

CMPT 3830: Machine Learning

Work Integrated Learning-1

**Project Report: Phase 1**

**ClusterCatalyst**

**In collaboration with**



**Submitted By:**

|  |  |  |
| --- | --- | --- |
| Name | Student ID | Email |
| Sirjana Chauhan | 3109309 | schauhan09@norquest.ca |
| Rajinder Kaur | 3100808 | rkaur808@norquest.ca |
| Jasmeet Kaur | 3103824 | jkaur824@norquest.ca |
| Celine PANICKER | 3098918 | cpanicker@norquest.ca |

**Submission: Date: [27 February 2025]**

**Winter 2025**

**Table of Contents:**

A table of content is required with section numbers, names with correct page numbers. An example is shown below.

Contents

[List of Figures: 3](#_Toc176599393)

[List of Tables: 4](#_Toc176599394)

[1. Project Phase: 5](#_Toc176599395)

[2. Team Members’ Name with specific roles 6](#_Toc176599396)

[3. Reporting Period: [Specify the reporting period, e.g., Month/Year to Month/Year] Key breakdown of each part submission. 6](#_Toc176599397)

[4. Project Overview: Overview with the problem statement and solution approach you followed. 7](#_Toc176599398)

[5. Dataset 8](#_Toc176599399)

[5.1 Exploratory Data Analysis (EDA) Highlights: 9](#_Toc176599400)

[5.2 Visualization: 11](#_Toc176599401)

[6. Challenges Encountered: 17](#_Toc176599402)

[7. Stakeholder Engagement: 18](#_Toc176599403)

[8. Lessons Learned: 19](#_Toc176599404)

[9. Future Recommendations: 19](#_Toc176599405)

[10. Impact on the Community: 20](#_Toc176599406)

[11. Project Conclusion: 21](#_Toc176599407)

[12. Acknowledgments: 21](#_Toc176599408)

[13. Appendices: 22](#_Toc176599409)

[14. References 24](#_Toc176599410)

# List of Figures:

Figure 1 Boxplot of Price (Before and After Outlier Removal)

# Figure 2 Line Graph of Price Trend Across Model Years by Stock Type

# Figure 3 Voilin Graph of Mileage Distribution by Stock Type

# Figure 4 Bar Graph of Vehicle Age Distribution

# Figure 5 KDE plot of Price Distribution by Drivettrain

# Figure 6 Heatmap of Feature Correlation Heatmap

# Figure 7 Bar Chart Average Mileage per Make

# Figure 8 Histogram of price Distribution by Brands

# Figure 9 Bar chart of Average Mileage by Fuel Type

# Figure 10 Pie Chart of Market Share of Top 10 Vehicle makes

# List of Tables:

# Table 1 Reporting period

# Project Phase:

Here, we have problem in which we are utilizing clustering techniques to identify

similarities and differences between vehicle makes based on factors like price, mileage, and age. Currently, we are focusing on building clustering model.

**Present Accomplishments:**

**Data Cleaning & Preprocessing:**

The project team handled data inconsistencies and cleaned up any missing information as well as duplicate cases.

**Exploratory Data Analysis (EDA):**

A comprehensive analysis of price, mileage, year and model distribution characteristics took place.

Vehicle sales behavior patterns became visible through pattern recognition using visual tools.

**Encoding Categorical Variables**:

Applied one-hot encoding for categorical variables with a limited number of categories.

Frequency encoding was implemented for variables with numerous categories.

**Future Accomplishments:**

**Exploring different machine learning models**

* We have gone through different clustering techniques such as K-Means, Hierarchical Clustering, and DBSCAN.
* We found that K-Means clustering is best in this case to grouping similar vehicle makes.
* Doing PCA to select the features

**Choosing a model:**

* We will check scalability. So, the model can handle large datasets.
* Clusters should be useful and meaningful based on our problem of Go Auto dataset.
* Use different methods to determine clustering effectiveness.

**Applying Model**

* We have already done preprocessing on dataset by handling missing values and normalizing numerical columns.
* Now, we will implement K-Means algorithm to cluster vehicles based on their features.

**Model Evaluation:**

* In this, we are measuring clustering effectiveness.
* Also, we are Analyzing cluster distributions to get groupings of vehicle makes.

**Model Optimization:**

* In this, we are using different testing techniques to improve clustering.
* We will adjust the number of clusters based on our situation.
* To show our findings, we will use visuals.

