Automobile Imports-85 Dataset

Statistical Methods Project Overview

March 23, 2025

Introduction

- In this project, we are exploring the **Automobile Imports-85 Dataset**.
- The dataset contains various attributes of automobiles including technical specifications, performance metrics, and price.
- Our goals:
  - Provide a comprehensive overview of the dataset.
  - Understand the theoretical aspects of the methods we plan to use.
  - Propose analyses based on hypothesis testing and predictive modeling.

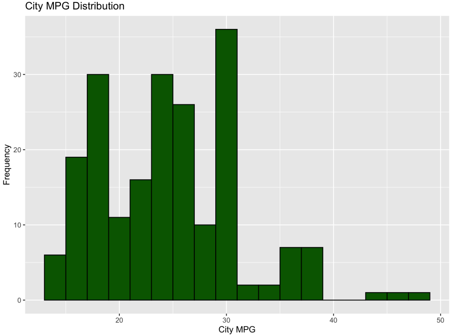
General Description of the Dataset

- The dataset is sourced from the UCI Machine Learning Repository.
- It contains **205** instances representing different automobiles. Some records are incomplete.
- The attributes include both categorical and numerical variables. Examples include:
  - **Symboling**: A risk rating for the automobile (ranging from -3 for very safe to $+3$ for very risky).
  - **Normalized Losses**: A normalized measure of the insurance losses, which indicates repair costs.
  - **Make**: The manufacturer of the automobile (e.g., alfa-romero, audi, bmw).
  - **Fuel Type** and **Aspiration**: Attributes describing the fuel used and engine aspiration (e.g., standard, turbo).
  - **Engine Size** and **Horsepower**: Key performance metrics.
  - **Price**: Often used as a target variable for regression modeling.

Detailed Summary of Attributes

- **Symboling**:
  - Risk factor with values typically ranging from -3 (low risk) to +3 (high risk).
- **Normalized Losses**:
  - Represents relative insurance losses; higher values imply higher repair/maintenance costs.
- **Make**:
  - The car manufacturer, which may influence design, performance, and pricing.
- **Fuel Type** and **Aspiration**:
  - Fuel type (e.g., gas, diesel) and whether the engine is naturally aspirated or turbocharged.
- **Engine Attributes**:
  - Includes engine size (in cubic centimeters), horsepower, and other performance metrics.
- **Performance Metrics**:
  - Such as city and highway MPG, which reflect fuel efficiency.
- **Price**:
  - A key continuous variable often used for regression analysis.

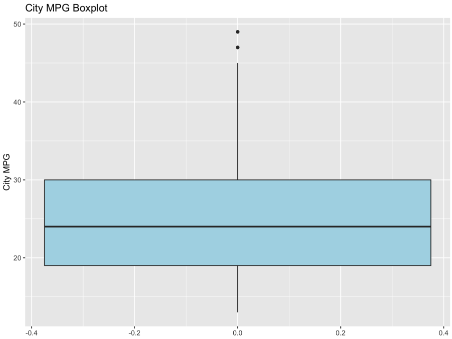
City MPG: Histogram



Interpretation:

- The histogram shows most vehicles have city MPG in the 20–30 range.
- A few cars achieve higher city MPG (over 40), indicating exceptional fuel efficiency.
- This distribution helps us spot the central tendency and potential outliers.

City MPG: Boxplot



Interpretation:

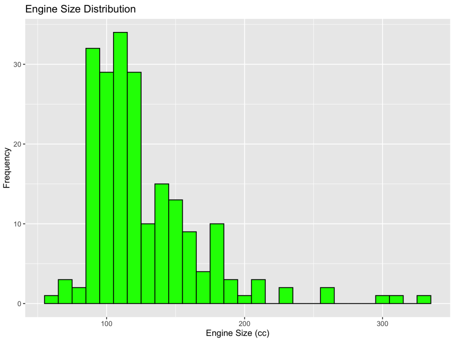
- The median city MPG is around mid-20s.
- Outliers appear in the upper range (above 40 MPG).
- This visualization complements the histogram by showing data spread and outliers.

