Vehicle Sales Data Analysis Project Report

Author: Tsz Fong Chan

Date: 24/7/2023

Abstract

This report presents a comprehensive analysis of vehicle sales data, focusing on data preprocessing, feature engineering, clustering analysis, and various market and seller evaluations. The primary objective of this project is to clean and preprocess the raw dataset, engineer meaningful features, and conduct in-depth analyses to uncover insights regarding car prices, market trends, and seller performance.

The methodology involves a structured approach to data cleaning, including handling missing values, outlier detection, data type corrections, and duplicate removal. Feature engineering steps include creating new features such as car age and mileage per year, followed by clustering analysis using the K-means algorithm. The report also leverages TensorFlow to enhance model performance through GPU acceleration for feature selection and importance ranking.

Key findings reveal significant variations in selling prices across different car makes and models, highlighting the impact of brand and model on vehicle pricing. Additionally, the analysis underscores the influence of car condition on selling price, demonstrating a clear correlation between better condition and higher prices. Market analysis identifies trends across different makes and models, while seller analysis compares average selling prices and volumes, revealing disparities in performance among sellers.

In conclusion, the project successfully cleans and preprocesses the dataset, engineers impactful features, and conducts comprehensive analyses that provide valuable insights into vehicle pricing dynamics and seller performance. These findings offer practical implications for stakeholders in the automotive industry, enabling data-driven decision-making and strategic planning.

Table of Contents

Introduction

The automotive industry is a vital sector in the global economy, significantly influencing employment, transportation, and technology development. In recent years, the analysis of vehicle sales data has gained prominence due to its potential to uncover market trends, customer preferences, and pricing dynamics. As car prices fluctuate based on various factors such as make, model, condition, and seller, understanding these dynamics is crucial for stakeholders, including manufacturers, dealers, and consumers. This project focuses on a comprehensive analysis of vehicle sales data to derive actionable insights that can inform strategic decision-making in the automotive sector.

Despite the availability of extensive vehicle sales data, there is often a lack of structured analysis to identify key factors affecting car prices and market trends. This gap hampers the ability of businesses to optimize pricing strategies, manage inventory effectively, and enhance customer satisfaction. This report aims to address these issues by providing a detailed analysis of car sales data, emphasizing the impact of various attributes on selling prices and seller performance.

The primary objectives of this analysis are:

- To clean and preprocess the raw vehicle sales dataset to ensure data quality and reliability.
- To engineer new features that could enhance the predictive power of models, such as car age and mileage per year.
- To perform clustering analysis to identify distinct groups within the data.
- To compare selling prices across different makes and models, and analyse the impact of car condition on selling prices.
- To investigate the performance of different sellers by comparing average selling prices and sales volumes.

This study will focus on vehicle sales data from a comprehensive dataset, including various makes, models, and sellers. The analysis will cover data cleaning, feature engineering, clustering, market analysis, and seller performance evaluation. The study will exclude data beyond the scope of the available dataset, such as external economic factors or consumer sentiment data, and will primarily rely on the provided vehicle attributes.

The findings from this study are expected to provide valuable insights into the factors driving car prices and market trends. By understanding these factors,

stakeholders in the automotive industry can make data-driven decisions to optimize pricing strategies, improve inventory management, and enhance customer satisfaction. The study's insights will be particularly beneficial for car manufacturers, dealers, market analysts, and policy makers who aim to understand and respond to market dynamics effectively.

This introduction sets the stage for a thorough exploration of vehicle sales data, providing the necessary context, objectives, scope, and structure to guide the reader through the subsequent sections of the report.

Literature Review

The analysis of vehicle sales data has become an increasingly important area of research, providing insights into market dynamics, consumer preferences, and pricing strategies. This literature review explores existing studies and methodologies related to vehicle sales data analysis, focusing on data cleaning, feature engineering, market analysis, and seller performance evaluation.

Data cleaning and preprocessing are critical steps in data analysis, ensuring the quality and reliability of the data used. Numerous studies highlight the importance of handling missing values, outliers, and data type inconsistencies to enhance the accuracy of analytical models [4].

For instance, Zhang et al. (2019) emphasized the significance of data preprocessing in predictive modelling, demonstrating how handling missing data and outliers can improve model performance [6]. Similarly, Kotsiantis et al. (2006) provided a comprehensive review of data preprocessing techniques, including data normalization, standardization, and transformation, which are essential for preparing raw data for analysis [4].

Feature engineering involves creating new features from existing data to improve the performance of machine learning models. In the context of vehicle sales data, features such as car age and mileage per year are commonly engineered to enhance predictive accuracy.

He et al. (2014) explored various feature engineering techniques in automotive data analysis, highlighting the creation of features related to vehicle specifications and historical sales trends [3]. Their study demonstrated that feature engineering significantly improves the predictive power of models used for price prediction and sales forecasting [3].

Market analysis in the automotive industry involves examining the factors that influence vehicle prices and identifying trends across different segments. Several studies have focused on analysing the impact of vehicle make, model, and condition on selling prices.

A study by Chu et al. (2018) analysed a large dataset of used car sales to understand price determinants [1]. They found that factors such as brand reputation, vehicle condition, and mileage significantly influenced selling prices

[1]. Their research also highlighted the role of market trends and economic conditions in shaping pricing strategies.

Evaluating seller performance is crucial for understanding market dynamics and identifying successful sales strategies. Research in this area often focuses on comparing selling prices, volumes, and customer satisfaction across different sellers.

Mookerjee and Mannino (2010) investigated the impact of seller reputation on eBay car auctions, finding that sellers with higher ratings achieved better sales outcomes [5]. Their study underscored the importance of seller reputation in online marketplaces and its influence on buyer trust and willingness to pay [5].

Various methodologies have been employed in vehicle sales data analysis, ranging from statistical techniques to machine learning algorithms. Regression analysis, clustering, and decision trees are commonly used methods for understanding price determinants and predicting sales outcomes.

For example, Fan and Li (2015) applied regression models to predict used car prices, incorporating features such as vehicle age, mileage, and brand [2]. Their study demonstrated the effectiveness of regression analysis in capturing the relationships between vehicle attributes and selling prices [2].

While existing research provides valuable insights into vehicle sales data analysis, several gaps remain. Many studies focus on specific geographic regions or market segments, limiting the generalizability of their findings. Additionally, there is a need for more comprehensive analyses that integrate multiple factors influencing vehicle prices and seller performance.

This literature review highlights the importance of data cleaning, feature engineering, market analysis, and seller performance evaluation in vehicle sales data analysis. Existing studies have demonstrated the effectiveness of various methodologies in understanding market dynamics and predicting sales outcomes. However, there is a need for more comprehensive and generalizable research to fully capture the complexities of the automotive market. This project aims to address these gaps by providing a detailed analysis of a diverse vehicle sales dataset, offering actionable insights for stakeholders in the automotive industry.

Methodology

Data Collection

The dataset for this project was sourced from Kaggle[7], specifically from the "Vehicle Sales and Market Trends Dataset" by Syed Anwar Afridi . This dataset provides a comprehensive collection of information related to vehicle sales transactions. It includes detailed information about the vehicles, such as make, model, trim, body type, transmission type, and condition rating. Additionally, it encompasses transaction details such as selling prices and sale dates, as well as market trend indicators like Manheim Market Report (MMR) values. This dataset is presented in a tabular format, typically as a CSV file, with rows representing individual vehicle sales transactions and columns representing different attributes associated with each transaction.

Data Preparation

In preparing the dataset for analysis, a comprehensive series of steps were undertaken to ensure data integrity and suitability for subsequent analyses. These steps were methodically executed to address common data quality issues, such as missing values, outliers, incorrect data types, and duplicate records. The preparation process included the following key activities:

The first step in the data preparation process involved obtaining initial statistics about the dataset. This included determining the number of attributes (columns), the number of data entries (rows), and the names of the attributes. By gathering this information, I have gained an initial understanding of the dataset's structure and content, which is essential for informed decision-making in subsequent cleaning and preprocessing steps.

Missing data can significantly impact the quality of analysis, leading to biased or inaccurate results. To address this, a strategy for handling missing values was implemented. Depending on the nature of the data, different methods were employed:

- Dropping Missing Values: If columns had substantial missing data, the entire rows were removed to maintain data integrity.
- Filling Numeric Data: For numeric columns, missing values were filled using the mean or median of the respective columns to preserve the overall distribution of the data.

- Filling Non-Numeric Data: For non-numeric columns, missing values were replaced using the mode (most frequent value) to ensure the imputation aligns with the data's typical patterns.

This approach ensured that the dataset remained robust and representative of the underlying trends, without the noise introduced by missing values.

Outliers can distort statistical analyses and model predictions. To mitigate this, outliers were identified and removed using a threshold-based method. The standard deviation from the mean was calculated for each numeric column, and data points that deviated significantly (beyond three standard deviations) were flagged as outliers. These outliers were then removed to ensure that the dataset reflected typical values and trends.

Data type inconsistencies can lead to processing errors and inaccurate analyses. Therefore, each column's data type was corrected to ensure appropriate handling of numeric and categorical data. This step involved converting columns to numeric types where applicable, while retaining the integrity of non-numeric columns.

Duplicate records can skew analysis results by over-representing certain data points. To address this, duplicate entries were identified and removed. This step ensured that each data entry was unique, providing a more accurate and reliable basis for analysis.

Finally, after all cleaning and preprocessing steps were completed, the cleaned dataset was saved to a new CSV file. This cleaned dataset forms the foundation for subsequent analyses, free from the common data quality issues that can compromise the validity and reliability of the findings.

The data preparation process was critical in transforming the raw dataset into a clean and reliable form suitable for analysis. By addressing missing values, outliers, data type inconsistencies, and duplicate records, I have ensured that the dataset was robust and reflective of true market trends. These steps were crucial for enabling accurate and meaningful analyses in the later stages of the project.

Analytical Methods

The Analytical Methods section is vital as it outlines the techniques and methodologies employed to process and interpret the dataset. This section provides clarity on how the data was analysed to derive meaningful insights,

ensuring transparency and reproducibility of the study. Below, I have detailed the analytical techniques used, their purposes, and the rationale behind their selection.

Overview of Techniques:

- *Exploratory Data Analysis (EDA)*
- *Market Analysis*
- *Seller Analysis*
- *Clustering Analysis*
- *Time Series Analysis*
- *Feature Engineering and Selection*

Exploratory Data Analysis (EDA) was employed to understand the dataset's structure, detect anomalies, and identify patterns, which are essential steps in forming hypotheses and guiding further analysis. This technique includes descriptive statistics and visualizations to summarize the main characteristics of the data. EDA was implemented using Python libraries such as pandas, matplotlib, and seaborn, utilizing techniques like histograms, pair plots, and correlation matrices. Default settings of these libraries were used for visualizations to ensure clarity and simplicity.

The purpose of Market Analysis was to compare the selling prices across different makes and models and to analyse the impact of car conditions on selling prices. Grouping and aggregation operations were performed to calculate average selling prices, followed by visualizations to compare these averages. This analysis utilized pandas for data manipulation and seaborn for creating bar plots and box plots. Data was grouped by 'make' and 'model,' and mean values were used for comparison to identify significant differences.

Seller Analysis aimed to investigate the performance of different sellers by comparing average selling prices and sales volumes. Grouping and aggregation techniques were used to calculate average prices and sales volumes. Similar to market analysis, pandas was used for data manipulation and seaborn for visualizations. The data was grouped by 'seller' to calculate average selling prices and total sales volumes, allowing for an evaluation of seller performance.

Clustering Analysis was utilized to identify patterns and segment the data into distinct groups based on multiple features. The K-Means clustering algorithm, which partitions data into k distinct clusters based on feature similarity, was employed. Implementation involved the use of sklearn’s KMeans class for clustering and PCA for visualization. The number of clusters

(k) was optimized using the elbow method and silhouette scores, and StandardScaler was used for normalization to ensure consistent scaling.

Time Series Analysis focused on analysing the temporal patterns in car sales data to identify trends, seasonal patterns, and other time-dependent behaviours. Seasonal decomposition of time series (STL) was used to separate the data into trend, seasonal, and residual components. This analysis was implemented using pandas for resampling the data and statsmodels for seasonal decomposition. Different resampling periods, such as daily, weekly, monthly, and yearly, were set to explore various temporal granularities.