# Team Members’ Name with specific roles

|  |  |
| --- | --- |
| Task | Owner |
| Data Collection & Cleaning | Rajinder Kaur |
| Exploratory Data Analysis | Sirjana Chauhan, Jasmeet Kaur |
| Encoding Categorical Variables | Celine Panicker, Rajinder Kaur |
| Demo-1 (EDA Presentation) | All Team Members |
| Phase-1 report (EDA Results) | All Team Members |
| Clustering Model development | Sirjana Chauhan, Rajinder Kaur, Jasmeet Kaur |
| Model Evaluation & Improvements | All Team Members |
| Demo-2 (Model Presentation) | All Team Members |
| Phase – 2 Report (Model Insights) | All team Members |
| Final Visualization in Looker Studio | Celine Panicker, Sirjana Chauhan, Jasmeet Kaur |
| Final Report Submission | All Team Members |
| Final Presentation to Go Auto | All Team Members |

# Reporting Period:

# Table 1 Reporting period

|  |  |  |
| --- | --- | --- |
| Phase | Timeframe | Tasks Completed |
| Data Collection & Preprocessing | January 2025 (Week 1 to Week 2) | Gathered and cleaned vehicle data (price, mileage, age, listing type).Handled missing values.Encoded categorical variables, including frequency encoding for vehicle makes.Scaled numerical features (price, mileage, age). |
| Clustering Analysis | January 2025 (Week 3 to Week 4) | Applied the K-Means clustering algorithm.Determined the optimal number of clusters using the elbow method.Analyzed results and visualized clusters based on price, mileage, and age.Generated preliminary insights for vehicle comparison. |
| Insight Generation & Recommendations | February 2025 (Week 1) | Analyzed cluster characteristics and identified patterns.Developed recommendations for cross-selling similar vehicles.Presented findings with visualizations (scatter plots). |
| Client Demo & Feedback | February 2025 (Week 2) | Presented Demo 1 to the client.Collected feedback on project (listing types, encoding, market trends). |
| Post-Demo Refinements & Next Steps | February 2025 (Week 3, Ongoing) | Plan to restructure dataset with separate Active and Sold columns.Reevaluate encoding methods. |

# Project Overview: Overview with the problem statement and solution approach you followed:

# Overview of the problem statement: The task is to identify similarities and differences between vehicle makes (brands such as Toyota, Honda, Ford, etc.) using clustering techniques.

# Real-World Use Case

# Imagine a customer comes to a dealership looking for a Toyota Camry, but it's out of stock. The dealership can use your analysis to recommend a Honda Accord or Nissan Altima because they are in the same cluster (similar price, mileage, and age).

# The primary factors for this clustering include:

# Price: The cost of vehicles from various brands.

# Mileage: How much the vehicles have been driven.

# Age: The year of manufacture of the vehicles.

# The goal of the project is to group vehicle makes based on these attributes, which will provide insights into how different brands compare with one another. Additionally, the clustering model will offer recommendations to dealerships for cross-selling similar vehicles from different brands, which can help with inventory management and customer satisfaction.

# Solution Approach

# Data Collection & Preprocessing The first step involved gathering relevant data from the client’s inventory system. This included vehicle price, mileage, age, and listing type (Active or Sold). Data preprocessing was performed to clean and prepare the dataset for analysis:

# Handling missing values: Filling or removing missing data points.

# Encoding categorical variables: We used frequency encoding for certain features like vehicle make, which will be revisited based on client feedback.

# Normalizing the data: To ensure that all features (price, mileage, age) contribute equally to the clustering process.

# Clustering Analysis We applied clustering techniques to group similar vehicles together. The clustering method chosen was K-Means because of its efficiency and scalability for this dataset. We used the following steps:

# Feature selection: we are doing PCA to find Features. However , focused on the main variables - price, mileage, and age.

# Choosing the number of clusters: Based on the dataset size and analysis of the elbow method, we determined an optimal number of clusters for vehicle makes.

# Insights Generation & Recommendations After performing the clustering, we will analyze the groupings to understand the similarities and differences between the vehicle makes. For example:

# Vehicles with similar price points will be clustered together, helping dealerships identify which brands could be cross-sold.