City MPG Statistics

City MPG

- Mean: 25.21951
- Median: 24
- Variance: 42.79962
- Standard Deviation: 6.542142

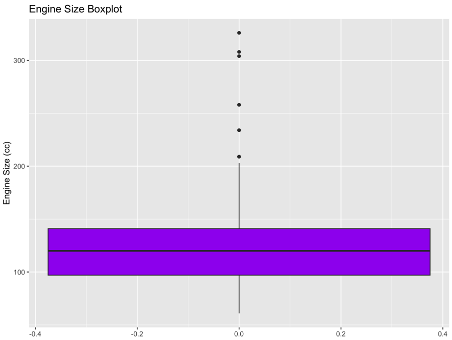
Engine Size: Histogram



Interpretation:

- Distribution centers around 100–150 cc for many vehicles.
- A smaller number of cars have significantly larger engines (above 200 cc).
- Helps identify how engine sizes cluster or spread across the dataset.

Engine Size: Boxplot



Interpretation:

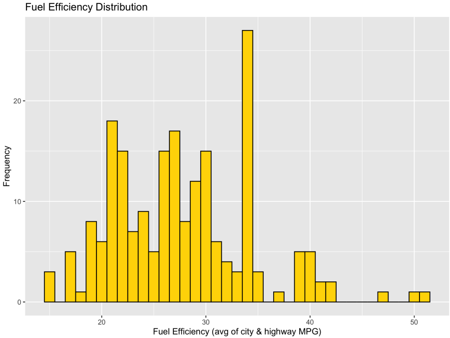
- The boxplot indicates a median near 120 cc.
- The whiskers show the typical range, while any dots above the top whisker highlight very large engine sizes.
- Quickly reveals presence of potential outliers or skew in engine sizes.

Engine Size Statistics

Engine Size

- Mean: 126.9073
- Median: 120
- Variance: 1734.114
- Standard Deviation: 41.64269

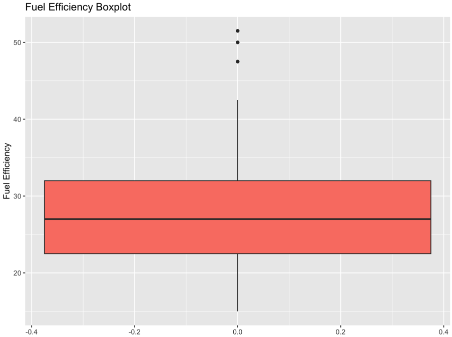
Fuel Efficiency: Histogram



Interpretation:

- Fuel efficiency (mean of city and highway MPG) often clusters in the mid-to-high 20s.
- Fewer vehicles demonstrate extremely high combined MPG (above 40).
- Reflects how well cars balance city vs. highway performance on average.

Fuel Efficiency: Boxplot



Interpretation:

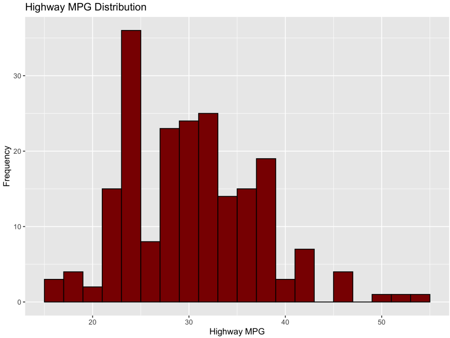
- The median combined MPG is in the upper 20s.
- Outliers surpassing 40 indicate vehicles with notable overall efficiency.
- This helps identify the typical range vs. exceptional performers.

Fuel Efficiency Statistics

Fuel Efficiency (Average of City & Highway MPG)

- Mean: 27.98537
- Median: 27
- Variance: 44.43606
- Standard Deviation: 6.666038

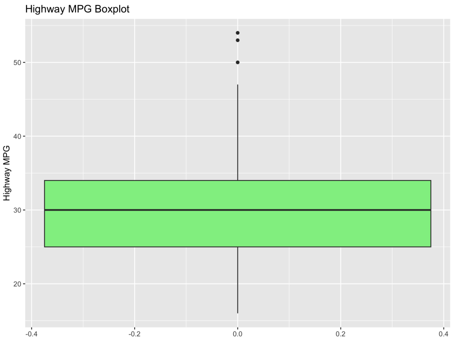
Highway MPG: Histogram



Interpretation:

- Most cars achieve highway MPG in the high 20s to low 30s.
- A small set of vehicles have very high highway MPG (above 40).
- Reflects general trends in fuel efficiency for open-road driving.

