CMPT 3830: Machine Learning
Work Integrated Learning-1

Project Report: Phase 1
Clustering-Based Vehicle Make Analysis for Cross-Selling Insights

In collaboration with

Submitted By:

Name	ID	Email
Manveen Kaur	3052214	mna723@norquest.ca
Sahil chand	3080196	Schand17@norquest.ca
Getachew Teila	3070114	gteila@norquest.ca
Love Maan	3050145	lsingh27@norquest.ca
Aashish Arora	3091126	aarora26@norquest.ca

Submission: Date: October 15,2024
Fall 2024

Contents

List of Figures:

List of Tables:

1. Project Phase:

Summarzing our previous phases, EDA has been successfully done on our data set. Steps include handling outliers , missing values , feature engineering and preparation of data for model(standardizing and encoding the data)

In this phase, we explored different machine learning models to cluster vehicles in the Go Auto dataset, focusing on models like K-means and DBScan. The goal was to find the best clustering model that groups vehicles based on their features, such as price, mileage, model year, and make.

We plan to begin by applying Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, which helped us retain only the most important information while simplifying the dataset for the models. Moreover methods like feature importance are kept in view for feature selection as by identifying the most influential features we can prioritize them during model training.

Model Selection:

K-means: Can be chosen for its simplicity and efficiency in clustering similar vehicles.
Initial results showed distinct clusters based on price and mileage, making it a good fit for understanding vehicle groupings.

OR

DB Scan: Can be selected for its ability to handle noise and find clusters of various shapes.
It was particularly useful for identifying outliers and niche vehicle segments that don't fit neatly into other clusters.

Model Evaluation:

We can use the Silhouette Score to measure how well the vehicles fit into their clusters.
Furthermore, the more methods taught in class would be used to evalate model based on different metrics.

Model Optimization:

- Hyperparameter Tuning:

For K-means, we are optimizing the number of clusters using the elbow method and silhouette analysis.

For DB Scan, we are looking at tuned parameters like epsilon (ε) and minimum samples to get the best balance between cluster density and noise handling.

- Cross-validation: We are also planning to apply cross-validation to ensure the models are robust and work well across different subsets of the data.

Techniques like grid search and random search are used to find the best hyperparameters.

Final Model Selection:

After evaluating both models, we will select the best-performing one based on metrics like the silhouette score and the business requirements of Go Auto.
The final step will involve fine-tuning the chosen model's hyperparameters and integrating the results into a recommendation system for vehicle groupings.

2. Team Members' Name with specific roles

NAMES	POSITIONS	ROLES
Aashish Arora	Model Evaluator	Model Performance Evaluation Validation and Testing Error Analysis Model Interpretability and Explainability Monitoring and Continuous Improvement & coding.
Manveen Kaur	Project Manager	Project Planning and Scope Management Scrum Master Team Coordination and Leadership Stakeholder Communication and Reporting Quality Assurance and Testing
Sahil Chand	Deadline Manager/ Code Developer	Progress Reporting and Documentation Risk Assessment and Deadline Performance Monitoring and Efficiency Optimization Writing Clean and Efficient Code Model Development and Implementation Performance Optimization and Scaling Continuous Learning and Technology Upgradation
Getachew	Lead Developer	Technical Leadership and Guidance Architecture Design and Implementation Code Review and Quality Assurance Integration of ML Models and Systems Troubleshooting and Problem-Solving Performance Optimization and Scalability Planning Mentorship and Skill Development for Team Members
Love Maan	Interpreter	Model Output Analysis and Interpretation Explaining Model Results to Stakeholders Bridging Communication Between Technical and Non-Technical Teams Holding Meetings Ensuring Model Transparency and Explainability

		Evaluation of Bias and Fairness in Predictions Collaborating with other teammates for Model Refinement.

3. Reporting Period: [Specify the reporting period, e.g., Month/Year to Month/Year]Key breakdown of each part submission.

Dates	Milestones	Details/Comment	Status
Sep 17	Team charter	We developed a Team Charter that defines our group roles, goals, and communication plans. Each member has specific responsibilities, such as data cleaning, model selection, and evaluation, while we work together on Go Auto's dataset to perform analysis and apply machine learning models	**Completed**
Sep 18 – Sep 24	Project Charter	We developed a Project Charter focused on applying clustering techniques to Go Auto's dataset. The main objective is to group vehicles based on factors like price, mileage, and age, providing insights for cross-selling strategies and inventory management. We outlined key tasks and data visualization, with a clear timeline and deliverables for each phase of the project	**Completed**
Sep 24 - Oct 07	Demo 1: Presentation	We described Go Auto's dataset, explained its structure, and identified key features like price, mileage, and model year. We conducted Exploratory Data Analysis (EDA) to clean the data, visualize trends, and handle missing values. The results from EDA helped uncover important patterns, such as the relationship between vehicle age and price, guiding our next steps in clustering.	**Completed**

4. Project Overview:

The aim of this project is to leverage machine learning techniques, specifically clustering, to analyze vehicle make based on key attributes such as price, mileage, and model year. This analysis will help Go Auto, a leading automotive retailer, optimize its inventory management and enhance cross-selling strategies. By identifying similarities across different vehicle brands, Go Auto can gain insights into which makes share common characteristics, enabling the company to offer more tailored recommendations to potential customers, ultimately improving both sales and customer satisfaction.

Problem Statement

Go Auto deals with a diverse range of vehicle make, each with varying attributes. Identifying patterns and similarities between these makes can be challenging but essential for improving cross-selling opportunities. The inability to cluster similar vehicles hinders Go Auto's ability to effectively recommend alternatives to customers, leading to potential missed sales opportunities. This project aims to address this challenge by using clustering methods to group vehicles based on shared attributes, helping Go Auto enhance their inventory and recommendation systems.

Solution Approach

To address the challenge of clustering vehicles based on attributes like price, mileage, and model year, we applied the following systematic approach:

1. Data Cleaning

Irrelevant columns such as listing_id, dealer_street, and listing_url were removed to retain only essential vehicle and dealer attributes for analysis.

2. Feature Engineering

A new feature, Age, was created by subtracting the vehicle's model_year from 2024, which helps capture a crucial aspect of the vehicle's life cycle for clustering. Transmission and price columns were also updated using mapping and logics to get rid of unwanted values.

3. Data Normalization

We applied Quantile Transformation to ensure numerical features like price and mileage were normalized. This normalization improved the clustering algorithm's performance by mitigating the effects of skewed data. Z score was used for methods of normal distribution.

4. Handling Duplicates and Missing Data

The dataset was checked for duplicates. Missing values in categorical features were filled using the mode, ensuring data completeness for clustering.

5. Encoding

Categorical variables such as make, model, and drivetrain_from_vin were converted into numerical values using Frequency and one hot Encoding, allowing machine learning algorithms to process them effectively.

6. Data Export

After all the necessary transformations and cleaning, the final dataset was exported and saved as CBB_Listings2g.csv, ready for clustering using algorithms like K-Means.

Future Steps

1. Cluster Validation and Optimization

After performing initial clustering, we plan to validate the results using methods like silhouette scores and elbow methods to determine the optimal number of clusters. We may experiment with different clustering algorithms, such as DBSCAN or more, to see which yields the best results for the dataset.