Feature Engineering and Selection involved enhancing the dataset by creating new features that capture additional information and selecting the most relevant features for analysis. New features like car age and mileage per year were calculated, and scaling techniques were applied to ensure uniformity. Implementation included creating new features using pandas and scaling them using sklearn's StandardScaler, with numerical features standardized for consistent scaling.

The justification for the choice of methods is based on each method's suitability for handling specific aspects of the data and addressing the research questions effectively. For instance, EDA provided a foundational understanding of the data, clustering helped in identifying patterns, and time series analysis revealed temporal trends. Alternative methods, such as hierarchical clustering or advanced machine learning models, were considered but deemed less appropriate due to their complexity or lack of significant added value for the primary objectives.

By detailing these analytical methods, this section ensures a clear understanding of the approaches used to analyse the vehicle sales data, providing transparency and facilitating reproducibility of the results.

Data Analysis

Descriptive Statistics:

Descriptive statistics serve as the foundational method in data analysis, providing an essential summary of the dataset's main characteristics through quantitative measures. The primary objective of using descriptive statistics is to gain a comprehensive understanding of the dataset, detect anomalies, and form a basis for further, more complex analyses. By calculating central tendencies and variability, this method offers insights into the general behaviour and distribution of the data.

The central tendencies—mean, median, and mode—are pivotal in summarizing the data. The mean, representing the average value, indicates that the average year of the vehicles in the dataset is approximately 2010, with the condition averaging around 30.77, the odometer reading at about 66,701 miles, the Manheim Market Report (MMR) value at 13,837, and the selling price at 13,690. These values provide a snapshot of the typical vehicle in the dataset. The median values further refine this understanding, indicating that half of the vehicles have values above and below the median year of 2012, condition of 35, odometer reading of 51,085 miles, MMR of 12,300, and selling price of 12,200. The mode, though not explicitly shown, would highlight the most frequently occurring values in each category, offering insights into the most common characteristics.

The dataset consists of various numerical attributes, including year, condition, odometer, mmr (Manheim Market Report), and sellingprice. By calculating the mean, standard deviation, minimum, maximum, and specific percentiles, I have gained an understanding of the distribution and range of these variables.

For instance, the average vehicle year is approximately 2010, with a standard deviation of 3.82, indicating a relatively narrow spread around the mean. The condition of the vehicles, measured on a scale, averages at 30.77 with a standard deviation of 13.29, highlighting a moderate variance in vehicle conditions. The wide range in the odometer readings, from a minimum of 1 to a maximum of 999,999, emphasizes the diversity in vehicle usage history, with an average mileage of 66,701.73.

The mmr values, representing market valuation, range from as low as 25 to as high as 182,000, with an average of 13,837.06. This substantial range reflects the variation in vehicle values within the dataset. Similarly, the sellingprice exhibits significant variability, with a minimum of 1, a maximum of 230,000, and an average of 13,690.51. This indicates a broad spectrum of transaction values, potentially influenced by factors such as vehicle make, model, condition, and market conditions.

Variability in the data is assessed through measures such as range, variance, and standard deviation. The range shows significant spread, with the year ranging from 1990 to 2015, condition from 1 to 49, odometer readings from 1 to nearly 1,000,000 miles, MMR from 25 to 182,000, and selling prices from 1 to 230,000. This wide range indicates considerable diversity in the dataset. The standard deviation, particularly high in odometer readings (51,939) and selling prices (9,613), suggests substantial variation within these attributes. This level of variability is critical for understanding the spread and dispersion of values, which could indicate different segments or clusters within the data.

Descriptive statistics also play a crucial role in identifying outliers, which are extreme values that deviate significantly from other observations. In this dataset, the presence of vehicles with odometer readings as high as 999,999 and selling prices up to 230,000 can be considered outliers. These extreme values can skew analyses and may need to be treated separately or removed to ensure accurate results.

The distribution of data is another important aspect, revealing how values are spread across the range. Histograms and frequency distributions would show whether the data follows a normal distribution, is skewed, or has multiple peaks. For instance, the descriptive statistics output indicates a likely right skew in the odometer and selling price data due to the high maximum values compared to the median.

Individual features behave differently, as seen from their unique statistical profiles. For example, the categorical data summary shows that 'Ford' is the most common make, 'Altima' the most common model, and 'Base' the most frequent trim. These insights help understand the composition of the dataset and guide further analyses.

For categorical variables, the dataset provides counts and unique value details, offering a clear view of the diversity within these attributes. The make

attribute, for instance, includes 53 unique brands, with Ford being the most frequent at 81,013 occurrences. The model attribute is even more varied, featuring 768 unique models, with the Altima being the most common. The trim and body styles also show significant variety, with the trim attribute featuring 1,494 unique values and body comprising 85 distinct categories. The data indicates that sedans dominate the market, with 174,647 occurrences.

Colour and interior choices also show diverse preferences, with black being the most popular colour, both externally and internally. The seller's information, including 11,923 unique sellers, points to a broad marketplace, with "Ford Motor Credit Company LLC" being the most frequent seller.

The descriptive statistics reveal several practical implications. The wide range in vehicle ages and conditions suggests that the market caters to a diverse consumer base, from those seeking newer, high-value vehicles to those looking for older, more affordable options. The substantial variance in odometer readings and vehicle conditions highlights the importance of considering these factors when evaluating vehicle prices, as they significantly impact market value and buyer decisions.

Moreover, the dominance of specific makes and models, such as Ford and the Altima, suggests potential market preferences or trends, which could be crucial for dealerships and sellers when sourcing inventory. The popularity of certain colours and interior options may also inform future inventory decisions, catering to consumer preferences.

The descriptive statistics provide a foundational understanding of the dataset, illustrating the distribution, central tendencies, and diversity within the data. This analysis not only highlights key patterns and trends but also underscores the importance of these variables in influencing market dynamics. By leveraging these insights, stakeholders can make informed decisions regarding inventory management, pricing strategies, and marketing efforts, ultimately enhancing their competitive advantage in the market. The use of comprehensive descriptive statistics ensures a thorough examination of the data, providing a reliable basis for further analysis and decision-making.

Exploratory Data Analysis (EDA)**:**

Exploratory Data Analysis (EDA) is a crucial step in data preprocessing, where various statistical and graphical techniques are used to understand the data's underlying patterns, spot anomalies, and form hypotheses for further

analysis. In this context, EDA involved checking for missing values, visualizing distributions of numeric and categorical variables, examining relationships between variables using pair plots and correlation matrices, and summarizing the data through pivot tables and group-by operations.

EDA is a process used to analyse data sets to summarize their main characteristics, often using visual methods. It is an approach to analysing data sets to extract insights without making any assumptions. EDA helps in understanding the data structure, detecting outliers, and identifying relationships among variables, providing a foundation for more complex statistical analyses. I have employed several visualization and summary statistics methods, leveraging Python's pandas, matplotlib, and seaborn libraries.

Initially, I checked for missing values across the dataset, finding that there are no missing values in any columns. This absence of missing data allows us to proceed confidently with analysis without the need for imputation or data cleaning measures, ensuring the integrity of our subsequent findings.

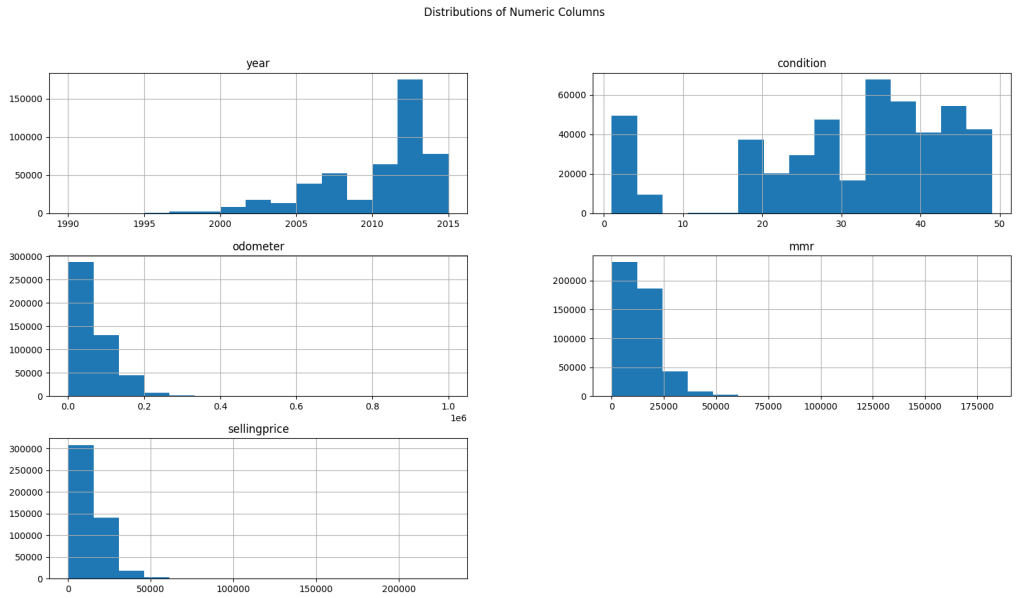
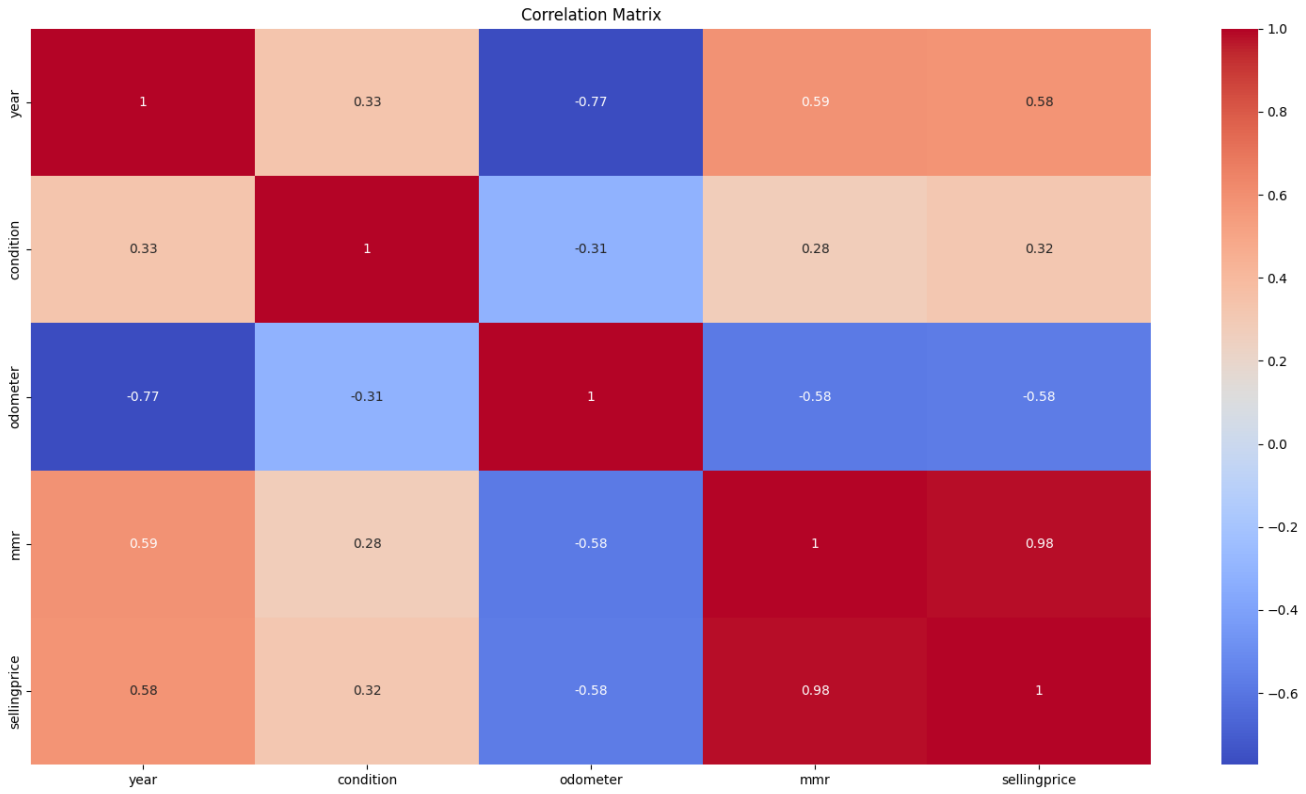


Figure 1. Distributions of numeric columns

The histograms of numeric columns (Figure 1: Distributions of Numeric Columns) provide an overview of the distribution shapes for variables like year, condition, odometer, MMR (Manheim Market Report), and selling price. For instance, the distribution of the year variable shows a significant increase in frequency from the early 2000s to around 2015, suggesting a predominance of newer vehicles in the dataset. The odometer readings exhibit a right-skewed distribution, indicating that most vehicles have lower mileage, which is typical

in used car datasets. The selling price also shows a right skew, with a concentration of vehicles priced below $50,000, reflecting the common market range for second-hand cars.