# We also visualized the results using scatter plots to show the relationships between price, mileage, and age.

# Feedback Incorporation After Demo 1, we received valuable client feedback:

# Listing Type: The client requested that the listing type be split into two columns (Active and Sold) rather than using a single categorical variable.

# Frequency Encoding: The client questioned why we used frequency encoding, suggesting a re-evaluation of encoding methods.

# Reevaluate Encoding: The client suggested for the number of features sjould not exceed more. Unfortunately, we did not check the number of features. The number od feature are 37 in total after encoding.

# Based on this feedback, we plan to modify our approach:

# Split the listing type into Active and Sold columns.

# Reassess the encoding strategy to avoid unnecessary complexity. We are planning to reduce features as much as possible after encoding. The total features are 37 after encoding. We will reapply the Label encoding instead of Feature encoding if it will help in reducing the features.

# Adjust our clustering analysis to account for market changes, especially the post-COVID price surge.

# Dataset

# The dataset, compiled by Go Auto’s Business Intelligence Team using Canadian Black Book (CBB) APIs, from Edmonton dealerships over the past 30 days.

* The dataset contains information about vehicles sold in **Edmonton**.
* It is used to analyze sales patterns and recommend vehicles from different brands to dealerships.

# The dataset is cleaned and preprocessed for ****clustering analysis****.

# Key Data Components:

# Vehicle Info: Make, Model, Year, Mileage, Price, Features, Engine, Fuel Type

# Dealership Info: Name, Type, Location, Contact Details

# Sales & Listing Data: Stock Type (New/Used), Price History, Days on Market, Price Change.

* The dataset has 145114 datapoints and 46 attributes.
* The dataset has numerical columns and categorical columns.
* **Size:** 145,114 rows, 46 columns.
* **Types of Attributes:**
* dealer\_city categorical
* dealer\_postal\_code categorical
* stock\_type categorical
* vin categorical
* mileage numerical
* price numerical
* msrp numerical
* model\_year numerical
* make categorical
* model categorical
* series categorical
* style categorical
* exterior\_color categorical
* interior\_color categorical
* wheelbase\_from\_vin numerical
* drivetrain\_from\_vin categorical
* engine\_from\_vin categorical
* transmission\_from\_vin categorical
* fuel\_type\_from\_vin categorical

# 5.1 Exploratory Data Analysis (EDA) Highlights:

The team performed EDA procedures on Go Auto's database with information about vehicles and outlets. The main objective was to detect patterns in the data and search for anomalous values to ready the data for clustering.

**Data Cleaning & Preprocessing:**

• Checked for duplicates and found zero duplicate rows.

• Identified imputed them using mode (most frequent value).

• Applied data type conversion, converting object columns into categorical values for better processing.

**Outlier Detection & Handling:**

• Used Interquartile Range (IQR) method to identify extreme values in numerical columns like price and mileage.

• Applied capping techniques to replace extreme outliers beyond acceptable bounds.

**Figure 1 Boxplot of Price (Before and After Outlier Removal)**

**Feature Selection:**

From the 46 original columns, selected 20 most relevant features, including:

* Key numerical variables: price, mileage, MSRP, model year
* Categorical features: make, model, drivetrain, fuel type

**Key Trends, Outliers, and Patterns**

• Prices for used cars continue to increase because both buyer interest and asset value loss rates remain low.

• New car prices dropped during the period from 2014 through 2016 yet the underlying reason behind this trend is unknown.

• FWD leads the market but 4WD and RWD hold positions throughout different price categories.

• BMW alongside Merceds-Benz reach their highest price point at $120,000 yet Nissan and their economy vehicles stay at a lower end of the spectrum.

**Correlation & Feature Importance:**

• MSRP shows an extensive statistical relationship (0.80) with price values which explains its essential role in establishing vehicle prices.

• The statistical analysis reveals a moderate relationship between vehicle dimensions and pricing (0.60) because car size impacts price points.

• The data indicates that vehicle price remains unaffected by mileage therefore other elements govern pricing.

**Insights Discovered:**

Through clustering dealerships can choose vehicles for recommendations by grouping them based on their price ranges alongside mileage and drivetrain compatibility.

• The peak driving efficiency of CNG cars stems from commercial-oriented drivers but electric and hydrogen vehicles demonstrate the lowest gas efficiency because they remain relatively new.

• Ford together with Chevrolet along with Ram controls the market strategies for inventory and sales which determines dealership stock decisions. The research results will lead to improved feature engineering together with clustering approaches that enhance decision-making effectiveness.