2. Dimensionality Reduction

To enhance computational efficiency and focus on the most critical features, techniques like Principal Component Analysis (PCA) would be applied. This would reduce noise and make the clustering process more effective.

3. Incorporating Additional Features
We would explore incorporating new features, such as geographic location or dealership-specific data, to refine the clustering analysis. This could provide more granular insights into regional preferences and trends.

4. Customer Segmentation Integration
In future iterations, we plan to integrate customer segmentation data to match specific customer profiles with clustered vehicle makes. This would allow Go Auto to tailor marketing and cross-selling strategies even further.

5. Recommendation System Development
Based on the clusters formed, we can develop a recommendation system that suggests similar vehicles to customers based on the features they value most, enhancing cross-selling opportunities.

5. Dataset

Vehicle Information	Dealership Information	Vehicle Listing Information	VIN-Based Attributes	Additional Data Points
• Make, Model, Year, Mileage, Price, MSRP (Manufacturer's Suggested Retail Price) • Vehicle Identification Number (VIN) and Style • Certified Status, Leather, Navigation, and Exterior Colo	• Dealer ID, Dealer Name, Location (City, Province, Postal Code) • Dealer Type and Stock Type (e.g., New, Used)	• Listing ID, URL, First Date on Market, Days on Market • Number of Price Changes and Price History	• Information extracted from VIN, such as Engine Type, Transmission, Drivetrain, and Fuel Type	• Distance to Dealer, Listing Dropoff Date, and Location Scor

Figure: Category Table for Data Attributes

Categorical attributes: listing_id, listing_heading, listing_type, listing_url, dealer_id, dealer_name, dealer_street, dealer_city, dealer_province, dealer_postal_code, dealer_url, dealer_email, dealer_phone, dealer_type, stock_type, vehicle_id, vin, uvc, make, model, series, style, certified, has_leather, has_navigation, exterior_color, exterior_color_category, interior_color, interior_color_category, drivetrain_from_vin, engine_from_vin, transmission_from_vin, and fuel_type_from_vin.

Numerical attributes : days_on_market, mileage, price, msrp, model_year, wheelbase_from_vin, number_price_changes, distance_to_dealer, location_score, price_analysis, and price_history_delimited.

Total Datapoints: 145,114 rows and 46 columns, representing vehicles listed for sale.
Attributes: The dataset includes both numerical and categorical features, which are essential for clustering analysis. Below is a breakdown of the types of attributes:

Numerical Attributes:
- `price`: Vehicle price, with a mean of 47,866 and a standard deviation of 90,702. Values range from 0 to 13,095,320.
- `mileage`: Mileage on the vehicle, ranging from 0 to 1,000,008 miles, with an average of 45,985 miles.
- `days_on_market`: Days the vehicle has been listed, with an average of about 47 days.
- `vehicle_id`: Unique identifier for each vehicle.
- `model_year`: Year of the vehicle's model, with most vehicles being from 2020 to 2024.

Categorical Attributes:
- `make`, `model`: The make and model of the vehicle, with 45 unique vehicle makes and over 700 different models.
- `dealer_name`, `dealer_city`, `dealer_province`: Information about the dealers, with most located in Edmonton, Alberta.
- `drivetrain_from_vin`, `engine_from_vin`, `transmission_from_vin`: Information extracted from the vehicle's VIN, detailing the drivetrain (e.g., AWD), engine type, and transmission.
- `certified`: Whether the vehicle is certified (binary, mostly 0).
- `exterior_color`, `interior_color`: Colors of the vehicle, with `black` being the most frequent for both.

Unique Features:
- `listing_id`: Unique identifier for each listing, ensuring there are no duplicate listings.
- `listing_type`: Denotes whether a vehicle has been sold or is still listed.
- `dealer_url`: Link to the dealer's website.

Data Issues and Processing:
- Missing Data: Some attributes, particularly in color categories and dealer-related columns, have missing values.
- Duplicates: No duplicate rows were found in the dataset.
- Data Transformation: Categorical variables were encoded, and the dataset was cleaned by replacing missing or inconsistent values (e.g., empty spaces were replaced with NaN).

	listing_id	listing_heading	listing_type	listing_url	listing_first_date	days_on_market	dealer_id	dealer_name	dealer_street	dealer_city
145109	8c4d2cd4-db92-11ee-ab06-77676d6b474c	2024 Volvo XC90 Recharge T8 Ultimate Bright Th...	Sold	https://www.volvocarsedmonton.com/inventory/ne...	3/5/2024 0:00	99	11132447	Volvo Cars Edmonton	1205 101 St SW	Edmonton
145110	8b564927-db92-11ee-8456-d37a8da2f6cb	2024 Volvo XC90 Recharge T8 Ultimate Bright Th...	Sold	https://www.volvocarsedmonton.com/inventory/ne...	3/5/2024 0:00	96	11132447	Volvo Cars Edmonton	1205 101 St SW	Edmonton
145111	474ecb38-0c14-11ef-a6d5-81e7286561b2	2024 Volvo XC90 Recharge T8 Ultimate Bright Th...	Sold	https://www.volvocarsedmonton.com/inventory/ne...	5/6/2024 0:00	22	11132447	Volvo Cars Edmonton	1205 101 St SW	Edmonton
145112	7208325d-e725-11ee-bf7d-151c90416979	2024 Volvo XC90 Recharge T8 eAWD PHEV Ultimate...	Sold	https://www.volvocarsedmonton.com/inventory/ne...	3/20/2024 0:00	71	11132447	Volvo Cars Edmonton	1205 101 St SW	Edmonton
145113	70f556a5-6bff-4836-97cc-c859d6b88f15	Pre-Owned 2024 Volvo XC90 Recharge Plug-In Hyb...	Sold	https://www.southgateaudi.com/used/Volvo/2024-...	7/20/2024 0:00	4	11204118	Southgate Audi	1235 101 Street SW	Edmonton

```
df.shape
```

```
(145114, 46)
```

Figure: Pandas Frame depicting tail of dataset along with shape.

5.1 Exploratory Data Analysis (EDA) Highlights:

We conducted a comprehensive Exploratory Data Analysis (EDA) to gain a deeper understanding of the dataset, identify key trends and patterns, and address potential data issues, such as outliers and missing values.

Conducted Comprehensive EDA on the Collected Data:

To start the EDA, we loaded the dataset and ensured all columns were visible for detailed analysis. We examined the structure of the data to detect missing values, duplicates, and inconsistencies. The steps we took include:

- **Dropping Uninformative Features**

We dropped the columns which might not be needed (dealer_email','dealer_id' , 'listing_first_date', 'days_on_market', 'dealer_type', 'dealer_street', 'series', 'dealer_url', 'exterior_color', 'interior_color', 'wheelbase_from_vin', 'listing_dropoff_date','listing_id','days_on_market', 'vehicle_id' , 'vin','certified', 'number_price_changes','price_history_delimited','location_score','price_analysis','listing_url ','dealer_phone','Column1','uvc').

Our problem statement did not require us to use these columns, instead of the exterior and interior color, the category columns for both were selected to ensure not losing any information.

