CMPT 3830: Machine Learning
Work Integrated Learning-1

Project Report: Phase 1
Project Title: The Data Pit Crew

In collaboration with

Submitted By:

Name	ID	Email
Gurleen kaur	3105332	gkaur@norquest.ca
Jashraj vashisht	3104596	jvashisht@norquest.ca
Tanish dhawan	3108427	tdhawan27@norquest.ca
Mankaran singh	3106501	msingh501@norquest.ca
Navanjot singh	3108144	nsingh144@norquest.ca

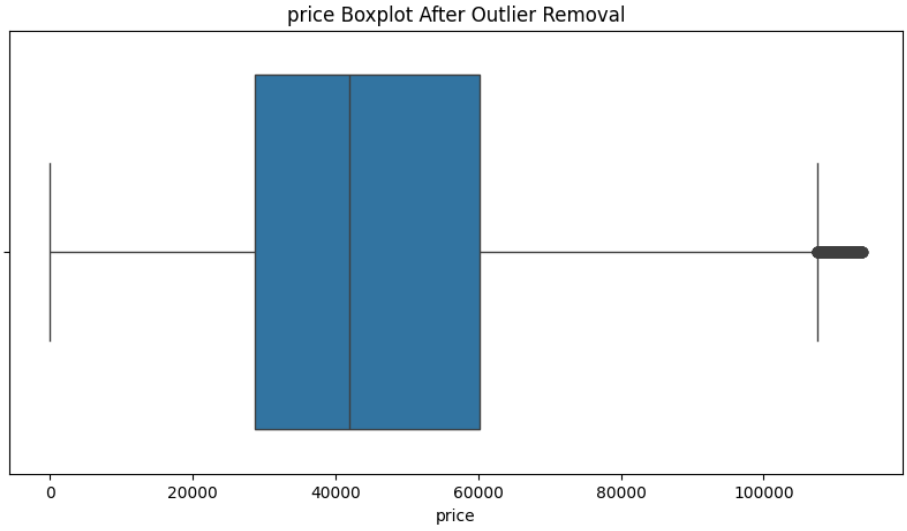
Submission: Date: 27 February 2025
Winter 2025

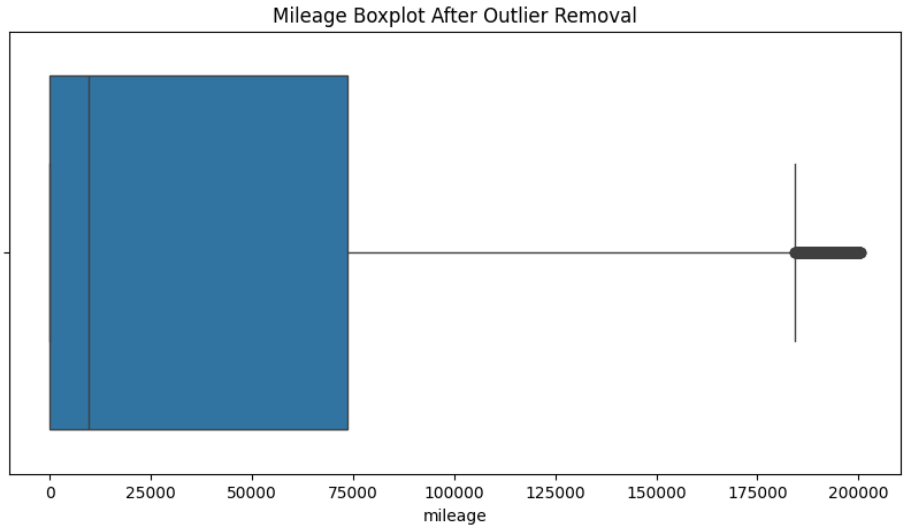
Table of Contents:
A table of content is required with section numbers, names with correct page numbers. An example is shown below.

