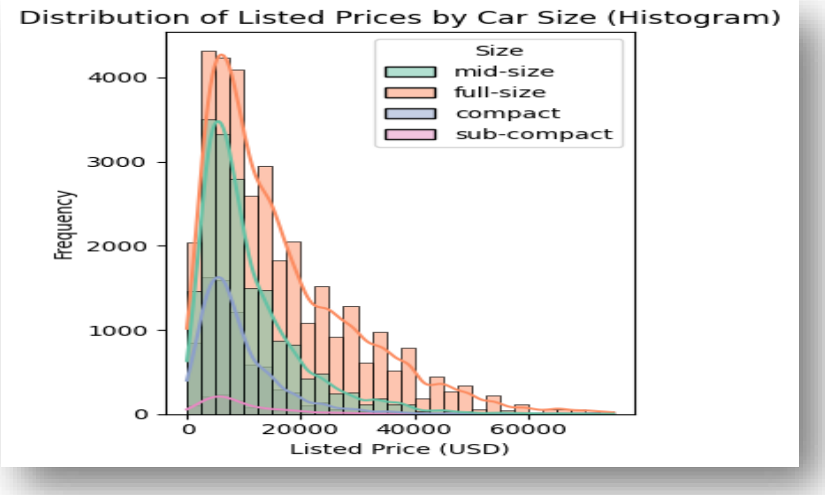


Figure16:Histplot

EDA:

⇒ Full-size cars have the highest average price of around $16,000.
⇒ Compact and Sub-compact cars are the cheapest with average prices $9000.
⇒ Maximum price for all sizes goes up to $75,000.

Findings:

⇒ Bigger cars like Full-size and Mid-size are generally more expensive.
⇒ Smaller cars like Compact and Sub-compact are more affordable.