- **Data Cleaning and Transformation**

Missing values in key categorical columns, such as exterior_color_category and interior_color_category, were filled using the mode. No duplicates were found during the check. Prices were adjusted to ensure no discrepancies, with msrp values of 0 replaced by corresponding price values, and prices exceeding msrp capped at the msrp level. Very low prices (below 10% of msrp) were replaced with the median price based on similar msrp groups.

Figure: PRICE VS MSRP (Before fixing the attribute)

Figure: Price Vs MSRP (after fixing)

- **Creating the age column**

We engineered a new feature, Age, by subtracting the model_year from 2024. This feature helps capture the age of vehicles, which plays a crucial role in clustering them based on their lifespan and depreciation trends.

- **Categorical data Encoding**

Transmission codes were mapped (e.g., '7' to 'M' for manual) to meaningful labels. For high cardinality columns like make and model, we applied Frequency Encoding, while for fewer categories like listing_type and drivetrain_from_vin, One-Hot Encoding was used to simplify data preparation.

- **Outlier Detection and Removal**

Outliers were detected using two methods: Z-score for normally distributed columns (e.g., model_year, make, and style), and the IQR method for skewed data like price, mileage, and distance_to_dealer. Outliers were removed to ensure data quality and consistency.

- **Data Standardization**

Using *StandardScaler*, we standardized numerical columns like price, mileage, and the newly created Age column, ensuring that all variables were on the same scale, crucial for clustering algorithms.

Identified Key Trends, Outliers, and Patterns:

During the EDA, we uncovered important trends and patterns in the dataset:

- **Pricing and Mileage Trends**:
  Vehicles with high prices generally showed high mileage, although certain luxury models defied this trend, with low mileage but high prices. Adjusting the prices allowed us to better analyze these patterns.

- **Categorical Data Insights**:
  Ford was the most frequent make in the dataset, highlighting potential inventory concentration. Automatic transmissions were the most common (A), aligning with current market trends.
- **Outliers**:
  Extreme values in pricing and mileage were identified, corrected, or removed during the outlier detection phase, making the dataset cleaner and more robust for analysis.
- **Visualisations**
  Different graphs helped us identify distribution of data and trends in different attributes of the data set which helped us reach conclusive outcomes regarding the kind of encoding to be used for them and what outlier process.

Discovered Insights:

Through our EDA, we uncovered several key insights:

- 'price_history_delimited' and 'price_analysis' are still being massaged regarding the aspect of a car being listed twice in dataset (Case of Re-listing the car).
- Another column ' distance_to_dealer' does not specifically mention about the points between which the distance id calculated
- Column1 was just an empty column with no values.
- Our figures attached still portray Outliers but they were discovered not to be actual/true outliers as some cars do have a mileage upto 200,000 which is pretty common and they cannot be put as outliers.
- By identifying and correcting price discrepancies, we can improve Go Auto's pricing strategy, helping them optimize for profitability while maintaining competitive pricing.
- Strong correlations between vehicle age (as calculated in the Age column) and mileage were observed. However, clustering will help detect outliers—such as older vehicles with lower mileage—that could be marketed as premium options.
- The dataset showed that most listings were concentrated in Edmonton, Alberta, suggesting that regional clustering could help Go Auto balance their inventory to better meet demand in other locations.
  Our EDA process revealed important trends, outliers, and patterns in the data, while ensuring that the dataset was properly cleaned and prepared for further analysis. By addressing issues like skewness and missing data, we ensured the dataset was optimized for clustering, and we gained valuable insights into key variables such as vehicle age and make.

5.2 Visualization:

1. Developing Interactive Visualizations to Represent EDA Findings

To analyze and present the findings from the Exploratory Data Analysis (EDA), we developed a series of interactive visualizations using Plotly, and static visualizations with Seaborn and Matplotlib. These visualizations helped uncover trends, relationships, and distributions in both categorical and numerical data, guiding insights for Go Auto's cross-selling strategies and inventory optimization.

- Box Plots:
  Box plots were used to identify the spread, median, and outliers of numerical variables such as price, mileage, and model_year. For example, the box plot for price by make allowed us to detect underpriced or overpriced brands, helping Go Auto in pricing strategies.
- Histograms and KDE Plots:
  Histograms were combined with Kernel Density Estimates (KDE) to show the distribution of continuous variables such as price and mileage. These helped reveal the concentration of listings in specific price and mileage ranges, highlighting vehicle segments that may be oversupplied or in demand.
- Bar Charts:
  Bar charts provided frequency counts of categorical variables, such as make, exterior_color_category, and interior_color_category. These visualizations helped assess which vehicle makes or colors were most common, giving insights into potential overstocked or undersupplied categories.
- Scatter Plots:
  Scatter plots visualized relationships between continuous variables, such as price and mileage. This helped detect clusters or trends, such as higher-mileage vehicles being priced lower, allowing Go Auto to identify potential pricing optimizations.
- Interactive Visualizations:
  Using Plotly, we developed interactive versions of box plots, bar charts, and scatter plots. These interactive features allowed stakeholders to explore data dynamically by zooming, hovering over data points to get more details, and filtering categories (e.g., filtering by vehicle make to see price distribution or mileage trends).

2. Visualizations Developed

- Box Plot: Price Distribution by Make:

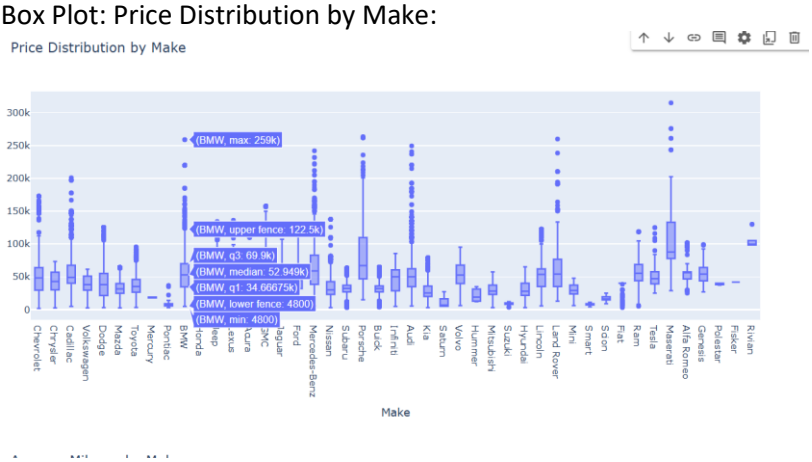


Figure: Boxplot depicting Price distribution by make.

This box plot showcased the price spread for different vehicle makes, highlighting outliers. This helped Go Auto identify which makes had more price variability, guiding potential discounting or marketing efforts.

- Bar Chart: Average Mileage by Make:

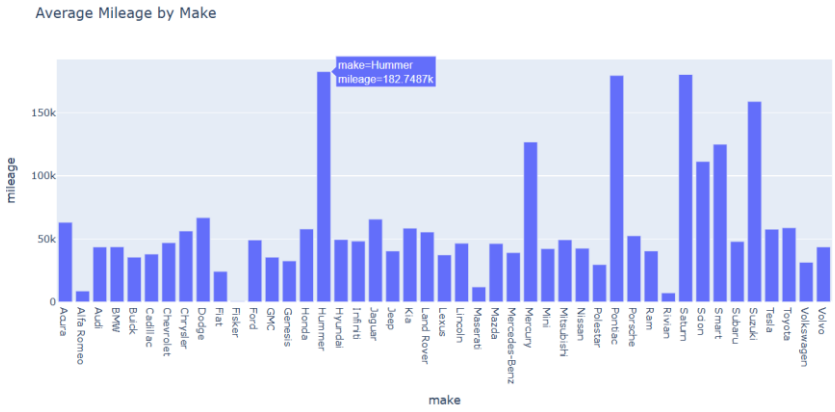


Figure: Bar Chart depicting Average Mileage by Make.

