# CRISP-DM Analysis on Used Cars Dataset

## 1. Business Understanding

The goal of this analysis is to understand what factors influence the price of used cars. This will help used car dealerships optimize their inventory and pricing strategies. By identifying key attributes that affect pricing, dealerships can make informed decisions on which cars to stock, how to price them, and what features to highlight in sales efforts. Understanding consumer preferences and the factors that drive car prices is crucial in a highly competitive market. This approach follows the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, which provides a robust framework for analyzing and solving business problems using data mining techniques (Shearer, 2000).

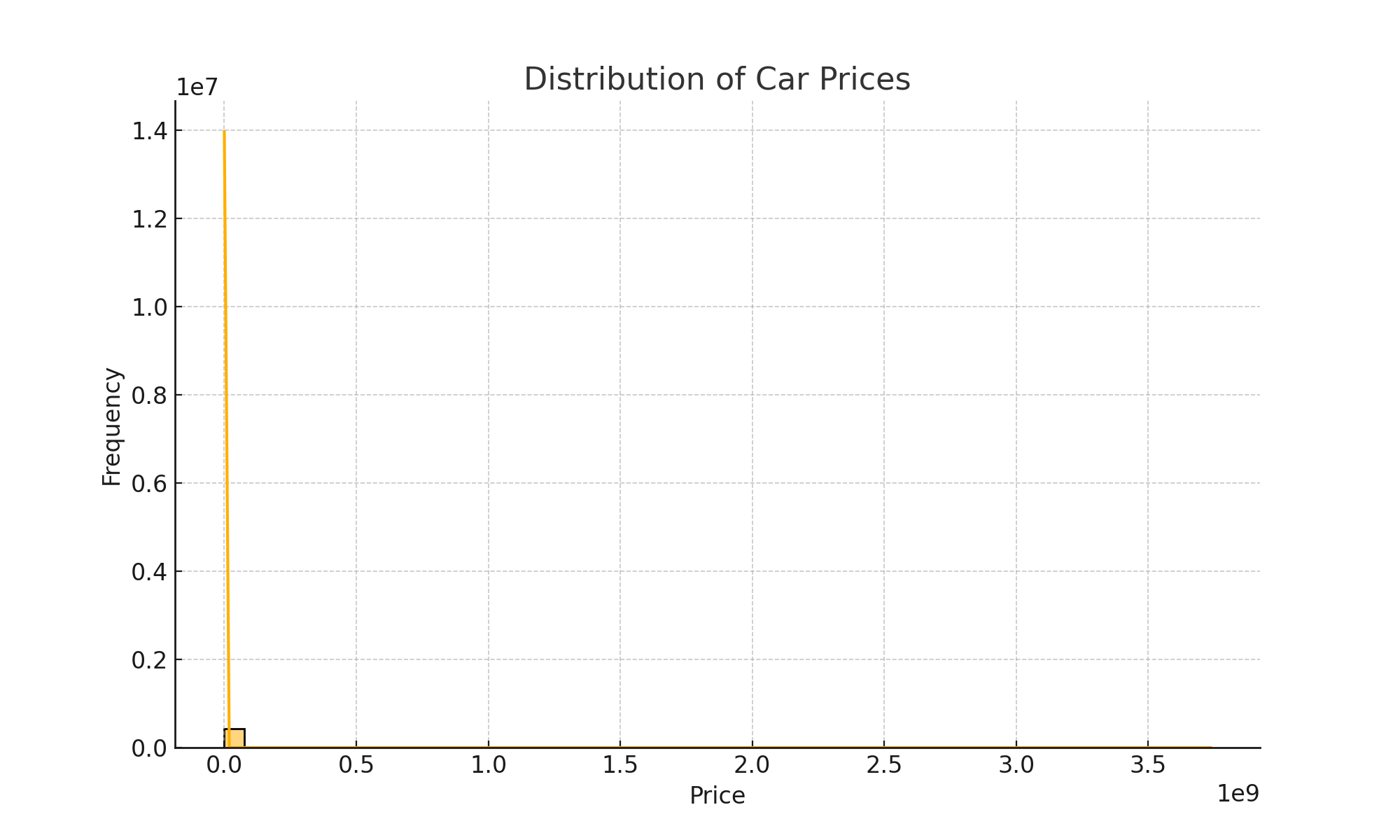
## 2. Data Understanding and Preparation