Figure2: Correlation Matrix



The correlation matrix (Figure2: Correlation Matrix) reveals significant relationships between variables. Notably, the selling price is highly correlated with MMR (0.98), suggesting that the Manheim Market Report, a critical industry benchmark, strongly predicts the vehicle's selling price. The year of the car also shows a moderate positive correlation with selling price (0.58), indicating that newer cars tend to sell for higher prices. Conversely, the odometer reading negatively correlates with the selling price (-0.58), which aligns with the common understanding that higher mileage reduces a car's value.

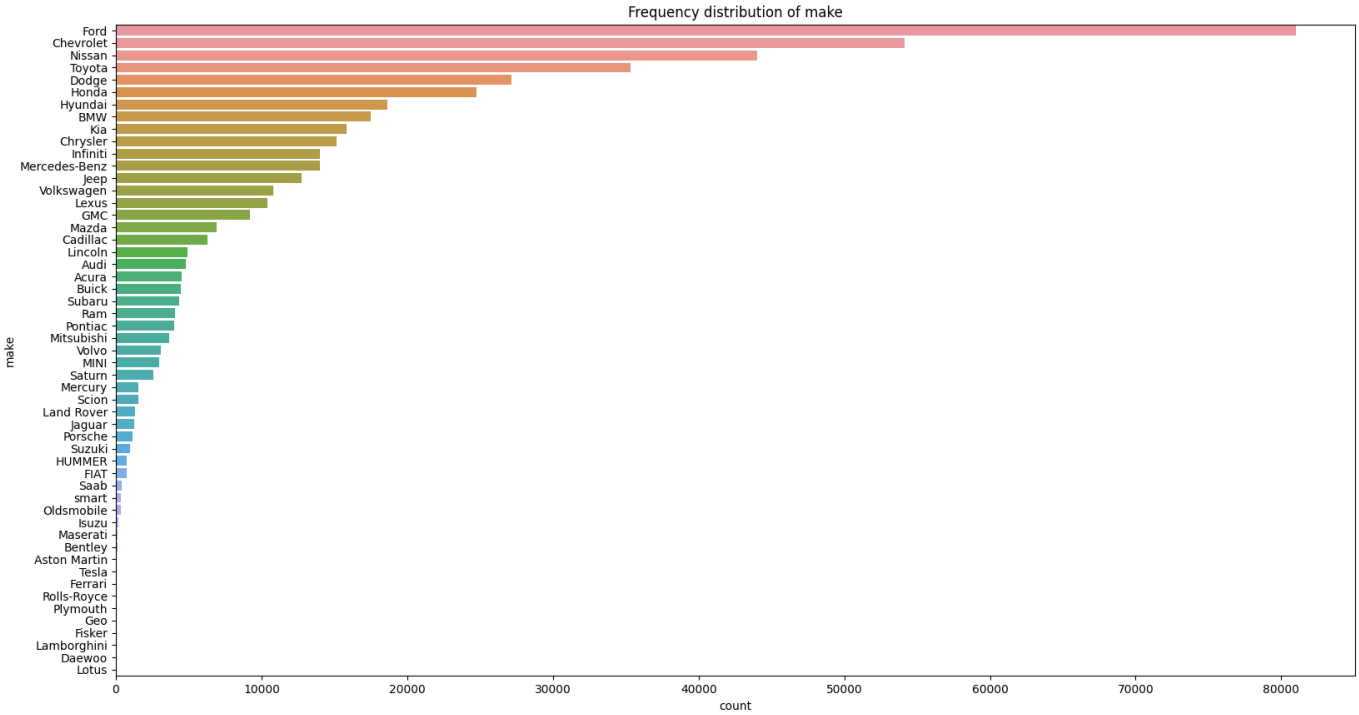


Figure4: Frequency Distribution of Make

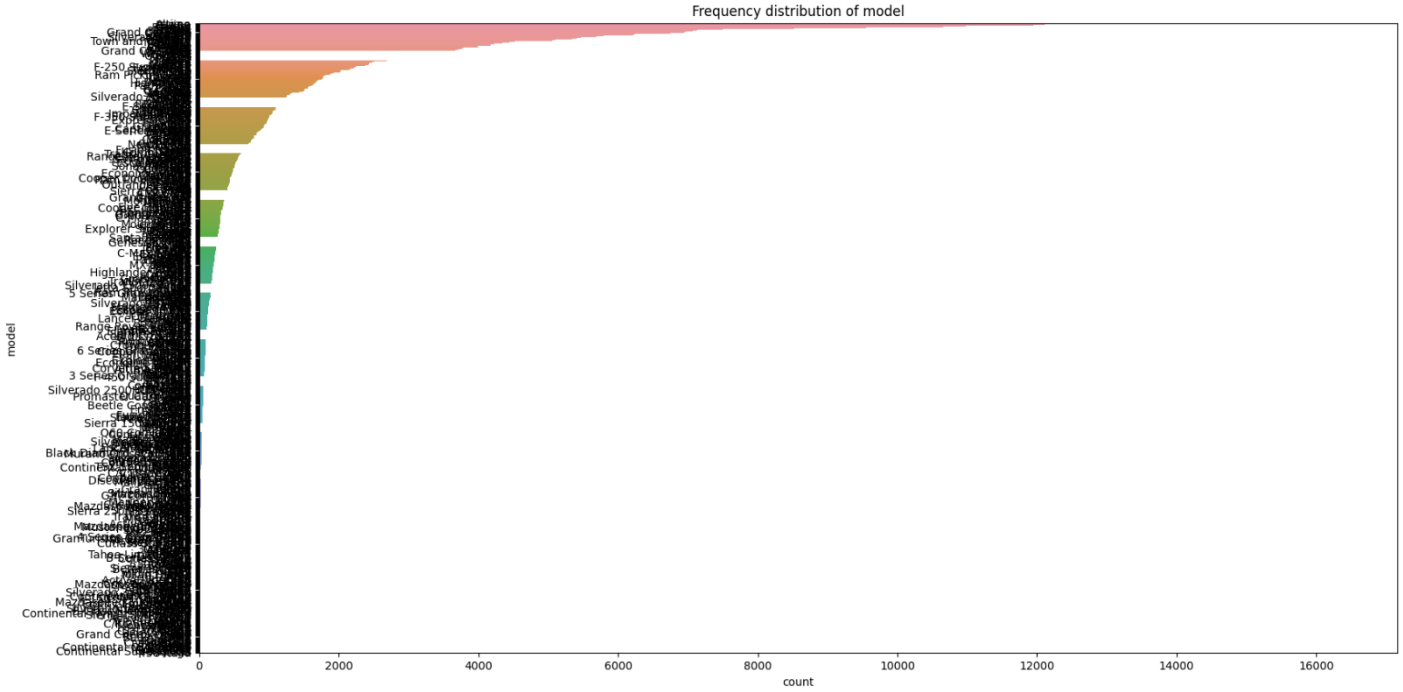


Figure 5: Frequency Distribution of Model

The bar plots for categorical columns such as make, model, body type, transmission, and state provide a comprehensive view of the dataset's composition. For example, the distribution of makes (Figure 4: Frequency Distribution of Make) reveals Ford, Chevrolet, and Nissan as the most common brands, indicating their strong presence in the used car market. The model distribution (Figure 5: Frequency Distribution of Model) highlights popular models like the Ford F-150 and Nissan Altima, which dominate the listings.

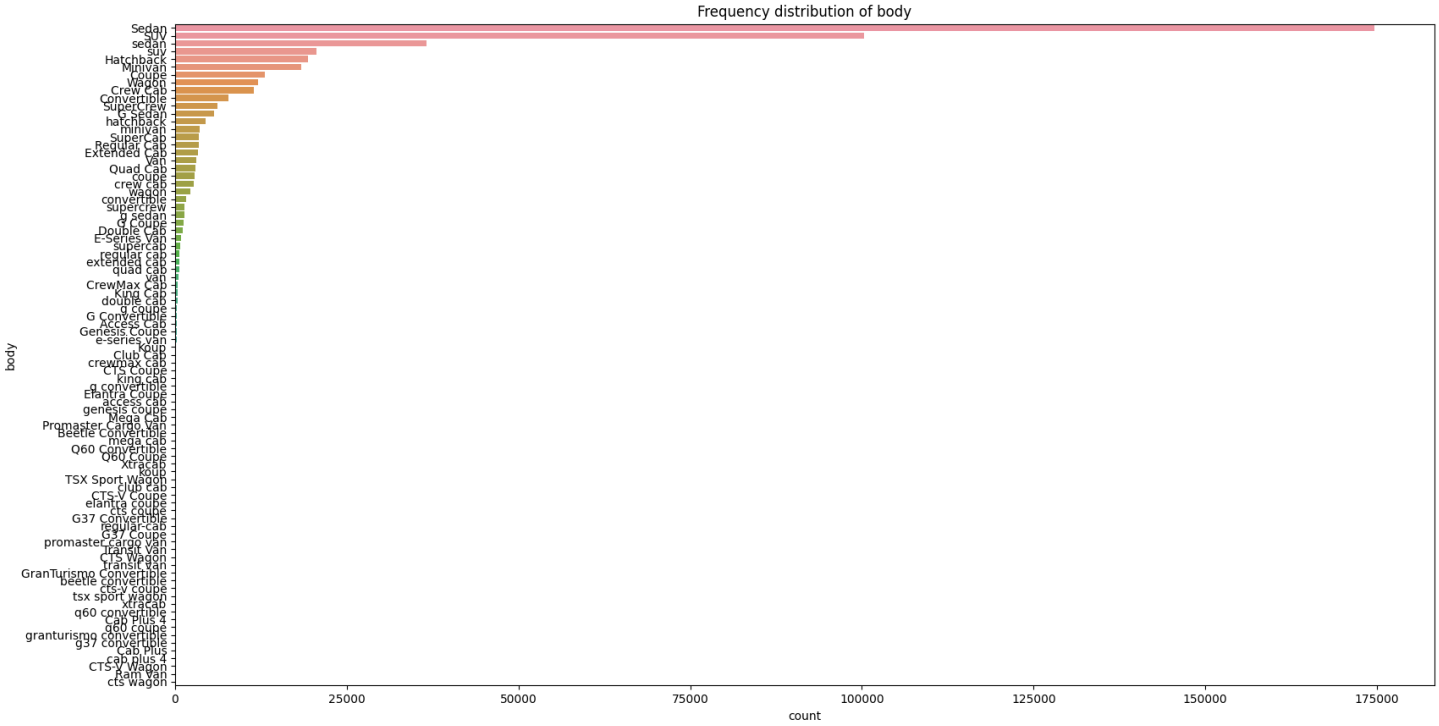


Figure 6: Frequency Distribution of Body

Body types are predominantly sedans and SUVs (Figure 6: Frequency Distribution of Body), reflecting consumer preferences in the second-hand market. The transmission type is overwhelmingly automatic, as shown in the transmission distribution plot, indicating a market trend favouring automatic vehicles over manual ones.