A bar chart visualized the average mileage for each vehicle make, helping assess the overall condition of different brands in Go Auto's inventory. Makes with higher average mileage might require targeted marketing efforts to move older inventory.

- Bar Chart: Vehicle Count by Make:

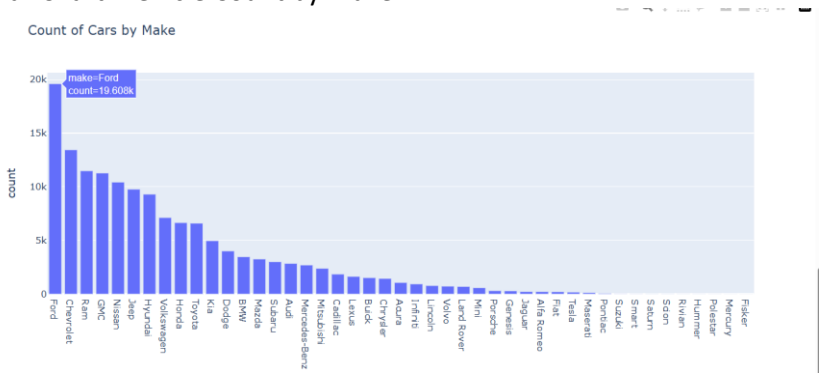


Figure: Bar Chart depicting Cars Count By Make

This chart showed the count of vehicles per make, allowing Go Auto to identify which brands were over- or under-represented in their inventory, thus guiding inventory management and cross-selling strategies.

- Scatter Plot: Price vs. Mileage:

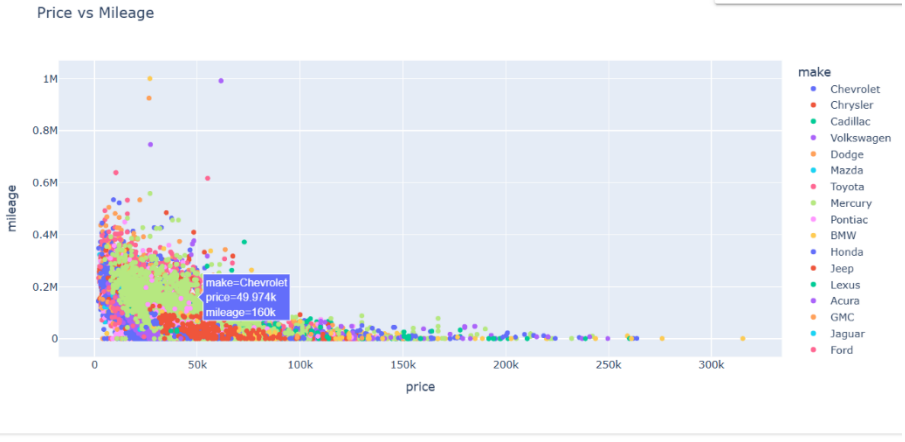


Figure: Scatter Plot Price Vs Mileage

- The scatter plot illustrated the relationship between vehicle price and mileage. It helped

visualize clusters where vehicles with higher mileage were priced lower, revealing opportunities for promotional campaigns targeting high-value, low-mileage vehicles.

- Frequency Plots:

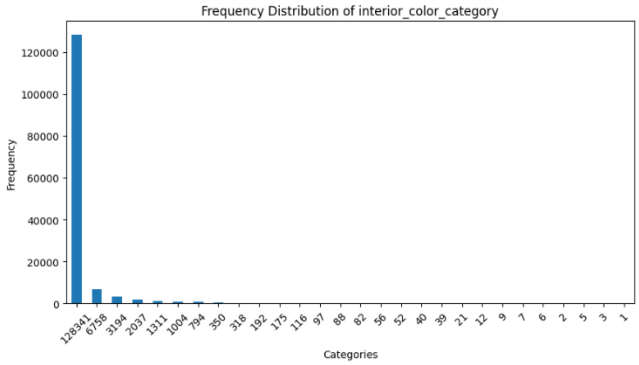


Figure: Bar Chart with Distribution of interior

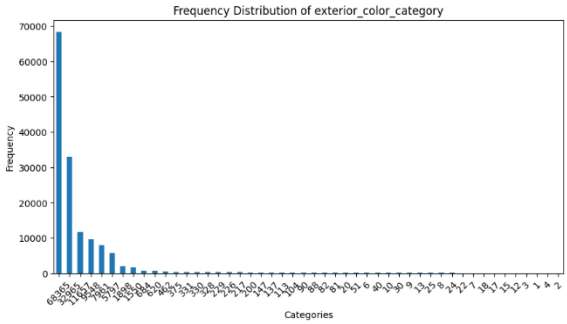


Figure: Bar Chart (Distribution of Exterior)

Frequency plots of categorical variables such as make, exterior_color_category, and interior_color_category provided insights into which vehicle features were most common, allowing Go Auto to focus on promoting certain categories more effectively.

3. Ensuring Visualizations Were Intuitive and Conveyed Actionable Insights

To ensure that the visualizations were clear, intuitive, and provided actionable insights:

- Use of Color and Legends:
  Visualizations were color-coded for clarity. For example, in scatter plots, different vehicle makes were represented by distinct colors, allowing for easy differentiation between data points.
- Descriptive Labels and Titles:
  Each chart was equipped with descriptive titles, axis labels, and legends, ensuring that the purpose and content of each visualization were clear. This allowed stakeholders to quickly understand the data without needing a deep technical background.
- Interactivity:
  Plotly's interactive features such as zooming, hovering, and filtering allowed stakeholders to dive deeper into the data, making it easier to explore relationships and trends dynamically. This interactivity encouraged stakeholders to explore different vehicle attributes and identify insights that would not be apparent in static visualizations.
- Actionable Insights:
  The visualizations were designed with business goals in mind. For example, the scatter plot showing price vs. mileage revealed opportunities for promoting vehicles with low mileage but competitive pricing. Similarly, box plots highlighting price outliers guided potential adjustments in pricing strategies.