6. ***Condition vs Price***

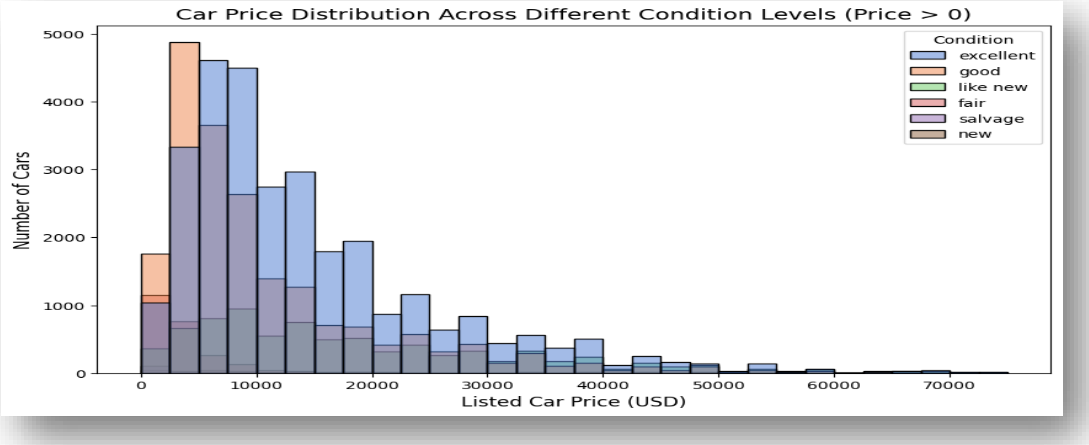


Figure17:Histplot

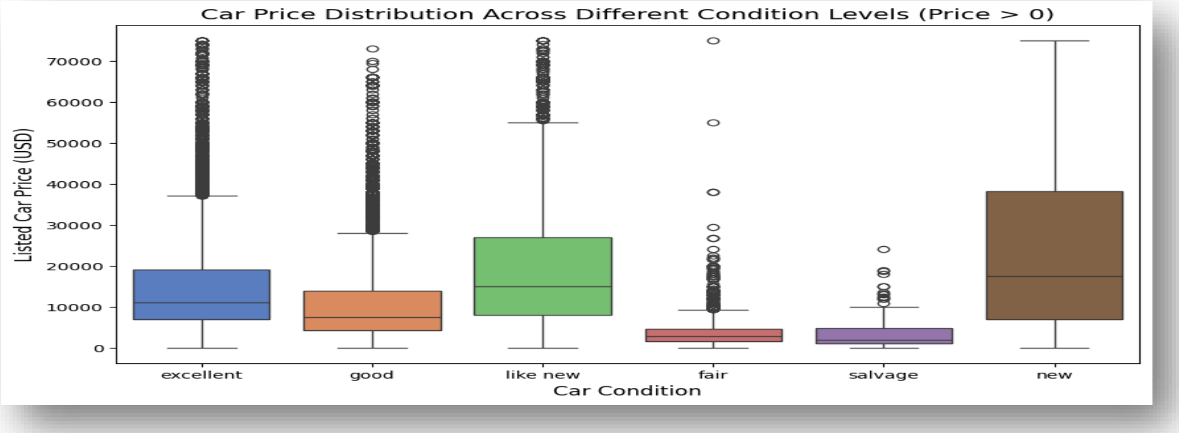


Figure18:Boxplot

EDA:

- ⇒ Excellent condition cars are priced around $14,000, followed by Good condition cars around $10,000.
- ⇒ Like new and New cars are priced around $20000.

Findings:

- ⇒ higher prices for New or Like New cars.
- ⇒ Most of the used cars in the market fall under Excellent and Good condition

7. ***Fuel Type vs Price***

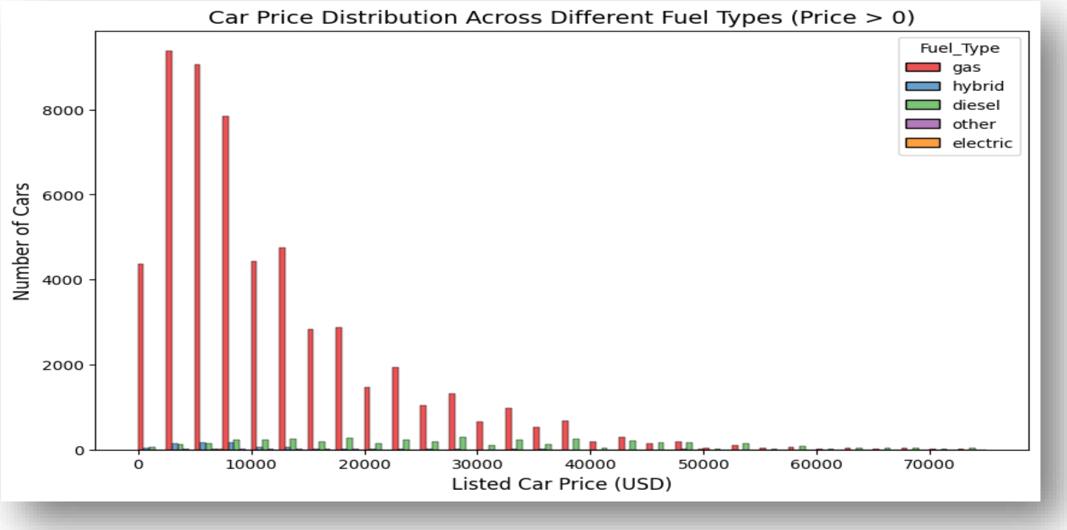


Figure19:Histplot

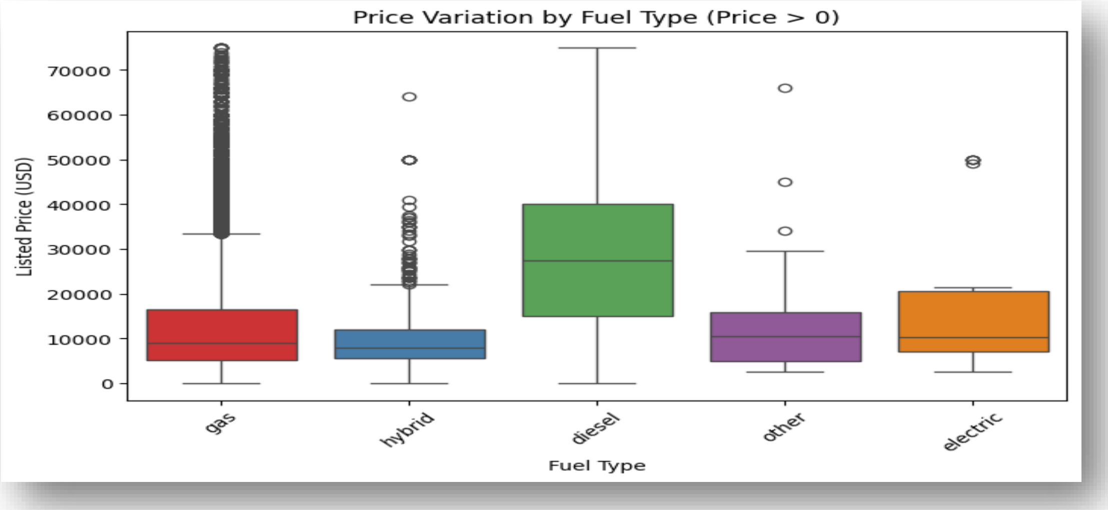


Figure20:Boxplot

EDA:

⇒ Diesel cars have the highest average price of around $29,000, followed by Electric cars with an average price of $17,800.

⇒ Gas cars dominate the market and have an average price of $12,500.

Findings:

⇒ Diesel cars are the most expensive on average.
⇒ Electric cars, although limited in availability, are relatively higher in price.
⇒ Gas cars dominate the market.

Multivariate Analysis

Price vs Selected Feature

Features Used for Multivariate Analysis
Size_N
Condition_N
Fuel_Type_Encoded
Odometer
Age_Of_Car
Transmission_Encoded
Drive_Encoded
Cylinders
Title_Status_Encoded

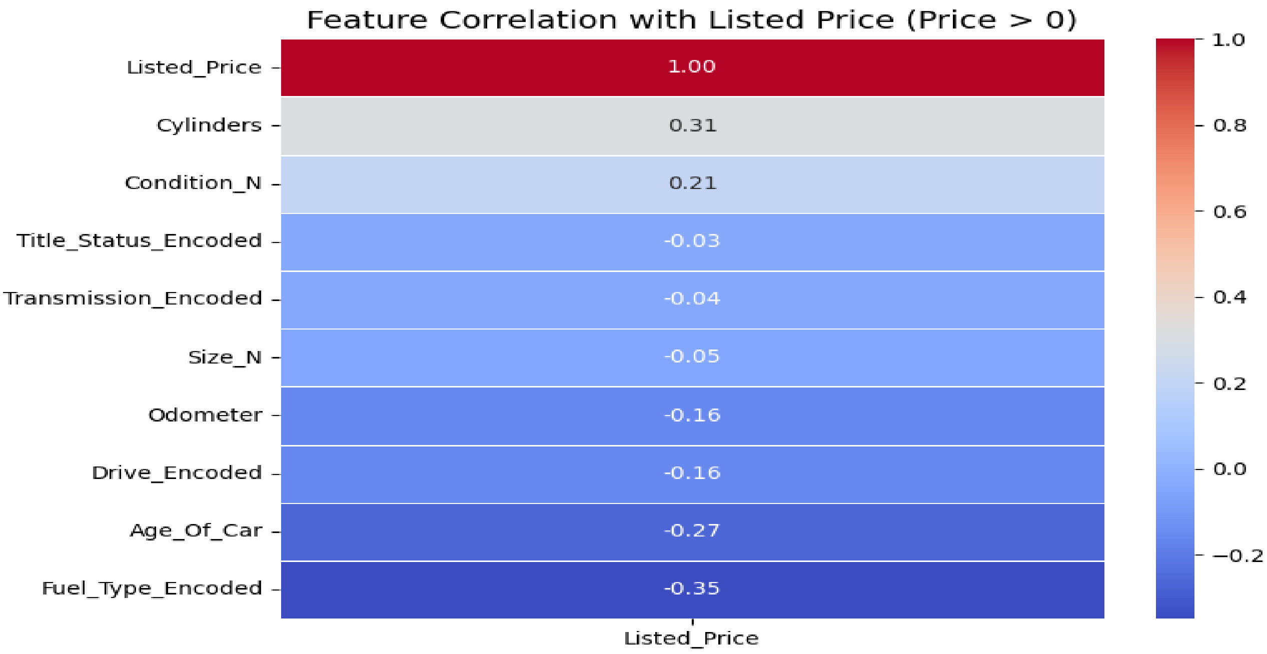


Figure21:Heatmap