Contents

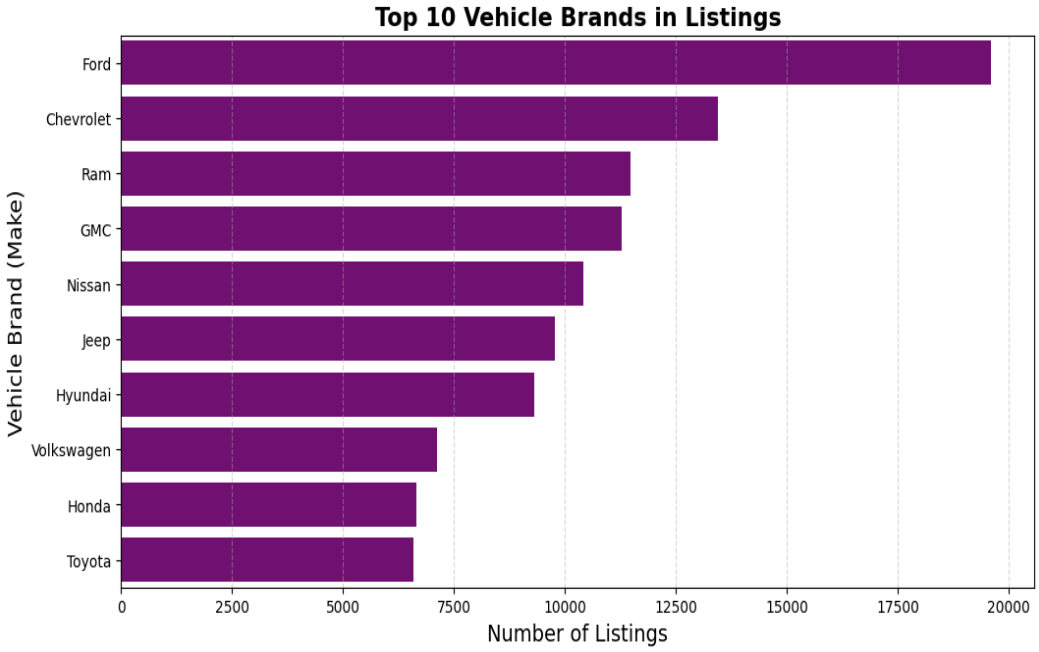
List of Figures:
Figures in each page (if any) MUST be listed in this section.

1.