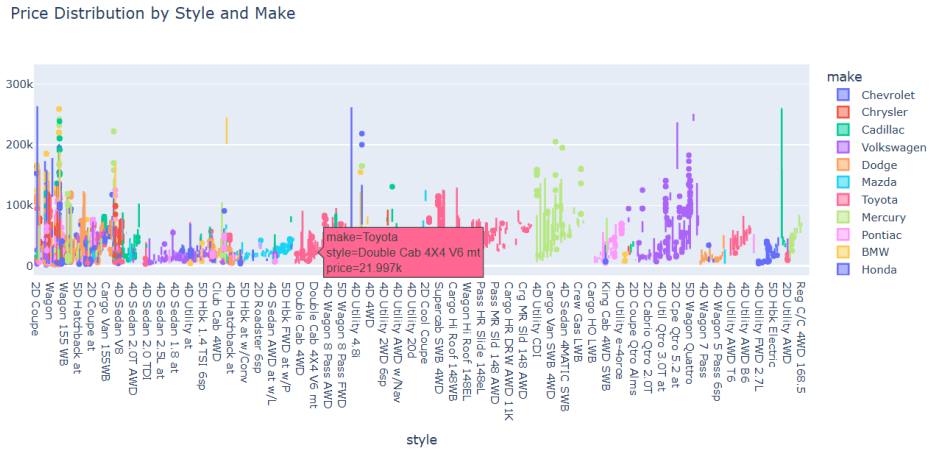


Figure: Price distribution by style and make

6. Challenges Encountered:

Throughout the project stage, we faced a range of technical and methodological obstacles, as well as some team-related challenges, all of which were overcome through working together. Nevertheless, these obstacles impacted the project's timeline.

Identifying Skewness and Outliers:

One of the primary obstacles was managing imbalanced data and selecting suitable techniques for detecting outliers. We applied IQR and Z-score techniques to eliminate anomalies. Columns were typically regarded as skewed and addressed with the IQR method if the mean was notably greater than the median. Nevertheless, the difficulty lied in the fact that eliminating approximately 35,000 anomalies may not completely reflect the actual scenario, especially as we are dealing with car information where exceptional figures (for instance, exorbitant prices) could be valid entries. Boxplots offered some insights, but additional thorough analysis is necessary to determine if these outliers indicate actual errors or legitimate data points.

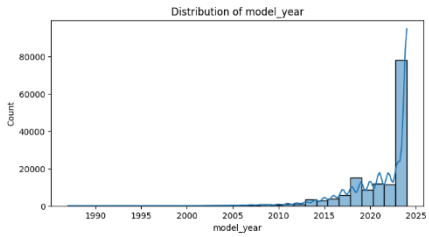


Figure: Showing distribution of model year

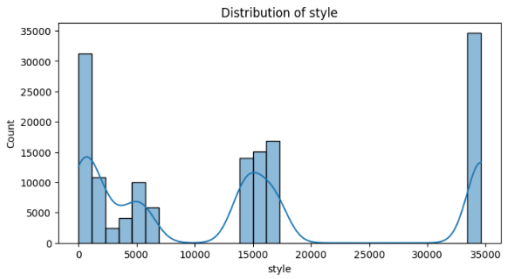


Figure: Showing distribution of style

Displaying Numeric Columns:

By utilizing Seaborn and Matplotlib, we generated boxplots and histograms to examine the spread of quantitative attributes such as price, model_year, make, mileage, and distance_to_dealer.An issue that arose was understanding these graphics within the framework of the dataset. For instance, histograms displayed the general price and mileage distribution, while boxplots identified outliers, including some that may not actually be anomalies because of the car industry's characteristics (e.g., luxury cars with high prices). Balancing the removal of outliers with preserving crucial vehicle data was an ongoing issue.

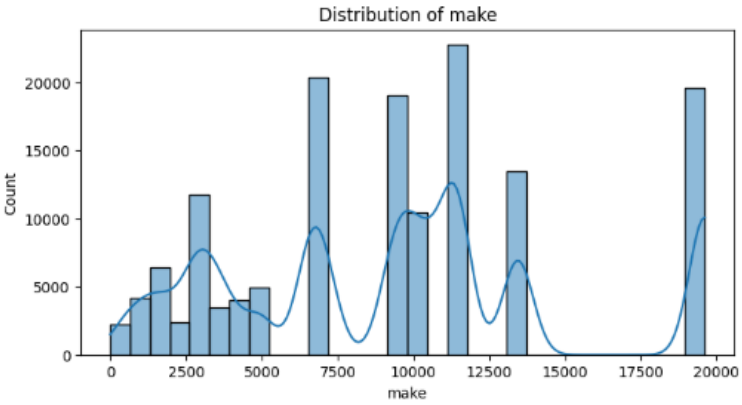


Figure: Distribution of Make

Problems with the price column:

There were discrepancies in the price column, with some prices higher than the msrp or unreasonably low. In order to solve this issue, we set limits on prices at the manufacturer's suggested retail price (msrp) and substituted extremely low prices (below 10% of the msrp) with the median price of vehicles in the same price range. This modification increased uniformity, but we needed to be careful not to unintentionally eliminate valid pricing inconsistencies.

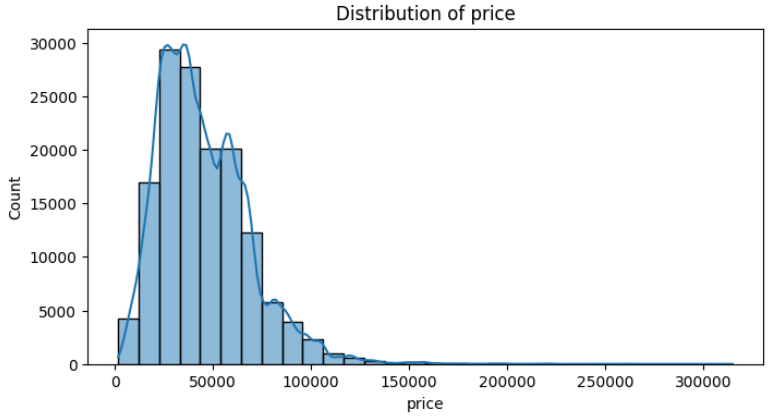


Figure: Distribution of Price

Categorizing Data in a Different Form:

Choosing how to encode information was another difficulty that arose. At first, One-Hot Encoding was thought of, but it produced a vast number of columns, causing the dataset to become very sparse and hard to manage. Frequency Encoding was utilized on columns such as make, model, style, exterior_color_category, and interior_color_category as a solution. This decreased the number of dimensions but could potentially over-simplify the information. Additional testing of encoding techniques will need to be conducted to determine if utilizing more advanced methods such as target encoding or feature hashing can enhance model accuracy without overwhelming the dataset.

Mapping of transmission.

The transmission_from_vin column contained unclear entries that needed to be mapped. We hypothesized that automatic transmissions (A) were indicated by more common values, whereas manual transmissions (M) were associated with less common values. While this idea follows the

typical patterns of the market, there is a possibility that certain data points were misclassified. Future work may be directed towards verifying this by utilizing more dealership data.

Distance to the dealership and vehicles that have been listed again:

The distance_to_dealer column was difficult to interpret in determining the proximity of a car to the dealership and what to do with relisted vehicles. Because certain vehicles may be relisted several times (showing extended inventory or changing demand), further filtering may be required to accurately analyze this information

Feature Engineering for Age:

The Age column was created by deducting the vehicle's model_year from the current year (2024). This characteristic was essential for categorizing vehicles according to their age, a significant factor in pricing and resale plans. Nevertheless, this straightforward method presupposes that all vehicles are included in the current year, which may not be true at all times. A system with greater flexibility may be necessary to regularly update the age.

