CMPT 3830: Machine Learning
Work Integrated Learning-1

Project Report: Phase 1

"Days on Market Predicator for Vehicle Sales Optimization"

In collaboration with

Submitted By:

Name	ID	Email
Rohit	3095056	rrohit56@norquest.ca
Spandan Dahal	3092592	sdahal92@norquest.ca
Abhinav Datt	3097289	adatt@norquest.ca
Jatin Dandyan	3095052	jdandyan@norquest.ca

Submission: Date: 15/10/2024

Fall 2024

Table of Contents:

Contents

List of Figures:

List of Tables:

1. Project Phase:

1. Introduction:

We have completed the Exploratory Data Analysis (EDA) and data preprocessing phases, including handling missing values, outliers, and encoding categorical variables like make, model, and fuel type. Now, we are focusing on developing a Linear Regression model to predict the number of days a vehicle will stay on the market, helping Go Auto optimize inventory and sales strategies.

2. Machine Learning Model Exploration: Linear Regression Model

- We have chosen Linear Regression as our primary model for predicting the number of days a vehicle will stay on the market. Linear regression is a simple yet effective model for understanding the linear relationship between variables such as price, mileage, and model year.
- The target variable, days on market, is regressed against the independent variables, including model year, make, price, mileage, and other relevant features.

3. Applying the Linear Regression Model:

- After preparing the dataset, we applied the Linear Regression model, using the standardized features as input variables.
- The model was trained on a subset of the data (training set) and evaluated on the remaining portion (test set) to assess its generalization capabilities.
- We are focusing on interpreting the coefficients of the linear regression model to understand how each feature impacts the number of days a vehicle will stay on the market.

4. Model Evaluation:

- We are using the following metrics to evaluate the performance of the Linear Regression model:

  - Mean Squared Error (MSE): This measures the average squared difference between predicted and actual values of days on market. A lower MSE indicates better accuracy.
  - R-squared (R^2): This metric shows how well the independent variables explain the variance in days on market. A higher R-squared value means that the model explains a larger proportion of the variance.

5. Model Optimization:

- Feature Selection: We are experimenting with removing features that do not significantly contribute to the model’s performance. By reducing irrelevant features, we aim to improve the model’s accuracy and interpretability.
- Cross-Validation: To avoid overfitting, we are applying k-fold cross-validation. This ensures the model generalizes well and isn’t overly biased by any subset of data.

2. Team Members' Name with specific roles

Rohit	1. Rohit is responsible for cleaning the dataset, fixing errors, and handling missing data to ensure the quality of the data used for the project. 2. He is also in charge of selecting the most important features, such as vehicle price and mileage, that will be used to predict the days on market for each vehicle.
Spandan Dahal	1. Spandan is tasked with building the machine learning model that will predict how long a vehicle will stay on the market. 2. He will also focus on refining and optimizing the model to ensure it delivers the most accurate and reliable predictions possible.
Abhinav Datt	1. Abhinav's role is to design clear and effective visualizations, including charts and dashboards, to present the data insights and predictions. 2. He will use tools like PowerBI or Looker Studio to create these visualizations, making the results easy to understand for stakeholders.
Jatin Dandyan	1. Jatin ensures that the project runs smoothly by coordinating tasks, organizing meetings, and setting deadlines for the team. 2. He monitors the team's progress, resolves any issues that arise, and ensures that the project stays on track to meet its goals.

3. Reporting Period:

Dates	Milestone	Details/Comment
Sep 17	Team charter	• Finalize the team charter for the project, outlining roles, responsibilities, and communication protocols. Ensure accountability and assign final roles for the project. Set up Trello for progress tracking and begin populating tasks.
Sep 24	Project Proposal Template	• Use the provided Project Proposal Template to draft the full project proposal. Define the scope, deliverables, tasks, and risk management. Outline a clear project plan with a timeline for milestones, resources, and budget.
Oct 8	Demo 1	• Focus on presenting data analysis and handling feature encoding. Analyze key features for prediction and finalize encoding processes (e.g., one-hot

		encoding for categorical variables). Present findings to the team and review insights for further improvement

4. Project Overview:

Problem Statement:

The primary challenge of this project is to predict the number of days a vehicle will remain on the market before it is sold. This prediction is crucial for optimizing inventory management and sales strategies for Go Auto, a vehicle dealership company. By accurately predicting the "days on market" for each vehicle, dealerships can better manage their inventory, dynamically adjust pricing, and enhance decision-making processes regarding sales and marketing efforts.

Solution Approach:

To address the problem, we began by thoroughly cleaning the dataset, handling missing values, and resolving any data inconsistencies to ensure the quality of the information. We then applied feature selection techniques to focus on the most relevant attributes, such as vehicle price, mileage, and model year, which are key to predicting the days a vehicle will stay on the market. We used one-hot encoding to convert categorical data, such as car makes and models, into a numerical format for the model. After preprocessing, we are currently in the process of building a linear regression model to predict the days on market. Throughout the project, we have created visualizations to identify trends and insights from the data, helping us refine the model for better accuracy and more effective predictions.

5. Dataset

The Go Auto dataset contains 145,114 rows (or data points) and 46 columns, representing detailed information about vehicle listings across various dealerships. This dataset is rich with attributes related to vehicle features, dealer information, and sales activity. Here's a breakdown of the key components of the dataset:

1. Number of Data Points:

    - Rows: The dataset has 145,114 rows, where each row represents a unique vehicle listing.
    - Columns: There are 46 columns that describe various aspects of the vehicles, dealers, and market performance.

2. The Go Auto dataset contains **12 numerical columns** and **34 categorical columns** across 145,114 rows.

5.1 Exploratory Data Analysis (EDA) Highlights:

✓ Data Collection and Exploration:

As part of the Exploratory Data Analysis (EDA), we analysed the relationships between these features to identify patterns and potential anomalies. This analysis helped us understand which features would have the most significant impact on predicting the number of days a vehicle stays on the market.

From the 46 columns, we selected the following key features for our project:

- Model Year
- Make
- Model
- Mileage
- Price
- MSRP
- Transmission Type
- Fuel Type
- Days on Market (target variable)

These features were chosen because they are the most relevant factors that can influence how long a vehicle remains on the market before it is sold.