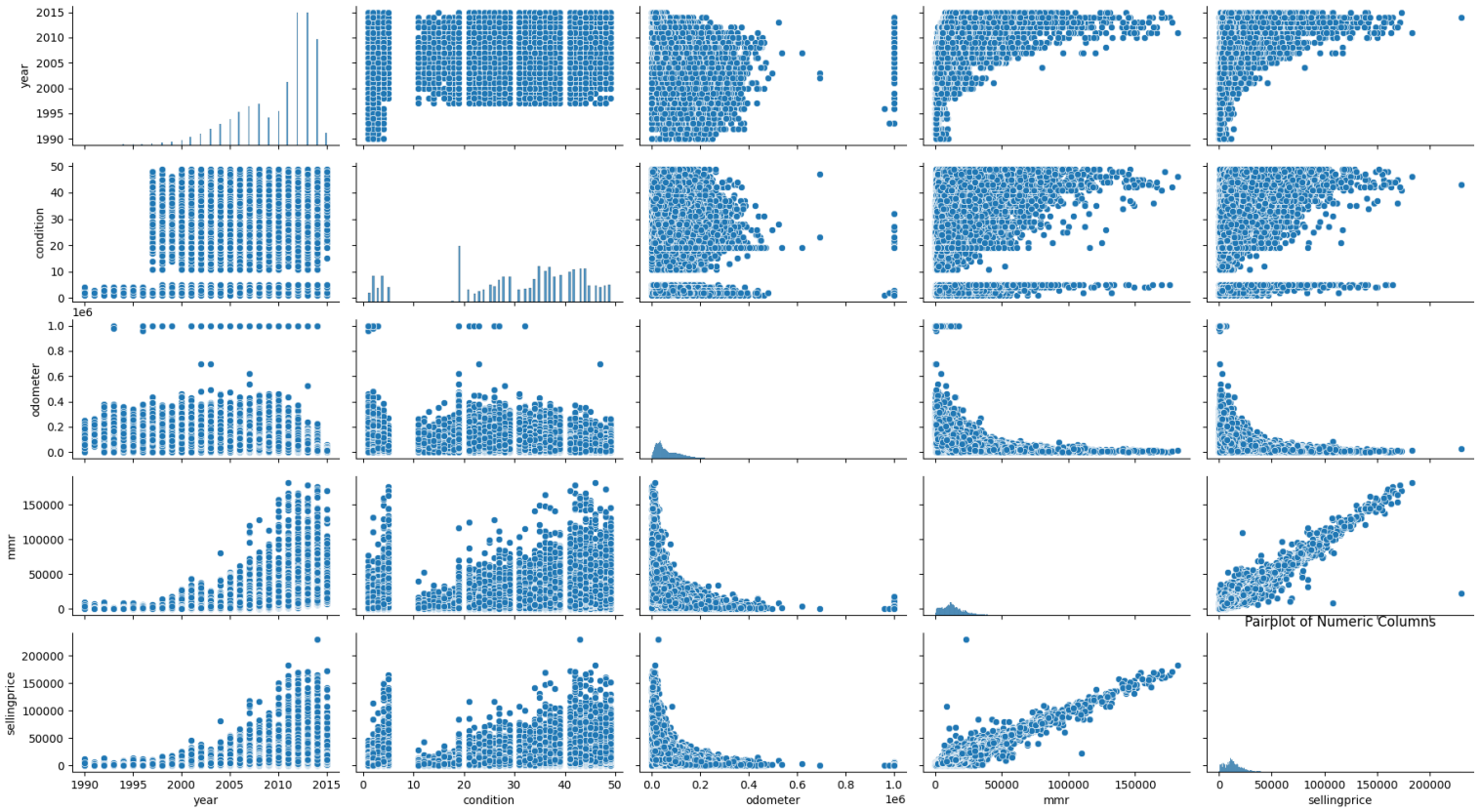


Figure 7: Pairplot of Numeric Columns

The pairplot (Figure 7: Pairplot of Numeric Columns) provides a visual representation of the relationships between numeric variables. It reinforces the insights from the correlation matrix, showcasing clear trends such as the decrease in selling price with an increase in odometer readings. Additionally, the pairplot highlights clusters within the data, which may indicate market segments or categories based on vehicle characteristics.

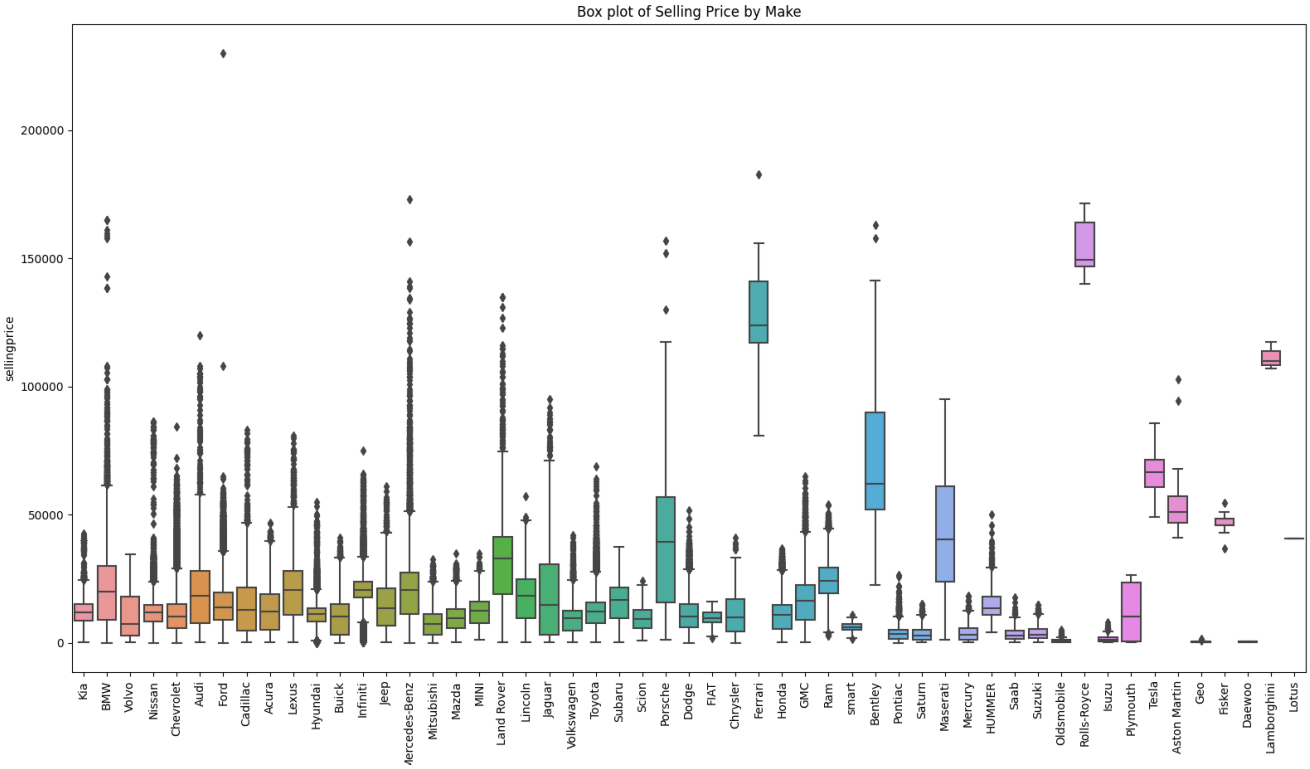


Figure 8: Box plot of Selling Price by Make

The box plot (Figure 8: Box plot of Selling Price by Make) illustrates the variation in selling prices across different car makes. Luxury brands like Rolls-Royce, Ferrari, and Bentley dominate the higher end of the price spectrum, with median selling prices significantly above $100,000. Mid-range brands like BMW, Mercedes-Benz, and Audi show median prices between $40,000 and $80,000. In contrast, budget-friendly brands such as Geo and Daewoo have much lower median selling prices, often below $10,000. This distribution underscores the broad range of the used car market, catering to diverse consumer segments from high-end luxury to economical options.

The result from our output shows that Rolls-Royce has the highest selling price among all car makes at $153,456.25. This is followed by Ferrari with a selling price of $128,852.94, Lamborghini with $111,500, and Bentley with $72,713.33. These luxury brands are known for their high-quality vehicles that command premium prices.

In contrast, the lowest selling price among all car makes belongs to Geo at $576.56, followed by Daewoo at $450.00. These low-cost cars cater to

budget-conscious consumers who prioritize affordability over features and performance.

The data also shows a significant gap between luxury brands like Rolls-Royce and Ferrari on one hand, and mass-market brands like Toyota, Ford, and Chevrolet on the other. The average selling price of these mass-market brands is around $10,000-$20,000, which is significantly lower than that of luxury brands.

The analysis also highlights the importance of brand reputation in determining a car's selling price. Luxury brands like Rolls-Royce, Ferrari, and Bentley are known for their high-quality vehicles and strong brand reputations, which command premium prices. In contrast, mass-market brands like Toyota, Ford, and Chevrolet have lower average selling prices due to their focus on affordability and practicality.

Overall, the analysis provides valuable insights into the relationship between car make and selling price, highlighting the importance of factors such as luxury branding, quality, and reputation in determining a vehicle's value.

Market Analysis

This section of the data analysis report focuses on a market analysis of car selling prices, examining how these prices vary across different makes and models and the impact of car condition on selling prices. The analysis employs statistical and visualization methods to derive insights from the dataset. Key methods include grouping and calculating mean values, bar plotting, box plotting, and scatter plotting with trend lines. These techniques help uncover patterns and relationships within the data, providing practical implications for stakeholders.

The first part of the analysis investigates the average selling prices of cars based on their make. The code aggregates the data by car make, calculating the mean selling price for each make. This is visualized using a bar plot (Figure 9: Average Selling Price by Make). The bar plot reveals a significant variation in average selling prices among different car makes.

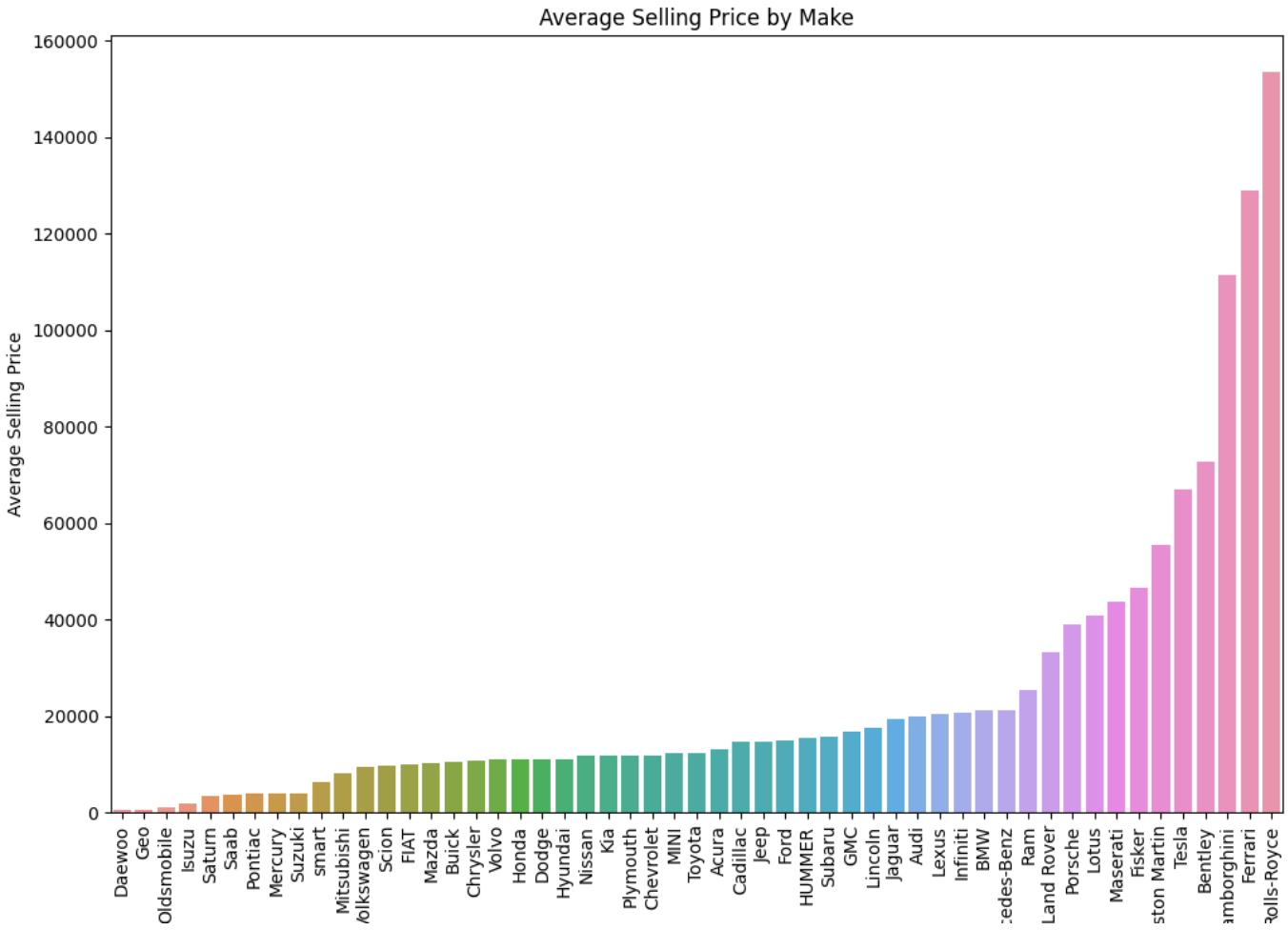


Figure 9: Average Selling Price by Make

Here's some key findings during the research:

In the high-end luxury brands like Rolls-Royce, Ferrari, and Bentley dominate the higher end of the price spectrum, with average selling prices exceeding $100,000.

