M2 Project Deliverable 1
Project Overview Data Acquisition Report

Team
Snehita Moturu - G01388464
Pravallika Avula - G01388664
Sai Roopesh Diddi- G01353614

Prof. Charles Lynch

AIT664 - 001: Information: Representation, Processing & Visualization

George Mason University

17 September 2023

The focus area for analysis:

Understanding the relationship between car characteristics and mileage per gallon (mpg), a crucial aspect in determining fuel economy and environmental effect, is the main emphasis of this research analysis. We specifically want to look into the relationships between car pricing and characteristics like fuel type, engine size, horsepower, and mileage.

Abstract

The automobile industry has had a huge impact on many countries since its introduction in the United States in 1895. There have been numerous improvements and diverse car models produced over the years. When consumers consider purchasing a car, they typically consider affordability, luxury, and safety.

Initial Requirements

To analyze a dataset, it is important to perform data cleaning, which involves eliminating missing values and anomalies. In addition, it is essential to assess the affordability of automobiles, taking into account various features such as the number of doors, cylinders, horsepower, and fuel system. Businesses must carefully examine the multitude of factors that impact car pricing before developing a car model. Due to the increasing prices of new cars making them less accessible to buyers, the global used car industry has experienced significant growth. Various websites offer algorithms for predicting automobile prices, although their accuracy may vary.

We've acquired a complete dataset containing information about automobiles, including the model, fuel type, horsepower, and miles per gallon. The dataset has been verified to match the requirements for regression analysis, allowing us to forecast automobile prices based on these characteristics. Furthermore, we have developed research questions that direct our analysis as we explore the relationships between features and their influence on car costs.

When analyzing the factors that impact a car's price and longevity, we will take into consideration various car manufacturers such as Alfa-Romero, Audi, Chevrolet, Mercury, Mitsubishi, Nissan, Plymouth, Toyota, Volkswagen, and Volvo. We will also categorize cars based on their fuel type, whether it's diesel or gasoline. Additionally, we will analyze the continuous range of horsepower values, which typically falls between 52 and 300, as well as the highway MPG (Miles Per Gallon), which ranges from 14 to 56. Lastly, we will examine the continuous price range of cars, which typically spans from 5113 to 45800. By carefully examining these factors, we can make informed decisions when purchasing a car and gain a better understanding of the automobile industry.

Hypothesis

As we delve into the correlation between car prices and attributes such as mileage per gallon, car make, fuel type, engine size, and horsepower, it is crucial to address missing values and outliers in the dataset.

Additionally, we will investigate how car tuning can potentially reduce the price of a vehicle. Our analysis will place a strong emphasis on categorical variables while ensuring that both numerical and categorical attributes are thoroughly examined.

Questions to be addressed

What factors have the most significant influence on car prices?
Among different car manufacturers, which ones are in higher demand?
Is it accurate to say that gasoline-fueled cars are more efficient than diesel-powered ones?
Were there specific price ranges for cars that experienced increased demand?
Does the presence of missing data affect both car pricing and demand?

Data Source

This dataset is sourced from the 1985 Ward's Automotive Yearbook and was contributed by Jeffery C. Schlimmer to UCI's Machine Learning Repository. It encompasses three distinct categories of information:
Automobile Specifications: This category comprises details about various car characteristics.
Insurance Risk Rating Assignment: It includes the assignment of insurance risk ratings to the vehicles.
Normalized Losses Comparison: This section involves normalized losses in comparison to other cars in the dataset.
Additionally, the dataset contains an analysis of different car models, specifically focusing on the average highway mpg and city-mpg of each car. This dataset serves as a valuable resource for automobile manufacturers, aiding them in designing cars that align with market demand and preferences.

Attributes:

Displacement:	Continuous
mpg:	Continuous
cylinders:	Integer
horsepower:	Continuous
weight:	Continuous
acceleration:	Continuous
model_year:	Integer
origin:	Integer
car_name:	Categorical

Conclusion:

In summary, the information in Module 2 gives the requirements, hypothesis, and a brief explanation of the target area we have chosen. Data preprocessing and exploratory data analysis (EDA), which were covered in Part 2 of the Analysis project, will be included in the following phase.

References

[1] *Auto-mpg dataset*. (2017, July 2). Kaggle. https://www.kaggle.com/datasets/uciml/autompg-dataset

[2] Quinlan,R.. (1993). Auto MPG. *UCI Machine Learning Repository*. (n.d.). https://doi.org/10.24432/C5859H.