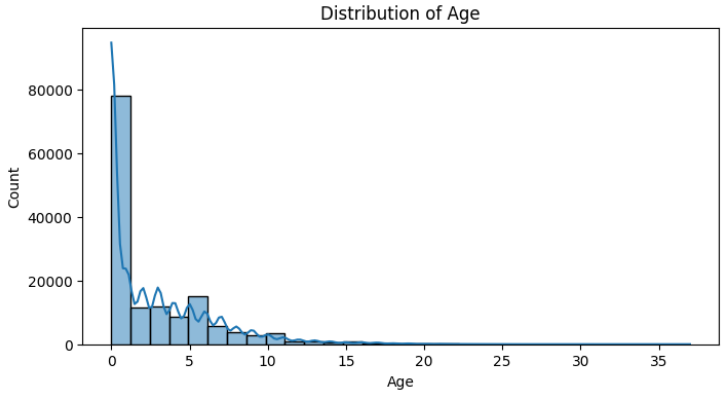


Figure: Showing Distribution Of Age

Problems with coordinating team members:

Apart from the technical difficulties, we faced problems within the team, specifically debates regarding the utilization of varying approaches. Some team members favored different encoding methods, while others preferred traditional data cleaning techniques. The variance in methodology caused delays in the project as we dedicated extra time to coordinating on a cohesive strategy. Consequently, various tasks, such as the submission of the report, were postponed past the initial deadline. Better communication and more specific delegation of tasks might have prevented these delays.

7. Stakeholder Engagement:

In the most recent meeting with stakeholders, we shared our Exploratory Data Analysis (EDA) results as part of a regular assessment for Go Auto. The conversation included both our project on clustering vehicle makes for improving cross-selling tactics and other issues being tackled by the team. The objective was to collect information on how to make the most of the dataset for different business goals and tackle the challenges we faced during the analysis.

The largest obstacle with the dataset up to now:

We pointed out the difficulty of dealing with categorical variables containing numerous distinct categories, like car brand and model, along with addressing outliers in the price dataset. Stakeholders recognized that the intricacy of these characteristics might affect the clustering model and stressed

the significance of taking into account business-specific anomalies (such as luxury brands) instead of just statistical anomalies.

Difficulties arise when using One-Hot Encoding for a large number of categories.
Stakeholders were especially curious about our approach to dealing with categorical variables that had a large amount of distinct values, such as make and model. We stated that One-Hot Encoding created an excessive number of columns, prompting us to switch to Frequency Encoding for a more easily manageable dataset.

Significance of the "Dealer Distance" Function:
A conversation took place regarding the possible implementation of the distance_to_dealer function. Stakeholders raised concerns about the value this feature provides for grouping vehicle brands, especially in relation to cross-selling tactics. We recognized that even though distance_to_dealer may not have as much importance in clustering vehicle brands, it could still be valuable for analyzing regional demand or improving logistics. Further analysis may be necessary to understand the complete impact of the feature.
Price Analysis Confusion:
There was continued uncertainty regarding the price_analysis attribute. Stakeholders had queries about how to interpret it and its role in the clustering process. Despite thorough conversation, there was no unanimous agreement on the optimal method to manage this characteristic. Further investigation is needed to ascertain the extent of its importance in the analysis.

Recommendations and insights from stakeholders:

Analysis at the level of Vehicle Identification Numbers to identify discrepancies.

Another suggestion was to look into discrepancies in records for both listings and sales of the same VIN number. Stakeholders proposed examining the frequency of a car being relisted or transferred among dealerships. This could aid in pinpointing vehicles that are difficult to sell and could benefit from targeted cross-selling tactics or price changes.

Additional points brought up by interested parties:

Categorizing Features into Groups:

Stakeholders had inquiries on our strategy for clustering categorical features like make, model, series, and style. Grouping these critical features in meaningful ways for obtaining cross-selling insights poses a challenge. They recommended taking into account the business environment while determining how to cluster these characteristics and possibly developing various models according to varying groupings.

Importance of features in clustering:

Stakeholders were curious about the features we considered most significant for clustering. We mentioned that our focus was on attributes such as price, mileage, model_year, and make, as these factors are strongly connected to customer preferences and inventory control. Nevertheless, we intend to investigate more characteristics like design and powertrain to enhance the groupings even more.

Selling patterns again.

A stakeholder raised a question about whether we had investigated vehicles being resold numerous times in a brief timeframe. Dealerships frequently transfer cars to different locations in order to enhance their sales opportunities, and analyzing these trends could assist in recognizing which vehicles may benefit from alternate cross-selling approaches. We recognized that although this was not a priority but definitely something to be considered ahead

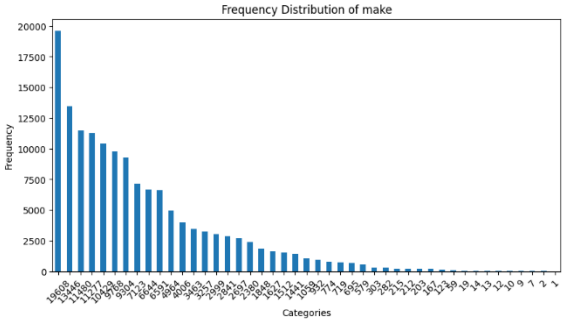
8. Lessons Learned:

During the project, various key lessons were learned, showcasing achievements and identifying areas for growth to inform future projects.

Managing data and manipulating features:
An important lesson learned was the significance of properly managing imbalanced data and extreme values. Applying both the IQR and Z-score techniques to eliminate outliers proved effective, however, it was discovered that outliers in vehicle data could sometimes be valid data points (e.g., expensive luxury cars). In the future, a more sophisticated method for detecting outliers will be required.
We also improved the process of feature engineering, specifically when developing the Age column. At first, the formula did not update automatically, but we fixed it by employing the today method to accurately calculate Age according to the present year. This modification boosted the precision of our clustering and overall analysis.

Moreover, there were difficulties with the price section. Initially, we set a price limit at the manufacturer's suggested retail price and substituted extremely low prices with the median value of comparable vehicles. Nevertheless, we came across instances in which both the price and MSRP were zero. To tackle this, we imposed a specific rule to replace these instances with the median value of the column. This maintained consistency throughout the dataset, but we acknowledge the necessity for additional improvements when dealing with special pricing instances.

Methods of encoding information:
Another thing we learned was how difficult it can be to encode categorical data. In the beginning, One-Hot Encoding resulted in an excessive number of columns, causing challenges in dataset management. As a resolution, we opted for Frequency Encoding to streamline the dataset, which underscored the necessity of choosing encoding methods that strike a balance between accuracy and usability



Visualizations and Identifying Outliers:
Using Seaborn, Matplotlib, and Plotly to generate visualizations aided in our enhanced comprehension of the data. Boxplots, scatter plots, and histograms gave us useful information, but we discovered that certain visualizations, like bar charts with numerous

distinct categories, can become messy. In the future, our main goal is to create visualizations that are clearer and more concise, especially when working with high-cardinality variables, to enhance the accessibility and informativeness of the analysis.

Enhancing team collaboration and optimizing work processes:

One major takeaway from this project was the necessity for enhanced team coordination and workflow organization. At the beginning of the project, disputes regarding data management and approach hindered our advancement, resulting in postponed submissions. In order to avoid similar problems in the future, we have put in place more rigid regulations for team cooperation, including consequences for failing to meet deadlines. Furthermore, we discovered that addressing technical obstacles, like effectively managing computational resources for interactive visualizations, could have been improved through task division and workflow optimization. If we had designated specific tasks earlier in the project, it would have helped us meet deadlines better and minimized duplication of efforts. This framework will guarantee team members stay synchronized, responsible, and on course to achieve project objectives.

