USED CAR PRICE ANALYSIS REPORT

BY THAIER HAYAJNEH

EXECUTIVE SUMMARY

This report explores the factors that influence used car pricing, providing valuable insights for used car dealers aiming to optimize their inventory and pricing strategies. By analyzing a dataset of 426,000 used cars, we have identified essential attributes that significantly impact car prices and developed predictive models to assist in inventory management. This analysis follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to ensure a structured and comprehensive approach.

OBJECTIVES

The primary objective of this analysis is to identify the factors that make a car more or less expensive. This understanding will help used car dealers make informed purchasing, pricing, and inventory marketing decisions. Specifically, the goals are to:

- Identify the main features that affect used car prices.
- Develop predictive models to estimate car prices based on these features.
- Provide actionable recommendations for inventory management and pricing strategies.

UNDERSTANDING THE DATASET

Dataset Overview:

- The dataset has 426,880 entries and 18 columns
- Columns include information about the car's region, price, year, manufacturer, model, condition, cylinders, fuel type, odometer reading, title status, transmission, VIN, drive type, size, car type, paint color, and the state it's registered in.

Data Types and Missing Values:

- Most columns have non-null entries, but some data, such as condition, cylinders, drive, size, type, and paint_color, are missing in several columns.
- The year and odometer columns (Important for analysis) have some missing values but are mostly complete.

Statistical Summary:

- Price: The price column, the target variable, ranges from 0 to about 3.7 billion, which suggests some outliers or incorrect entries since a price of 3.7 billion for a used car is unrealistic. But data only uses 426K.
- Year: The year of the cars ranges from 1900 to 2022. This range might include some historic or antique vehicles. TBD
- Odometer: The mileage of the cars varies from 0 to 10 million miles, with an average of around 98,043 miles. The maximum value seems super high, which can be outliers or errors.

COLUMNS DESCRIPTION

1. id: Unique identifier for each listing
2. region: Geographic region where the car is listed
3. price: The listing price of the car
4. year: Year of manufacture
5. manufacturer: Car manufacturer
6. model: Car model
7. condition: Condition of the car (e.g., new, like new, etc.)
8. cylinders: Number of cylinders in the engine
9. fuel: Type of fuel used (e.g., gas, diesel, etc.)
10. odometer: Odometer reading
11. title_status: Title status of the car (e.g., clean, salvage, etc.)
12. transmission: Type of transmission (e.g., automatic, manual)
13. VIN: Vehicle Identification Number
14. drive: Type of drive (e.g., 4wd, fwd, RWD)
15. size: Size category of the car
16. type: Type/category of the car (e.g., sedan, SUV, truck)
17. paint_color: Color of the car's paint
18. state: State where the car is listed

```
RangeIndex: 426880 entries, 0 to 426879
Data columns (total 18 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   id            426880 non-null  int64
 1   region        426880 non-null  object
 2   price         426880 non-null  int64
 3   year          425675 non-null  float64
 4   manufacturer  409234 non-null  object
 5   model         421603 non-null  object
 6   condition     252776 non-null  object
 7   cylinders     249202 non-null  object
 8   fuel          423867 non-null  object
 9   odometer      422480 non-null  float64
 10  title_status  418638 non-null  object
 11  transmission  424324 non-null  object
 12  VIN           265838 non-null  object
 13  drive         296313 non-null  object
 14  size          120519 non-null  object
 15  type          334022 non-null  object
 16  paint_color   296677 non-null  object
 17  state         426880 non-null  object
dtypes: float64(2), int64(2), object(14)
```

DATA PREPROCESSING

Data pre-processing or cleaning is an important step that needs to be considered when conducting a data analysis because it improves the data quality and makes it easier for machine learning algorithms to use and interpret. These were the results obtained after conducting data cleaning on the vehicle dataset:

- Price Filtering: rows with prices below 100 and above 100,000 dollars were removed. This will eliminate free/swap entries and outliers.
- Odometer Filtering: All entries with odometer readings above 300,000 miles were removed. This takes care of errors or very high-mileage cars less likely to be relevant for typically used car buyers.
- Year Data: Any missing rows in the year information was dropped since the car's age is important for pricing analysis.
- Attributed of Missing Values: Missing values such as manufacturer, model, fuel, and transmission were replaced with the most common value (mode) in each column.

Cleaned Data Characteristics:

- The cleaned dataset now contains 383,449 entries. All important fields (price, year, manufacturer, model, fuel, transmission) have complete data.
- The price ranges from 101 to 99,999 dollars, and the years range from 1900 to 2022.
- The odometer readings are more consistent, ranging from 0 to 299,999 miles.

Data after preprocessing

	id	price	year	odometer
count	3.834490e+05	383449.000000	383449.000000	383449.000000
mean	7.311486e+09	18860.984885	2011.094104	91585.928105
std	4.389369e+06	14377.477373	9.461509	61645.662169
min	7.301583e+09	101.000000	1900.000000	0.000000
25%	7.308103e+09	7499.000000	2008.000000	38048.000000
50%	7.312620e+09	15500.000000	2013.000000	86912.000000
75%	7.315252e+09	27777.000000	2017.000000	134617.000000
max	7.317101e+09	99999.000000	2022.000000	299999.000000)

```
Data columns (total 18 columns):
 #   Column        Non-Null Count
---  ------        --------------
 0   id            383449 non-null
 1   region        383449 non-null
 2   price         383449 non-null
 3   year          383449 non-null
 4   manufacturer  383449 non-null
 5   model         383449 non-null
 6   condition     237274 non-null
 7   cylinders     227606 non-null
 8   fuel          383449 non-null
 9   odometer      383449 non-null
 10  title_status  376239 non-null
 11  transmission  383449 non-null
 12  VIN           236830 non-null
 13  drive         266832 non-null
 14  size          107240 non-null
 15  type          301299 non-null
 16  paint_color   270370 non-null
 17  state         383449 non-null
dtypes: float64(2), int64(2), object
```

EDA

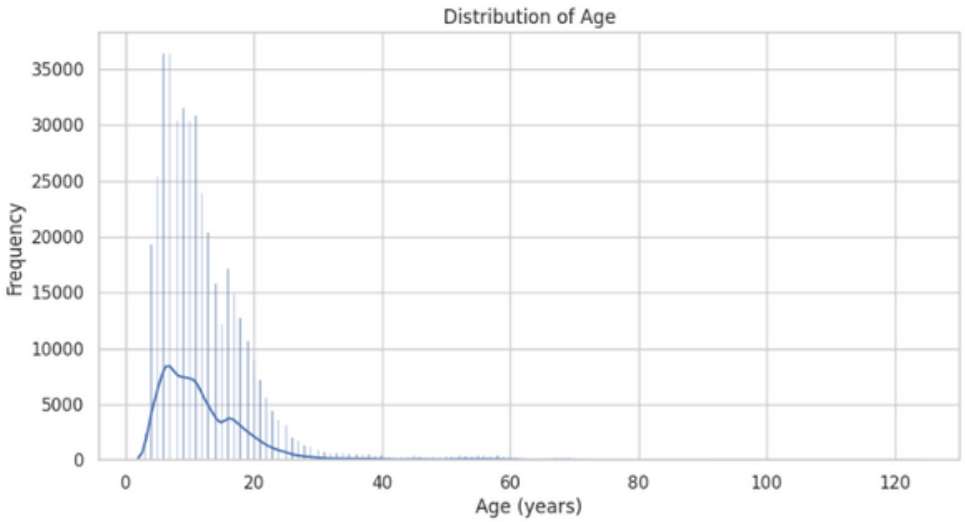
EXPLORATORY DATA ANALYSIS

Exploratory data analysis involves summarizing the main characteristics of a dataset, often through visual methods, to gain a deeper understanding of the data you're working with. EDA helps uncover patterns, identify outliers or errors, and understand relationships between variables. This knowledge is essential for making informed decisions about data cleaning, feature engineering, and model selection, ultimately leading to more accurate and reliable results.