# Visualization:

Our exploratory data analysis findings required visualization through interactive and static elements that illustrated donation hotspots as well as high-traffic routes alongside distribution patterns. The visualizations served to generate plain understandings which stakeholders could use for taking decisions.

Our interactive visualizations were generated through Matplotlib together with Seaborn and Plotly which enabled users to observe trends and patterns in the data. The selection of visualizations for the dataset relied on their specialization in representing distinct dataset elements.

# We applied simplicity along with real-time capability and business-oriented analytics to develop visualizations that helped decisions requiring simplicity and usefulness.

# Clear and Understandable Visuals

# The appropriate visualization types included bar charts for categorization and scatter plots for pattern recognition as well as heatmaps for relationship analysis.

# Titles and labels with legends were incorporated into the visual representation for better interpretation.

# We used colored shading to identify patterns such as dark hues representing greater sales levels.

# Visualizations and Actionable insights:

# Figure 2 Price Trend Across Model Years by Stock Type

# Used car prices have steadily increased, showing consistent demand and slow depreciation.

# New car prices dropped from 2014 to 2016, possibly due to high production or market

# changes.

# Figure 3 Mileage Distribution by Stock Type

# A diagram of a mileage distribution AI-generated content may be incorrect.

# The distribution of used car miles extends . The mileage of used vehicle is extremely higher.

# The data for new vehicles shows 0 mileage because these cars have minimal kilometers

# Figure 4 Vehicle Age Distribution

# A graph of a vehicle age distribution AI-generated content may be incorrect.

# Vehicle Age: Majority of vehicles are 1 year old or newer, indicating high turnover. Older vehicles are still available but gradually decrease in number.

# Figure 5 Price Distribution by Drivettrain

# 

# A graph of a number of colored lines AI-generated content may be incorrect.

# FWD: Highest peak, indicating it’s the most common drivetrain.

# 4WD & RWD: Lower, wider curves, showing they are less frequent but spread across various price ranges.

# Price Range: Higher density for FWD doesn’t mean it’s cheaper, just more prevalent in that range.

# Figure 6 Feature Correlation Heatmap

# A diagram of a heatmap AI-generated content may be incorrect.

# MSRP & Price (0.80): Strong correlation; MSRP significantly affects price.

# MSRP & Wheelbase (0.60): Larger vehicles tend to have higher MSRPs.

# Mileage & Price (Weak): Price isn’t solely determined by mileage; other factors matter.

# Model Year & Mileage (Weak): New cars don’t always have low mileage, and older cars aren’t always high-mileage.

# Figure 7 Average Mileage per Make

# A graph of a number of miles per make AI-generated content may be incorrect.

# High-Mileage Brands: Hummer, Saturn, and Scion show more usage.

# Low-Mileage Brands: Alfa Romeo, Fisker, Rivian, and Porsche likely have newer or limited-use cars.

# Luxury vs. Economy: Luxury brands (BMW, Mercedes, Lexus) have moderate mileage, while economy brands (Toyota, Honda, Ford) show higher mileage.

# Trends: High mileage suggests durability, while low mileage indicates limited use or quicker resale.

# Figure 8 price Distributation by Brands

# A group of graphs showing different colors AI-generated content may be incorrect.

# Price Range: Most brands peak between $30,000–$60,000.

# Nissan: Right-skewed distribution, with most cars being lower-priced.

# Chevrolet, Ford, and Ram: Multiple peaks, indicating varying models or trim levels.

# GMC: Wider spread, suggesting both budget and high-end models.

# Luxury Models: Some brands show a small peak near $120,000, likely for luxury or high-performance vehicles.

# Figure 9 Average Mileage by Fuel Type

# A graph of a mileage by fuel type AI-generated content may be incorrect.

# CNG Vehicles: Have the highest average mileage, likely due to commercial use.

# Electric & Hydrogen Cars: Have the lowest mileage, possibly due to being newer models.

# Clustering Insight: This data helps in clustering vehicles by fuel efficiency and aids dealerships in cross-brand recommendations.