The dataset contains various attributes of used cars, including price, year, manufacturer, model, condition, cylinders, fuel type, odometer, title status, transmission, drive type, size, type, paint color, and location information like state and region. Significant data cleaning was performed to address missing values and inconsistencies. Missing values in critical columns like 'year' and 'odometer' were filled with median values, while categorical variables were filled with 'Unknown'. Columns with excessive missing data, such as 'VIN' and 'size', were removed to streamline the analysis. The cleaned dataset was then prepared for exploratory data analysis and modeling, adhering to the data preparation steps outlined in the CRISP-DM process (Shearer, 2000).

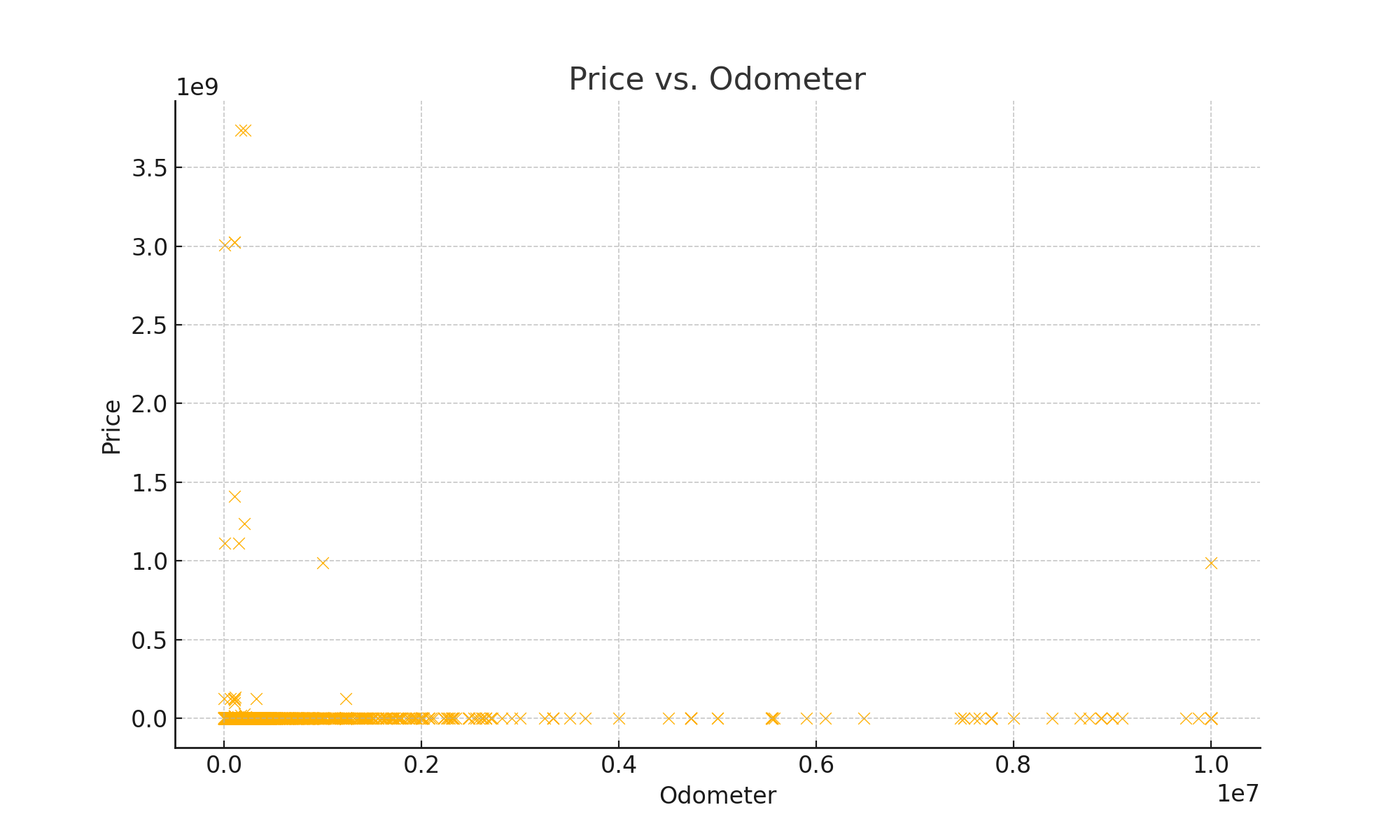
## 3. Exploratory Data Analysis (EDA)