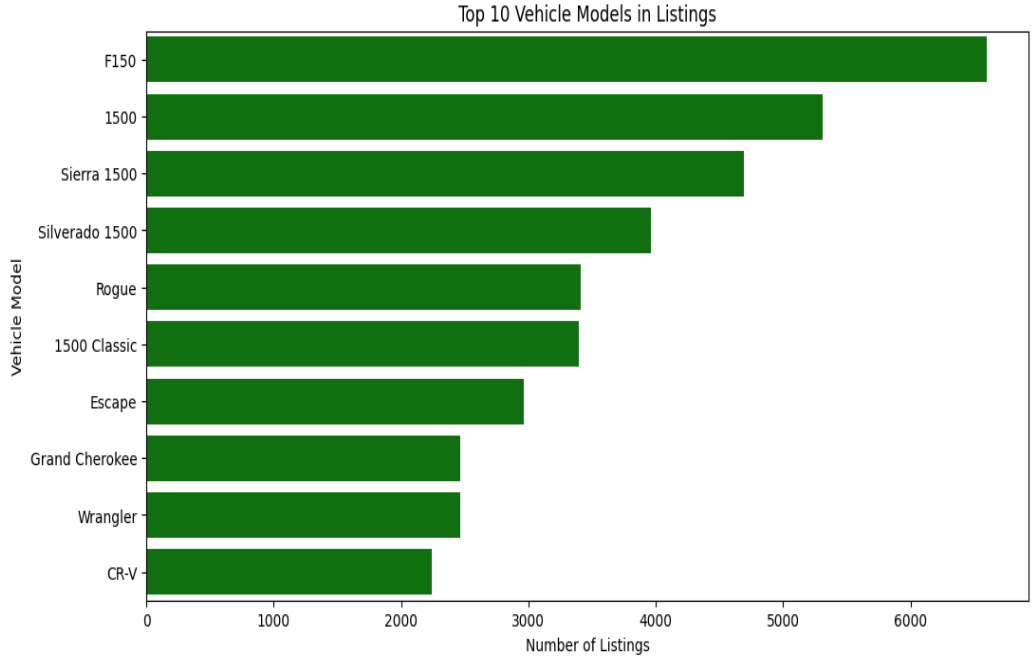




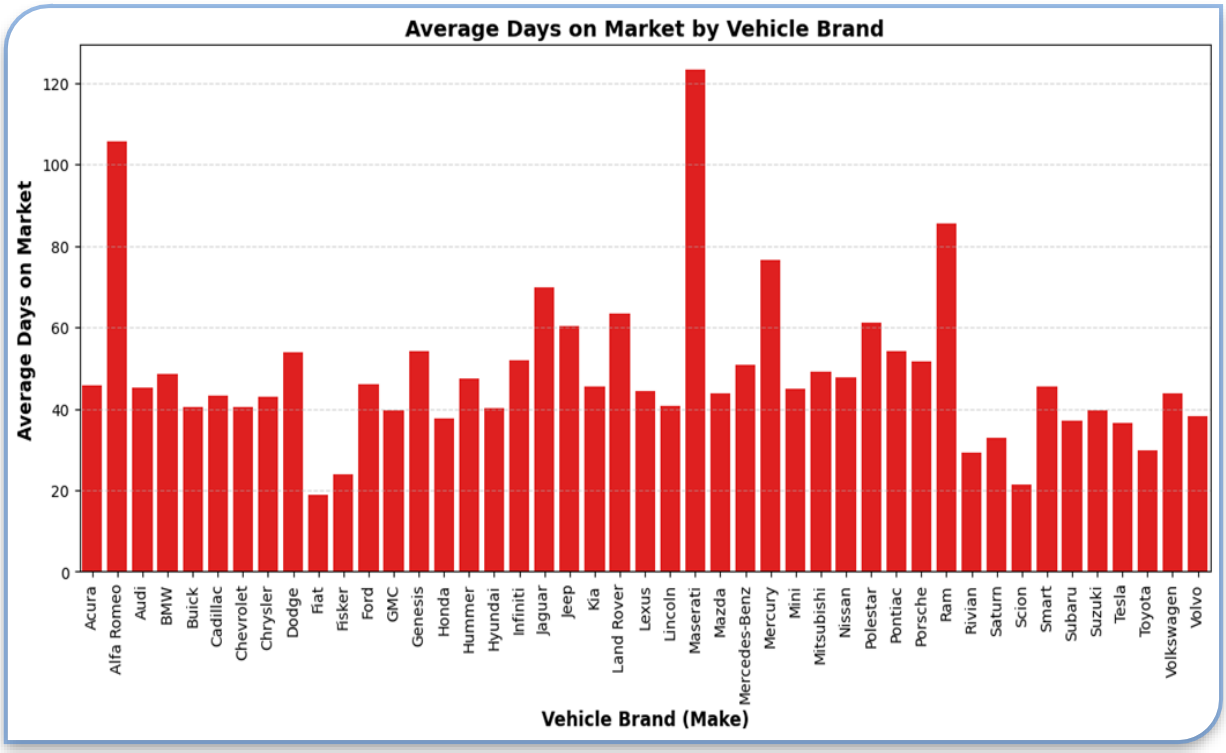
2.