# Figure 10 Vehicle Count By Make

# 

# Leading Brands: Ford (18.6%), Chevrolet (12.7%), and Ram (10.9%) hold the largest number count in dataset.

# Notable Competitors: GMC, Nissan, Jeep, and Hyundai also have significant appearance in data.

# Smaller Players: Toyota, Honda, and Volkswagen maintain a smaller yet impactful presence.

# Key Takeaway: Ford, Chevrolet, and Ram lead sales, shaping inventory and marketing strategies.

# Challenges Encountered:

* **Missing Data:** Some attributes had missing values.

series = 825

exterior\_color = 6049

interior\_color = 51663

**Handling Missing Data:** we used Mode imputation to replace missing values to all three categorical column.

We utilized mode imputation techniques for categorical features because it replaces absent data with the most common entry thus maintaining data distribution while avoiding systematic bias. Ready-made though computationally cheap and effective for data which is consistently missing at random.

* **Outliers Detection:** Outliers were detected in our dataset.

Number of outliers per column:

mileage 4066

price 3232

msrp 3009

model\_year 5517

wheelbase\_from\_vin 4025

**Handling Outliers:** The capping method was selected to address outliers because it restricts data points beyond defined limits to maintain proper data distribution. Outliers retain only a small impact on model and statistical results through capping techniques yet the analysis preserves most of the original data spread.

* **Encoding:** Encoding is used to convert non-numeric data (like text or categories) into numerical format, so machine learning algorithms can process it.
  + **One-Hot Encoding:**One-Hot Encoding creates a new binary column for each unique category in a feature. Each column contains 1 for the presence of a category and 0 for its absence.
  + **Applied to Columns**: listing\_type, stock\_type, transmission\_from\_vin
  + **Frequency Encoding:**
  + It transforms category variables in a dataset by replacing them with occurrence counts across the entire dataset. This conversion technique transforms descriptive data values through numerical value generation based on their appearance statistics
  + **Applied to columns**: 'dealer\_city', 'dealer\_postal\_code', 'vin', 'model\_year', 'make', 'model','series', 'style', 'exterior\_color', 'interior\_color', 'wheelbase\_from\_vin', 'drivetrain\_from\_vin', 'engine\_from\_vin', 'fuel\_type\_from\_vin'.

# Stakeholder Engagement:

* **Feedback:** Clients suggested revising listing type encoding and considering recent price trends. The client requested that the listing type be split into two columns (Active and Sold) rather than using a single categorical variable.
* **Frequency Encoding:** The client questioned about frequency encoding, suggesting a re-evaluation of encoding methods. And calculate the number of features after encoding . The client suggested to not exceed the features more.
* **Considering Recent Price Trends:** The client noted a significant bump in vehicle prices after-2021, influenced by post-COVID demand
* **Action Taken:**
  + We are creating two different csv files one with Active listing type and another with Sold listing type, and work on it separately.
  + We calculated the Features after Encoding which is 37. We are trying to use label encoding if it helps to decrease the number of features.
  + We considered the reason for increament of price in 2022and 2024, which gave us understanding about the prices.

# Lessons Learned:

* **Importance of data preprocessing in model accuracy:** A significant portion of the project's success stemmed from the careful data preprocessing and feature engineering steps. Data cleaning, encoding, and transforming the data were important in ensuring that the models were built on solid, high-quality data. Revisiting the listing type and considering its separation into "Active" and "Sold" helped clarify the data’s meaning and provided better insights. The feedback regarding the encoding strategy underscored the need for better understanding of the implications of different encoding methods. Although frequency encoding worked well in certain contexts, its complexity may suggest us to use Label encoding which might help us decrease the number of features.
* **Value of stakeholder feedback in refining analysis:** The client feedback was incredibly valuable in refining the analysis and ensuring that the results were directly aligned with the business needs. For instance, their suggestion to split the listing type into two csv file (Active and Sold) allowed for clearer insights and avoided unnecessary assumptions about listing states. Moving forward, maintaining more consistent communication with stakeholders throughout the project would ensure that the analysis remains tightly aligned with evolving needs and expectations.

# Future Recommendations:

**Apply Principal Component Analysis (PCA) for dimensionality reduction:**

* What Dimensionality Reduction seeks to accomplish is the reduction of dataset dimensions because multiple features present correlations. The principal component analysis simplifies datasets through uncorrelated principal components while maintaining essential information from the initial features.
* The reduction of dimensions through PCA enhances K-Means clustering results by enabling the creation of clusters that are better separated from each other.
* The primary component analysis technique contributes to data quality improvement and overfitting reduction by zeroing in on essential variables which represent most data variance while disposing of unimportant data components.
* PCA allows data researchers to transform their information into 2D or 3D spatial projections for better understanding of dataset patterns along with visual insights.
* **Explore alternative clustering techniques like K-Means:**

The dataset contains price and mileage and make features that form natural distinct groups between budget and luxury vehicles along with other clusters. The clusters produced by K-Means maintain clear distinctions between the items they group.

