
# Final Project Milestone 2: Draft Report

**Shreeya Ambre, Purva Prakash Kekan, Harsh Anadkat, Vinh Pham**

**College of Professional Studies, Northeastern University**

**ALY 6040: Data Mining**

**Professor Justin Grosz**

**June 17, 2024**

# Analysis of Automobile Costs

## Introduction

The automotive industry is highly dynamic, with car prices fluctuating based on a myriad of factors. For dealerships, auction houses, and buyers, accurately predicting the selling price of a car is crucial for making informed decisions. This project aims to develop a model that can predict the selling price of cars based on historical sales data, encompassing various attributes such as the car's make, model, year, condition, and odometer reading. By leveraging data analysis and machine learning techniques, we seek to provide insights that can aid stakeholders in the automotive market.

The business value of this project lies in its ability to provide a robust tool for accurately forecasting car prices, ultimately enhancing decision-making across the automotive market. By utilizing historical data and advanced analytical techniques, we can deepen our understanding of the factors driving car prices, enabling stakeholders to navigate the complexities of the automotive market with greater confidence and precision. This predictive capability not only aids in optimizing pricing strategies but also ensures fair market valuations, leading to increased trust and efficiency in car sales and purchases. For dealerships, it helps in optimizing pricing strategies to remain competitive while maximizing profits. Auction houses can set realistic reserve prices and estimate expected auction outcomes, enhancing their operational efficiency. Buyers, on the other hand, can benefit by making informed purchasing decisions, ensuring they pay a fair market value for vehicles.

## Data Cleaning & Exploration

To answer the business question of how to accurately predict the selling prices of cars based on their attributes, we will undertake a systematic approach. First, we collected and prepared the data, which include historical sales data with various attributes such as make, model, year, trim, body type, transmission type, VIN, state, condition, odometer reading, color, interior, seller, Market Monitor Retail (MMR) value, selling price, and sale date. Data cleaning involved handling missing

values using appropriate imputation methods like mean, median, and mode for different types of data, removing duplicates, and ensuring correct data types for each attribute.

Our dataset has a ‘sale date’ column that has time and date information in it, but it's not immediately useful. After formatting this column more manageably, we divided it into two new columns: one for the time and one for the date. By separating the data, we could more accurately examine trends over time and determine when car sales were at their peak.

Data integrity is essential. We look for any missing values and deal with them methodically.The missing values found in ‘sellingprice’, ‘odometer’, and ‘mmr’ variables had a very low percentage compared to the overall data. Removing these missing values wouldn’t harm or reduce the data quality.

We removed any rows where the 'sellingprice' column had missing values. The selling price is a crucial target variable for our analysis. Any missing values in this field would compromise the integrity and accuracy of our predictive models. Ensuring that this field is complete is essential for building a reliable model.

We removed any rows where the 'odometer' column had missing values. The odometer reading is vital for assessing the vehicle's usage, which significantly impacts its valuation. Accurate odometer readings are necessary for proper analysis and to avoid skewing the model with incomplete data.

We removed any rows where the 'mmr' column had missing values. The MMR provides a benchmark value for vehicle prices based on comprehensive market data. Missing values in this field would reduce the reliability of our analysis, as MMR is a critical reference point for determining vehicle value.

We used a histogram to illustrate the distribution of the vehicles' conditions in order to comprehend them. In order to preserve the integrity of the distribution as a whole, missing values

in the condition column were substituted with the median value. This process was essential for precise car appraisal.

The vehicle identification number must be unique. Because duplicate entries could distort our research, we eliminated them to make sure every automobile was represented uniquely.

The analysis may be hampered by blank or missing entries in category columns like make. The most frequent numbers within each category were used to fill up these gaps, guaranteeing consistency and dependability in our data.

Similarly for the model variable we used the approach of using mode to fill up the missing values to ensure the consistency in our data

We also made sure that the trim variable was filled correctly. We utilized the most popular trim level for each make and model combination for determining trim.

We used a similar strategy for the body, that is, utilizing the most popular body level for each make and model combination for determining body type making sure that each car has a certain body type.