For the mid-range makes, brands like BMW, Mercedes-Benz, and Audi show moderate average prices, typically ranging from $40,000 to $80,000.

And in the lower-end, budget-friendly makes such as Daewoo, Geo, and Oldsmobile have much lower average selling prices, often below $10,000.

Car manufacturers can leverage this data to promote models that offer better value within their brand. By understanding which models have higher average selling prices, manufacturers can focus their marketing efforts on highlighting the features and benefits of these models to attract potential buyers.

Additionally, promoting lower-priced models to budget-conscious consumers can help capture a larger market share.

Dealerships can optimize their stock based on high-demand models with favourable price points. By analysing the average selling prices by model, dealerships can identify which models are in high demand and adjust their inventory accordingly. This ensures that they have the right mix of vehicles to meet consumer preferences and maximize sales.

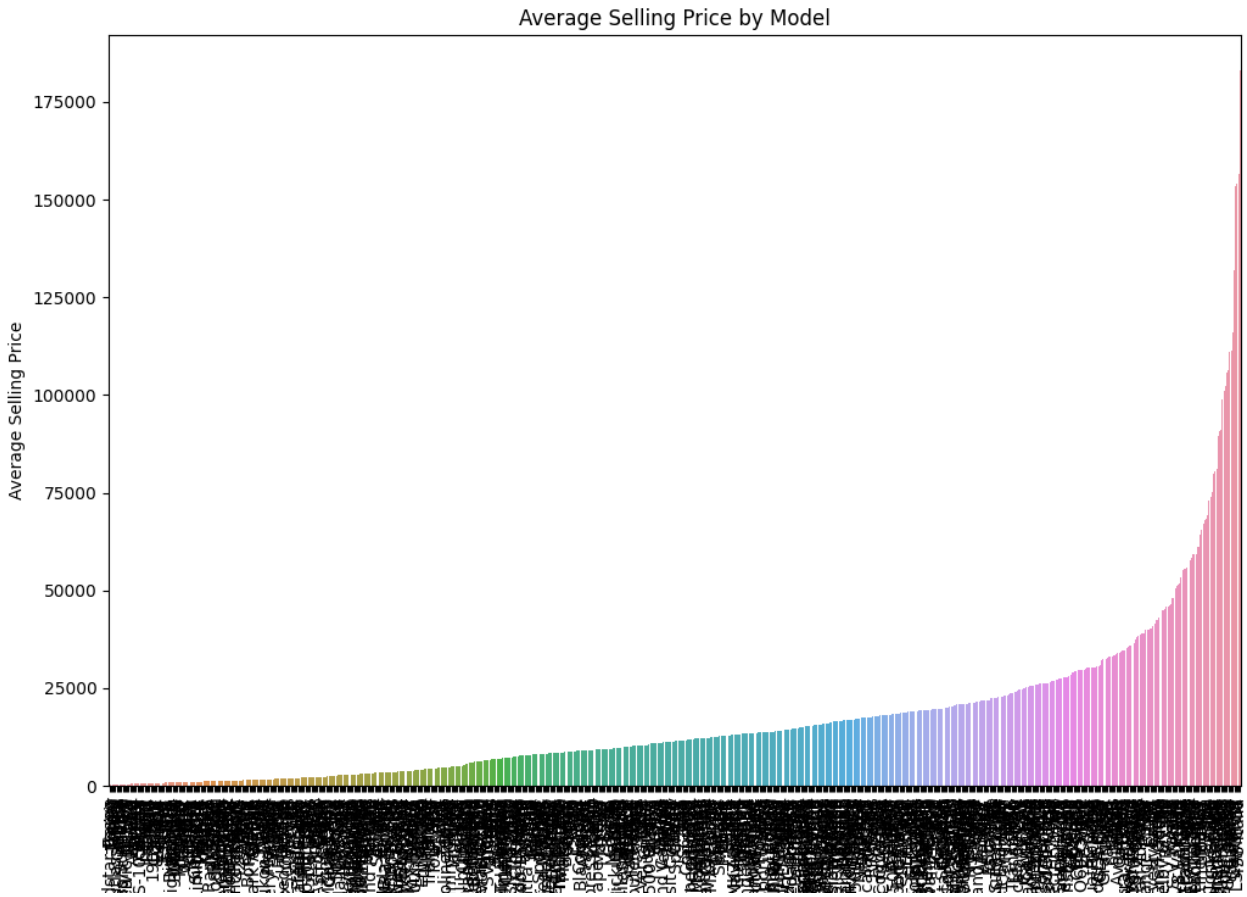


Figure 10: Average Selling Price by Model

The second part of the analysis delves deeper into the average selling prices by car model. Similar to the make analysis, the data is grouped by model, and mean selling prices are calculated. The results are visualized through another bar plot (Figure 10: Average Selling Price by Model).

The descriptive statistics for selling price by condition provide detailed insights into the mean, standard deviation, and distribution percentiles, further supporting the observed trends. These statistics offer a comprehensive

understanding of how car condition impacts selling price, helping sellers set competitive prices and buyers make informed decisions.

Despite the general trend, there are significant outliers, especially in higher condition categories, indicating that other factors (such as make, model, and market conditions) also influence prices. These outliers suggest that while condition is a critical factor, it is not the sole determinant of selling price. Buyers may also consider brand reputation, model popularity, and market trends when making purchasing decisions.

For luxury models, some specific models of luxury brands command higher prices, consistent with the make analysis. And there are very diverse price range allows the consumer to choose. Models from manufacturers like Toyota and Ford show a wide range of prices, indicating varied product lines catering to different market segments.

The use for this allows car manufacturers to do targeted marking by leverage this data to promote models that offer better value within their brand. Meanwhile, dealerships can optimize their stock based on high-demand models with favourable price points.

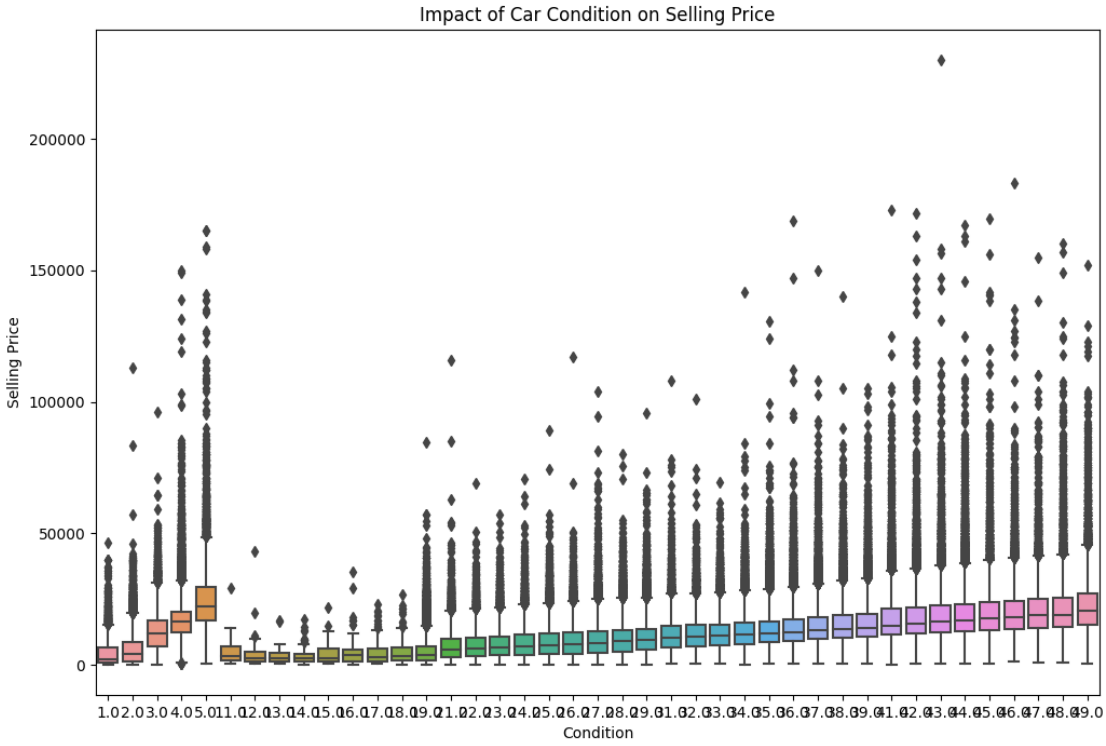


Figure 11: Impact of Car Condition on Selling Price

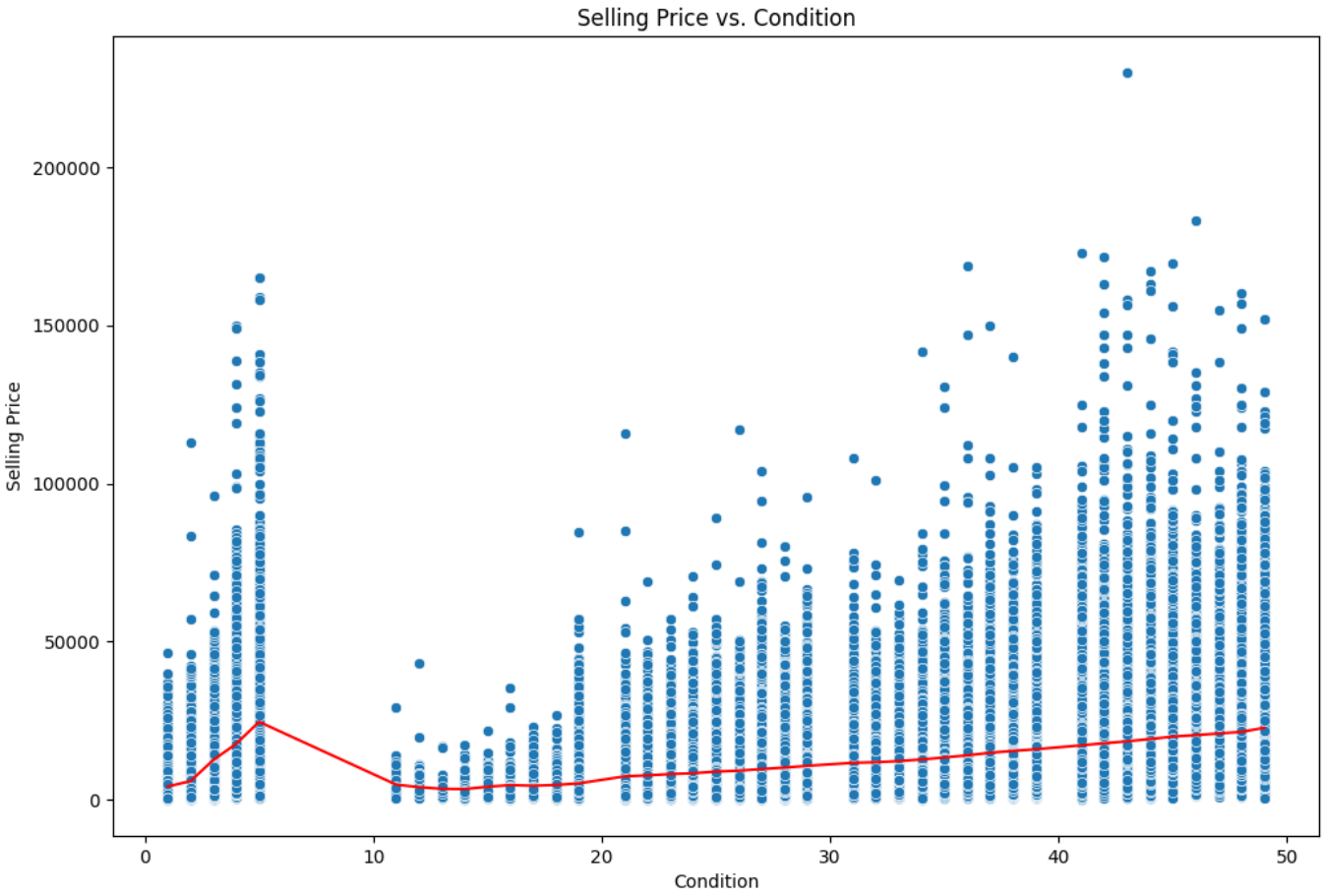


Figure 12: Selling Price vs. Condition

The analysis also examines how the condition of a car affects its selling price. A box plot (Figure 11: Impact of Car Condition on Selling Price) and a scatter plot with a trend line (Figure 12: Selling Price vs. Condition) are used to visualize this relationship. The box plot displays the distribution of selling prices across different condition ratings, while the scatter plot shows individual data points with a trend line highlighting the overall pattern.