K-Means operates efficiently on big datasets without affecting performance because it offers scalability features when running through large datasets. Each K-Means cluster has a prominent centroid point which simplifies the interpretation of central characteristics among different vehicle subsets.This method succeeds in grouping spherical clusters so it functions optimally with the easily distinguishable vehicle categories found in the data set. The application of K-Means segmentation produced efficient categories of vehicles which strengthened analytical processes and decision-making ability.

* **Automate preprocessing steps for efficiency:** Data preprocessing is often a time-consuming task but automating it can dramatically improve the efficiency of the project, especially when handling large datasets. Developing and implementing an automated data pipeline that handles common preprocessing tasks such as: Identifying and dealing with missing values, duplicates, and errors. Automating the generation of new features based on existing data (e.g., encoding categorical variables, generating interaction terms). Automatically scaling numerical features to ensure they fit within appropriate ranges for machine learning models. Automating preprocessing steps will not only save time but also reduce the risk of human error, ensuring more consistent and reproducible analysis across iterations.

# Impact on the Community:

* **Improved vehicle inventory management** **for dealerships:** By analyzing and streamlining vehicle inventory data, dealerships can optimize stock levels and ensure they are carrying the right mix of vehicles that match customer demand. Improved inventory management will lead to reduced overstocking and understocking, ensuring dealerships are more responsive to market trends. This will translate to reduced waste and better utilization of resources within the dealership network, contributing to a more sustainable approach to car sales and inventory management. Dealerships can focus on serving the community's needs by providing timely access to the cars customers are seeking.
* **Better cross-selling insights** **leading to customer satisfaction:** Through advanced analysis, dealerships can identify patterns in customer behavior, leading to better cross-selling opportunities. By understanding what products or services are most relevant to each customer, dealerships can offer targeted promotions, financing options, or additional services. This improves customer satisfaction by personalizing the shopping experience and ensuring that customers receive valuable recommendations that meet their needs. It also fosters stronger customer relationships and encourages loyalty, ultimately leading to a more supportive and engaged automotive community.
* **Data-driven decision-making** for car pricing and sales: Data-driven insights enable dealerships to set competitive and profitable car prices based on real-time market trends, demand, and competitor pricing. This fosters fair pricing, building trust with customers. By adapting to market shifts, dealerships ensure vehicles are available at appropriate prices while optimizing sales and inventory strategies to lower costs, potentially making cars more affordable. Efficient stock management improves vehicle availability, enhances accessibility, and reduces excess inventory, promoting sustainability. Ultimately, these insights create a more efficient, customer-focused automotive industry, benefiting both businesses and the community.

# Project Conclusion:

This phase successfully implemented clustering analysis to identify vehicle make similarities based on price, mileage, and age. The feedback received from stakeholders was instrumental in refining the model to provide meaningful insights. The project has effectively met its initial goals by helping dealerships better understand which brands can be grouped together for cross-selling opportunities. The integration of machine learning has contributed to **more efficient inventory management**, **enhanced customer recommendations**, and **better pricing strategies** for Go Auto’s dealerships.

We have future plans to execute Principal Component Analysis (PCA) to minimize data dimensions which will optimize feature choice and enhance data visual presentation. The analysis will use cluster techniques alongside predictive modeling to identify vehicle sales patterns that help decision-making for automobile dealerships.

The platform created allows us to perform sophisticated analysis for extracting relevant insights and making data-driven recommendations throughout the upcoming phases of work.

# Acknowledgments:

# We express appreciation to every participant who supported the Go Auto Dataset project.

# The project member Srijana Chauhan , Rajinder Kaur and Jasmeet Kaur together with Celine Panicker devoted themselves to data preprocessing and encoding.