✓ Data Preprocessing: Handling Unrealistic Values and Outliers in Price and MSRP

During the data preprocessing phase, we identified anomalies in the price and msrp columns, where some values ranged between 0 and 1000, which is unrealistic for vehicle listings. Additionally, there were discrepancies between the price and msrp values for certain vehicles, where prices were significantly higher than the msrp, indicating potential outliers.

To address these issues, we applied the following steps:

1. Calculating Mean Prices for Similar Vehicles: We grouped vehicles based on key attributes such as make, model, model year, transmission, and fuel type. For each group, we calculated the average price and msrp, excluding unrealistic values in the range of 0 to 1000. This allowed us to generate more reasonable estimates for vehicles with missing or outlier values.

2. Replacing Unrealistic Values: We replaced all price and msrp values between 0 and 1000 with the corresponding group averages. This ensured that these values aligned with other vehicles of similar characteristics, maintaining consistency in the dataset.

3. Handling Major Price Outliers: In some cases, we identified extreme outliers where the price was disproportionately higher than the msrp.

Vehicle Example:

- VIN: 1GT49YEY8RF277865
- Make: GMC
- Model: Sierra 3500
- Price: 12,908,925
- MSRP: 120,699
- Model Year: 2024

In this instance, the price was unreasonably high compared to the msrp. To correct this, we manually adjusted the price by replacing the last few digits, making it more aligned with the msrp. This manual adjustment ensured the values made sense within the context of the vehicle’s attributes.

4. Final Clean-up:

➢ After handling outliers and replacing unrealistic values, we removed any temporary columns used during this process, keeping the dataset clean and ready for model training. This step ensured that no irrelevant data remained in the dataset and that all price and msrp values were realistic and consistent.

By addressing these outliers and discrepancies manually and through calculated averages, we improved the overall quality and reliability of the dataset, making it more suitable for machine learning model development.

✓ Handling Missing Values in Price and MSRP

After addressing the outliers in the price and msrp columns, we identified missing values in these columns. To ensure the dataset was complete for model training, we used the fillna function to replace missing values with the mean of the respective columns.

- Price: Missing values in the price column were replaced using fillna(df['price'].mean()), which filled the gaps with the mean price.
- MSRP: Similarly, missing values in the msrp column were replaced using fillna(df['msrp'].mean()), ensuring consistency by filling in the mean msrp.

➢ Why Use the Mean with fillna?

Using the mean with the fillna method is effective because:

1. It maintains the overall distribution of the data without introducing bias.
2. It allows us to retain all records and ensure no data is discarded.

5.2 Visualization:

a. Correlation Heatmap:

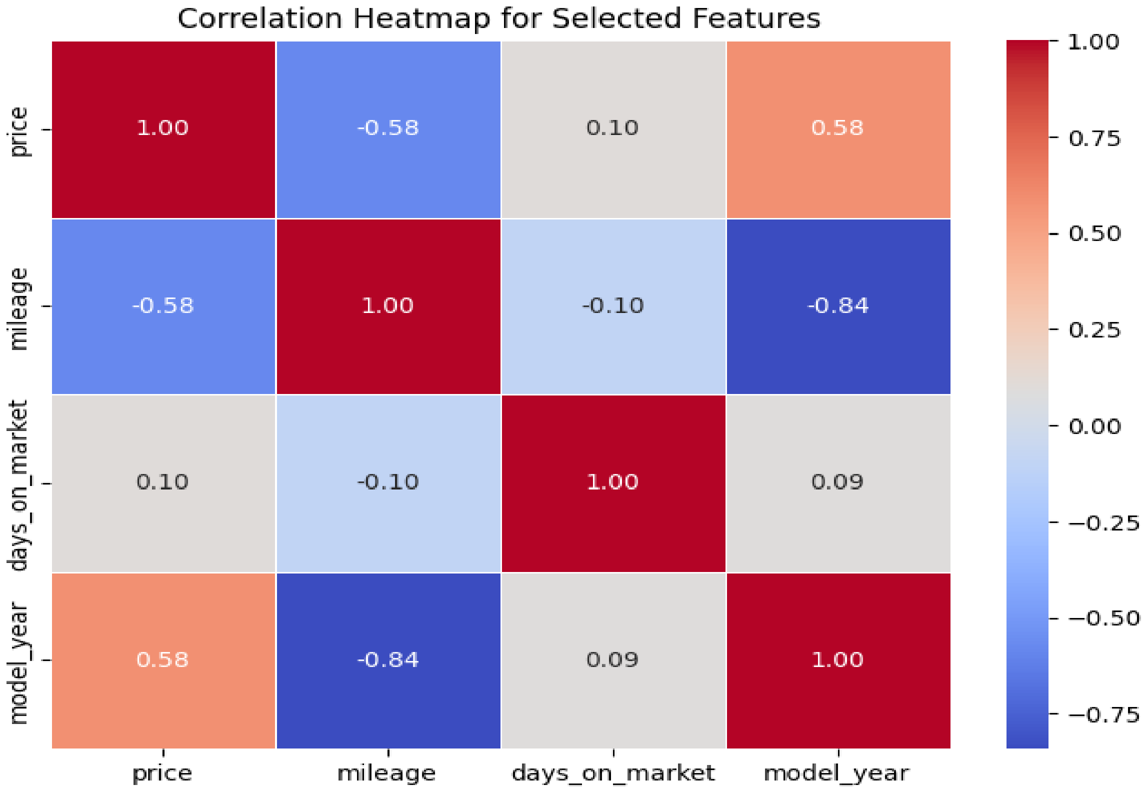
The heatmap above is an interactive visualization that represents the correlation between selected numerical features from the dataset, including price, mileage, days_on_market, and model_year. This visualization was developed as part of the Exploratory Data Analysis (EDA) to identify relationships between important vehicle attributes.

Key Findings from the Heatmap:

- Days on Market and Price (+0.10 correlation):
  There is a weak positive correlation between price and days_on_market, suggesting that higher-priced vehicles may stay on the market for a slightly longer period, but the effect is minimal.

- Days on Market and Mileage (-0.10 correlation):
  Similarly, the correlation between days_on_market and mileage is weak and negative, indicating that vehicles with higher mileage might sell slightly quicker, but the impact is minor.