Highway MPG: Boxplot



Interpretation:

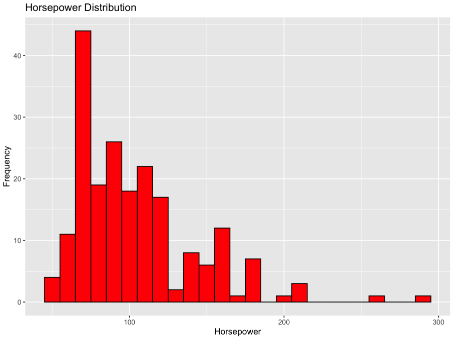
- Median highway MPG around 30.
- Outliers are above 45 MPG, showing highly efficient models.
- This boxplot contrasts city vs. highway distributions, revealing how highway MPG tends to be higher overall.

Highway MPG Statistics

Highway MPG

- Mean: 30.75122
- Median: 30
- Variance: 47.4231
- Standard Deviation: 6.886443

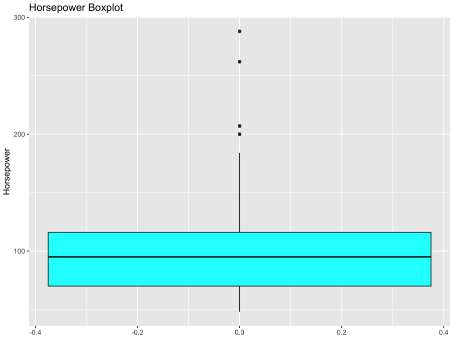
Horsepower: Histogram



Interpretation:

- The horsepower distribution is centered around 80–120 HP for many cars.
- Some vehicles reach 150+ HP, reflecting higher-performance models.
- Highlights the variety of power outputs in the dataset.

Horsepower: Boxplot



Interpretation:

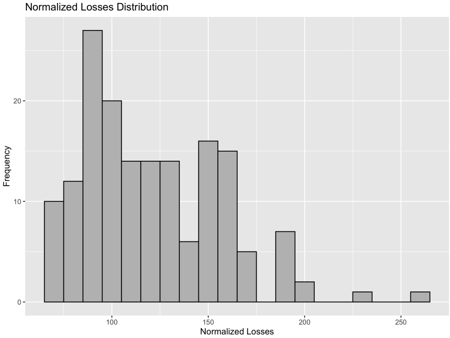
- Median horsepower near 95 HP.
- Whiskers show typical range, with outliers above 160 HP.
- This reveals a skew towards lower horsepower but with some high-end performance cars.

Horsepower Statistics

Horsepower

- Mean: 104.2562
- Median: 95
- Variance: 1577.231
- Standard Deviation: 39.71437

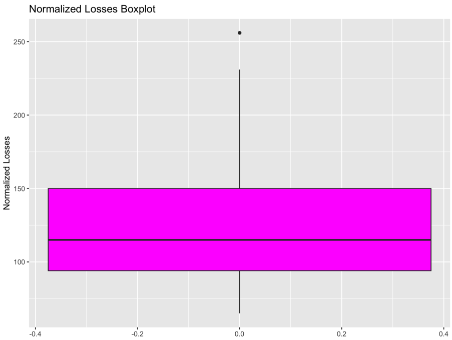
Normalized Losses: Histogram



Interpretation:

- The majority of normalized losses fall between approximately 80 and 150.
- A secondary cluster exists in the 150–200 range, with a few extremely high values over 200.
- Reflects varying insurance repair costs, indicating certain cars may be costlier to insure.

Normalized Losses: Boxplot



Interpretation:

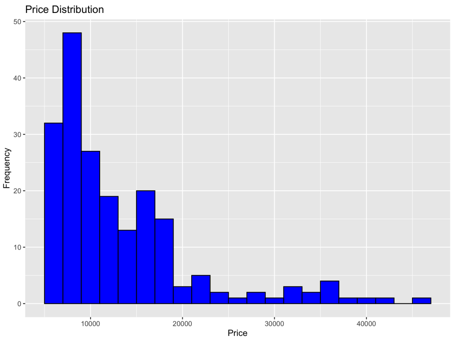
- The boxplot indicates the median near 115.
- One outlier above 250 suggests a notably higher insurance loss estimate.
- Highlights the spread and skew in repair cost risk for different vehicles.