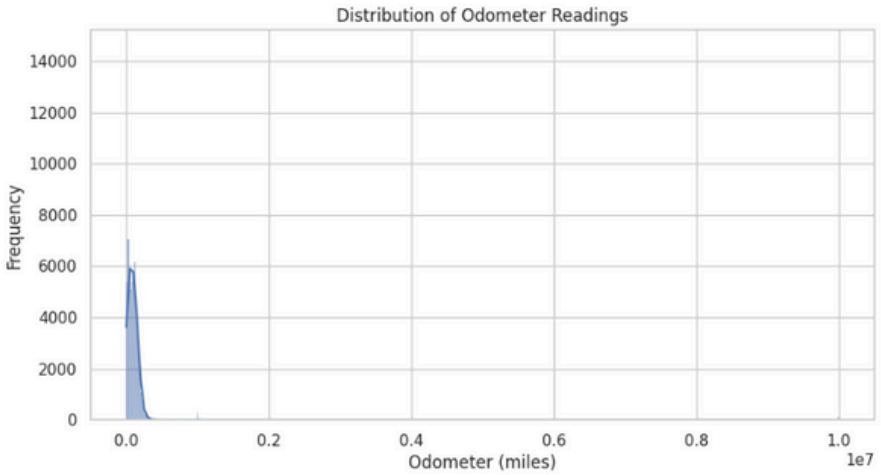
A thorough Exploratory Data Analysis (EDA) was conducted to uncover key features and relationships. This process involved visualizing distributions, correlations, and patterns within the dataset to gain a comprehensive understanding of the factors affecting used car prices.

Histograms

The histogram for the age of vehicles shows a right-skewed distribution. This means many vehicles in this dataset are relatively new, with a peak frequency at younger ages. The frequency decreases for older vehicles, suggesting fewer ancient vehicles are in the dataset. The presence of cars at 0 indicates that many are nearly new. This distribution can imply that newer vehicles are more commonly available or preferred in the dataset you're analyzing.

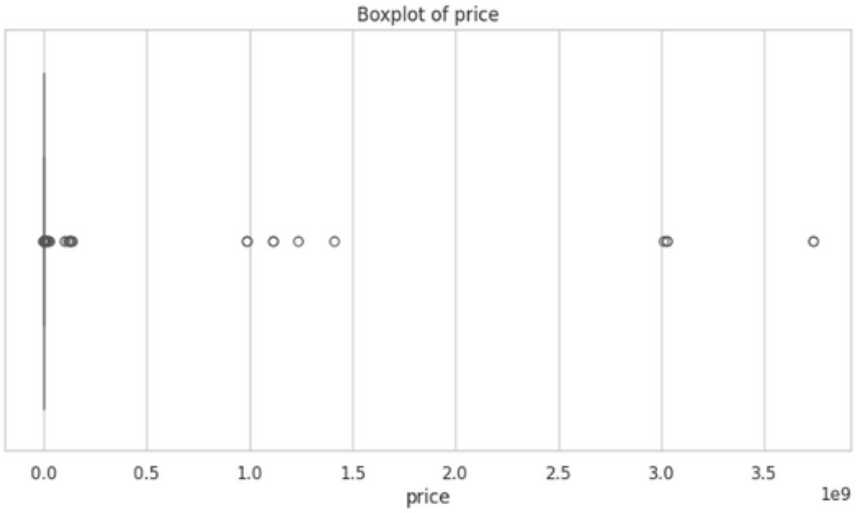


The histogram for odometer readings shows a sharp peak near the lower end of the scale, which rapidly declines as mileage increases. This indicates that most vehicles in the dataset have low to moderate mileage, with few vehicles exhibiting high mileage. The rapid drop-off in frequency as mileage increases suggests that higher-mileage cars are less common, possibly due to lower desirability, reduced availability, or the nature of the dataset focusing more on newer or less-used vehicles.

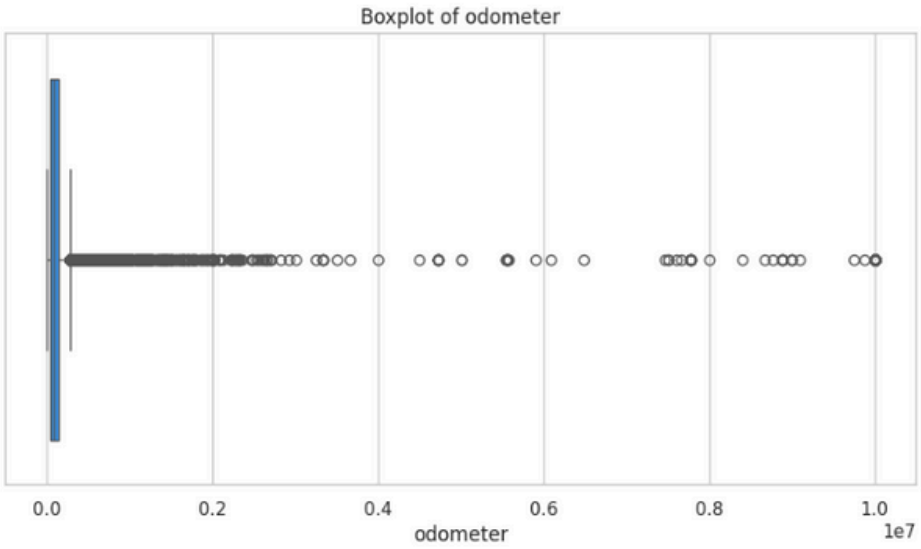


Boxplots

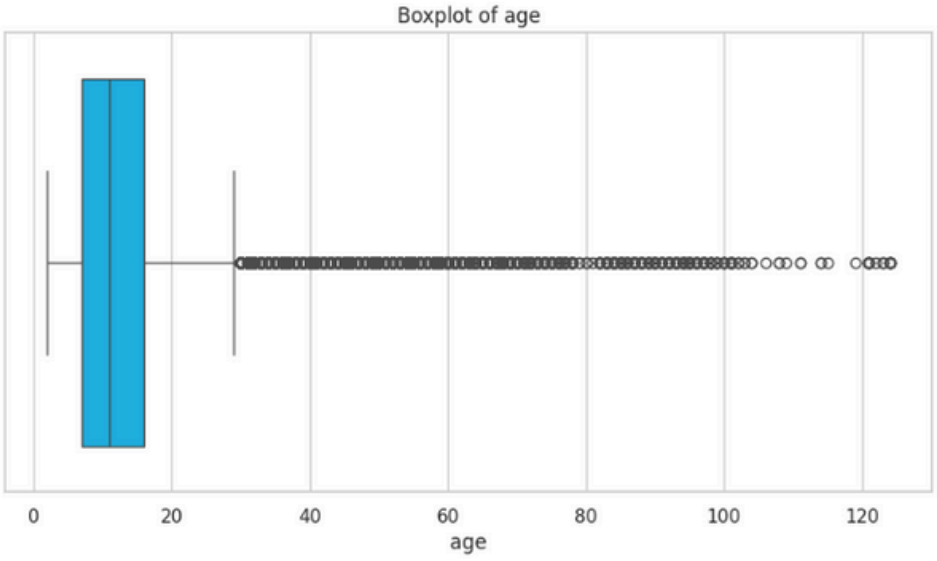
The boxplot of the price reveals significant outliers, with a few extreme values far exceeding the general price range of most used cars. These outliers suggest the presence of high-end or incorrectly priced listings. The bulk of the data is concentrated at the lower end, indicating that most used cars are priced within a relatively affordable range.