Top 10 Vehicle Brands in Listings
Ford
Chevrolet
Ram
GMC
Nissan
Jeep
Hyundai
Volkswagen
Honda
Toyota
Vehicle Brand (Make)
0
2500
5000
7500
10000
12500
15000
17500
20000
Number of Listings

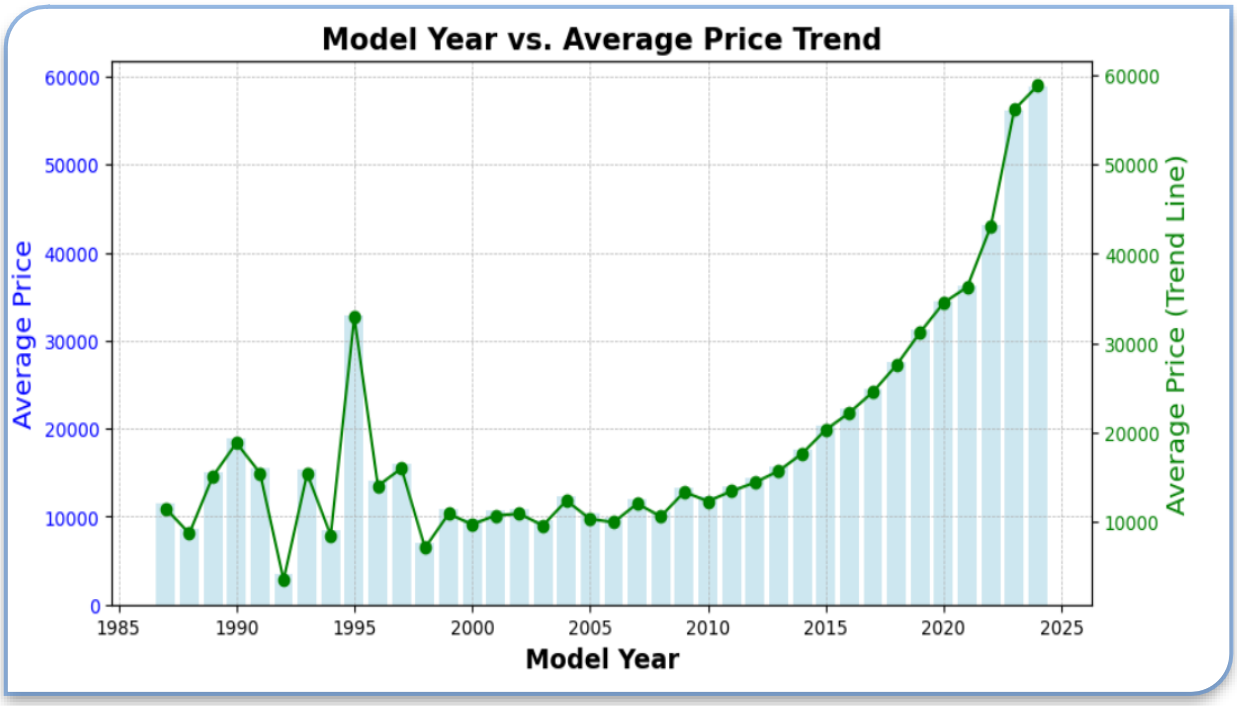
3.


Top 10 Vehicle Models in Listings
F150
1500
Sierra 1500
Silverado 1500
Rogue
1500 Classic
Escape
Grand Cherokee
Wrangler
CR-V
Vehicle Model
0
1000
2000
3000
4000
5000
6000
Number of Listings