- Days on Market and Model_year (+0.09):
  The correlation between model_year and days_on_market is +0.09, indicating a very weak positive correlation. This suggests that newer vehicles (with higher model years) tend to stay on the market slightly longer, but the effect is minimal.



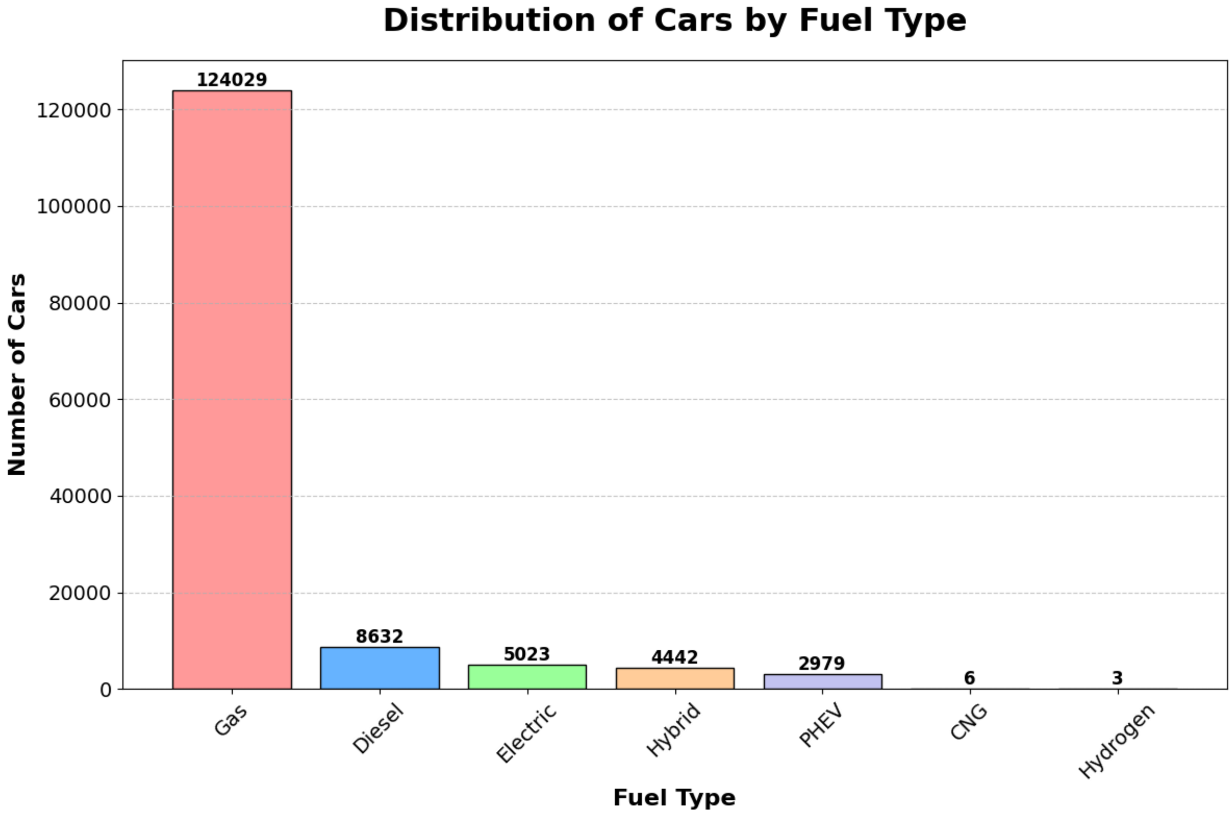
b. Fuel Type Bar Graph:

This bar chart shows the number of vehicles by fuel type in the dataset.

- Gasoline Vehicles: The majority, with 124,029 cars, dominate the dataset.
- Diesel: There are 8,632 diesel cars, the second-largest category.
- Electric: 5,023 electric vehicles reflect growing interest, though still smaller than gas and diesel.
- Hybrid and PHEV: 4,442 hybrid and 2,979 plug-in hybrids indicate some demand for fuel-efficient options.
- CNG and Hydrogen: These fuel types are rare, with only 6 CNG and 3 hydrogen vehicles.

➢ Conclusion:

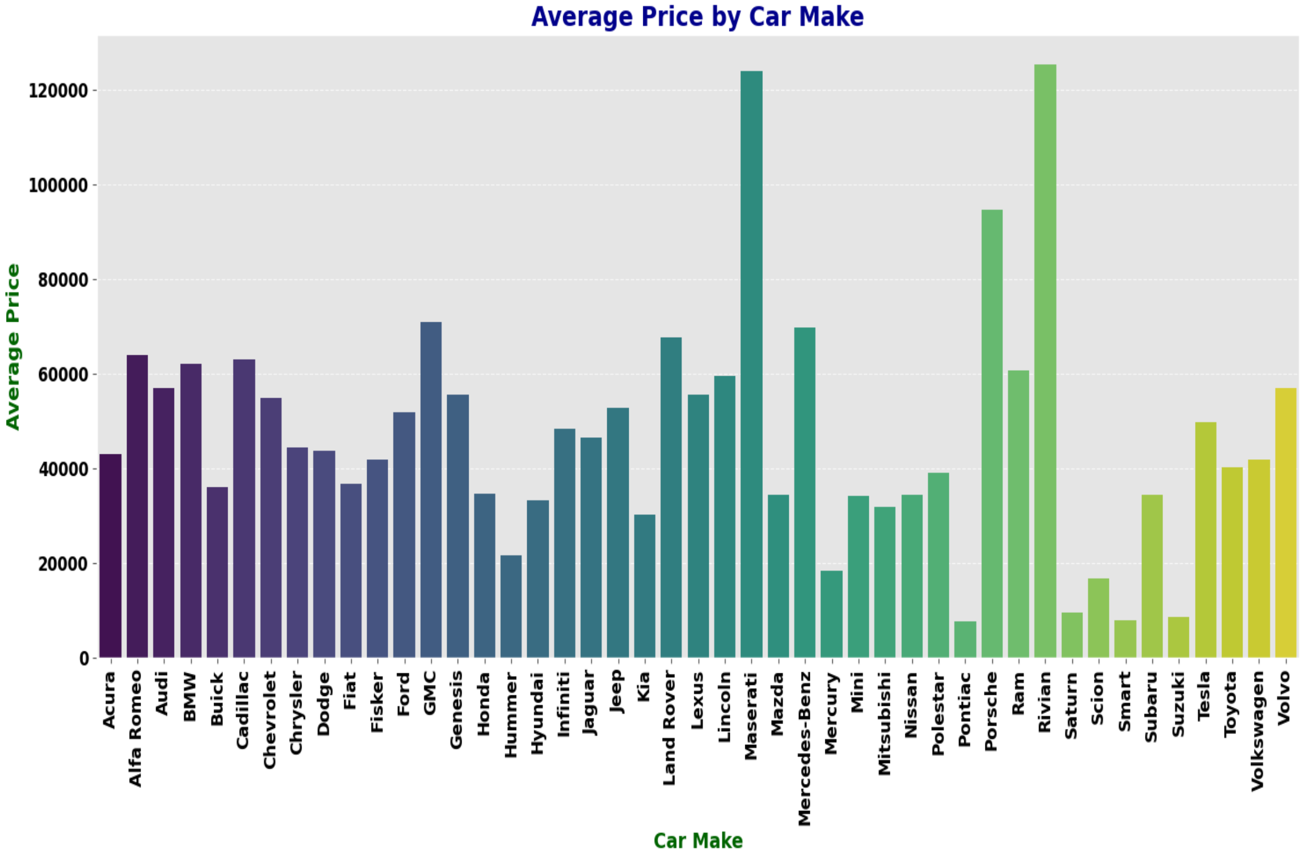
Gasoline vehicles are the largest group, with other fuel types present in much smaller numbers, highlighting the continued dominance of traditional fuel in the market.