EDA was conducted to explore the relationships between different features and car prices. Visualizations helped in identifying trends and patterns. Key findings from the analysis include:

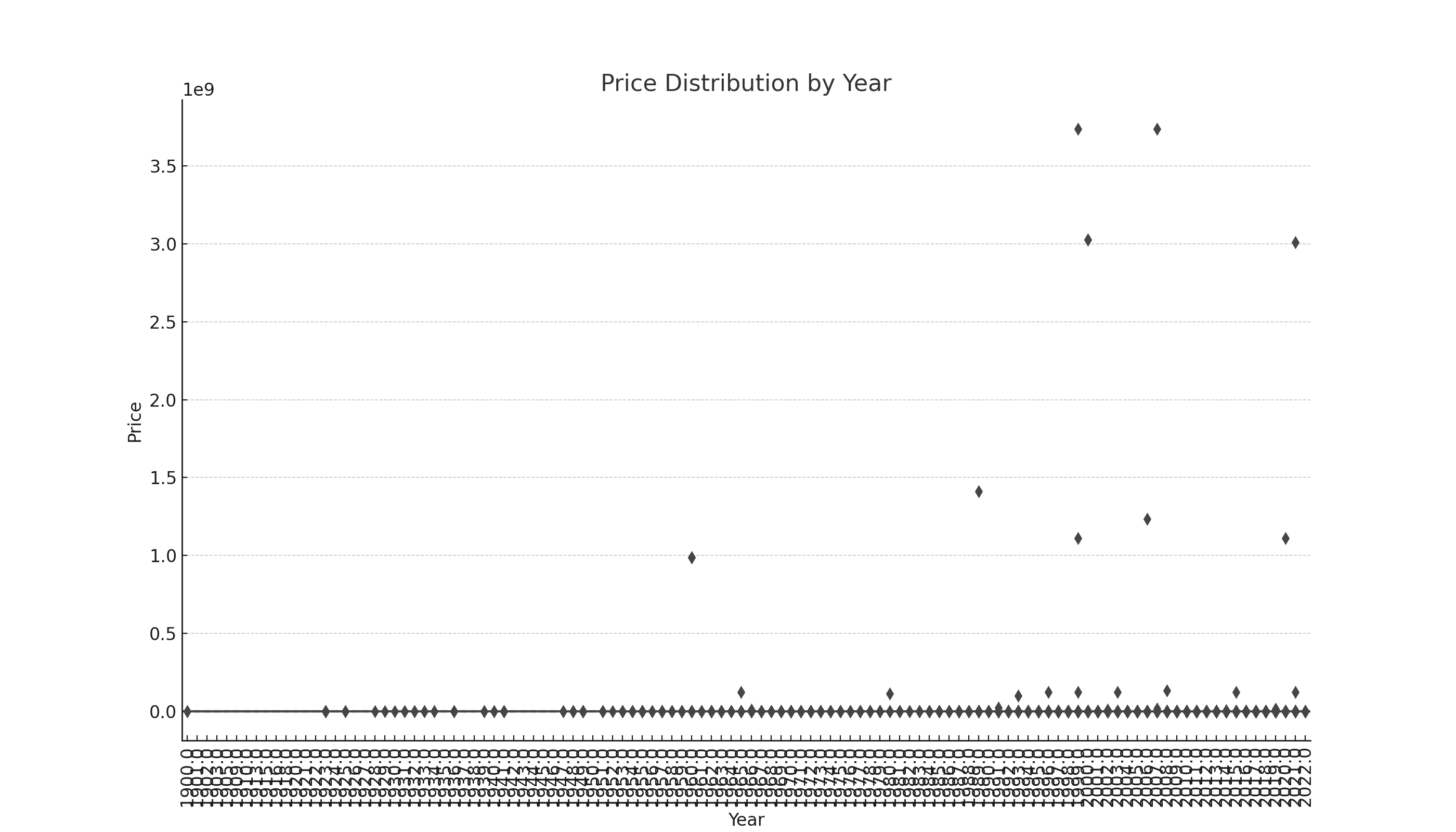
* Price Distribution: The distribution of car prices is right-skewed, with most cars priced below $20,000. There are several outliers, with some cars priced extremely high, likely luxury or rare models. This skewness suggests a market dominated by lower-priced vehicles, with a small segment of high-value cars that could skew average pricing metrics (Tukey, 1977).