The variable transmission had very less percentage of missing values. We substituted "Unknown" for any missing values in the column. The transmission type is an important feature that influences vehicle preference and pricing. Replacing missing values with "Unknown" ensures that we retain all records while explicitly acknowledging the absence of this data.

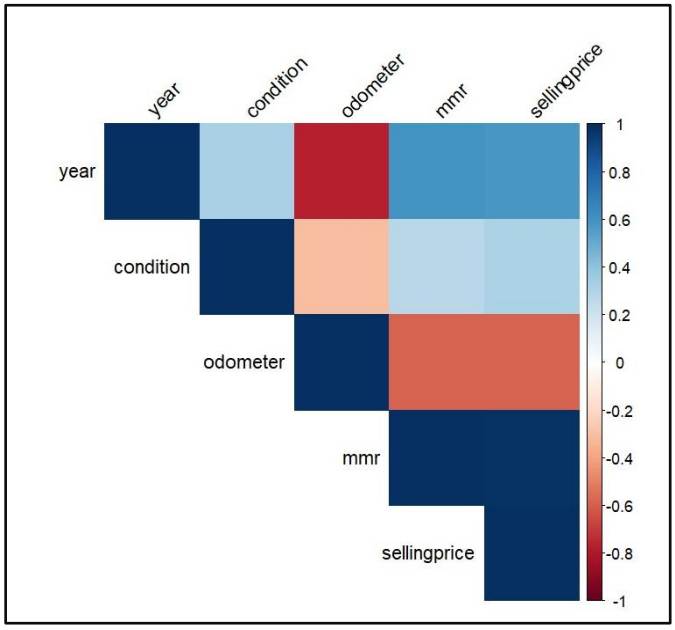
We removed any rows where the 'color' column had missing values. The percentage of missing values in the 'color' column was very low. To ensure data integrity and quality, we opted to remove these rows entirely, as their absence does not significantly impact the dataset size.

We removed any rows where the 'interior' column had missing values. Similar to the 'color' column, the 'interior' column had a very low percentage of missing values. Removing these rows helps maintain the overall integrity of the dataset without substantially reducing its size.

Now that our data is clear and comprehensive, we explore the data further to find patterns. For each of the numerical variables in our dataset, we compute summary statistics. This provides us with a high-level summary of the data by displaying the values of each variable—the mean, standard deviation, minimum, maximum, and median.

To see the distribution of important numerical data like year, condition, odometer, mmr, and selling price, we plotted histograms. We can better comprehend the distribution and core patterns of these variables thanks to these visuals.

Finally, in order to determine the links between numerical variables, we computed and displayed a correlation matrix. Important relationships in our auto sales dataset are revealed by the correlation matrix, which is crucial for making wise business decisions. The year of manufacture, condition, mileage on the odometer, MMR value, and selling price are important factors. Some noteworthy observations include that newer cars are more valuable because they have greater selling prices and lesser mileage. Better-maintained cars also typically have lesser mileage and command higher costs, underscoring the need for upkeep. There appears to be an emphasis on selling low-mileage autos as higher mileage is correlated with lower selling prices. MMR's dependability in determining competitive prices is supported by the positive association it has with selling price.



By establishing higher prices for more recent, well-kept vehicles with little mileage and competitive prices for vehicles with high mileage, we utilize these insights to improve our pricing methods. To optimize income, we concentrate on promoting and selling well-maintained, low-mileage cars. We can maintain our profitability and competitiveness by matching our prices to MMR values. All things considered, these insights aid in pricing optimization, inventory control, and market positioning improvement, all of which raise profitability and satisfy customers. This aids in our comprehension of the variables that could affect the cars' selling price.

## Metrics of Interest

**1. Selling Price:** The selling price is the dependent variable, or the target variable, that we aim to predict in this analysis. It represents the final amount at which a car is sold, and it is influenced by a multitude of factors such as the car's condition, make, model, age, and mileage, among others. Understanding the factors that influence the selling price is crucial for developing accurate predictive models. This insight allows stakeholders to adjust their strategies accordingly. Analyzing the selling price also helps in identifying market trends and understanding seasonal fluctuations, which can be critical for planning inventory and sales strategies.