# We are grateful to Go Auto for offering the dataset as well as important information to our research.

# We express gratitude towards our CMPT 3830 instructor together with our mentors who provided continuous guidance.

# All of us from Go Auto extend our gratitude to the classmates who provided feedback throughout the project.

# The work behind this project involved multiple team members and we are excited about additional data evaluations.

# Appendices:

**Sprint 1**: Team Charter Development

**Timeframe:** January 16 - January 23, 2025

**Focus:** Establishing team structure and defining project objectives.

**Tasks Completed:**

* Developed the Team Charter, outlining project goals, team roles, and responsibilities.
* Defined short-term and long-term objectives for the project.
* **Established team name:** *ClusterCatalyst*.
* Determined logistics such as communication channels and meeting schedules (virtual meeting twice a week on Monday and Thursday at 6:00PM) to go over the updates on the project, issues faced and additional work that would be done.

**Meeting Minutes:**

**Date:** January 20, 2025

**Duration:** 1 hour

**Attendees:** Sirjana Chauhan, Rajinder Kaur, Jasmeet Kaur, Celine Panicker

* Discussed key elements to be included in the Team Charter.
* Drafted an initial version of the Team Charter.
* Assigned official roles to each team member.
* Finalized team meeting frequency and communication protocols.

**Sprint 2**: Project Charter Development

**Timeframe:** January 24 - January 30, 2025

**Focus:** Defining the scope, objectives, and key deliverables for the project.

**Tasks Completed:**

**Developed the Project Charter, outlining:**

* Problem Statement: Vehicle Make Similarities using Clustering.
* Project Goals & Objectives: Identifying vehicle make similarities to enhance dealership cross-selling strategies.
* Methodology & Tools: Using clustering techniques such as K-Means to group vehicles based on price, mileage, and age.
* Stakeholder Expectations: Providing Go Auto with actionable insights.
* Established the timeline and major milestones for the project.
* Reviewed risk assessment and mitigation strategies.

**Meeting Minutes:**

**Date:** January 27, 2025

**Duration:** 1 hour

**Attendees:** Sirjana Chauhan, Rajinder Kaur, Jasmeet Kaur, Celine Panicker

* Discussed and finalized project objectives.
* Identified key datasets and potential preprocessing steps.
* Assigned responsibilities for each phase of the project.

**Sprint 3**: Demo 1 - Exploratory Data Analysis (EDA) and Initial Model Implementation

**Timeframe:** January 31 - February 13, 2025

**Focus:** Conducting EDA, preprocessing data, and implementing initial clustering models.

**Tasks Completed:**

* Conducted Exploratory Data Analysis (EDA):
* Identified missing values and handled outliers by using encoding.
* Analyzed key patterns in vehicle pricing, mileage, and age.
* Preprocessed dataset:
* Encoded categorical variables.
* Normalized numerical features.
* Implemented Visualization:
* Predict the best insights and popularity of make, model in the dataset after encoding.
* Visualized data distributions and trends.
* Prepared Demo 1 Presentation for Go Auto.

**Meeting Minutes:**

**Date:** February 10, 2025

**Duration:** 2 hours

**Attendees:** Sirjana Chauhan, Rajinder Kaur, Jasmeet Kaur, Celine Panicker

* Reviewed encoding methods and find best approach.
* Identified key findings and insights from EDA.
* Finalized presentation content and visualizations.

**Sprint 4**: Project Phase 1 Report Preparation

**Timeframe:** February 14 - February 27, 2025

**Focus:** Refining clustering models, incorporating client feedback, and finalizing Phase 1 documentation.

**Tasks Completed:**

* Incorporated client feedback from Demo 1:
* Refactored listing type encoding (split Active and Sold into separate columns).
* Adjusted clustering parameters to reflect market trends.
* Addressed concerns related to frequency encoding.
* Project Phase 1 Completed

**Tasks To be Done**

• Conduct model optimization:

• Fine-tuned clustering hyperparameters.

• Re-evaluated model performance using Silhouette Score.

• Developed final visualizations and insights for Go Auto.

• Summarized dataset analysis, clustering results, and recommendations.

• Documented challenges encountered and future improvements.

**Meeting Minutes**

**Date:** February 24, 2025

**Duration:** 2 hours

**Attendees:** Sirjana Chauhan, Rajinder Kaur, Jasmeet Kaur, Celine Panicker

* Reviewed all sections of the Phase 1 report.
* Ensured alignment with project goals and stakeholder expectations.
* Discussed next steps for refining clustering models in later phases. Scrum logs and meeting notes.

# References

No External References were used.