* - Mileage Impact: A negative correlation exists between odometer readings and car prices, indicating that cars with higher mileage tend to be less expensive. This relationship is consistent with the expectation that more usage results in lower vehicle value. Understanding this trend is vital for dealerships to price cars according to their mileage (Sheather, 2009).



* - Yearly Trends: Newer cars generally have higher prices, with older cars showing greater price variability, reflecting varying conditions and values. The box plot of car prices by year shows that as the cars get older, the spread in pricing widens, which may be due to factors like condition, rarity, or specific model demand (Sheather, 2009).



A graph of a box plot

Description automatically generated with medium confidence

A diagram of red and blue squares

Description automatically generated

A graph of a graph of cars

Description automatically generated with medium confidence

l

A graph of cars with numbers and colors

Description automatically generated with medium confidence

## 4. Modeling Attempts and Results

Multiple modeling attempts were made, including Linear Regression and Random Forest. The Linear Regression model showed poor performance, failing to capture complex patterns due to its limitations in handling non-linear relationships and interactions among features. A more sophisticated Random Forest model demonstrated potential for capturing complex interactions between variables but required further optimization for practical deployment. Random Forest, a popular ensemble method, is effective in handling large datasets with complex interactions (Breiman, 2001). The model's performance highlighted that car year, odometer, and manufacturer were significant predictors of price.

## 5. Conclusions and Recommendations

Based on the analysis, it is recommended that dealerships focus on the following factors to optimize their pricing strategies:

* - Car Age and Mileage: Prioritize inventory with lower mileage and newer models, as these factors significantly impact car prices. Cars with fewer miles and newer manufacturing years command higher prices, aligning with consumer expectations of reliability and performance (Breiman, 2001).
* - Key Features: Highlight cars with favorable conditions, popular manufacturers, and appealing features that are shown to increase value. Dealerships should ensure these attributes are prominently displayed in listings to attract buyers.
* - Further Modeling: Continued refinement of predictive models, including feature engineering and the use of advanced algorithms such as Gradient Boosting Machines or Neural Networks, will help improve price predictions and enhance inventory decision-making (Friedman, 2001). Incorporating additional data, such as market demand trends or specific feature availability, could further refine price optimization models.

Overall, this analysis provides a comprehensive view of the used car market and offers actionable insights that dealerships can apply to their pricing strategies. By understanding and leveraging key factors that affect vehicle value, dealerships can better align their inventory with market expectations and enhance profitability.

## References

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of data warehousing, 5(4), 13-22.

Sheather, S. (2009). A modern approach to regression with R. Springer Science & Business Media.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.