9. Future Recommendations:

Some more refined areas for improvement that align closely with our work and problem statement regarding vehicle clustering :

1. **Exploring More Advanced Clustering Algorithms:**

While basic clustering techniques like K-Means provide valuable insights, they come with limitations, especially when dealing with data that doesn't form clear, spherical clusters. In future projects, it would be beneficial to explore more advanced clustering algorithms, such as:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This method is particularly useful for datasets with noise or outliers and can identify clusters of varying shapes.

Hierarchical Clustering: This technique could allow us to build a hierarchy of clusters, making it easier to visualize relationships between smaller and larger clusters, which could be useful for comparing vehicle makes.

2. **Using Dimensionality Reduction Techniques:**

Working with many features can introduce redundancy or noise into the dataset, which may hinder the performance of clustering algorithms. To improve this, applying dimensionality reduction techniques could help:

Principal Component Analysis (PCA): This technique would reduce the number of features while retaining the most important information, leading to more efficient clustering and potentially revealing hidden patterns.

3. **Improving Collaboration and Version Control:**

While we collaborated effectively as a group, there is room to enhance our workflow:

Version Control with Git: Using tools like Git and GitHub for version control would improve collaboration by allowing us to track changes, avoid conflicts, and manage different versions of our codebase.

Automated Documentation: Implementing automated documentation tools to track changes and steps taken by each team member would make it easier to maintain clear communication and consistency throughout the project lifecycle.

10. Impact on the Community:

Our project has the potential to bring positive change to the automotive industry and the broader community. By clustering vehicle makes based on key attributes such as price, mileage, and age, we can improve the efficiency of vehicle management, increase customer satisfaction, and even support more sustainable practices in the auto sales sector. The key impacts of the project include:

1. **Improved Vehicle Inventory Management:**

The clustering model helps dealerships like Go Auto optimize their inventory by identifying similar vehicle makes and models. This means that dealerships can better organize their inventory, ensuring that similar vehicles are grouped together, making it easier for customers to find what they're looking for. By grouping similar vehicles, dealerships can also identify which vehicles are in high demand and make data-driven decisions on which models to stock more frequently.

2. **Enhanced Customer Experience:**

One of the major impacts of this project is improving the experience for customers. By using the clusters, dealerships can offer more personalized recommendations, suggesting vehicles that closely match a customer's preferences. This improves customer satisfaction as buyers are more likely to find vehicles that meet their needs, saving them time and effort in their search. Additionally, customers can have access to better pricing insights, as they can easily compare similar vehicles in a specific cluster.

3. **Empowering Local Communities:**

Local dealerships can benefit greatly from the insights provided by this project. By using clustering techniques to optimize their offerings, smaller, local businesses can become more competitive in the market. This, in turn, supports the local economy by helping these businesses thrive and continue providing jobs within the community.

4. **Data-Driven Decision Making for Consumers:**

For the broader community of car buyers, our project provides better access to data-driven insights. Customers can use the results of the clustering model to compare vehicles based on real-world data, making more informed decisions about their purchases. This empowers consumers to find the best value for their money, leading to smarter and more satisfying buying decisions

11. Project Conclusion:

In conclusion, this project successfully achieved its goal of clustering vehicle makes based on important factors like price, mileage, and age. We thoroughly cleaned and prepared the data, ensuring it was ready for analysis. The exploratory data analysis (EDA) gave us valuable insights into the relationships between different vehicle features, which helped guide our clustering process. By grouping similar vehicles, we provided dealerships with a tool to better organize their inventory and offer personalized recommendations to customers. This not only improves the efficiency of sales but also enhances customer satisfaction. Additionally, the project went beyond its original goals by identifying ways to promote vehicle options and laying the groundwork for future improvements,

such as using more advanced clustering methods and real-time data. Overall, the project was a success, delivering practical results that can help dealerships operate more effectively and benefit the community.

12. Acknowledgments:

We would like to extend our heartfelt thanks to our instructor, Md Mahbub Mishu for their invaluable guidance and support throughout this project. Their feedback and advice were instrumental in helping us navigate the challenges and stay on track, ensuring the successful completion of our work.
We would also like to sincerely thank Tiago from Go Auto for his insightful contributions and industry expertise. His input provided us with a deeper understanding of the automotive market, which greatly informed our analysis and the goals of the project.
Also Ms.Kaylee, From the WIL department of Norquest College, for giving us this wonderful opportunity to work on this project.
We are grateful for the assistance of all team members for their dedication and hard work in each phase of the project. Their collaboration was essential to the success of our efforts.

13. Appendices:

Scrum reports showing two sprints have been attached below
Sprint 1:September 10 – September 23,2024
Sprint 2: September 24 – October 7,2024

14. References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12, 2825-2830.
   https://scikit-learn.org/stable/

2. McKinney, W. (2010). *Data structures for statistical computing in Python*. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56). SciPy.
   https://pandas.pydata.org/

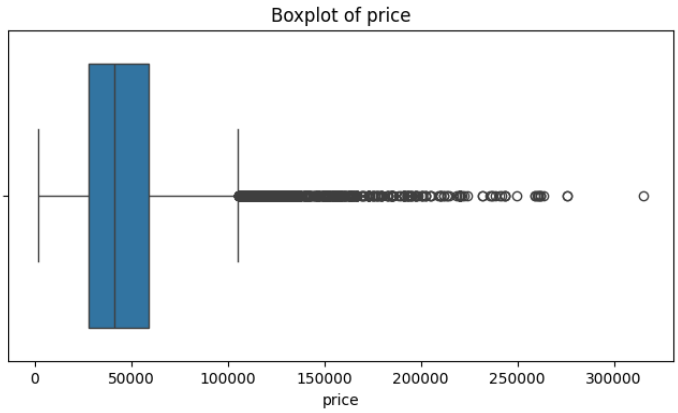
3. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2020). *SciPy 1.0: fundamental algorithms for scientific computing in Python*. Nature Methods, 17(3), 261-272.
   https://doi.org/10.1038/s41592-019-0686-2

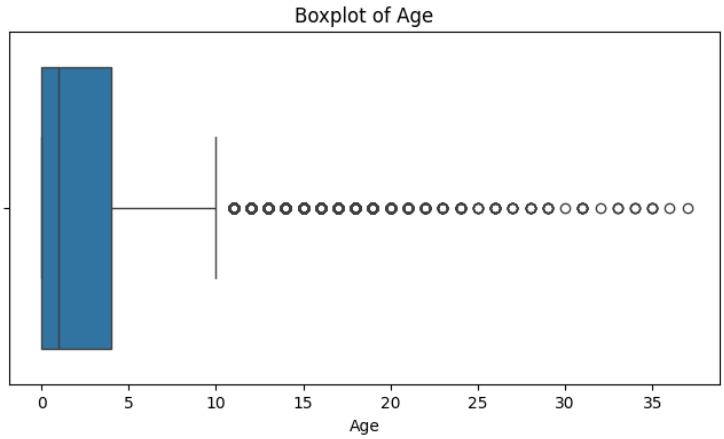
4. Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90-95.