**2. Odometer Reading:** The odometer reading is a key indicator of how much the car has been used, reflecting its total mileage. Higher mileage generally implies more wear and tear, which typically reduces the car's value. Including the odometer reading in the analysis helps in understanding how usage affects the pricing of the vehicle. This metric is particularly important for both buyers and sellers as it directly influences the perceived value and reliability of the vehicle.

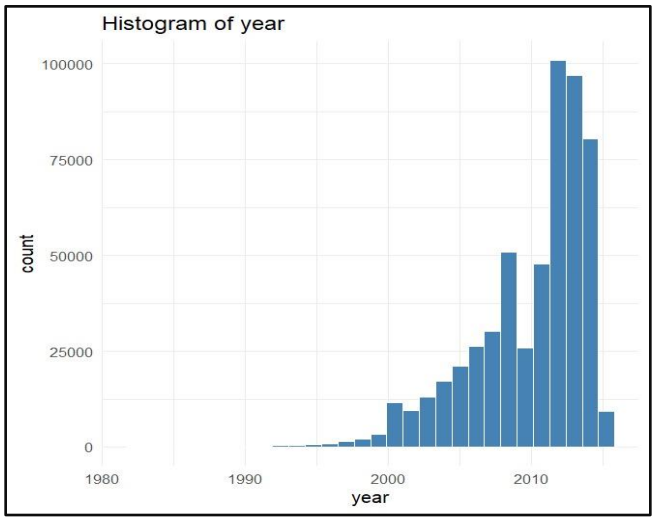
**3. Car Condition:** The condition of a car is another critical metric that significantly impacts the selling price. Car condition is typically assessed on a scale from poor to excellent and encompasses various factors such as the car's physical appearance, mechanical state, and overall performance. A car in excellent condition is likely to fetch a higher price compared to one in fair or poor condition. Including car condition in the analysis helps capture the qualitative aspects that directly influence a car's market value.

These are all interesting metrics and were chosen because they have a direct impact on the value of a car. The EDA indicated strong relationships between these metrics and the selling price, making them crucial for accurate predictions because they provide further in-depth insight into these variables and their correlations with each other.

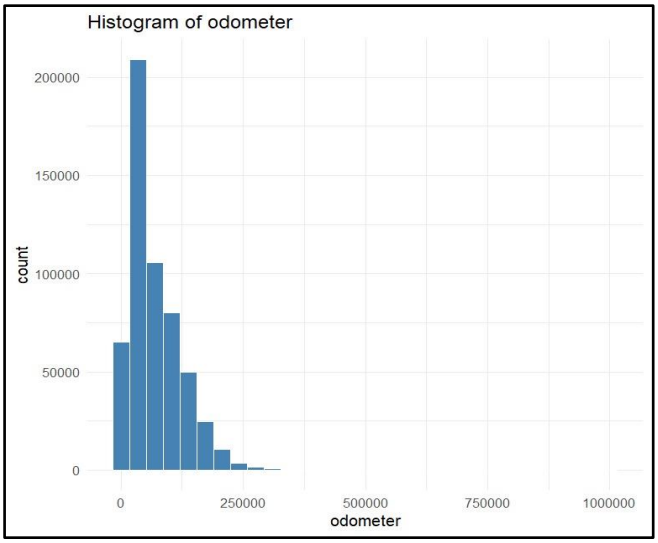
## Data Analysis and Visualization

1. **Distribution by Year**

   This histogram illustrates the distribution of car counts for three decades spanning the years 1990 to 2010. Throughout the years there has been an increasing trend, however, the count increased exponentially after 2010, reaching it peak at 10000 and then falling dramatically towards the end of the decade.