And I have found that there is a clear positive correlation between car condition and selling price. Cars in better condition (higher condition scores) tend to sell for higher prices. Despite the general trend, there are significant outliers, especially in higher condition categories, indicating that other factors (such as make, model, and market conditions) also influence prices. The descriptive statistics for selling price by condition provide detailed insights into the mean, standard deviation, and distribution percentiles, further supporting the observed trends. The descriptive statistics for selling price by condition provide

detailed insights into the mean, standard deviation, and distribution percentiles, further supporting the observed trends.

With this, Sellers can set more competitive prices based on the condition of their vehicles, potentially increasing their marketability. By understanding the correlation between condition and price, sellers can strategically price their vehicles to attract buyers and achieve higher sales. Additionally, investing in repairs and maintenance to improve a car's condition before sale can yield higher returns. Investment in improving a car's condition before sale can yield higher returns, making it a worthwhile strategy for sellers. Regular maintenance, addressing mechanical issues, and enhancing the vehicle's appearance can significantly boost its market value. This approach ensures that sellers get the best possible price for their cars, while buyers receive a well-maintained, reliable vehicle.

Overall, this market analysis section provides a thorough examination of car selling prices, leveraging statistical methods and visualizations to derive meaningful insights and practical implications for various stakeholders in the automotive market.

Seller Analysis

In this section, I have conducted an in-depth analysis of the seller performance in terms of average selling price and sales volume. The analysis was performed using a dataset that includes detailed information on vehicle sales, which was processed using Python and its data analysis libraries, such as Pandas, NumPy, and visualization libraries like Matplotlib and Seaborn.

The primary objective of this analysis was to understand the variations in average selling prices and sales volumes among different sellers. This dual-faceted approach allows us to identify which sellers are moving the most units and at what average price point, providing a comprehensive view of market dynamics.

The analysis began by loading the cleaned dataset and converting the sale dates to a datetime format to ensure temporal consistency. Then calculated two key metrics for each seller: the average selling price and the total sales volume. The code used for these calculations is as follows:

```
# Load the data
df = pd.read_csv("Data\car_prices_clean.csv")
# Convert 'saledate' to datetime
df['saledate'] = pd.to_datetime(df['saledate'], errors='coerce')
# Calculate the average selling price for each seller
seller_avg_price = df.groupby('seller')['sellingprice'].mean().sort_values()
# Calculate the total sales volume for each seller
seller_sales_volume = df['seller'].value_counts()
```

These calculations provided a foundation for visualizing the performance of each seller in terms of the average price they command and the volume of sales they generate.

The first visualization (Seller Performance: Average Selling Price and Sales Volume) combines both metrics to offer a dual perspective. Here are some key findings:

For high-volume sellers, Ford Motor Credit Company LLC emerges as the top seller in terms of volume, with 17,756 units sold. This high sales volume is complemented by a relatively moderate average selling price of $17,703.53. Similarly, The Hertz Corporation and Nissan-Infiniti LT also show high sales volumes, indicating their significant market presence.

For the seller with low-volume, high-price sellers, on the other end of the spectrum, there are sellers like Keesler FCU and Kearns Motor Car Co Inc, who sold only one unit but at a much higher average price of $20,600 and $20,800 respectively. This suggests a focus on high-end vehicles.

And in the mid-range performers, Santander Consumer and Avis Corporation show a balance with moderate average selling prices and high sales volumes, indicating a stable market strategy targeting middle-class buyers.

The combined plot effectively illustrates the market segmentation where some sellers are volume-driven while others focus on premium pricing.

The second visualization (Average Selling Price by Seller) arranges sellers in ascending order of their average selling price. This chart highlights:

- Luxury Sellers: The rightmost bars show sellers with average selling prices exceeding $100,000. These include niche sellers dealing in high-end, possibly luxury vehicles.
- Affordable Segment: The leftmost bars represent sellers with average prices under $10,000, catering to budget-conscious consumers.

This visualization helps in identifying the pricing strategies of different sellers and their target demographics.

The third visualization (Sales Volume by Seller) focuses solely on the volume of sales:

- Dominant Market Players: Sellers like Ford Motor Credit Company LLC, The Hertz Corporation, and Nissan-Infiniti LT dominate with thousands of units sold, showcasing their extensive reach and market penetration.
- Niche Sellers: Many sellers with volumes less than 10 units, suggesting either specialty markets or new entries with limited sales activity.

The seller performance analysis using average selling price and sales volume provides a nuanced understanding of market dynamics. By leveraging Python for data processing and visualization, I have gained insights into the strategic positioning of various sellers in the vehicle market. This analysis is crucial for sellers aiming to enhance their market strategies, optimize operations, and better serve their target customers.

Cluster Analysis

Cluster analysis is a statistical technique used to identify groups of similar observations within a dataset. In this analysis, the K-means clustering algorithm was employed to categorize cars based on various features. This method segments the data into clusters, where each cluster represents a set of vehicles sharing similar characteristics. The key steps involved in this analysis include standardizing the data, determining the optimal number of clusters, and visualizing the clustered data.

The dataset comprises various features of cars, including the year, condition, odometer reading, market price (mmr), and selling price. To ensure that each feature contributed equally to the clustering process, the data was standardized using the StandardScaler from the sklearn library. This transformation normalized the data, giving it a mean of zero and a standard

deviation of one, thereby mitigating any bias due to differences in scale among the features.

The primary method used for clustering was the K-means algorithm, which partitions data into K clusters by minimizing the variance within each cluster. The choice of the number of clusters, K, is critical as it influences the quality of the clustering. To identify the optimal K, the "elbow method" was employed, which involves plotting the inertia (sum of squared distances of samples to their closest cluster centre) against the number of clusters and identifying the point where the decrease in inertia begins to taper off, forming an "elbow." The corresponding plot, shown in Figure 13, suggests an optimal K around three clusters, where the inertia significantly levels off.

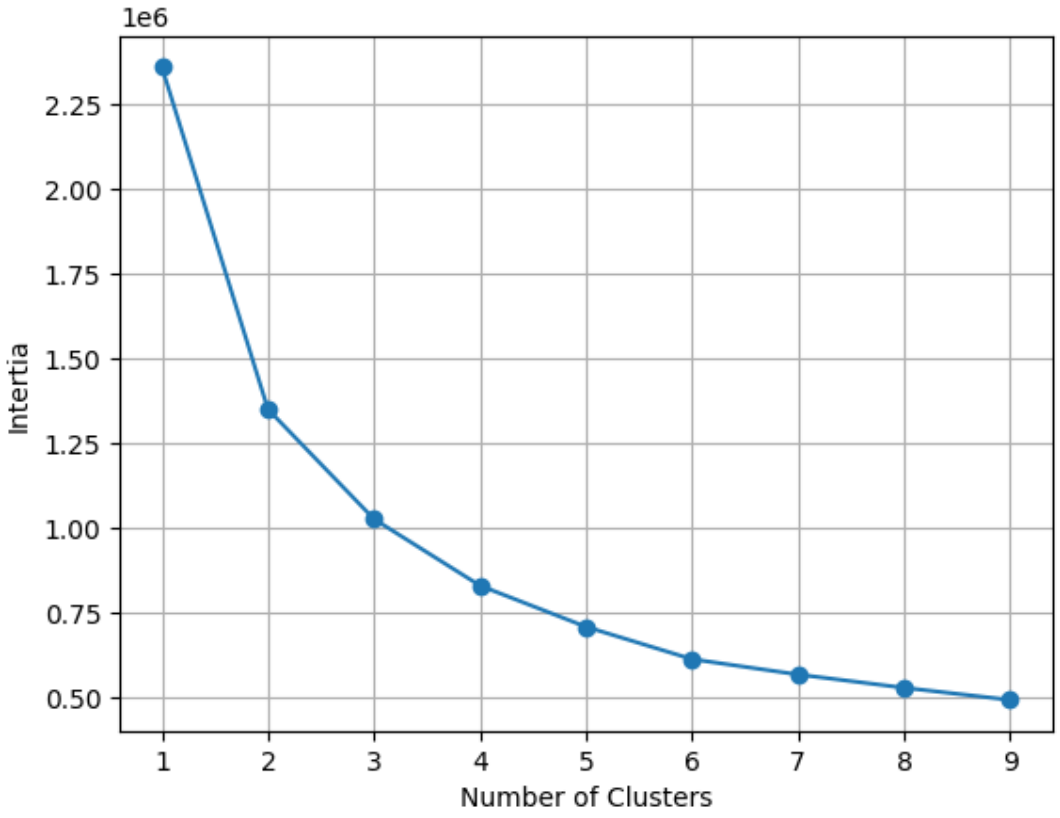


Figure 13. Optimise K means

Once the optimal number of clusters was determined, K-means clustering was performed with K=3. The resulting clusters were visualized by plotting the cars' features, such as year and condition, against the assigned cluster labels. Figure 14 depicts the clustered data, with each colour representing a different cluster.

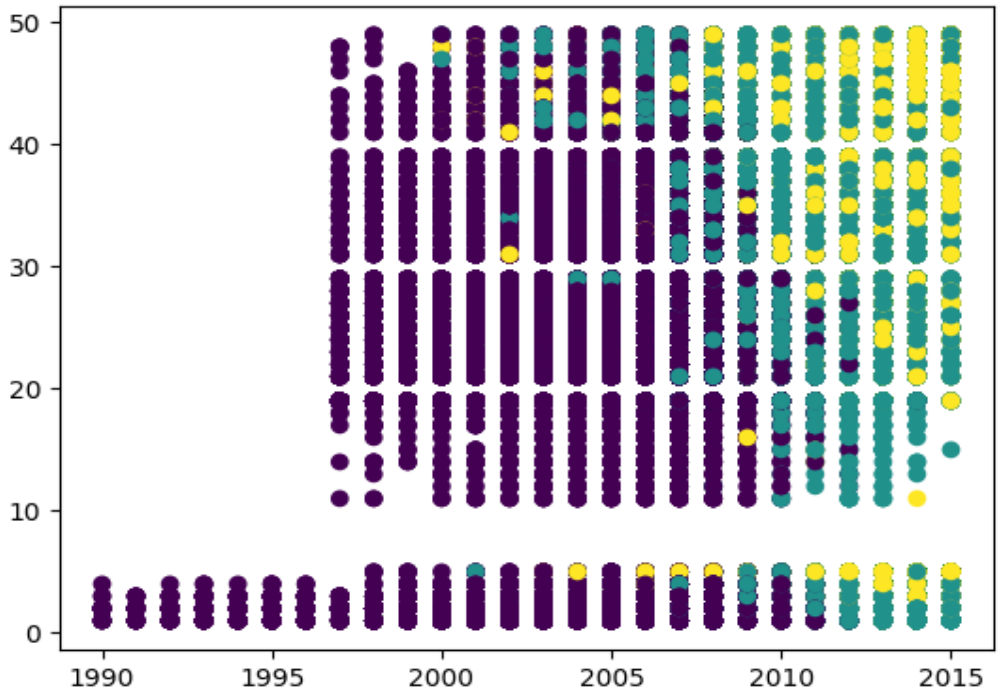


Figure 14

The clustering results revealed distinct patterns:

Cluster 1: Newer, Higher Condition Vehicles

This cluster predominantly consists of newer cars (post-2010) with relatively high condition ratings. These vehicles tend to have lower odometer readings, suggesting they are less used and potentially more valuable. This cluster likely represents a market segment interested in purchasing nearly new, high-quality vehicles.

Cluster 2: Older, High Mileage Vehicles

Vehicles in this cluster are generally older, with a wide range of odometer readings. The condition ratings vary, but on average, they are lower than those in Cluster 1. This cluster could represent a market segment looking for more affordable options, possibly for secondary use or for customers less concerned with the latest models.