The boxplot of the odometer readings shows a wide range of values with many outliers. The majority of the vehicles have lower odometer readings, suggesting they have not been driven extensively. The presence of outliers indicates that some cars have unusually high mileage, which could significantly impact their resale value.



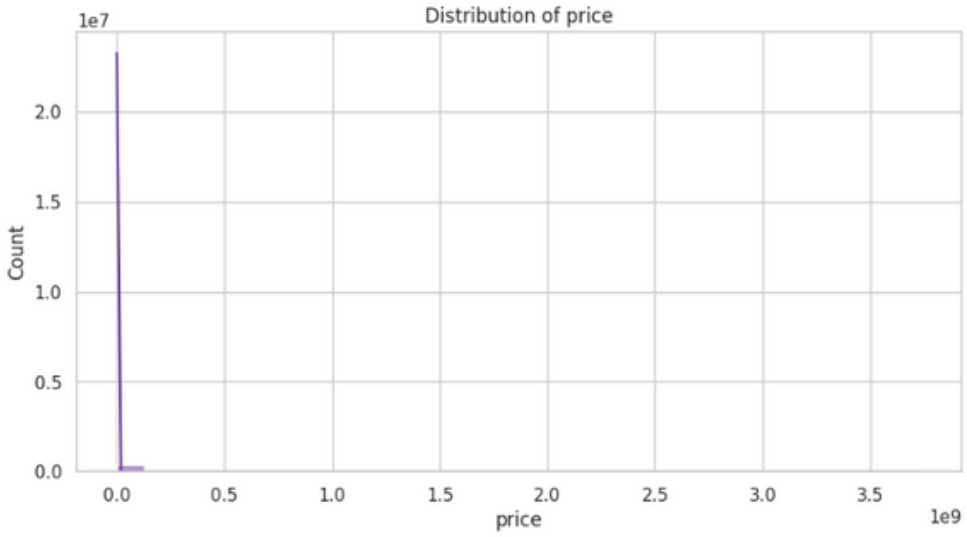
The boxplot of the age of the vehicles shows that most cars are relatively young, with ages clustered between 0 and 20 years. There are a few outliers representing much older vehicles. This distribution highlights the tendency for newer vehicles to dominate the used car market, likely due to their better condition and higher demand.



Histograms for Distribution

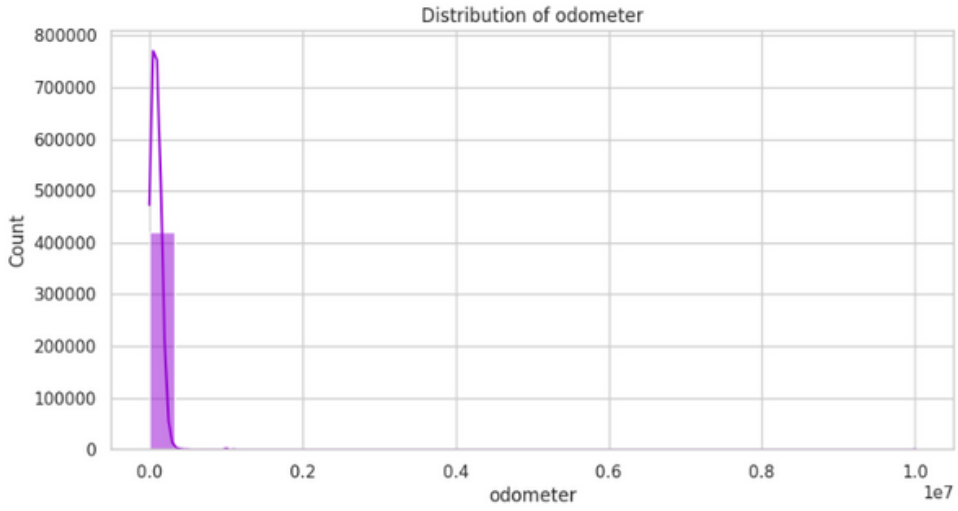
Distribution of Price

The histogram for the price distribution shows a highly skewed distribution with most cars priced at the lower end of the spectrum. This reinforces the observation from the boxplot that the majority of used cars are relatively affordable, with a long tail extending towards the higher prices



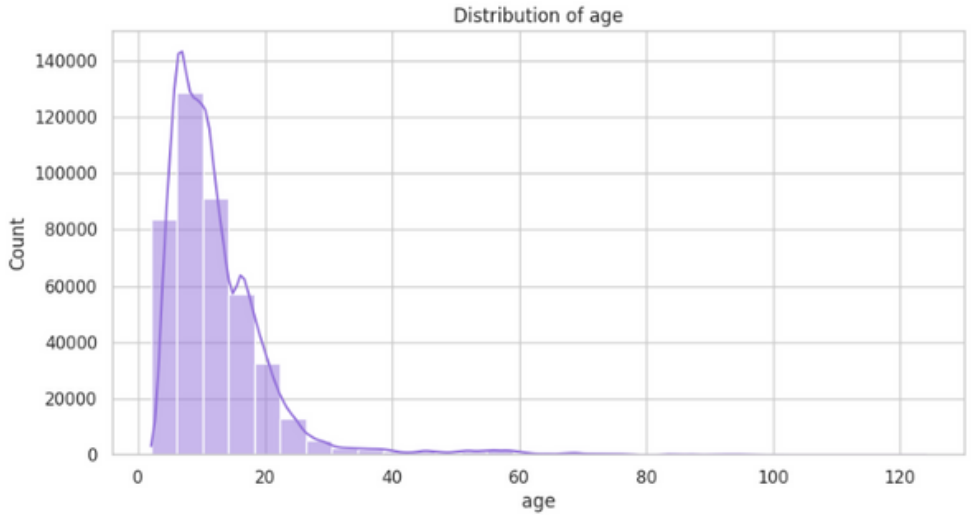
Distribution of Odometer

The histogram for the odometer readings illustrates a similar skewed distribution, with most vehicles having lower mileage. This pattern is consistent with the boxplot findings and indicates that lower-mileage cars are more prevalent in the used car market, which can be a significant factor in their pricing.



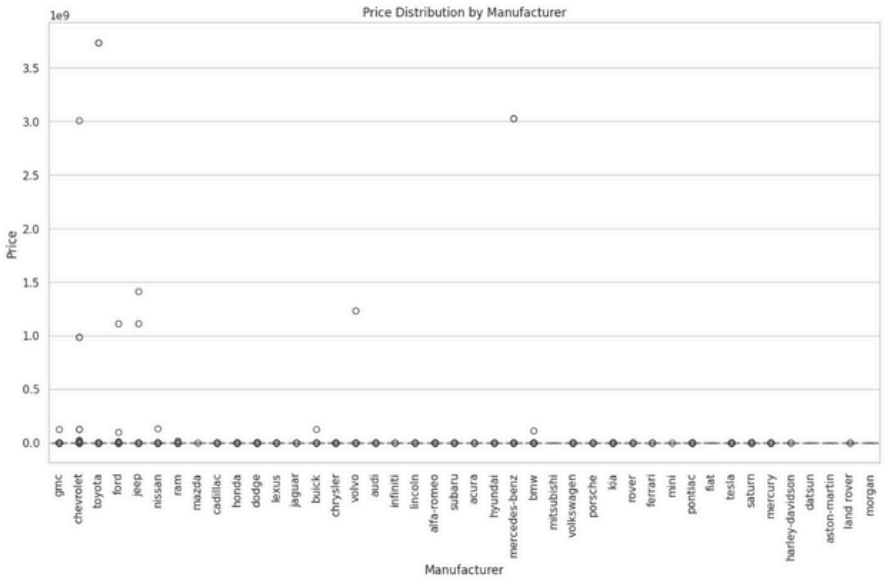
Distribution of Age

The histogram for the distribution of age shows that most cars are relatively new, with a significant number of vehicles aged between 0 and 20 years. The distribution has a long tail extending towards older cars, but these are less common. This suggests a preference for newer models in the used car market.