Normalized Losses Statistics

Normalized Losses

- Mean: 122
- Median: 115
- Variance: 1256.147
- Standard Deviation: 35.44217

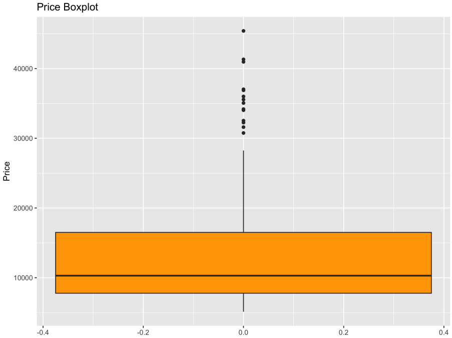
Price: Histogram



Interpretation:

- Most cars range from \$5,000 to \$15,000, peaking near \$10,000.
- A handful of luxury or high-performance vehicles exceed \$30,000.
- Strongly right-skewed distribution, typical for car prices in a diverse dataset.

Price: Boxplot



Interpretation:

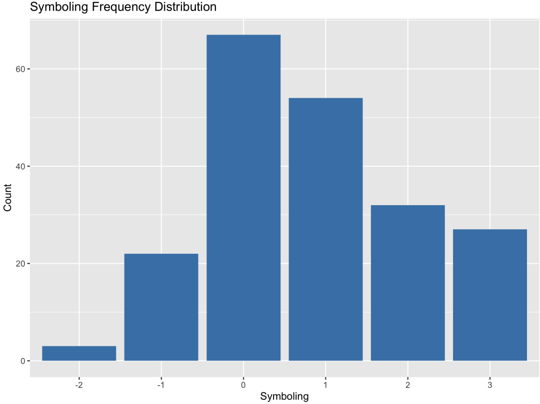
- Median price is around \$10,000.
- Many outliers exceed \$20,000, reflecting expensive, possibly luxury models.
- Conveys both the typical price range and a significant tail for high-end cars.

Price Statistics

Price

- Mean: 13207.13
- Median: 10295
- Variance: 63155863
- Standard Deviation: 7947.066

Symboling Frequency



Interpretation:

- Symboling **0** and **1** are the most common risk ratings.
- Fewer cars exhibit negative symboling (safer) or higher positive symboling (riskier).
- Indicates that moderate risk levels dominate the dataset.

Symboling Statistics

Symboling

- Mean: 0.8341463
- Median: 1
- Variance: 1.550789
- Standard Deviation: 1.245307

Frequency Distribution:

- -2: 3
- -1: 22
- 0: 67
- 1: 54
- 2: 32
- 3: 27

Expected Relationship: Engine Size, Horsepower, and Price

Hypothesis: Larger engine sizes and higher horsepower values will correlate with higher prices.

- **Engineering Rationale:**
  - Bigger engines and more horsepower often require more expensive manufacturing.
  - Performance-oriented vehicles (e.g., sports cars) typically have higher retail prices.
- **Market Dynamics:**
  - Consumers often pay a premium for powerful or luxury vehicles.
  - Insurance costs, maintenance, and brand positioning can also elevate final price.

Expected Relationship: Fuel Type, Aspiration, and MPG

Hypothesis: Fuel type (gas vs. diesel) and aspiration (turbo vs. standard) will influence fuel efficiency and performance metrics.

- **Fuel Type:**
  - Diesel engines often have higher fuel economy but may sacrifice acceleration.
  - Gas engines can offer quicker acceleration, but potentially lower MPG.
- **Aspiration:**
  - Turbocharged engines deliver more power for a given engine size.
  - However, they can also consume more fuel if driven aggressively.
- **Outcome for City/Highway MPG:**
  - We expect to see differences in both city and highway MPG across these categories.

Expected Relationship: Symboling, Insurance, and Other Categorical Factors

Hypothesis: Symboling (risk rating) correlates with repair costs, and other categorical attributes may show dependencies.

- **Symboling and Costs:**
  - Higher symboling (2 or 3) could imply higher insurance costs or risk of accidents.
  - Lower symboling (-2 or -1) indicates safer cars with reduced repair/maintenance costs.
- **Body Style, Drive Wheels, etc.:**
  - These categorical attributes may show associations via chi-square tests.
  - Example: Body style (sedan, hatchback, convertible) vs. price range or fuel type.
- **Statistical Exploration:**
  - Conducting chi-square tests can reveal dependencies among categories (e.g., *Fuel Type* vs. *Make*).

Expected Relationship: Normalized Losses and Repair Costs

Hypothesis: Higher Normalized Losses correspond to cars that potentially incur higher repair/maintenance costs, influencing overall affordability.