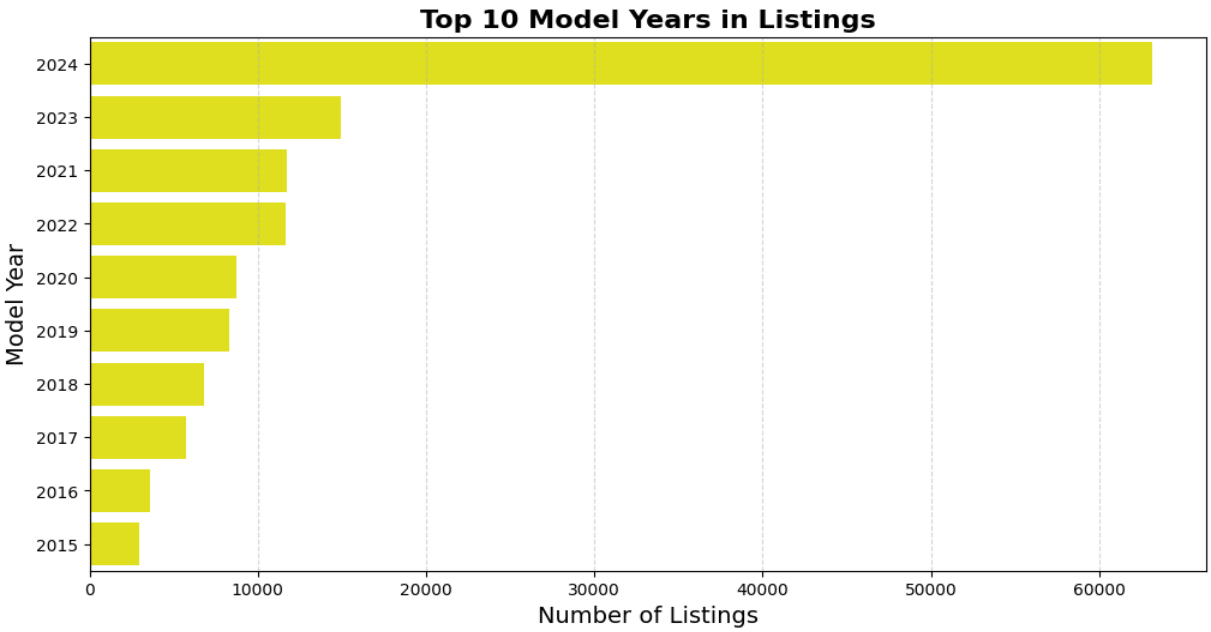
4.


Average Days on Market by Vehicle Brand
Average Days on Market
0
20
40
60
80
100
120
Acura
Alfa Romeo
Audi
BMW
Buick
Cadillac
Chevrolet
Chrysler
Dodge
Fiat
Fisker
Ford
GMC
Genesis
Honda
Hummer
Hyundai
Infiniti
Jaguar
Jeep
Kia
Land Rover
Lexus
Lincoln
Maserati
Mazda
Mercedes-Benz
Mercury
Mini
Mitsubishi
Nissan
Polestar
Pontiac
Porsche
Ram
Rivian
Saturn
Scion
Smart
Subaru
Suzuki
Tesla
Toyota
Volkswagen
Volvo
Vehicle Brand (Make)

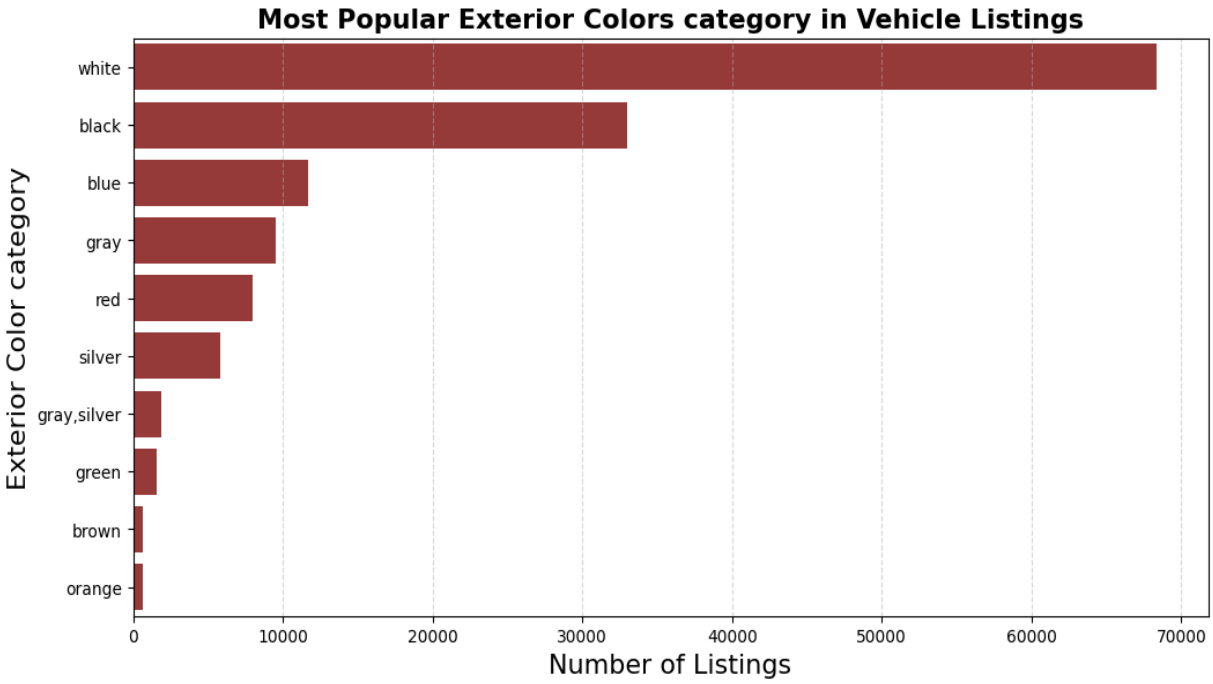
5.