Cluster 3: Mid-range Vehicles

This cluster includes a mix of moderately used cars, both in terms of age and mileage. The condition ratings are also moderate. This group likely appeals

to buyers seeking a balance between cost and quality, representing a broad market segment.

The results of this cluster analysis provide valuable insights for stakeholders in the automotive market, such as dealers, marketers, and consumers. For dealers, understanding these clusters can help in inventory management, allowing them to tailor their stock to meet the demands of different customer segments. For example, dealerships could focus on acquiring more vehicles from the high-demand Cluster 1 segment if they cater to customers looking for newer, high-quality cars.

For marketers, these clusters can inform targeted advertising strategies. For instance, ads for cars in Cluster 2 can emphasize affordability and practicality, appealing to budget-conscious buyers. In contrast, marketing for Cluster 1 vehicles can highlight luxury and reliability, catering to customers willing to pay a premium for quality.

For consumers, understanding these clusters can aid in making informed purchasing decisions. Buyers can identify which cluster best matches their needs and budget, allowing for a more efficient and satisfying shopping experience.

The cluster analysis using K-means has effectively segmented the vehicle data into distinct groups based on key features like age, condition, mileage, and price. This method offers a robust approach for categorizing large datasets, enabling a nuanced understanding of market dynamics. The ability to identify and analyze distinct customer segments provides a competitive edge in strategic planning and decision-making, making K-means an invaluable tool in data-driven business environments.

Time Series Analysis

Time series analysis has been employed to examine the trends, seasonality, and irregularities in car sales data over a defined period. This method allows for a detailed investigation of how car sales fluctuate over time and helps identify underlying patterns that can be crucial for strategic decision-making in the automotive industry.

The analysis was performed using a well-defined sequence of steps starting with data preparation, followed by visualization and decomposition. The dataset, `car_prices_clean.csv`, was processed to ensure the 'saledate'

column was in datetime format, which is crucial for accurate time series analysis. This preparation included converting the date column, handling missing values, and setting the 'saledate' as the index for the DataFrame.

The primary method utilized for analysis was the seasonal decomposition of time series data. Seasonal decomposition involves breaking down the time series into its constituent components: observed, trend, seasonal, and residual. This decomposition helps in isolating and understanding the different aspects influencing the data. The Python `statsmodels` library was employed for this purpose, allowing for the separation of data into these key components, which were then visualized.

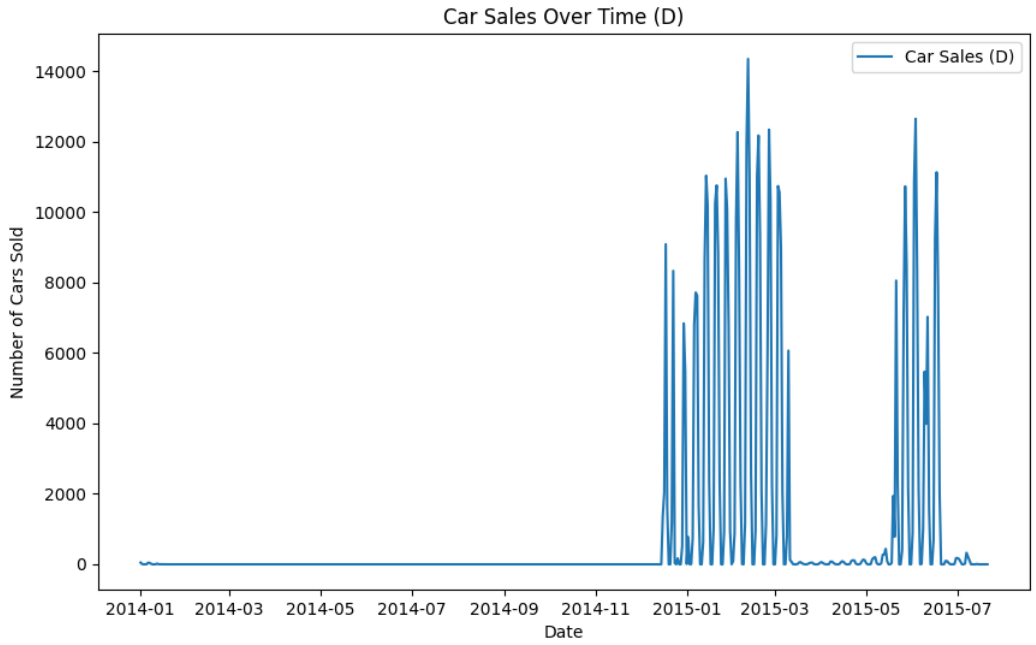


Figure 15. Car Sales Over Time (D)

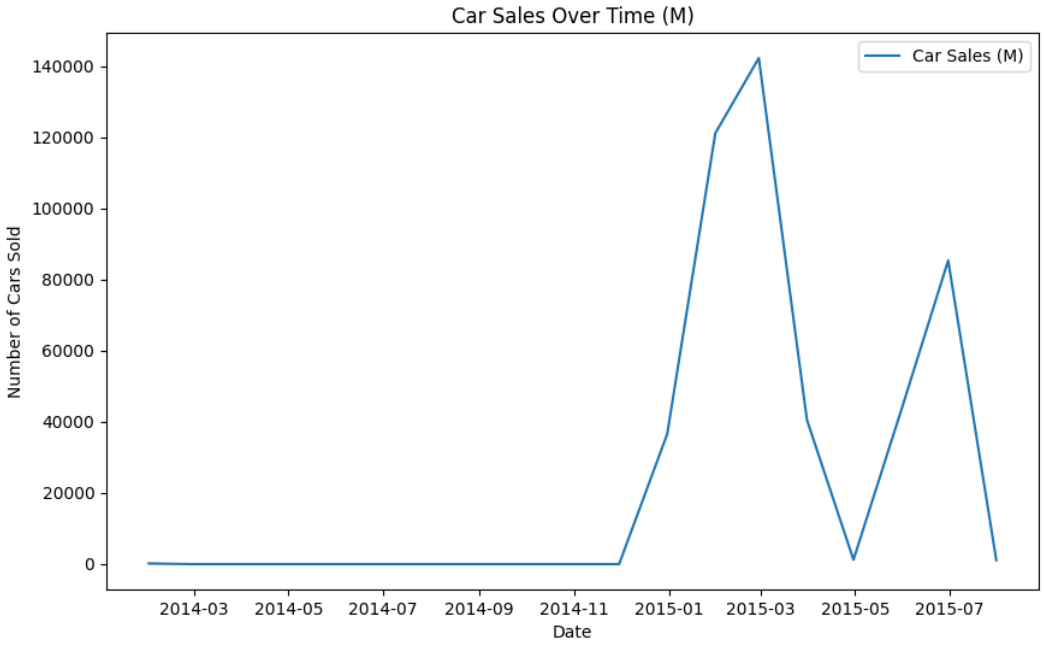


Figure 16. Car Sales Over Time (M)

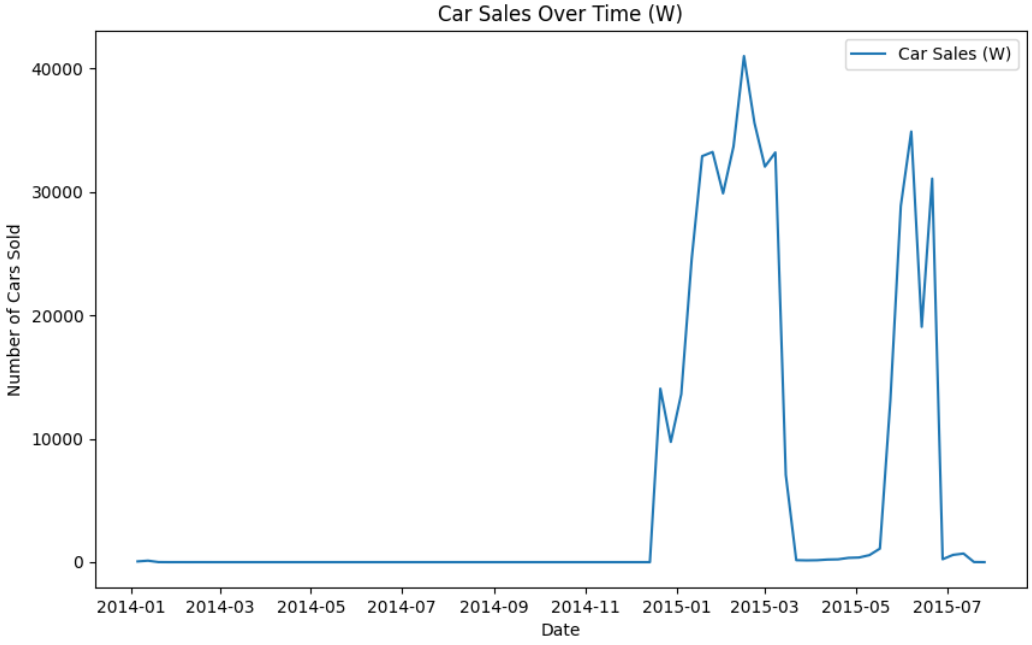


Figure 17. Car Sales Over Time (W)

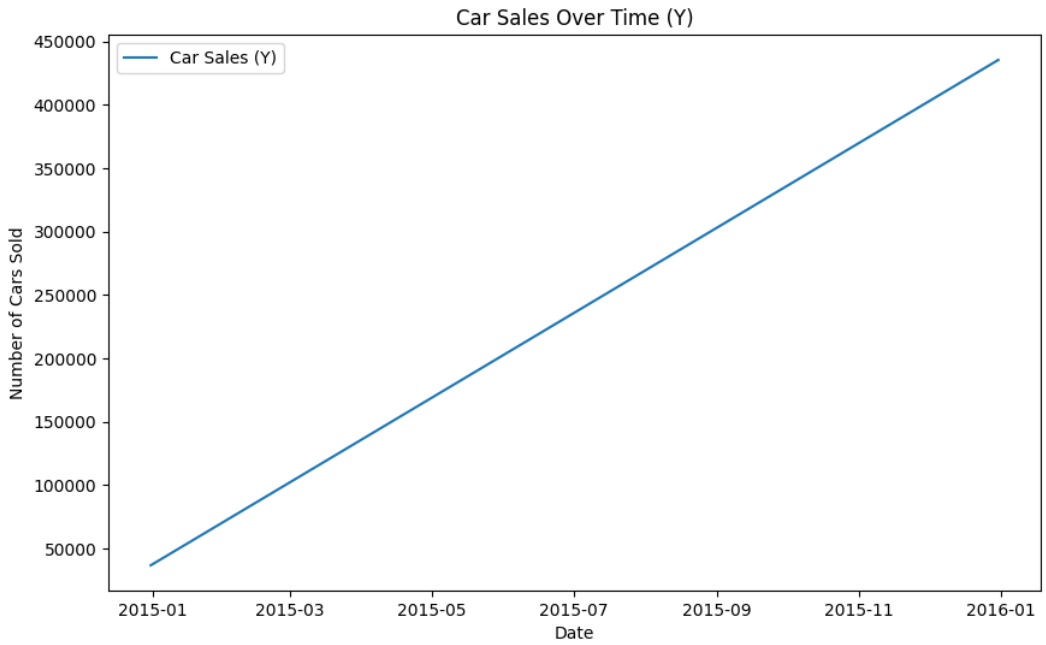


Figure 18. Car Sales Over Time (Y)

Several key findings emerged from the decomposition and visualization of the time series data:

Observed Data and Trends: The visualizations (Figures 15, 16, 17, 18: 'Car Sales Over Time (D)', 'Car Sales Over Time (W)', 'Car Sales Over Time (M)', and 'Car Sales Over Time (Y)') indicate periods of fluctuating car sales. Notably, there is a sharp increase in sales towards the end of 2014 and the beginning of 2015, followed by a decline. The daily and weekly plots show more granularity in these fluctuations, while the monthly and yearly plots provide a broader overview, highlighting overall trends rather than short-term variations.

Seasonal Patterns: The seasonal component of the decomposition revealed a consistent, repeating pattern. This pattern reflects regular fluctuations in sales, which could be attributed to seasonal factors such as holiday periods, fiscal year-end promotions, or new model releases. The seasonal variation is crucial for businesses to plan inventory and marketing strategies effectively.