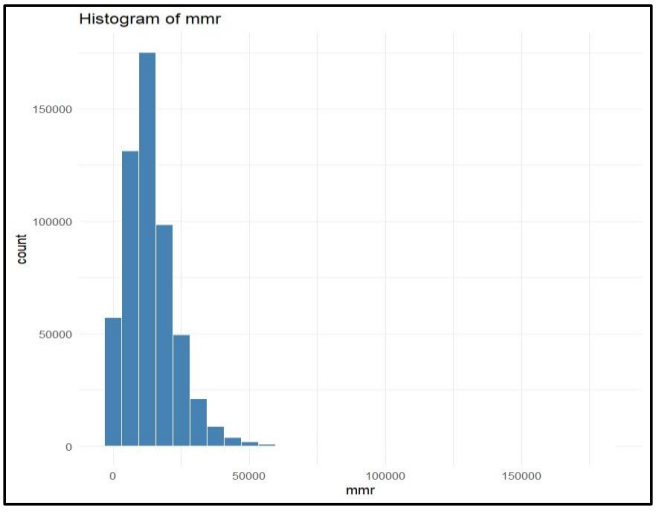


2. **Distribution by Odometer Reading**




This bar plot represents the odometer reading by car counts. As seen, more than 200000 cars account for a reading that is between 100000 to 150000. This seems to be the most frequent odometer reading for most cars. The count drops as the reading increases further.

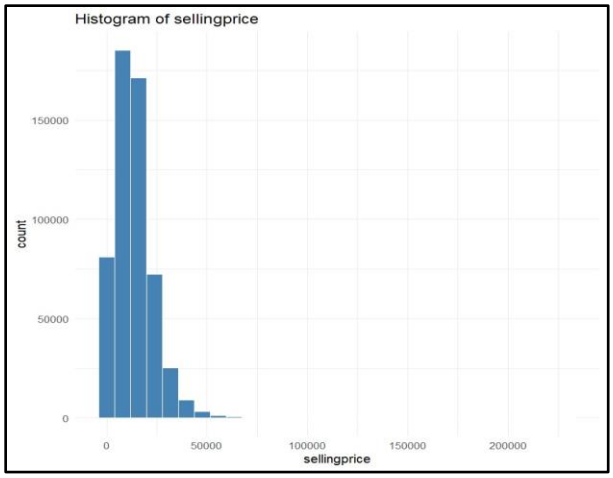
3. **Distribution by MMR**




This histogram illustrates the MMR values for cars and as seen, most MMR values were recorded between 0 to 50000. The most frequently recorded value is 3000 whereas 55000 has the least frequency. The graph observes a declining trend with an increase in MMR values.

4. **Distribution by Selling Price**

   This graph displays the selling price for cars by count. It can be observed that most of the cars were sold for a price range between 1000 to 3000. This suggests that most of the cars were sold for a better price whereas only a small proportion was sold at a much lesser price.