Model Year vs. Average Price Trend
Average Price
Average Price (Trend Line)
Model Year
0
10000
20000
30000
40000
50000
60000
1985
1990
1995
2000
2005
2010
2015
2020
2025

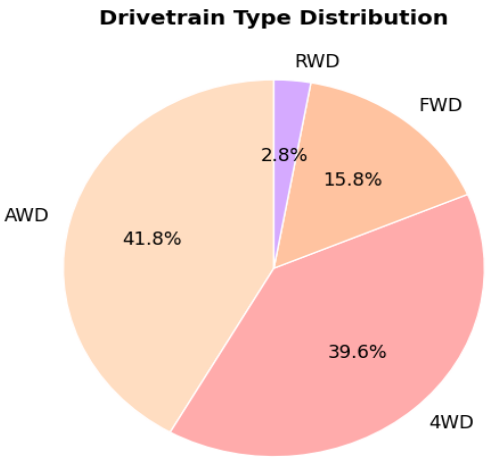
6.


Top 10 Model Years in Listings
Model Year
Number of Listings
2024
2023
2021
2022
2020
2019
2018
2017
2016
2015
0
10000
20000
30000
40000
50000
60000

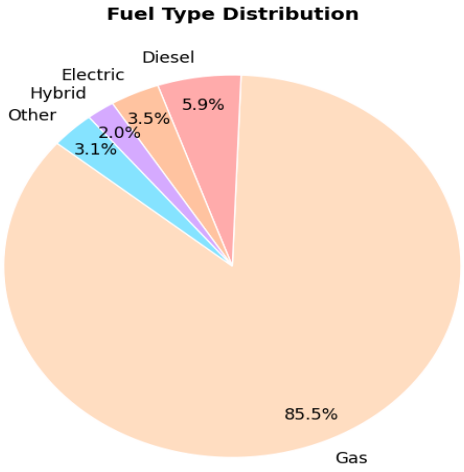
7.


Most Popular Exterior Colors category in Vehicle Listings
Exterior Color category
white
black
blue
gray
red
silver
gray,silver
green
brown
orange
0
10000
20000
30000
40000
50000
60000
70000
Number of Listings

8.


Drivetrain Type Distribution
RWD
FWD
2.8%
15.8%
AWD
41.8%
39.6%
4WD

9.



List of Tables:

We did not used tables in our project

1. Project Phase:

In this phase, we focused on preparing and understanding the dataset before applying machine learning techniques. Our main tasks included cleaning the data, filling in missing values, detecting and fixing outliers, and performing exploratory data analysis (EDA) to find patterns. This step is very important because it helps make sure the data is good for further analysis.