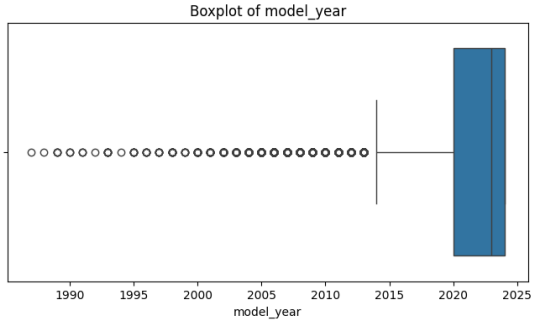
5. Pedregosa, F., Gaël, V., & Gramfort, A. (2011). *Preprocessing data. In Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

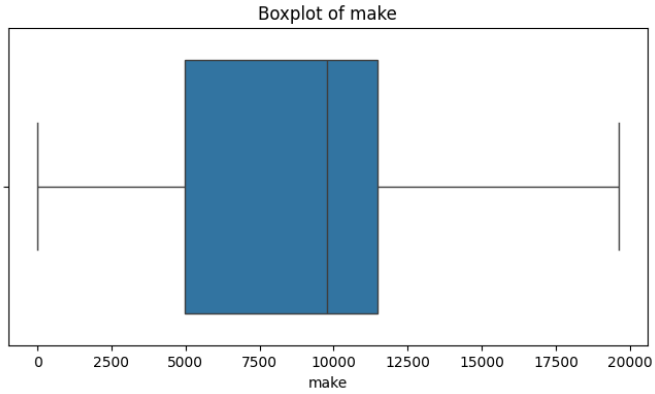
https://scikit-learn.org/stable/modules/preprocessing.html

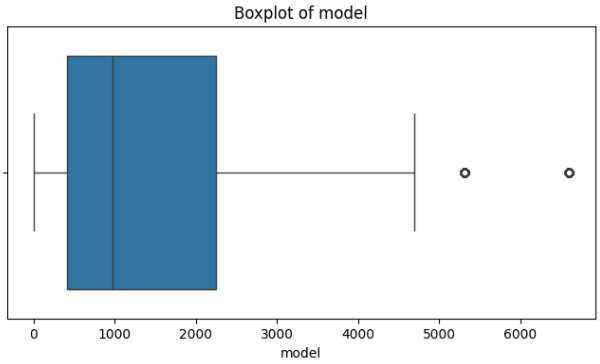
6. Plotly Technologies Inc. (2015). *Collaborative data science*. Plotly. https://plotly.com/python/

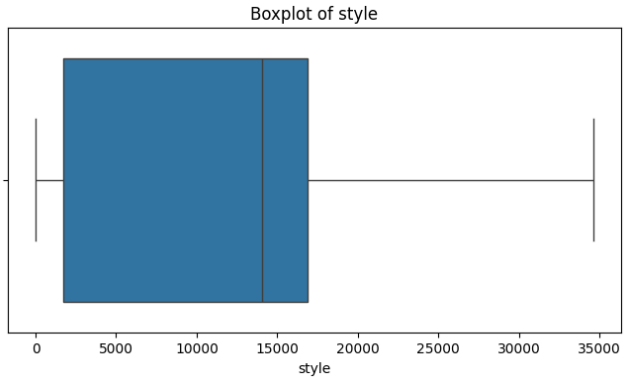




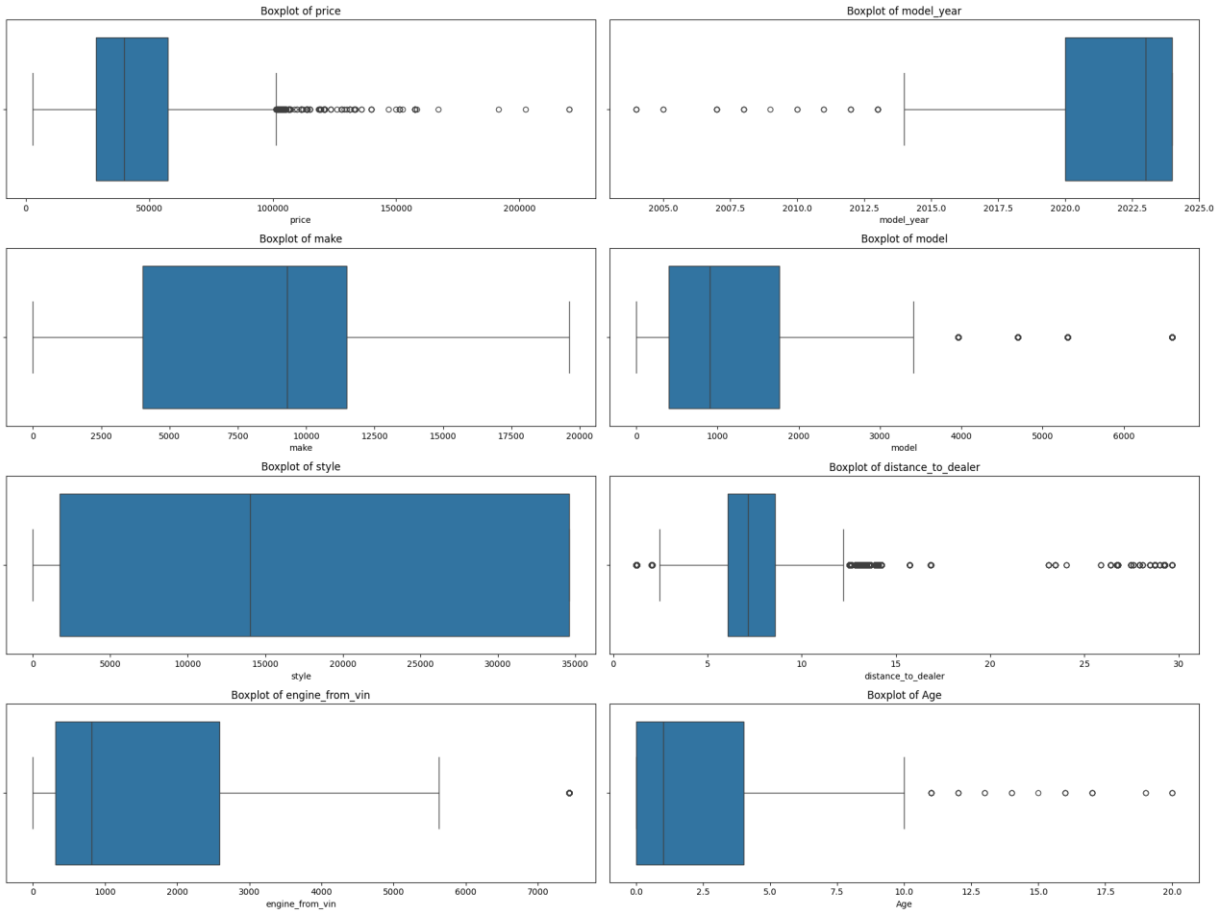








FIGURES SHOWING BOX PLOTS BEFORE REMOVING OUTLIERS



FIGURES AFTER REMOVING OUTLIERS