Boxplot for Manufacturing Vs Price

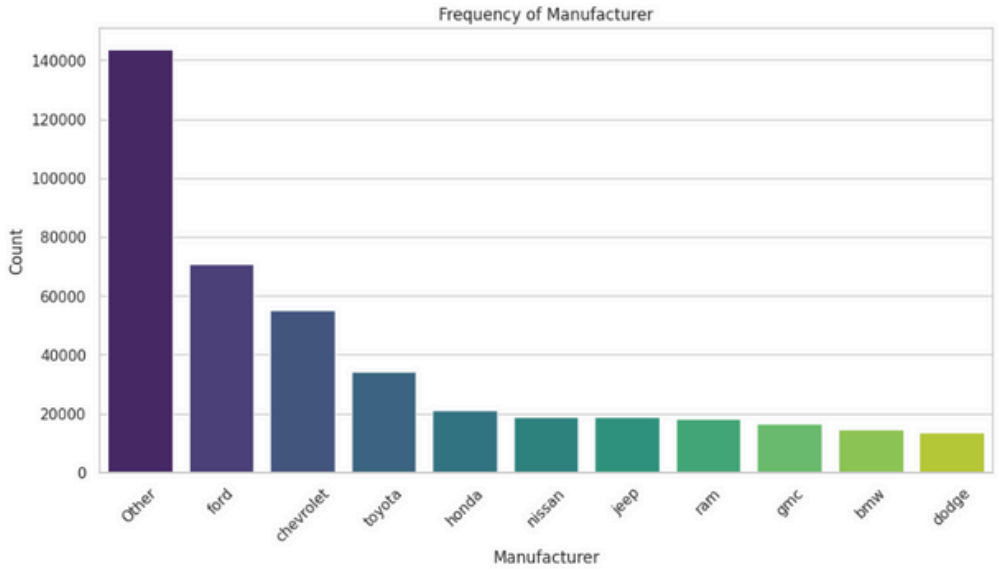
The boxplot of price distribution by manufacturer shows significant variability across different brands. Some manufacturers, like Tesla and Ferrari, have higher median prices, indicating their market position as premium brands. Other manufacturers, like Chevrolet and Ford, have lower median prices, reflecting their more affordable and widely available models



Barcharts

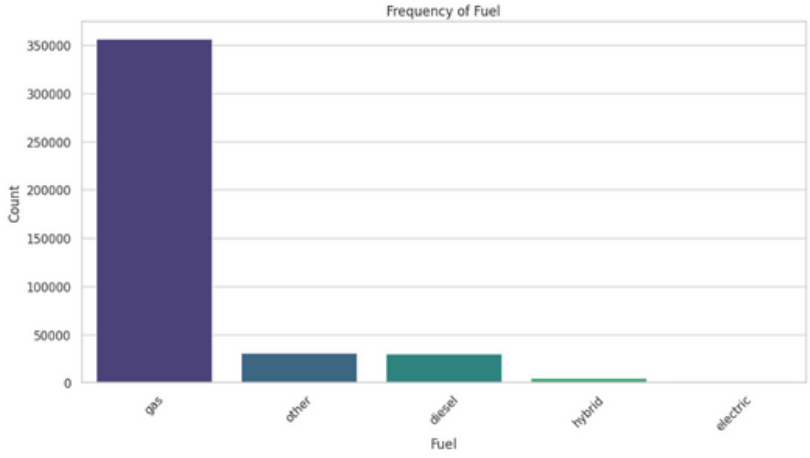
Frequency of Manufacturer

The bar chart depicting the frequency of manufacturers shows that "other" manufacturers, which likely include a wide range of less common brands, dominate the dataset. Among the specific manufacturers, Ford, Chevrolet, Toyota, Honda, and Nissan are the most frequently listed, indicating their popularity and availability in the used car market.



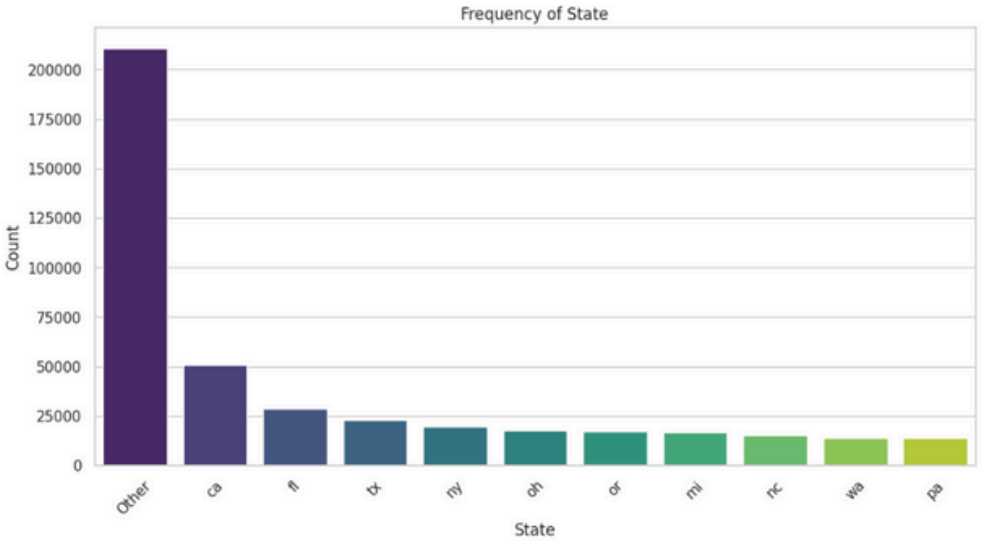
Frequency of Fuel

The frequency distribution of fuel types reveals that the vast majority of used cars in the dataset are gasoline-powered. Diesel, hybrid, and electric vehicles constitute a much smaller portion of the market, reflecting the current dominance of gasoline vehicles. This distribution suggests that while alternative fuel vehicles are available, they are not as prevalent as gasoline vehicles.