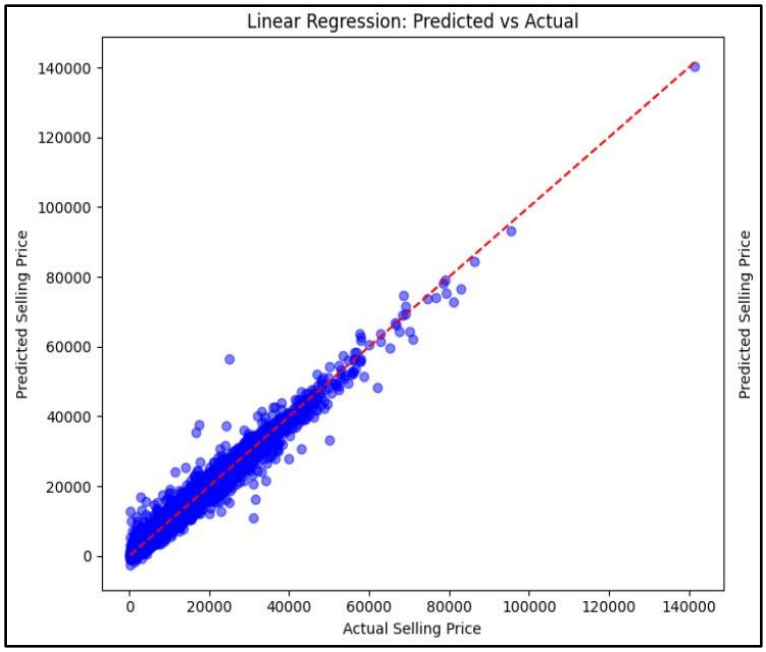

## Predictive Modeling

1. **Linear Regression Model**

The linear regression model developed in this analysis aims to predict the selling price of cars based on a comprehensive dataset of vehicle attributes. The scatter plot visualizes the model's predicted selling prices against the actual selling prices. The close alignment of the blue data points along the red dashed line, which represents a perfect prediction scenario, indicates a high degree of accuracy in our model's predictions. The calculated metrics further substantiate this: a Mean Squared Error (MSE) of 2,693,294.37 and an R-squared ($R^2$) value of 0.9705. These metrics imply that the model explains approximately 97.05% of the variance in the selling price, demonstrating

its robustness. The MSE and $R^2$ values suggest that it is highly effective in capturing the relationship between the selected variables and the selling price.

The selection of variables for the linear regression model was driven by their anticipated influence on the selling price, based on domain knowledge and initial exploratory data analysis. Key variables such as odometer reading, car condition, make and model, and trim and features were chosen because they encapsulate essential aspects of a car's value from both a buyer's and seller's perspective. The odometer reading reflects the vehicle's usage, where higher mileage typically depreciates the car's value. Car condition provides an assessment of the vehicle's physical and mechanical state, directly impacting its market price. The make and model were included because different brands and models have varying market perceptions and resale values. Lastly, specific trim levels and additional features can significantly enhance or detract from a vehicle's appeal and value. These variables were selected to capture the multifaceted nature of car pricing and to ensure that the model provides a comprehensive and accurate prediction of the selling price.

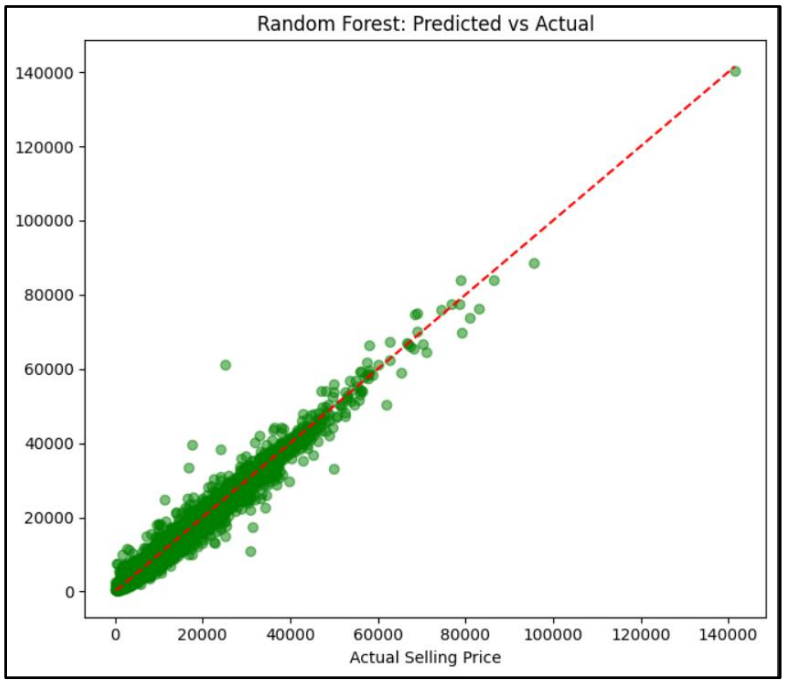



The linear regression model offers valuable insights into factors affecting selling prices, enabling stakeholders to make data-driven decisions. For instance, positive coefficients for certain models and trims suggest that these features are highly valued in the market, whereas negative coefficients indicate features that may detract from the car's value. Understanding these relationships allows stakeholders to make data-driven decisions. For dealerships, this model can optimize pricing strategies by accurately predicting the value of cars based on their attributes. This ensures competitive pricing, which can attract more customers and increase sales volume while maintaining profitability. For buyers, the model aids in making informed purchasing decisions. By

understanding the factors that influence a car's price, buyers can better assess whether a vehicle is priced fairly, helping them negotiate better deals and avoid overpaying.