c. Average price by Car Make:

➢ The bar chart displays the average price of vehicles by car make, showing significant price variation between brands.

3. Maserati and Rivian stand out with the highest average prices, exceeding $120,000.
4. Porsche, Mercedes-Benz, and Land Rover also have relatively high average prices, positioning them among the luxury brands.
5. More affordable brands like Kia, Hyundai, and Ford are toward the lower end of the price range.

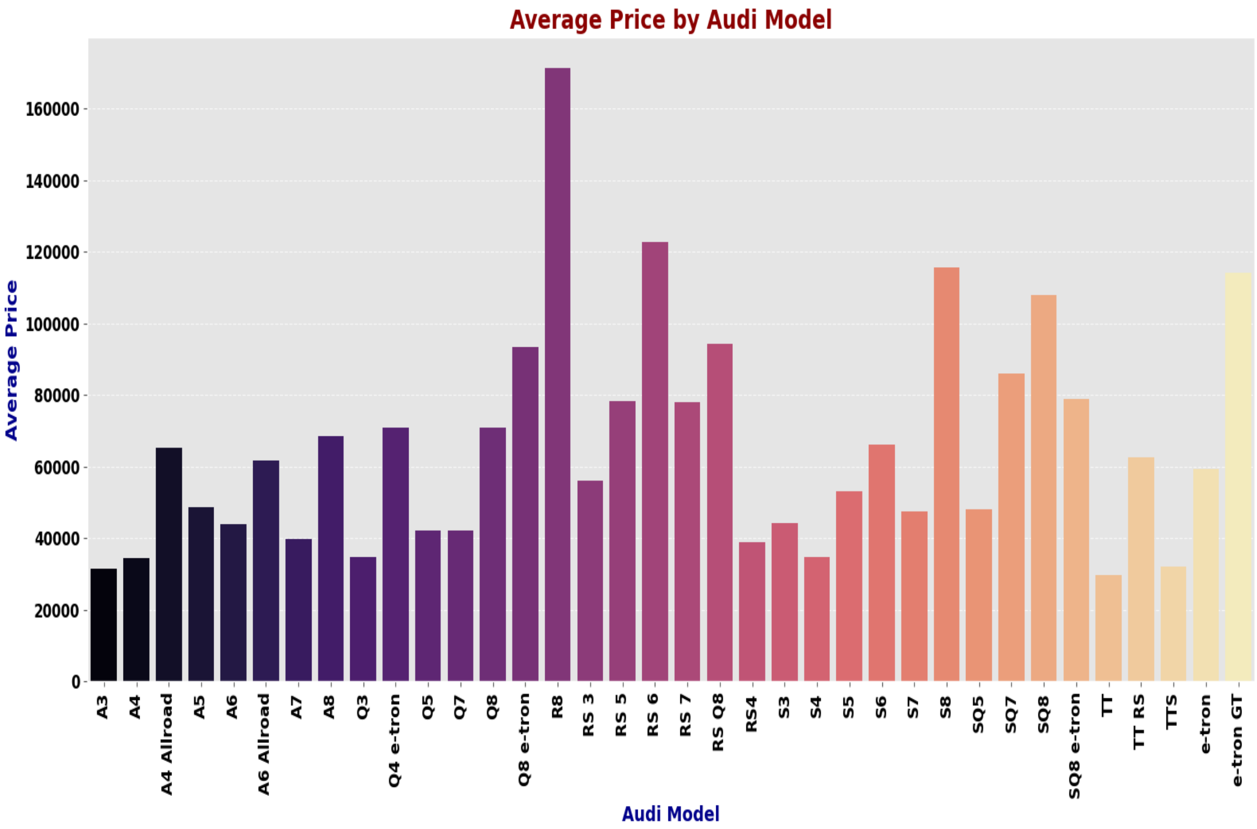


4. Average Price by Audi Model

The bar chart displays the average price for different Audi models, showing notable variations across the range of models.

- Audi R8 stands out with the highest average price, exceeding $160,000, making it the most expensive model in the lineup.
- Audi e-tron GT and RS 6 also have high average prices, placing them in the premium segment.
- More affordable models include Audi A3, S3, and TT, with average prices significantly lower compared to the high-end models.