- **Connection to Insurance:**
  - Normalized losses often reflect the average repair expense for a given model.
  - Vehicles with higher normalized losses may be pricier to insure or maintain.
- **Impact on Consumer Choice:**
  - Consumers might avoid models with very high expected repair costs.
  - In some cases, higher performance or luxury vehicles have higher normalized losses.
- **Statistical Exploration:**
  - We may see a correlation between Normalized Losses and Price or Symboling.

Expected Relationship: City MPG vs. Highway MPG

Hypothesis: Automobiles efficient in city driving often exhibit relatively higher highway MPG, but the improvement may vary by engine configuration and body style.

- **Driving Conditions:**
  - City driving tends to involve stop-and-go traffic, reducing efficiency.
  - Highway driving allows for steadier speeds, generally leading to better fuel economy.
- **Influencing Factors:**
  - Aerodynamics, transmission design, and engine tuning can cause varying gaps between city and highway MPG.
  - Some hybrid or diesel models show comparatively smaller differences between city and highway MPG.
- **Analysis Approach:**
  - Correlation tests may reveal strong or moderate relationships between city and highway MPG.
  - Segmenting by body style or engine type (turbo vs. non-turbo) could further clarify these patterns.

Hypothesis Testing: Detailed Overview

Our hypothesis testing will focus on evaluating associations between variables:

- **Null Hypothesis (H_0)**: Assumes no association between the variables under investigation (e.g., "Engine size is independent of price").
- **Alternative Hypothesis (H_1)**: Assumes there is a statistically significant association (e.g., "Engine size is associated with price").
- **Testing Procedure:**
  - For categorical variables, we will use the **Chi-square test of independence**.
  - For numerical relationships, correlation tests or regression analyses will be used.
- **Statistical Significance:**
  - A p-value less than $\alpha = 0.05$ will lead us to reject the null hypothesis.
  - Effect sizes, such as **Cramér's V**, will be computed to quantify the strength of associations.

Predictive Modeling: Decision Tree Analysis

Decision Trees will be used for classification and regression:

- **Algorithm:**
  - We plan to use the CART (Classification and Regression Trees) algorithm.
  - The tree splits data based on impurity measures (e.g., Gini index or information gain) for classification, and variance reduction for regression.
- **Model Building:**
  - Data will be split into training and testing subsets.
  - Cross-validation techniques will be used to prevent overfitting.
- **Interpretability:**
  - Decision trees provide intuitive rules that help identify the most influential attributes.

Predictive Modeling: Regression Analysis

Regression Analysis will help us predict continuous outcomes such as **price**:

- **Linear Regression:**
  - Models the relationship between one or more predictor variables and a continuous target variable.
  - Assumes linearity between predictors and the target.
- **Model Evaluation:**
  - Performance metrics include R-squared, RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error).
  - Diagnostic plots will be used to assess assumptions (e.g., normality of residuals, homoscedasticity).
- **Extensions:**
  - If the linear model is insufficient, we may explore polynomial regression or regularization techniques (e.g., Ridge, Lasso).

Proposal of Analyses

Our comprehensive analysis plan includes:

1. **Data Cleaning and Preparation:**
   - Handle missing values and encode categorical variables appropriately.
2. **Exploratory Data Analysis (EDA):**
   - Compute frequency distributions and visualize each attribute using bar plots, histograms, and scatter plots.
3. **Inferential Statistics and Hypothesis Testing:**
   - Use chi-square tests for categorical variables and correlation tests for numerical variables.
   - Formulate null and alternative hypotheses (e.g., "Engine size is associated with price").
4. **Predictive Modeling:**
   - **Decision Trees:** Build classification/regression trees using CART with cross-validation.
   - **Regression Analysis:** Develop linear or polynomial regression models to predict price, evaluating model assumptions and performance.

Conclusion and Future Work

- Our study aims to identify the key factors that influence automobile characteristics and pricing.
- Future work may include:
  - Incorporating ensemble methods or other advanced modeling techniques.
  - A deeper analysis of variable interactions and model refinements.
- Our findings will be compared with established automotive market theories to validate the insights.

References

Dua, D. and Graff, C. (2017). *UCI Machine Learning Repository: Automobile Data Set*. Available at: `https://archive.ics.uci.edu/ml/datasets/Automobile` (Accessed: March 13, 2025).

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.

AutoVista: Automobile Imports-85 Dataset Project. GitHub repository. Available at: https://github.com/piotrek1459/AutoVista