## 2. Random Forest Model

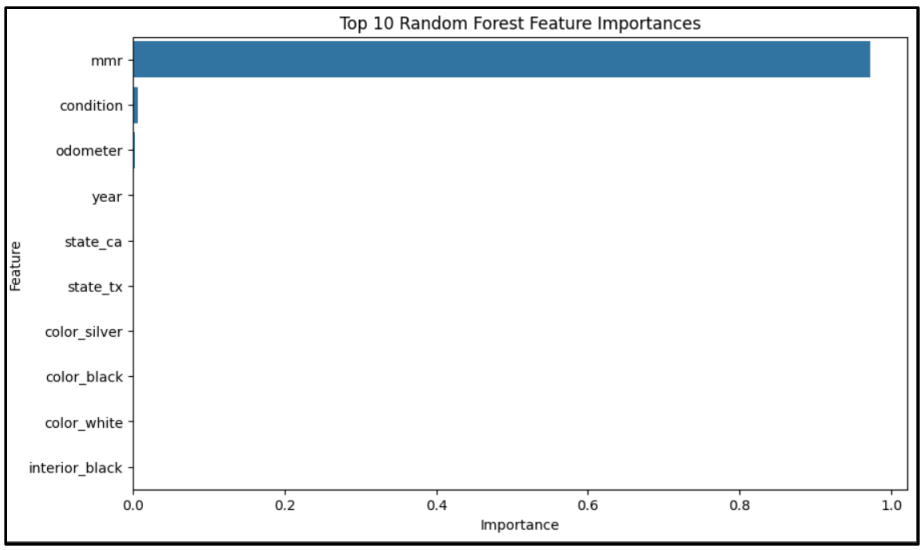
The Random Forest model, used for comparison, also shows promising results with an MSE of 2,436,633.75 and an $R^2$ of 0.9733. Although the Random Forest model slightly outperforms the linear regression model, the latter's interpretability and simplicity make it a valuable tool for understanding the influence of each variable on the selling price.




The figure displays a scatter plot that contrasts the actual car selling prices with the selling prices predicted by a Random Forest model. By showing how closely the model's predictions match the ideal situation, which is shown by the red dashed line, this plot helps assess the model's accuracy. The majority of the green data points cluster around this line, suggesting that selling price forecasts made by the Random Forest model are typically correct. A data point that approaches the red line indicates that the anticipated price and the actual selling price are almost the same, demonstrating the model's ability to successfully identify and understand underlying patterns in the data. For organizations, this alignment is critical since it indicates accurate forecasting capabilities, which are necessary for making smart price decisions and inventory management.

The output of the Random Forest model identifies the most important characteristics for car selling price prediction. The top predictor, "mmr" (Manheim Market Report), is highlighted in the bar

chart, underscoring its crucial function in pricing. MMR, which is derived from auction data, is essential for precise price estimation because it fully reflects market value components. "Condition" and "odometer" show up as important predictors after MMR. Odometer readings measure mileage and affect pricing because of wear, but condition represents the status of the car and directly affects its value. Though less obviously, other characteristics such as the year, the state (Texas, California), the color (silver, black, white), and the inside color (black) also hint at local tastes and aesthetic effects on pricing.




The capacity of the Random Forest model to draw attention to the significance of features offers insightful information to a range of automotive market participants. Dealerships and auction houses can set competitive and reasonable pricing for their goods by realizing that MMR is a crucial factor in determining selling prices. During sales presentations, highlighting the mileage and condition might aid in persuading prospective customers that the pricing is reasonable. Understanding that mileage and condition have a big impact on pricing helps individual buyers determine if an automobile is priced appropriately. To optimize resale value, sellers, on the other hand, can concentrate on keeping the vehicle in good condition and controlling the odometer readings.