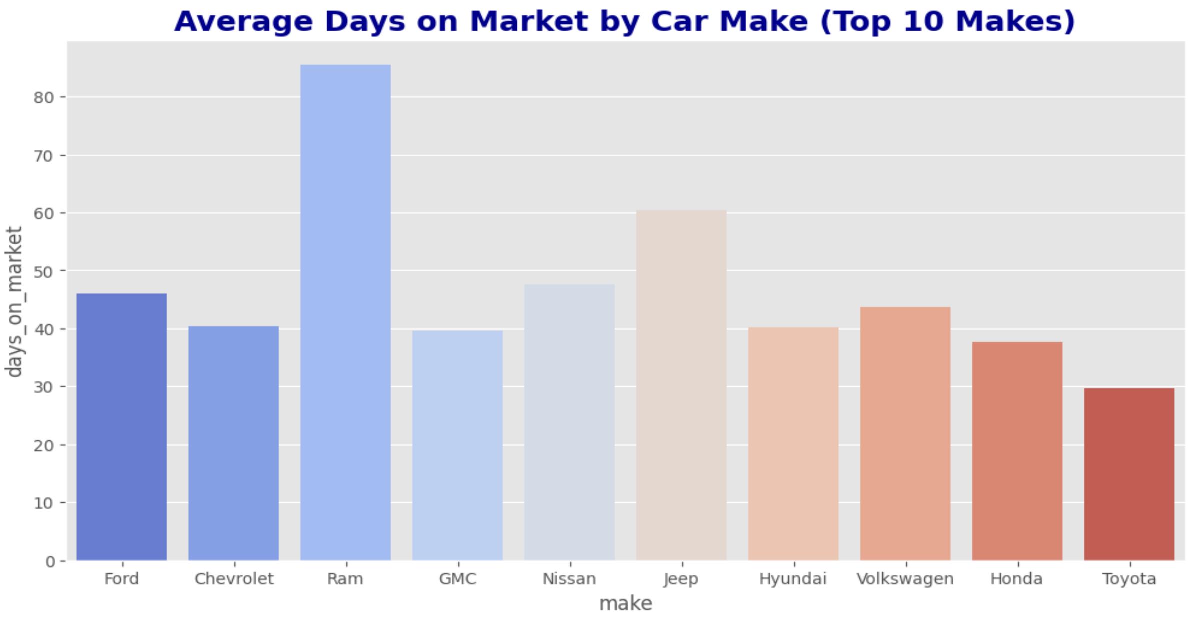
This chart highlights the price differences between Audi’s luxury sports cars and more economical models, providing insights into the pricing strategy across the brand's model lineup.



5. Average Days on Market by Car Make (Top 10 Makes):

The first chart shows the average days on market for the top 10 car makes, highlighting how long, on average, vehicles from each make stay on the market before being sold.

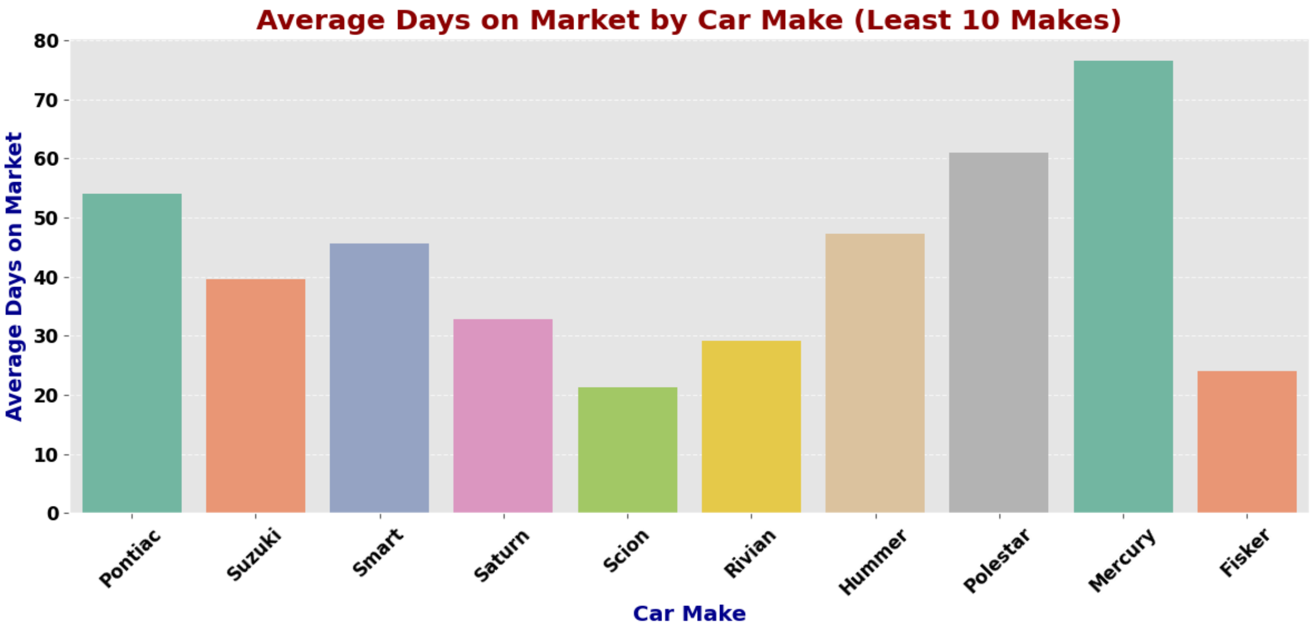
- Ram vehicles have the longest average days on market, exceeding 80 days.
- Toyota has the shortest time on the market, with vehicles averaging less than 30 days.



6. Average Days on Market by Car Make (Least 10 Makes):

This bar chart shows the average days on market for the least common car makes, indicating how long vehicles from these makes typically stay on the market before being sold.