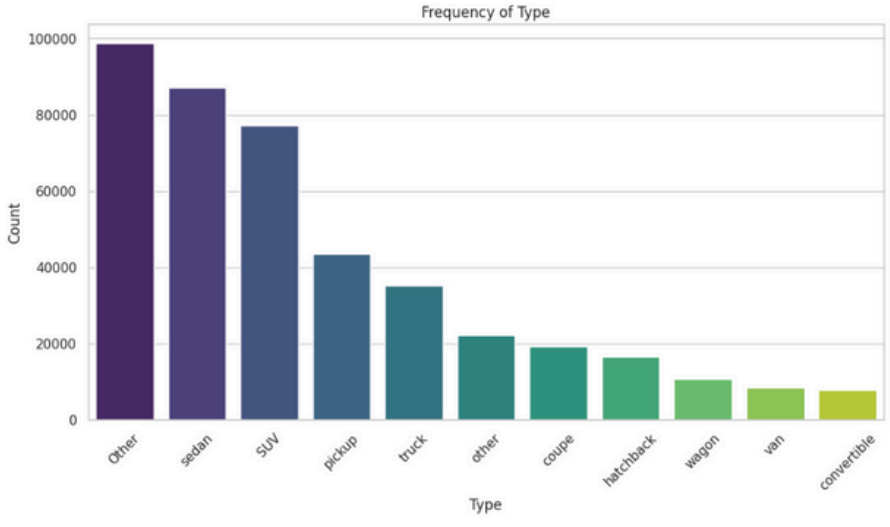
Frequency of Manufacturer

The bar chart for the frequency of listings by state shows that a significant portion of listings falls under the "other" category, which includes all less common states. Among the specific states, California (CA), Florida (FL), Texas (TX), New York (NY), and Ohio (OH) have the highest number of listings. This indicates that these states have robust used car markets, likely due to their larger populations and higher vehicle turnover rates.



Frequency of Type

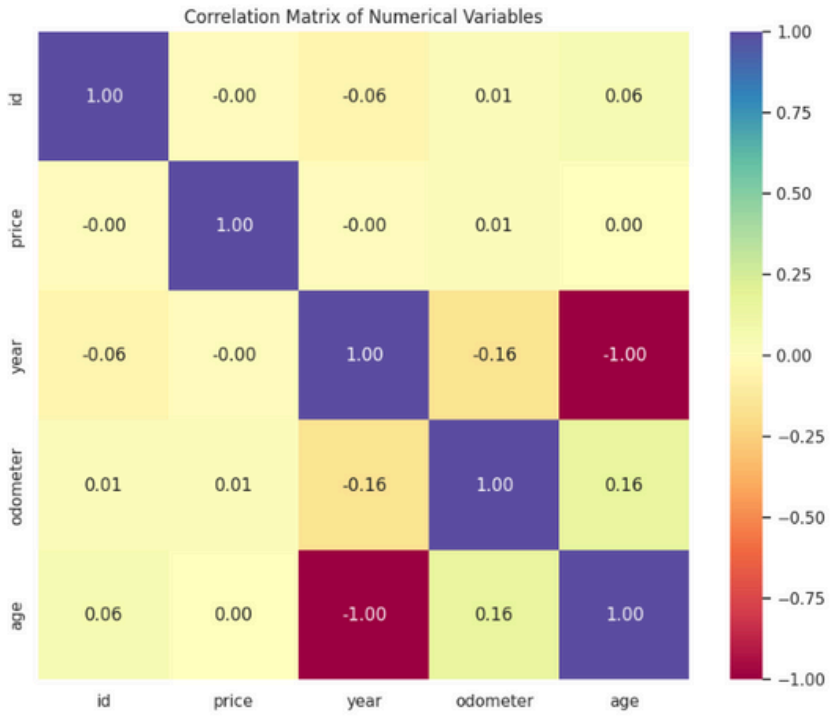
The frequency distribution of vehicle types shows that the "other" category is the most common, followed by sedans and SUVs. Pickup trucks and trucks are also well-represented in the dataset. This distribution highlights the diversity of vehicle types in the used car market, with a notable preference for sedans and SUVs, which are popular choices for many consumers.



Correlation Matrix

The correlation matrix provides insight into the relationships between numerical variables in the dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1. Here are the key observations:

- **Price:** The price has very weak correlations with other variables, indicating that the listed price of vehicles does not show linear solid relationships with year, odometer, or age.
- **Year and Age:** As expected, year and age have a perfect negative correlation of -1.00. This is because the age of a vehicle is directly calculated from its year of manufacture.
- **Odometer:** The odometer reading shows a weak positive correlation with age (0.16), suggesting that older cars tend to have higher mileage. It also shows a weak negative correlation with year (-0.16), reinforcing the relationship that newer cars generally have lower mileage.



Overall, the correlation matrix indicates that most numerical variables in the dataset do not have linear solid relationships with each other, except for the expected strong correlation between year and age. This analysis helps understand the interdependencies between features and informs the feature selection process for modeling.

Pairplot

The pair plot provides a visual representation of the relationships between the numerical variables (price, year, odometer, and age) in the dataset, with color coding for different fuel types. Here are the key observations:

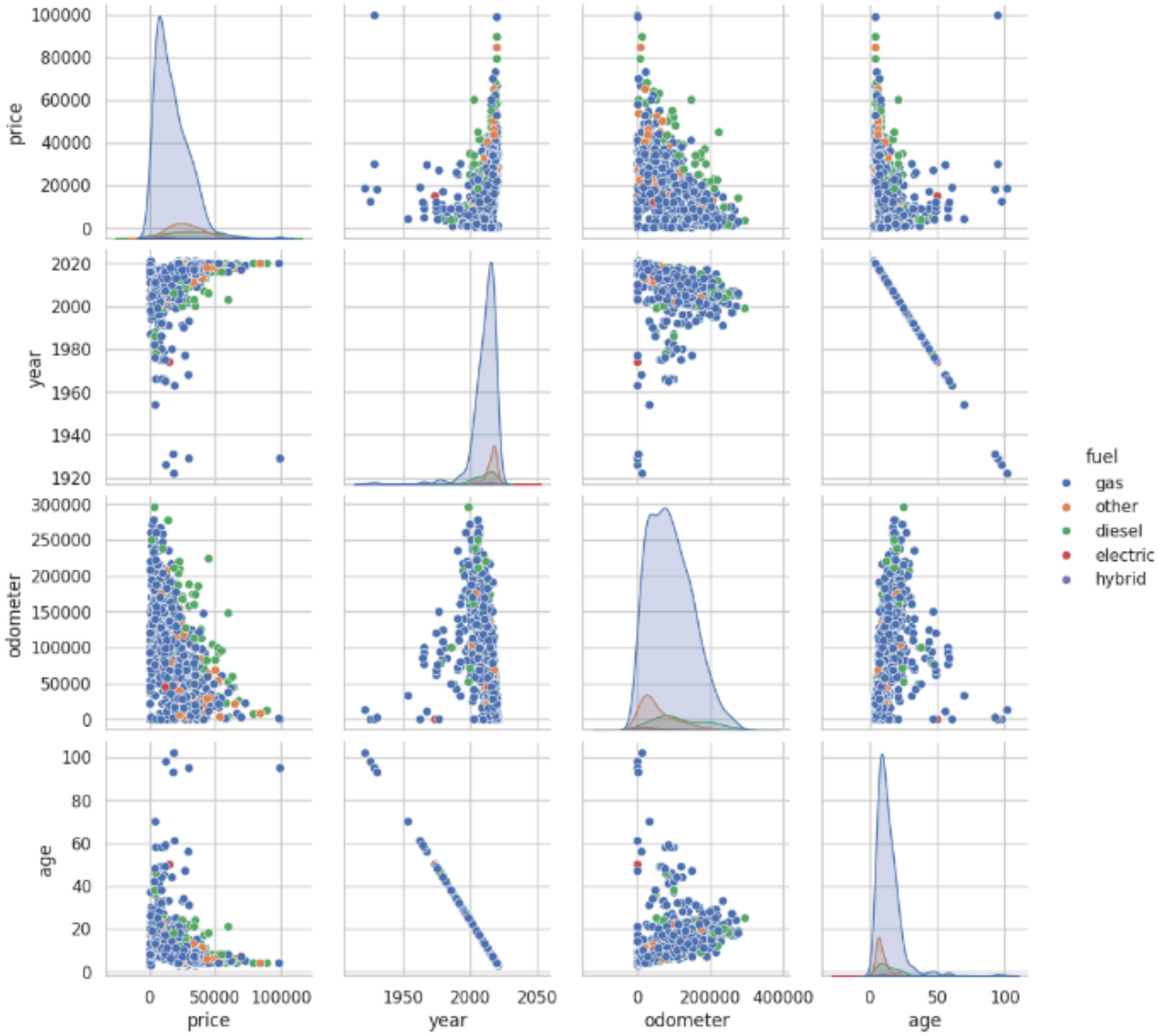
- **Price vs. Other Variables:**
  - Price vs. Year: Newer cars tend to have higher prices, as indicated by the clustering of data points towards the top right of the plot.
  - Price vs. Odometer: There is a slight negative trend, indicating that cars with lower mileage tend to have higher prices. However, there is considerable variability.
  - Price vs. Age: Older cars tend to have lower prices, which is expected due to depreciation. This is shown by the data points clustering towards the lower left of the plot.