2. Team Members' Name with specific roles

S NO.	NAME	STUDENT ID	ROLES	Responsibilities	EMAIL
1.	MANKARAN SINGH	3106501	Data Preparation Specialist	- Design initial dataset preparation steps - Study data to discover useful insights	msingh501@norquest.ca
2.	JASHRAJ VASHISHT	3104596	Team Leader	- Manage project schedule and assign tasks - Monitor key progress points - Communicate with the teacher for any doubts - Lead dataset preparation	jvashisht@norquest.ca
3.	GURLEEN KAUR	3105332	Data Visualization Specialist	- Design dashboards and visual graphs for analysis - Create interactive discovery tools	gkaur22@norquest.ca
4.	TANISH DHAWAN	3108427	Machine Learning Engineer	- Create and optimize machine learning systems - Use clustering methods to uncover key insights	tdhawan27@norquest.ca
5.	NAVANJOT SINGH	3108144	Documentation & Presentation Lead	- Write explanations and prepare presentation materials	nsingh144@norquest.ca

				- Document findings and lead demos & presentations	

3. Reporting Period:

Data Collection	**January 16, 2025**
Problem Definition	**January 20, 2025**
Team Charter Submission	**January 23, 2025**
Exploratory Data Analysis (EDA)	**January 30, 2025**
Data Preprocessing	**February 3, 2025**
Feature Engineering	**February 7, 2025**
Demo 1: EDA & Feature Engineering	**February 13, 2025**
Model Selection	**February 17, 2025**
Model Training & Evaluation	**February 20, 2025**
Model Deployment	**February 25, 2025**
Phase 1 Report Submission	**February 27, 2025**

4. **Project Overview:**
   This project aims to understand which cars are more popular in the market based on factors like price, mileage, year, and days on the market. By analyzing data, we hope to find trends that show why some cars sell faster than others. This will help car dealerships and buyers make better decisions

5. Dataset

- **Dataset Source**: The data comes from CBB Listings.
- **Dataset Size**: 145,114 rows (cars) and 46 columns (different details about each car).
- **Key Features Used**:
  - **Numerical Features**: Price, mileage, model year, days on market.
  - **Categorical Features**: Make (brand), model, stock type, exterior color, interior color.

Target Variable: Popularity is measured by how many days a car stays on the market. Fewer days mean a car is more popular

5.1 Exploratory Data Analysis (EDA) Highlights:

- **Missing Values**: We checked for missing values using the df.isnull().sum() function.
- **Handling Missing Values**: We used different methods to fill missing values. Categorical values were filled with the most common category (mode), and numerical values were filled with the average (mean).
- **Zero Values**: Some cars had zero values in important features like mileage. We fixed this by using averages based on the stock type.
- **Outliers**: We found and removed extreme values in price and mileage using the Interquartile Range (IQR) method to make sure they didn't affect our analysis.

5.2 Visualization:

- **Box Plots**: Used to show how we handled outliers in price and mileage.
- **Graphs and Charts**: Created to show relationships between features, such as how price changes with mileage or model year.
- **Feature Analysis**: We created graphs to see which car brands and models are the most popular.

6. Challenges Encountered:

- **Missing Data**: Some important information was missing, and we had to decide the best way to fill it.
- **Outliers**: Extreme values in the dataset had to be handled properly so they didn't affect the analysis.
- **Balancing Data**: Some brands and models had very few entries, making it harder to analyze them properly.

7. Stakeholder Engagement:

- We shared our findings with stakeholders to get feedback.
- They suggested adding more data points to improve accuracy
- Their insights helped us refine our analysis

8. Lessons Learned:

- **Data Cleaning is Essential**: Without handling missing values and outliers, the analysis can be misleading.

- **Visualization Helps Understand Data**: Graphs and charts make it easier to see patterns in the data.
- **Feature Selection Matters**: Choosing the right variables improves model accuracy.