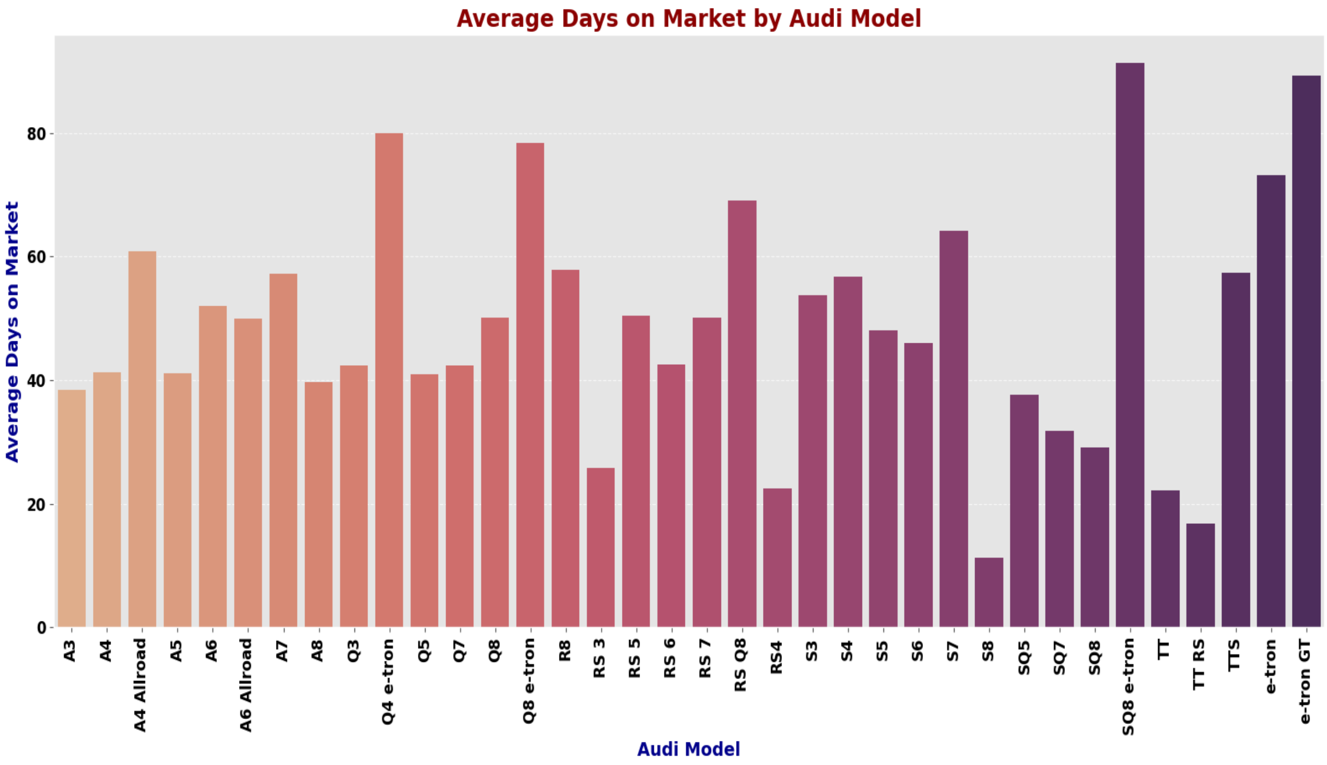
- Mercury cars have the longest average days on market, exceeding 70 days, followed by Polestar and Pontiac, which also take longer to sell.
- Scion and Fisker have the shortest average days on market, with their vehicles staying on the market for fewer than 30 days on average.



7. Average Days on Market by Audi Model:

The second chart breaks down the average days on market for different Audi models, showing significant variation in how long different models stay on the market.

- Audi SQ 8 e-tron and e-tron GT have the longest average days on the market, indicating these models take the most time to sell.
- Models like the Audi S8 and TT RS have much shorter average days on market, suggesting quicker sales.



6. Challenges Encountered:

During the project, we encountered several challenges that required problem-solving and adaptability:

1. Handling Outliers and Missing Values: One of the major challenges was dealing with outliers, especially in the price and msrp columns, where values were either unrealistically low or excessively high. We also had to address missing values in key columns, ensuring that our dataset was clean and complete for model training. This took considerable time and required careful decision-making.

2. Learning Encoding Techniques: At the start of the project, we had no prior experience with encoding categorical data. After researching and exploring different methods online, we learned about binary encoding and one-hot encoding. One-hot encoding,

though effective, presented its own challenge as it significantly increased the number of columns in the dataset. After applying it, we ended up with 796 columns, which made managing and processing the dataset more complex.

3. Standardization and Vehicle Age: While standardizing numerical columns, we noticed some irregularities in the model_year column. To address this, we decided to convert model_year to vehicle age (current year minus model year), which better aligned with the model's needs. We then performed standardization on the new vehicle age column, ensuring consistency across numerical features.

7. Stakeholder Engagement:

During our **Demo 1 presentation**, we shared our EDA findings with the stakeholders and discussed some potential challenges we identified. Specifically, we noted that there are many car models in the dataset, which we believe could create issues in future modelling. We are currently addressing this concern by exploring different encoding techniques to simplify the complexity of our project.

Additionally, we mentioned that while we plan to build one model for all car makes, we anticipate potential challenges in doing so. If it becomes too complex or ineffective, we will explore the possibility of creating separate models for each car make. The insights and additional information gathered from this session will be helpful as we continue refining our approach to ensure the model is robust and accurate.

8. Lessons Learned:

Throughout the project, we gained valuable insights into both the technical and collaborative aspects of data analysis and model building.

1. What Worked Well:
   - Data Preprocessing and Feature Engineering: We became proficient in handling outliers, missing values, and encoding, which are essential steps in any machine learning project. Learning how to convert categorical data into numerical forms (via binary and one-hot encoding) and feature engineering, such as transforming model_year to vehicle age, improved our understanding of how to make data more usable for model training.
   - Collaboration and Research: Working as a team and exploring solutions online was a major factor in overcoming our initial lack of knowledge in areas like encoding. We learned that consistent communication and dividing tasks based on strengths allowed us to work more efficiently.

2. Areas for Improvement:
   - Dealing with High-Dimensional Data: The use of one-hot encoding significantly increased the number of columns in our dataset, which introduced complexity. In future projects, we could explore alternative encoding methods, such as target encoding or dimensionality reduction techniques, to avoid bloating the dataset.
   - Model Selection and Experimentation: While working with the linear regression

model, we encountered some irregularities in the standardization process. We learned that careful attention to standardization is crucial for ensuring the best dataset quality for the model. In future projects, addressing these issues early will lead to smoother model development.

Overall, the project taught us the importance of planning, researching, and iterating, and we now have a better foundation for tackling similar challenges in future projects.

9. Future Recommendations:

1. Focus on Optimizing the Linear Regression Model: Given that we are committed to using the linear regression model, future iterations should focus on refining and optimizing it to ensure the best possible performance for this type of model.

2. Handling Large Dataset Complexity: The dataset is quite large, and we need to consider whether it's feasible to create a single model that works for all car makes, or if it's more effective to build separate models for each make. This approach could lead to better, more tailored predictions depending on the specific characteristics of each car brand.