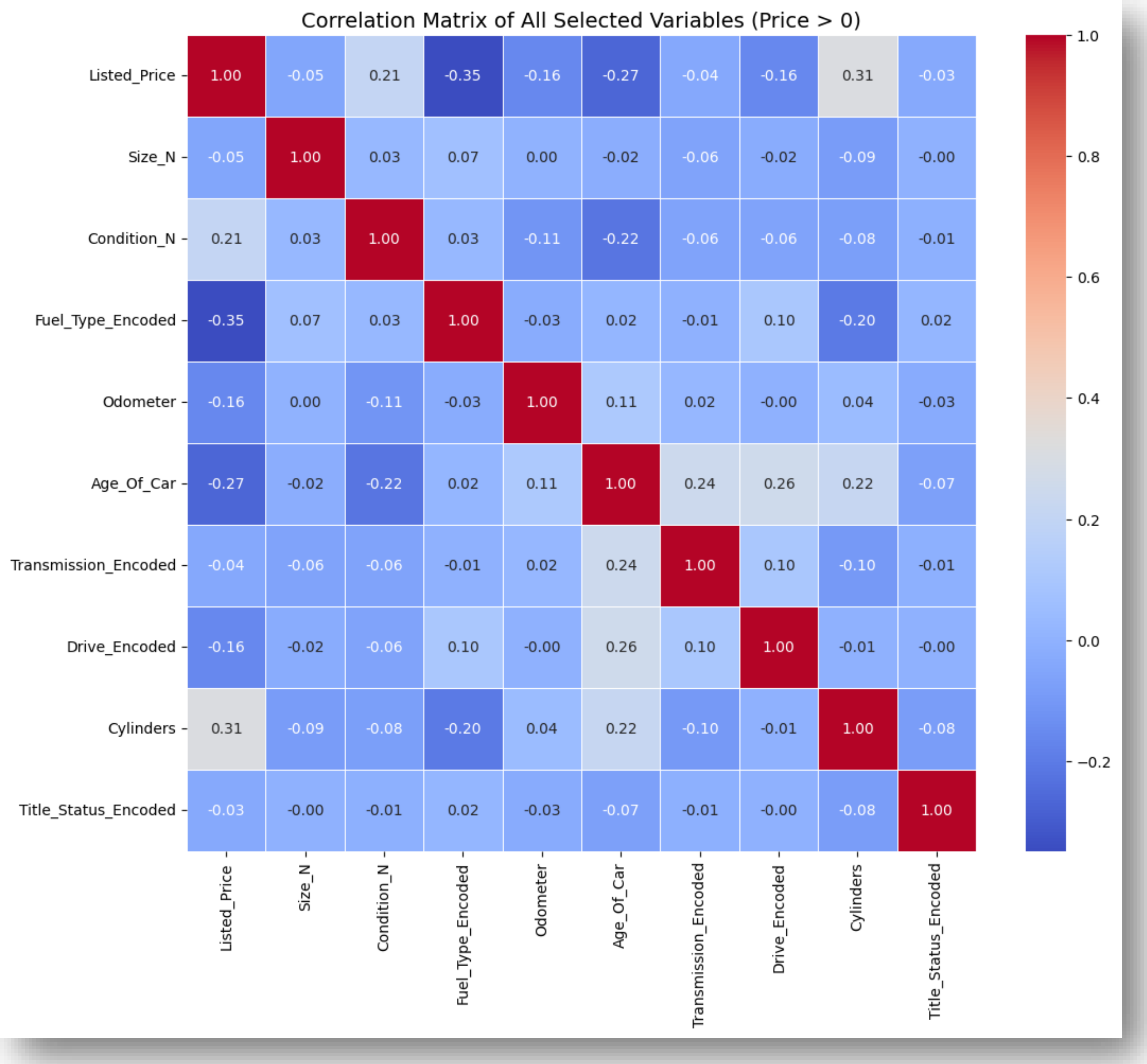


Figure22:Heatmap

EDA:

Correlation with Price:

⇒ **Cylinders (0.31)** and **Condition_N (0.21**) have a moderate positive relationship.
⇒ **Fuel_Type_Encoded (-0.35)** and **Age_Of_Car (-0.27**) have a strong negative correlation.
⇒ **Drive_Encoded (-0.16) and Odometer (-0.16)** have a mild negative correlation with Price.

Findings:

⇒ Cars with higher cylinders and better condition have a positive impact on price.
⇒ Older cars and specific fuel types are key features in reducing the price of a vehicle.
⇒ Features like odometer and drive-type have a mild impact on price, but their effect is less important compared to other variables.