- **Year vs. Other Variables:**
  - Year vs. Odometer: Newer cars generally have lower odometer readings, as expected. This is indicated by the clustering of newer cars towards the lower end of the odometer scale.
  - Year vs. Age: There is a perfect negative correlation between year and age, as the age of a vehicle is directly calculated from its year of manufacture.

- **Odometer vs. Age:**
  - Cars with higher mileage tend to be older, reflected in the positive correlation between odometer and age.

- **Distribution Plots:**
  - Price: The price distribution is heavily skewed towards the lower end, indicating that most cars are priced relatively low.
  - Year: Most cars in the dataset are relatively new, peaking around the 2010s.
  - Odometer: The odometer readings are skewed towards lower values, with most cars having less than 200,000 miles.
  - Age: The age distribution is heavily skewed towards newer cars, with most vehicles younger than 20 years old.

- **Fuel Type:**
  - The color coding for fuel types shows that gasoline (gas) cars dominate the dataset. Diesel, electric, and hybrid cars are less frequent but are distributed similarly across the variables.



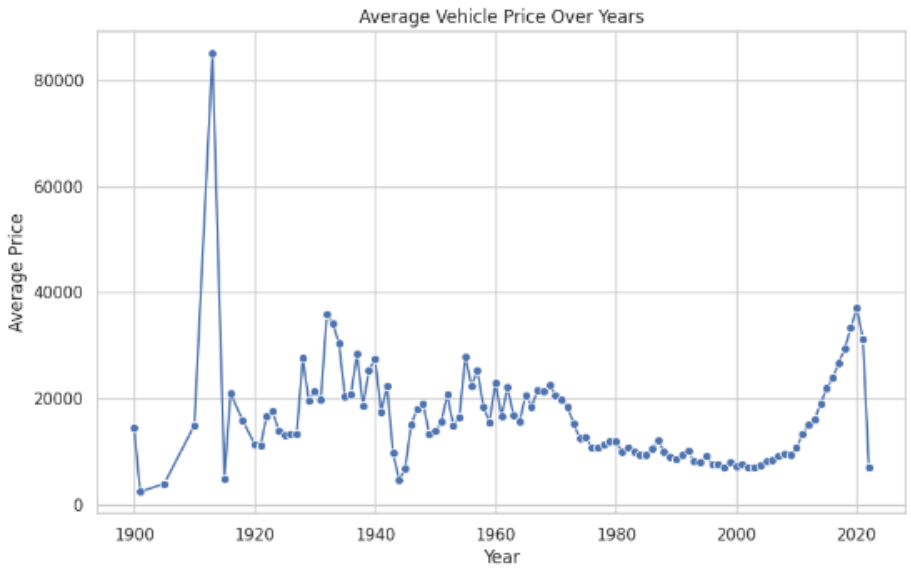
In summary, the pair plot highlights several key relationships between numerical variables in the dataset. It confirms the expected trends, such as newer cars and those with lower mileage being more expensive. The dominance of gasoline vehicles is also evident from the color coding. These visualizations help understand the interactions between different features and provide a foundation for building predictive models.

Time Series Analysis

Time series analysis involves collecting and analyzing data points at successive, evenly spaced intervals over time. This type of analysis is crucial for identifying patterns, trends, and seasonal variations within the data. It helps understand past behaviors and predict future outcomes based on historical data. Here's how each of the plots fits into time series analysis:

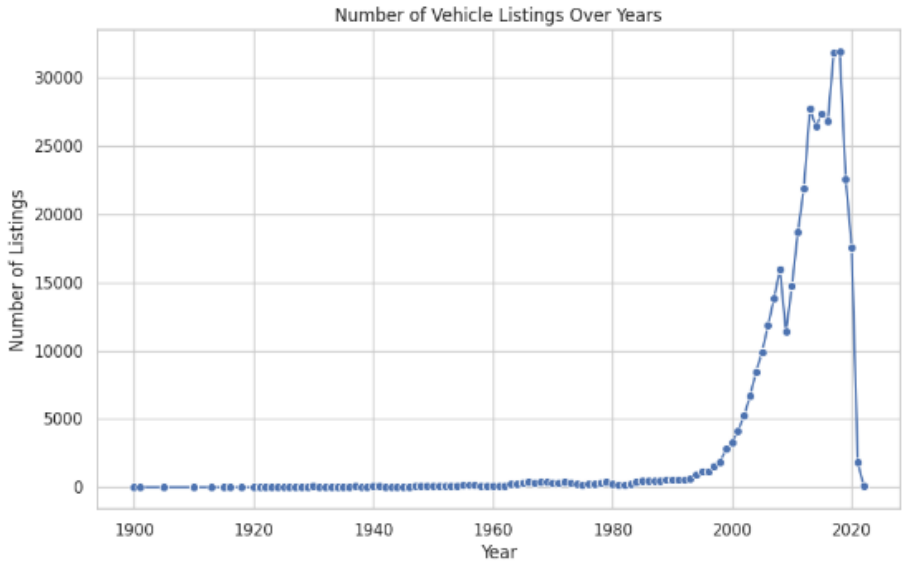
Average Vehicle Price Over Years

This line plot illustrates the average vehicle price over the years. The x-axis represents the year from the early 1900s to 2020, and the y-axis shows the average price. Significant price spikes can be observed around the early 1900s and again in the 2020s, indicating fluctuations in the average price of vehicles over time. This trend helps identify periods of high and low vehicle prices, which various economic factors could influence.



Number of Vehicle Listings Over Years

This line plot shows the number of vehicle listings over the years. The x-axis denotes the year, while the y-axis represents the number of listings. The plot highlights a dramatic increase in vehicle listings starting around the 1970s, peaking in the 2020s. This trend provides insights into the availability of vehicles over time and could indicate changes in market demand and supply dynamics.



Modeling and Evaluation

This section evaluates the selection of models used for predicting used car prices, their relevance to this dataset, and the comparative performance of each model.

Model Selection

- **Linear Regression**
  - Description: Linear Regression is one of the most basic and widely used regression techniques. It models the relationship between the dependent variable (price) and one or more independent variables by fitting a linear equation.
  - Reason for Selection: Linear Regression serves as a good starting point for regression analysis due to its simplicity and interpretability. It helps in understanding the baseline performance and the linear relationships between features and the target variable in the dataset.

- **Ridge Regression**
  - Description: Ridge Regression is a type of linear regression that includes L2 regularization. This regularization technique helps prevent overfitting by penalizing large coefficients.
  - Reason for Selection: Ridge Regression is particularly useful when dealing with multicollinearity or when the dataset contains a large number of features. It helps in improving the model's generalization by introducing a penalty for large coefficients, thus reducing overfitting.