## Model Comparison

For the Linear Regression model, we observe a Mean Squared Error (MSE) of 2,693,294.37 and an R-squared ($R^2$) value of 0.9705. The model's coefficients reveal how each feature impacts the

selling price. Notably, features such as ‘model interstate’, ‘trim rvb’, and ‘make airstream’ show high positive coefficients, indicating a significant positive influence on the car's selling price. On the other hand, features like ‘trim Unlimited Freedom Edition’, ‘trim Turbo S’, and ‘trims Drive30i’ have negative coefficients, suggesting they reduce the selling price. This model provides a clear interpretation of each feature's impact, which is useful for understanding the direct relationship between individual features and the selling price.

The Random Forest model demonstrates a superior performance with an MSE of 2,436,633.75 and an $R^2$ value of 0.9733. This model assigns importance scores to features, with ‘mmr’ emerging as the most critical predictor, having an importance score of 0.9722. Other significant features include ‘condition’ and ‘odometer’, though their importance scores are considerably lower. The Random Forest model, being a non-linear ensemble method, captures complex interactions between features, making it particularly effective for identifying key predictors without assuming a specific relationship form between the features and the target variable.

Comparing the two models, the Random Forest model exhibits lower MSE and a higher $R^2$, indicating superior prediction accuracy. While Linear Regression offers simplicity and direct interpretability, it assumes linear relationships and may not capture complex interactions, potentially leading to less accurate predictions. In contrast, the Random Forest model, with its ability to capture non-linear relationships, provides more accurate predictions and identifies the most critical factors affecting selling prices.

From a business perspective, the Random Forest model proves more advantageous due to its higher accuracy and capacity to capture complex relationships. The model's insight that mmr is the most crucial predictor allows businesses to focus on this feature when pricing cars, ensuring better pricing strategies and ultimately enhancing profitability and competitive advantage in the automotive market. However, the Linear Regression model's clarity and simplicity can be useful where interpretability is paramount, offering direct coefficients that facilitate specific business recommendations.

## Business Focus & Recommendations

Several strategic advice that businesses in the automotive sector can take into consideration can be derived from the investigation conducted on the dynamics of automobile pricing:

**Optimize Pricing Strategies:** To determine competitive prices based on variables like car condition, odometer readings, and market trends represented in MMR values, use predictive models like Linear Regression and Random Forest. With this strategy, cars are priced to maximize profitability and draw in buyers.

**Improve Inventory Management:** To reduce holding costs and maximize turnover rates, put data-driven inventory management techniques into effect. Dealerships and auction houses are able to prioritize high-value autos and modify inventory levels by precisely projecting selling prices.

**Customize Marketing Initiatives**: Utilize understandings of local inclinations and fashion trends found via data mining. To better connect with target client segments and increase sales conversion rates and market positioning, highlight cars with trendy features and colors.

**Encourage Openness and Trust:** Explain pricing choices in an open manner, stressing the role that impartial data plays in determining car values. Customers' trust is strengthened by this transparency, which raises consumer happiness and loyalty.

It is recommended to continuously develop models by updating and improving them in response to market input and continuous data collection. To keep up with changes in the industry and improvements in technology, include new factors and modify existing approaches.

## Conclusion

To sum up, this thorough investigation has provided profound insights into the variables influencing automobile selling prices. The key factors affecting automobile pricing were effectively revealed through the selected metrics and analytical approaches. Our examination of the vehicle sales data highlighted the critical impact of car condition and mileage on pricing. These findings underscore the importance of these factors in predicting car prices accurately.

We created robust predictive models that could significantly enhance decision-making in auto sales and purchases. These models will empower stakeholders across the automotive industry, from dealerships and auction houses to individual buyers and sellers, to make more informed and strategic decisions. The insights gained not only deepen our understanding of car pricing dynamics but also pave the way for further research into more complex and dynamic automobile pricing models, ultimately benefiting all parties involved in the automotive sector.