3. Improving Knowledge of Encoding Techniques: While we successfully implemented one-hot encoding, we realized that it greatly increased the complexity of our dataset. In the future, we should explore other encoding methods that can reduce complexity, such as target encoding or frequency encoding. Gaining more knowledge in this area will help us handle categorical data more efficiently.

10. Impact on the Community:

1. Optimizing Inventory Management: By accurately predicting how long vehicles will stay on the market, dealerships can make informed decisions about inventory management. This can lead to better resource allocation, helping dealerships optimize their stock, reduce overstock situations, and ensure a balanced supply of high-demand vehicles.

2. Improving Pricing Strategies: Dealerships can adjust their pricing strategies dynamically based on the predicted days on market, ensuring vehicles are priced appropriately for faster sales. This could benefit the community by offering better deals to consumers while allowing dealerships to maintain a steady cash flow.

3. Better Consumer Experience: Consumers benefit from a more responsive marketplace where cars are priced based on realistic sales timelines. This can lead to greater customer satisfaction, as buyers will find more accurately priced vehicles, reducing the risk of inflated prices or poorly maintained stock.

In summary, this project has the potential to positively impact the community by improving dealership operations, promoting sustainability, and enhancing the overall consumer experience in the automotive market.

11. Project Conclusion:

Overall, our project has been highly successful and has met, and in some cases, exceeded the initial goals we set. One of our primary objectives was to clean and prepare the dataset by handling missing values, which we effectively achieved. Through careful analysis and processing, we ensured the data was accurate and ready for model development.

Additionally, our Exploratory Data Analysis (EDA) led to valuable insights, and we created effective visualizations that highlighted key trends and correlations within the dataset. These visualizations were crucial in understanding relationships between features such as price, mileage, and days on market.

We also explored and implemented different encoding techniques to manage the complexity of the dataset, particularly with categorical variables, making our project more manageable.

In conclusion, the project has laid a strong foundation for developing an accurate prediction model and has contributed to improving data quality and decision-making processes. This work will not only enhance the project's future stages but also provide valuable insights for the automotive industry.

12. Acknowledgments:

We would like to express our sincere gratitude to several individuals and organizations who made this project possible:

- Md Mahbub Mishu, our instructor, and Caylee Kreller for their continuous support, guidance, and for helping us enrol in and successfully complete this project.
- Thiago Valentin, Data Scientist Lead at Go Auto and their representative, for providing invaluable insights and expertise throughout the project.
- Norquest College and the Go Auto team for offering us this opportunity to work on a meaningful and impactful project.
- A special acknowledgment to our team, Regression Rebels—Rohit, Spandan Dahal, Abhinav Datt, and Jatin Dandyan—for their dedication, teamwork, and hard work throughout the project. Each member's contributions is vital to the success of this project.

13. Appendices:

Scrum Report 1: September 9 – September 23, 2024

1. Project Overview

Project Name: Go Auto Project: Predicting Days on Market
Product Owner: Go Auto
Scrum Master: Spandan Dahal
Development Team:

- Data Preprocessing Lead: Rohit
- Model Developer: Spandan Dahal
- Visualization Expert: Abhinav Datt
- Project Coordinator: Jatin Dadhyan

Stakeholders: Go Auto management, business intelligence team.

Project Goals:

- Explore and understand the dataset, including handling missing values and outliers.
- Assign team roles and responsibilities and develop the Team Charter and Project Charter.

2. Sprint Planning Documentation

Sprint Overview:

- Sprint Duration: 2 weeks (September 9, 2024 - September 23, 2024)
- Sprint Goal: Complete dataset exploration, identify initial trends, and set up roles and responsibilities.

User Story	Story Points	Priority	Assigned To
US1: Team Formation and Role Assignment	3	High	Jatin
US2: Dataset Exploration and Initial EDA	8	Medium	Rohit, Abhinav
US3: Handle Missing Values and Data Cleaning	7	High	Rohit
US4: Identify Key Trends in Dataset	5	Medium	Abhinav, Spandan

User Stories:

- User Story 1: Team Formation and Role Assignment

a) Description: Jatin led the formation of the team, assigned roles, and created the Team Charter to facilitate effective collaboration.

b) Acceptance Criteria: Team members assigned roles, and a Team Charter was created.

c) Definition of Done: Roles assigned, Team Charter completed and approved by stakeholders.

- User Story 2: Dataset Exploration and Initial EDA

  a) Description: Rohit and Abhinav explored the dataset and conducted preliminary EDA to identify patterns and gain insights.

  b) Acceptance Criteria: Dataset explored and initial visualizations (e.g., bar charts) created to understand data distribution.

  c) Definition of Done: An EDA report generated, shared with the team for review.

- User Story 3: Handle Missing Values and Data Cleaning

  a) Description: Rohit ensured the dataset was free of inconsistencies by handling missing values effectively.

  b) Acceptance Criteria: Missing values addressed, cleaned dataset available for analysis.

  c) Definition of Done: Cleaned and validated dataset.

- User Story 4: Identify Key Trends in Dataset

  a) Description: Abhinav and Spandan worked on identifying key data trends, focusing on relationships between vehicle price, mileage, and days on market.

  b) Acceptance Criteria: Trends identified, and findings visualized.

  c) Definition of Done: Visualizations completed and presented for stakeholder review.

3. Sprint Execution Documentation

Daily Stand-ups: Example from Day 2:

- Rohit: Continued dataset exploration and began handling missing values. No blockers.
- Abhinav: Started working on initial visualizations of key features. No blockers.
- Spandan: Assisted in identifying key features for trend analysis. No blockers.
- Jatin: Monitored team coordination and ensured effective role assignment. No blockers.

Sprint Burndown Chart

Date	Story Points Remaining
Day 1	23
Day 4	16
Day 8	10
Day 10	6
Day 12	3
Day 14	0

4. Sprint Review

- Presented visualizations showing correlations between vehicle features such as price, mileage, and days on market.
- Discussed encoding techniques used and their impact on data preparation.

Feedback:

- Instructor: Ensure consistent handling of missing values throughout the dataset.
- Go Auto Representative: Provided detailed explanations of the data in the Excel sheet, such as the meanings of "A" and "M" in the transmission column, where "A" stands for automatic, and "M" stands for manual. They also explained that there were additional values ("6" and "7") where "7" represents automatic and "6" represents manual. This information helped the team accurately interpret and prepare the dataset.