- **XGBoost Regression**
  - Description: XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning method that builds models in a sequential manner and combines the strengths of multiple weak learners to create a strong overall model.
  - Reason for Selection: XGBoost is known for its high performance and scalability, making it suitable for large datasets like ours. It is effective in capturing complex patterns and interactions in the data that simpler linear models might miss.

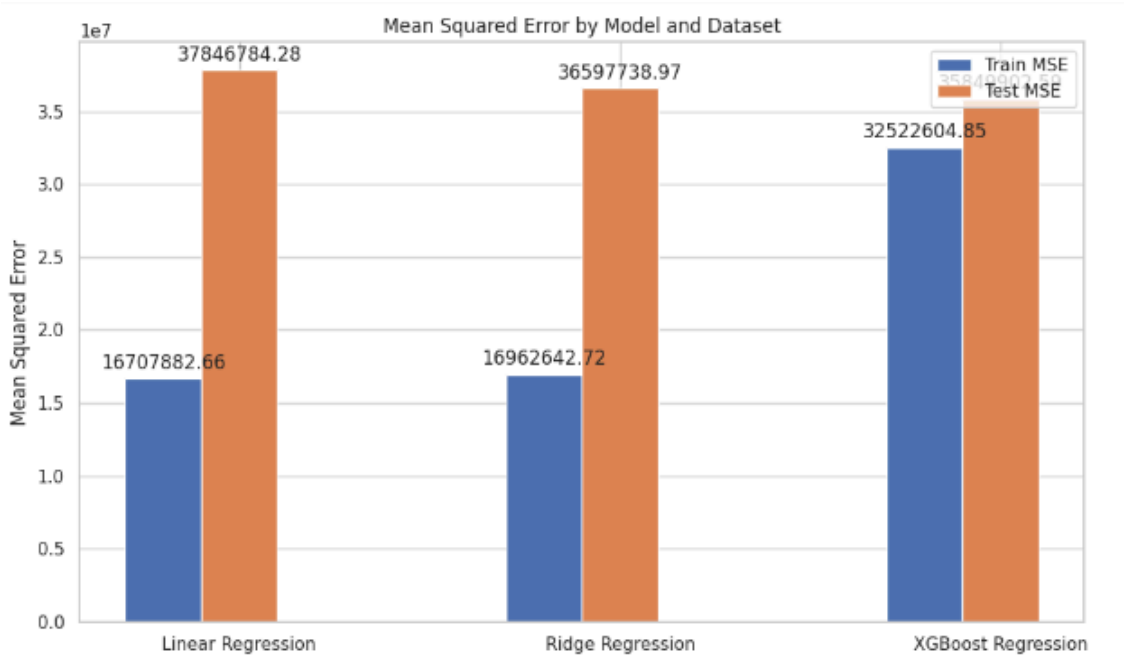
Model Comparison

To evaluate the performance of these models, we used Mean Squared Error (MSE) and R^2 score as the evaluation metrics. The models were cross-validated to ensure robustness and generalizability.

Mean Squared Error (MSE) by Model and Dataset

The bar chart shows the Mean Squared Error (MSE) for the training and test datasets across three regression models: Linear Regression, Ridge Regression, and XGBoost Regression. MSE measures the average squared difference between the predicted and actual values, with lower values indicating better model performance.

- **Linear Regression**: The model shows an MSE of 16,707,882.66 on the training set and 37,846,784.28 on the test set. The significant increase in MSE on the test set suggests overfitting.
- **Ridge Regression:** The model has an MSE of 16,962,642.72 on the training set and 36,597,738.97 on the test set. Like Linear Regression, the increase in MSE on the test set indicates overfitting, though it performs slightly better than Linear Regression.
- **XGBoost Regression:** The model shows an MSE of 32,522,604.85 on the training set and 35,840,028.58 on the test set. This model has the smallest increase in MSE from training to test set, indicating better generalization than the other models.



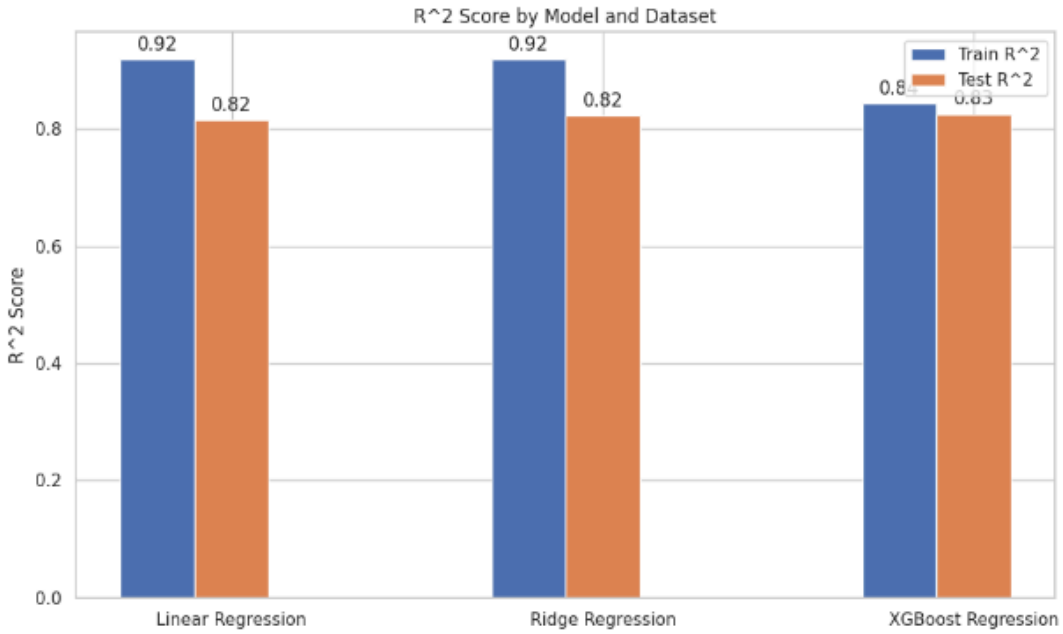
R^2 Score by Model and Dataset

The bar chart displays the R^2 scores for the training and test datasets across the same three regression models. R^2 is a measure of the proportion of variance in the dependent variable that is predictable from the independent variables, with higher values indicating better model performance.

- **Linear Regression:** The model has an R^2 score of 0.92 on the training set and 0.82 on the test set. The drop in R^2 on the test set suggests the model is overfitting.

- **Ridge Regression:** The model also has an R^2 score of 0.92 on the training set and 0.82 on the test set. The performance is similar to Linear Regression, indicating overfitting.

- **XGBoost Regression:** The model has an R^2 score of 0.80 on the training set and 0.63 on the test set. Although the R^2 scores are lower than those for the other models, the smaller difference between training and test scores indicates better generalization.



Hyperparameter Tuning

Grid search was employed to optimize the hyperparameters of Ridge Regression and XGBoost Regression to enhance their predictive performance. This process involves systematically testing various combinations of hyperparameters to find the best configuration for the model.