Machine Learning Approach

Supervised Machine Learning was used with *Multiple Linear Regression* to predict used car prices. This method was suitable because the target variable (Price) was continuous, and the model helped to find the relationship between different features like Odometer, Year, Condition, and Cylinders ***(Lee, 2020).***

Multiple Linear Regression was chosen because it is simple, easy to understand, and works well for predicting prices ***(Géron, 2019***). It helped to identify which factors most affect the car prices and provided useful information for business decisions.

Descriptive Statistics was used in *Exploratory Data Analysis (EDA)* to summarize and show patterns in the data. *Inferential Statistics* was used through the regression model to predict car prices and help Jason Motors Group (JMG) make better decisions ***(Sharda et al., 2019).***

Model and Performance Metrics

For the purpose of predicting used car prices, a **Multiple Linear Regression (MLR)** model was used. A dataset including 12,590 records for testing and 50,356 records for training was used to train the model.), this model was selected to illustrates the connection between various features and the target variable(Listed_Price).

Metrics Used

⇒ **R-squared**: Only 27% of the price variance can be explained by the model.
⇒ **Mean Absolute Error**: The model's estimates are generally $7,105 off.
⇒ **Root Mean Squared Error**: The RMSE of $10,011 indicates that certain forecasts are significantly off.
⇒ **Variance Inflation Factor**: VIF values are low (below 2).

Interpretation

⇒ **Intercept (12,118.46)**: The baseline car price when all features are zero, Starting point for the predicted price.
⇒ **Cylinders (2,130.28)**: Each additional cylinder increases the car price by $2,130.
⇒ **Fuel_Type (-5,972.38)**: Certain fuel types decrease the car price by $5,972.
⇒ **Age (-278.52)**: For each year older, the car price drops by $278.52.
⇒ **Odometer (-0.00576)**: Higher mileage slightly reduces the car price.

Pros:

⇒ Low VIF values (below 2) indicate no multicollinearity, meaning features provide unique information and coefficients remain stable. **(Kutner et al., 2005)**
⇒ The model gives a basic understanding of how key factors like Cylinders, Fuel Type,etc.affect car prices.

Cons:

⇒ Low R-squared value (0.27) shows the model explains only 27% of the price variation, indicating weak predictive power.

⇒ High MAE ($7,105) and RMSE ($10,011) suggest large prediction errors, reducing model accuracy and reliability.

Business Solution

⇒ **Sales Manager**: Utilize the MLR model to set competitive car prices by understanding the impact of car features like age, condition, and cylinders on pricing.

⇒ **Marketing Manager**: Analyse insights from regional data and other car features to target marketing, by focusing on popular car types and features in high-demand regions.

⇒ **CEO**: use data analytics to optimize operational decisions such as inventory management and pricing strategy **(Marr, 2016).**

⇒ I**mprove Efficiency**: Reduce manual pricing efforts by providing data-driven price that are consistent and accurate.

⇒ **Data-Driven Decisions**: Enable all stakeholders to make strategic decisions based on important pricing insights that will enhance overall business.

Recommendations for Improvement

Outlier Treatment

⇒ Remove or treat extreme values in features like Odometer and Price to reduce prediction errors**. (Han, Pei, & Kamber, 2011)**

Feature Engineering

⇒ Group car conditions as Good, Average, Poor, or create mileage categories (Low, Medium, High) to improve accuracy.

Add New Variables

⇒ Include additional features like car history and specific features (e.g., safety, model) to impact pricing.

Model Upgrade

⇒ Consider using advanced models like Support Vector Machine (SVM) or XGBoost for better prediction accuracy.

Customer Segmentation

⇒ Segment customers based on car preferences and budget using K-means clustering, targeting specific customer groups for pricing. **(Tan, Steinbach, & Kumar, 2019)**

Real-Time Price Monitoring

⇒ Machine learning to analyse market trends and prices in real time, to stay competitive **(Provost & Fawcett, 2013).**

References and Work Cited

⇒ Bishop, C. M. (2021). *Pattern recognition and machine learning*. Springer.

⇒ Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

⇒ Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. SAGE Publications.

⇒ Lee, W. (2020). *Python machine learning*. O'Reilly Media.

⇒ Géron, A. (2019). *Hands-On machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.

⇒ Sharda, R., Delen, D., & Turban, E. (2019). *Analytics, data science, & artificial intelligence*. Pearson.

⇒ Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied linear regression models* (4th ed.). McGraw-Hill/Irwin.

⇒ Marr, B. (2016). Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results. Wiley.

⇒ Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann

⇒ Tan, P. N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

⇒ Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.