Next Steps: Begin work on advanced visualizations and explore feature encoding.

5. Sprint Retrospective

What Went Well:

- The team worked well together, and roles were effectively assigned.
- Missing values were addressed promptly.

What Didn't Go Well:

- There were some delays during data exploration due to unclear trends.

Improvements: Better communication during EDA to minimize delays.

6. Product Backlog

Backlog Item	Priority	Story Points	Status
Complete Advanced Visualization	High	10	To Do
Apply Feature Encoding Techniques	High	8	To DO

Test Dataset Quality and Integrity	Medium	6	To Do
Model Selection and Experimentation	Medium	8	To Do

7. Definition of Done (DoD)

- Code Quality Verified: Code adheres to initial project standards and follows best practices for maintainability and readability.
- Testing Completed: Basic unit testing has been conducted to verify the integrity of data cleaning and preprocessing scripts.
- Team Review Conducted: All initial preprocessing scripts have undergone peer review, and suggested changes have been incorporated.
- Initial Documentation Created: Initial documentation on data cleaning, handling of missing values, and data structure has been completed for stakeholders and team reference.
- Successfully Integrated: All data preprocessing scripts have been integrated into the team workflow, and the cleaned dataset is ready for further analysis.

Scrum Report 2: September 24 – October 7, 2024

1. Project Overview

Project Name: Go Auto Project: Predicting Days on Market
Product Owner: Go Auto
Scrum Master: Spandan Dahal
Development Team:

- Data Preprocessing Lead: Rohit
- Model Developer: Spandan Dahal
- Visualization Expert: Abhinav Datt
- Project Coordinator: Jatin Dadhyan

Stakeholders: Go Auto management, business intelligence team.

Project Goals:

- Finalize visualizations and feature encoding.
- Prepare for Demo 1, showcasing progress.

2. Sprint Planning Documentation

Sprint Overview:

- Sprint Duration: 2 weeks (September 24, 2024 - October 7, 2024)
- Sprint Goal: Create advanced visualizations, complete encoding techniques, and prepare Demo 1.

User Story	Story Points	Priority	Assigned to
US1: Create Advanced Visualizations	8	High	Abhinav
US2: Learn and Implement Encoding Techniques	10	High	Rohit, Spandan
US3: Demo 1 Presentation Preparation	7	Medium	Jatin
US4: Finalize EDA Insights for Presentation	5	Medium	Rohit

User Stories:

- User Story 1: Create Advanced Visualizations

  a) Description: Abhinav worked on creating advanced visualizations like bar graphs and heatmaps to present data insights effectively.

  b) Acceptance Criteria: Key relationships between vehicle price, mileage, and days on market visualized.

c) Definition of Done: Visualizations completed and ready for Demo 1 presentation.

- User Story 2: Learn and Implement Encoding Techniques

  a. Description: Rohit and Spandan explored encoding techniques, including one-hot encoding and binary encoding, to convert categorical features to numerical formats.

  b. Acceptance Criteria: Categorical data encoded effectively for model preparation.

  c. Definition of Done: Encoded dataset ready for further modelling tasks.

- User Story 3: Demo 1 Presentation Preparation

  a. Description: Jatin coordinated the team to prepare Demo 1, including summarizing EDA findings and visualizations.

  b. Acceptance Criteria: Presentation slides prepared and reviewed by team members.

  c. Definition of Done: Demo 1 successfully presented to stakeholders.

- User Story 4: Finalize EDA Insights for Presentation

  a. Description: Rohit and Abhinav worked on finalizing the insights from EDA to include in Demo 1.

  b. Acceptance Criteria: EDA insights summarized and presented during the demo.

  c. Definition of Done: EDA insights finalized and presented.

3. Sprint Execution Documentation

Daily Stand-ups: Example from Day 2:

- Rohit: Explored and started implementing encoding techniques. No blockers.
- Abhinav: Worked on completing visualizations for Demo 1. No blockers.
- Spandan: Assisted with encoding and feature standardization. No blockers.
- Jatin: Coordinated Demo 1 preparation. No blockers.

Sprint Burndown Chart

Date	Story Points Remaining
Day 1	30
Day 4	22
Day 7	14
Day 10	10
Day 12	4
Day 14	0

4. Sprint Review

Demo:

- Presented visualizations showing correlations between vehicle features such as price, mileage, and days on market.
- Discussed encoding techniques used and their impact on data preparation.

Feedback:

- Instructor: Suggested exploring dimensionality reduction techniques for simplifying dataset complexity.
- Go Auto: Recommended making visualizations more straightforward for easier understanding.

5. Sprint Retrospective

What Went Well:

- Successful creation of visualizations and encoding.
- Demo 1 delivered on time.

What Didn't Go Well:

- Dataset complexity increased significantly due to one-hot encoding.

Improvements: Explore dimensionality reduction techniques to address high-dimensional data challenges.

6. Product Backlog

Backlog Item	Priority	Story Points	Status
Model Implementation and Testing	High	10	To Do
Further Visualizations Optimization	Medium	8	To Do
Develop Encoding Alternatives	Medium	6	To Do
Feature Engineering and Evaluation	High	9	To Do

7. Definition of Done (DoD)

- Code Quality Verified: Advanced visualizations and encoding techniques have been implemented following the project's coding standards.
- Testing Completed: Unit and integration tests have been performed for encoding techniques and visualizations to ensure correctness and consistency.
- Team Review Conducted: All visualizations and encoding implementations have been

reviewed by peers, and improvements based on feedback have been made.

- Detailed Documentation Provided: Documentation for the visualizations, encoding approaches, and the Demo 1 presentation has been completed to provide clarity for stakeholders and future reference.
- Successfully Integrated: Visualizations and encoded datasets have been fully integrated, and Demo 1 materials have been finalized and presented to stakeholders.

14. References

1. Invensis Learning. (2022). *Scrum tutorial for beginners | Scrum methodology | Scrum training* [Video]. YouTube. https://www.youtube.com/watch?v=HA5f54QNQWM

2. Go Auto & Canadian Black Book. (2024). CBB vehicle listings dataset. Go Auto.