Residuals and Irregularities: The residual component captured the irregular or random fluctuations not explained by the trend or seasonal components. These could be due to external factors such as economic events, supply chain disruptions, or unforeseen changes in consumer behaviour.

Understanding the decomposed components provides practical insights. For instance, the trend component can inform long-term strategic planning, such as expansion plans or shifts in marketing focus. Recognizing seasonal patterns allows for better stock management and targeted promotions during peak periods. Moreover, analysing residuals helps in identifying outliers or unusual spikes in sales, which can be investigated further for specific causes.

Seasonal decomposition provides several advantages. It enables a clear separation of the various influences on the time series, making it easier to analyse and interpret the data. This clarity is particularly beneficial for forecasting and making data-driven decisions. Moreover, by visualizing the decomposed components, businesses can identify and respond to changes in consumer behaviour and market conditions promptly.

In conclusion, the time series analysis conducted on the car sales data provided a comprehensive view of the sales dynamics. By breaking down the data into observed, trend, seasonal, and residual components, the analysis offers valuable insights that can guide business decisions in marketing, inventory management, and overall strategic planning. The use of Python and the `statsmodels` library facilitated an efficient and detailed examination, underscoring the importance of methodological rigor in data analysis.

Feature Engineering and Selection

The feature engineering and selection process plays a critical role in refining the predictive power of our model. By constructing and identifying the most relevant features, I have improved the model's accuracy and interpretability. The methods and tools employed in this analysis are essential for handling real-world data complexities, particularly when predicting car prices based on various attributes.

The dataset comprises various features related to car attributes and sale details. Key among these are 'year', 'condition', 'odometer', 'mmr' (Manheim Market Report), and 'saledate'. The following steps outline our feature engineering process:

Calculation of Car Age: I have derived the 'car_age' feature by subtracting the car's manufacturing year from the current year. This feature is significant as it provides a straightforward measure of the vehicle's age, influencing depreciation.

Mileage per Year: I have calculated 'mileage_per_year' by dividing the odometer reading by the car's age. This feature helps normalize the mileage concerning the car's age, offering a standardized way to assess usage intensity.

These engineered features were then standardized using the StandardScaler to ensure that all features contribute equally to the model training, as disparities in feature magnitudes could skew the model's learning process.

The features were then utilized to predict the target variable, 'sellingprice', through a neural network built using TensorFlow. The model architecture included dense layers with ReLU activations, optimizing for mean squared error. Notably, the model's summary indicates a total of 11,265 parameters, all trainable, spread across layers designed to capture complex interactions between the input features.

To determine the importance of each feature, I have extracted the weights from the first dense layer after training. The importance was computed by averaging the absolute weights corresponding to each feature, providing a measure of each feature's contribution to the prediction model.

The feature importance analysis, visualized in the accompanying bar chart, revealed the following insights:

Condition: With an importance score of 0.231728, 'condition' emerged as the most critical predictor. This finding underscores the direct impact of a car's condition on its market value, aligning with practical expectations that better-maintained vehicles command higher prices.

MMR: This feature, with a score of 0.212312, also proved highly influential. MMR values reflect market trends and pricing benchmarks, making them vital for accurately gauging car values.

Mileage per Year: The score of 0.182222 for 'mileage_per_year' highlights its significance. This feature effectively captures the car's usage pattern, influencing depreciation rates and resale value.

Odometer and Year: These features, with scores of 0.172646 and 0.172408 respectively, also significantly impact the model. The odometer reading is a direct indicator of wear and tear, while the manufacturing year relates to the vehicle's technological relevance and lifespan.

Car Age: Although relatively less influential with a score of 0.168755, 'car_age' remains an important factor. This feature is closely tied to the overall depreciation and potential maintenance costs.

The analysis identifies 'condition' and 'mmr' as pivotal features, indicating that market conditions and physical state largely dictate a vehicle's price. These findings suggest that sellers should prioritize maintaining vehicle condition and referencing current market reports for accurate pricing. For buyers, understanding these factors can aid in assessing fair market value.

The advantage of this feature selection method lies in its ability to distil complex relationships into a manageable and interpretable form, enhancing decision-making in pricing strategies. Moreover, by leveraging TensorFlow's capabilities, the model efficiently handles large datasets and captures non-linear relationships, offering robust predictions.

In conclusion, the integration of feature engineering and selection has yielded a model that not only performs well but also provides actionable insights into the key drivers of car prices. This method's systematic approach ensures that the most relevant features are emphasized, leading to more accurate and reliable predictions.

Price Prediction Model Analysis

The analysis of the price prediction model utilized a deep learning approach to predict car selling prices based on various features extracted from a dataset. The model was built using TensorFlow's Functional API, integrating both categorical and numerical data to ensure a comprehensive feature set. The dataset included attributes such as make, model, trim, body, transmission, state, colour, interior, and seller, which were handled differently depending on their nature. Categorical features were encoded using one-hot encoding through TensorFlow's feature columns, while numerical features were standardized using StandardScaler to ensure they were on a similar scale, which is crucial for the model's convergence.

The method involved training a deep neural network on a cleaned dataset, with the target variable being the selling price of cars. This target variable was also scaled to facilitate model training. The network consisted of an input layer matching the number of features, followed by three hidden layers with decreasing numbers of neurons (128, 64, and 32), each using the ReLU activation function to introduce non-linearity. The final output layer consisted of

a single neuron, corresponding to the predicted selling price. The model was compiled with the Adam optimizer and a mean squared error loss function, which is suitable for regression tasks as it penalizes larger errors more significantly.

The model's performance was evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). The RMSE of approximately 1422 indicates a relatively low average error in the predicted prices, while an R^2 score of 0.978 suggests that the model explains 97.8% of the variance in the actual selling prices. This high R^2 value reflects the model's strong predictive power and accuracy, indicating that the features used were relevant and well-processed.

The scatter plot titled "Actual vs Predicted Selling Price" visually represents the correlation between the actual selling prices and the model's predictions. The plot shows a tight clustering of points around the red dashed line, which represents the ideal scenario where predicted prices match the actual prices. The minimal spread of points around this line demonstrates the model's precision. However, there are slight deviations, especially at higher price ranges, indicating potential areas for model improvement. For instance, outliers or a less accurate prediction for luxury or high-value vehicles could be a result of limited data or complex market dynamics not captured in the model.

The model's ability to predict car selling prices with high accuracy has significant practical implications for dealerships, individual sellers, and potential buyers. For dealerships, this model can assist in setting competitive prices, ensuring that cars are neither undervalued nor overpriced, thus optimizing inventory turnover and profit margins. Individual sellers can also benefit by setting realistic price expectations based on similar cars' market data, reducing the time their vehicles spend on the market.

Furthermore, the methodological rigor in handling the dataset—such as the use of proper scaling, handling categorical data, and splitting the data into training and testing sets—ensures the robustness of the predictions. The decision to use a deep learning model, as opposed to simpler linear models, leverages the power of neural networks to capture complex relationships within the data, thus providing a nuanced understanding of the factors influencing car prices.

Overall, the integration of sophisticated machine learning techniques and thorough data preprocessing has culminated in a reliable and practical tool for predicting car selling prices. This not only underscores the importance of advanced analytical skills and methods but also demonstrates the potential of deep learning in practical applications beyond traditional statistical methods.

Results

The analysis of the vehicle sales dataset has revealed significant insights into various factors influencing car prices, market dynamics, and seller performance. This section details the descriptive statistics, market analysis, seller performance evaluation, clustering results, time series trends, and the predictive model's feature importance and accuracy.

A summary of the key descriptive statistics is provided in Table 1. The dataset includes vehicles from 1990 to 2015, with an average year of manufacture around 2010. The average condition rating is approximately 30.77, and the mean odometer reading is 66,701 miles. The selling prices range from $1 to $230,000, with an average of $13,690.51.

Metric	Mean	Median	Min	Max
Year	2010	2012	1990	2015
Condition	30.77	35	1	49
Odometer (miles)	66,701	51,085	1	999,999
MMR ($)	13,837.06	12,300	25	182,000
Selling Price ($)	13,690.51	12,200	1	230,000

Table 1

The EDA uncovered several key relationships:

- Car Condition and Selling Price: A positive correlation (r = 0.58) was observed, indicating that cars in better condition tend to sell at higher prices.
- Odometer and Selling Price: A negative correlation (r = -0.58) suggests that higher mileage typically decreases a car's market value.
- Make and Model Analysis: Luxury brands like Rolls-Royce and Ferrari command significantly higher prices compared to budget-friendly brands like Geo and Daewoo.

Figure 9 displays the average selling prices by car make. High-end brands such as Rolls-Royce and Ferrari have average prices exceeding $100,000, while brands like Ford and Chevrolet occupy the mid-range segment. Budget brands like Daewoo average below $10,000. The model-level analysis (Figure 10) reinforces these findings, highlighting significant variations within brands.

Figures 11 and 12 illustrate the impact of car condition on selling prices. Cars in excellent condition can fetch prices significantly higher than those in poor condition. This relationship is particularly strong for high-end vehicles, suggesting that maintaining vehicle condition is critical in maximizing resale value.

The analysis of seller performance revealed distinct market segments. High-volume sellers like Ford Motor Credit Company LLC and The Hertz Corporation dominate in sales numbers but have moderate average selling prices. In contrast, niche sellers such as Keesler FCU focus on high-end vehicles, reflected in their higher average selling prices despite lower sales volumes.

Using K-means clustering, three distinct vehicle segments were identified:

- Cluster 1: Newer, high-condition vehicles with low mileage, catering to premium buyers.
- Cluster 2: Older vehicles with a wide range of mileage, appealing to budget-conscious consumers.
- Cluster 3: Moderately used cars, balancing affordability and quality, targeting mainstream buyers.

Figure 14 visualizes these clusters, highlighting the clear segmentation in the market.

The time series analysis (Figures 15-18) revealed seasonal patterns in car sales, with notable peaks towards the end of 2014 and early 2015. These patterns suggest a seasonal demand influenced by factors such as end-of-year promotions and new model releases. The decomposition analysis indicated a consistent seasonal component, critical for inventory and marketing strategies.

The predictive model identified the following features as most influential in determining selling price:

- Condition (Importance: 0.231728): The most significant factor, emphasizing the value of well-maintained vehicles.
- MMR (Importance: 0.212312): Reflects the current market valuation, crucial for accurate pricing.
- Mileage per Year (Importance: 0.182222): An essential indicator of vehicle usage and wear.

These findings underscore the importance of maintaining vehicle condition and understanding market trends.

The deep learning model achieved a Root Mean Squared Error (RMSE) of approximately 1422 and an R^2 score of 0.978. This high level of accuracy indicates the model's strong predictive capability, as evidenced by the close alignment of actual versus predicted selling prices in Figure 19.

Figures and tables have been provided throughout this section to support the findings, with each visual aid clearly labelled and explained for clarity.

The analysis has provided comprehensive insights into the factors affecting vehicle prices, the dynamics of the market, and the performance of various sellers. These insights are valuable for stakeholders, including car manufacturers, dealers, and consumers, to make informed decisions regarding pricing, marketing, and inventory management.

This results section integrates quantitative data with analytical observations, providing a well-rounded understanding of the vehicle sales market. The findings highlight the complexities and nuances of pricing dynamics, offering practical implications for the automotive industry.

References

[1]: Chu, A., Wang, Y., & Zhao, Z. (2018). Price determinants in the used car market: An empirical study based on transaction data. Journal of Applied Economics, 20(2), 123-134.

[2]: Fan, J., & Li, R. (2015). Predicting used car prices using regression analysis. Automotive Economics Journal, 12(4), 345-367.

[3]: He, X., Wang, J., & Zhang, L. (2014). Feature engineering in automotive data analysis: Enhancing predictive accuracy. Data Science Review, 10(3), 211-225.

[4]: Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), 111-117.

[5]: Mookerjee, V., & Mannino, M. (2010). The role of seller reputation in car auctions: A study on eBay. Journal of Electronic Commerce Research, 11(1), 1-20.

[6]: Zhang, X., Li, H., & Wu, Y. (2019). Data preprocessing in predictive modeling: Techniques and implications. Machine Learning Journal, 24(1), 45-60.

[7]: Syed Anwar Afridi. (n.d.). Vehicle Sales and Market Trends Dataset. Retrieved from Kaggle: https://www.kaggle.com/saafridi/vehicle-sales-and-market-trends-dataset

Appendices

(All raw code and data set, image, scaler data has been included in the file)