9. Future Recommendations:

- Use **clustering algorithms** to group cars based on features and popularity.
- **Improve data collection** to get more details on features affecting car sales.
- **Test different machine learning models** to predict car popularity more accurately.

10. Impact on the Community:

- **Better Decision-Making for Buyers**: Helps customers understand which cars sell fast and which hold their value.
- **Optimized Pricing for Dealerships**: Car dealers can adjust prices based on market trends.
- **Improved Inventory Management**: Helps car sellers know which models are in high demand.

11. Project Conclusion:

We successfully prepared the dataset, handled missing values, fixed outliers, and performed data analysis. We identified trends in vehicle popularity based on key features like price and mileage. In the next phase, we will apply machine learning models to segment cars and predict demand

12. Acknowledgments:

We would like to thank our team members, stakeholders, and our instructors for their guidance and support throughout the project.

13. Appendices:

- Missing values checked by using : df.isnull().sum()

has_navigation	0
exterior_color	6049
exterior_color_category	34947
interior_color	51663
interior_color_category	58781
price_analysis	0

- To handle missing values, we mainly used mode and mean imputation

```
df['exterior_color_category'] = df['exterior_color_category'].fillna(df['exterior_color_category'].mode()[0])
df['interior_color_category'] = df['interior_color_category'].fillna(df['interior_color_category'].mode()[0])
```

- Based on stock type we handled zeros in mileage column

```
zeros_count = df[df['mileage'] == 0].groupby('stock_type').size()

print(zeros_count)
```

```
stock_type
NEW     10147
USED       28
```

- Checking outliers in price columns, code will count the outliers which are 3797

```
# Calculate Q1, Q3, and IQR
Q1_price = df['price'].quantile(0.25)
Q3_price = df['price'].quantile(0.75)
IQR_price = Q3_price - Q1_price

# Define lower and upper bounds for outliers
lower_bound_price = Q1_price - 1.5 * IQR_price
upper_bound_price = Q3_price + 1.5 * IQR_price

# Identify outliers
outliers_price = df['price'][(df['price'] < lower_bound_price) | (df['price'] > upper_bound_price)]

# Display the number of outliers
outliers_price_count = len(outliers_price)
print("outliers_price_count:", outliers_price_count)
```

```
outliers_price_count 3797
```

- Capping outliers by using IQR METHOD

```
# Calculate median price using the 'price' column from the dataframe 'df'
median_price = df['price'].median()

# Replace outliers with median
df['price'] = np.where(
    (df['price'] < lower_bound_price) | (df['price'] > upper_bound_price),
    median_price,
    df['price']
)

# Verify if outliers remain after replacement
price_data_after = df['price'].dropna()
outliers_price_after = price_data_after[(price_data_after < lower_bound_price) | (price_data_after > upper_bound_price)]

# Count remaining outliers
remaining_outliers_price = len(outliers_price_after)
print("remaining_outliers_price:", remaining_outliers_price)
```

```
remaining_outliers_price: 0
```

- CODE OR SYNTIX EXAMPLE FOR VISUALS

```
import matplotlib.pyplot as plt
import seaborn as sns

# Compute average days on market per brand (make)
avg_days_on_market = df.groupby("make")["days_on_market"].mean().sort_values()

# Plot bar chart with same color for all bars
plt.figure(figsize=(14, 6))
sns.barplot(x=avg_days_on_market.index, y=avg_days_on_market.values, color="red")  # Uniform color
plt.xlabel("Vehicle Brand (Make)", fontsize=12, fontweight="bold")
plt.ylabel("Average Days on Market", fontsize=12, fontweight="bold")
plt.title("Average Days on Market by Vehicle Brand", fontsize=14, fontweight="bold")
plt.xticks(rotation=90)  # Rotate labels for readability
plt.grid(axis="y", linestyle="--", alpha=0.5)
plt.show()
```

14. References

- *Google Colab*. (n.d.-b). https://colab.research.google.com/drive/1CUQibqRgAwYaKBBbStywTJFqEE_lQEwL?usp=sharing