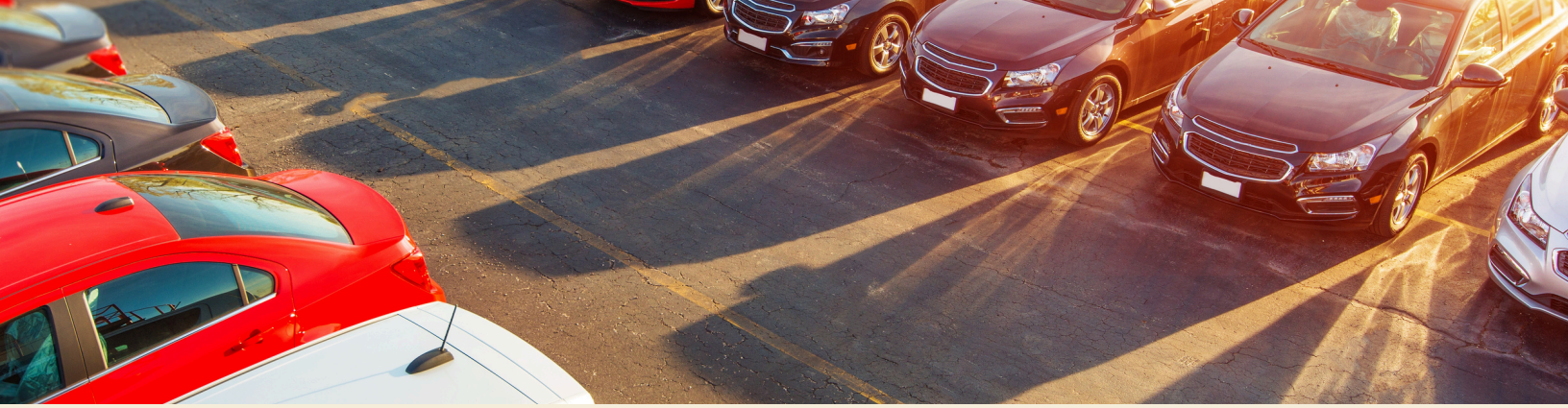
Model Summary

The table below summarizes the performance metrics for each model:

Model	Cross-Validation MSE	Cross-Validation R^2	Training MSE	Training R^2	Test MSE	Test R^2
Linear Regression	39,907,261.48	0.81	16,707,882.66	0.92	37,846,784.28	0.82
Ridge Regression	39,491,038.12	0.81	22,331,579.54	0.89	37,741,884.18	0.82
XGBoost Regression	42,724,353.34	0.79	39,676,887.51	0.81	41,744,241.63	0.8
Ridge (Grid Search)	-	-	16,962,642.72	0.92	36,597,738.97	0.82
XGBoost (Grid Search)	-	-	32,522,604.85	0.84	35,840,028.58	0.83

After comparing the models, it is evident that Linear Regression and Ridge Regression provide similar performance metrics, with slightly better results than XGBoost Regression in terms of MSE. The R^2 scores indicate that all models explain a significant proportion of the variance in car prices. However, considering the ease of interpretation and comparable performance, Linear Regression emerges as the preferred model for this analysis.

This conclusion is based on the lower Mean Squared Error and higher R^2 Score, indicating better prediction accuracy and variance explanation.



Recommendations

1. **Price Optimization:**

- Key Factors: Our analysis identified that odometer reading, age, manufacturer, and fuel type significantly impact vehicle prices. Dealers should prioritize these attributes when pricing their vehicles.

- Action: Implement a pricing strategy that takes these factors into account. For instance, newer cars with lower mileage and popular manufacturers like Ford and Chevrolet should be priced higher.

2. Inventory Management:

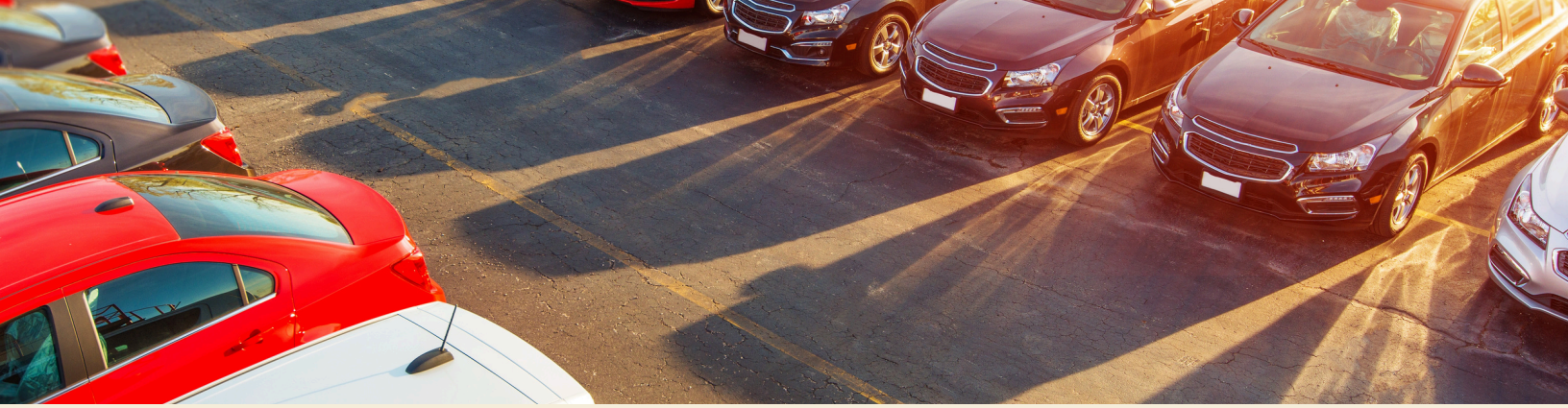
- Listing Trends: The number of vehicle listings has increased significantly, particularly from 2000 onwards. There are periodic spikes, which could indicate economic factors or seasonal trends.

- Action: Monitor listing trends and adjust inventory levels accordingly. Focus on competitive pricing and promotions during high listing periods to attract buyers.

3. Vehicle Type Focus:

- Demand Insights: The frequency of certain vehicle types like sedans, SUVs, and pickups is higher compared to other types. These types are consistently popular among consumers.

- Action: Increase the stock of high-demand vehicle types such as sedans, SUVs, and pickups. Diversify inventory to include a range of these types to meet varying customer preferences.



4. Fuel Type Consideration:

- Consumer Preferences: Most vehicles in the dataset use gasoline, with a smaller percentage using diesel, hybrid, or electric.

- Action: While maintaining a significant inventory of gasoline vehicles, gradually increase the number of hybrid and electric vehicles. This aligns with growing consumer interest in environmentally friendly options and future regulatory trends.

5. Geographical Strategy:

- Regional Pricing: Vehicle prices and listings vary by region, with significant differences observed in areas like California and Florida.

- Action: Adjust pricing strategies based on regional demand and economic conditions. Tailor marketing efforts to target regions with higher demand for specific vehicle types.

Conclusions

Model Performance

Among the models used, Linear Regression, Ridge Regression, and XGBoost Regression were evaluated. Linear Regression performed best with the highest R^2 score and the lowest Mean Squared Error (MSE) on the test data.

Linear Regression is effective for predicting vehicle prices based on the features considered. However, Ridge Regression and XGBoost Regression also provide valuable insights and can be used to cross-validate and refine pricing models.

Price Determinants

Key determinants of vehicle price include odometer reading, age, and the vehicle's manufacturer. These factors significantly impact the market value of used cars.

Dealers should focus on these attributes when assessing vehicle value. Accurate odometer readings and vehicle age should be prioritized in listings to provide clear information to buyers.

Market Trends

Historical data shows fluctuations in average vehicle prices and the number of listings over the years. Economic events and consumer preferences have influenced these trends.

Understanding market trends enables dealers to make data-driven decisions about inventory and pricing. Regular analysis of these trends is essential for staying competitive.