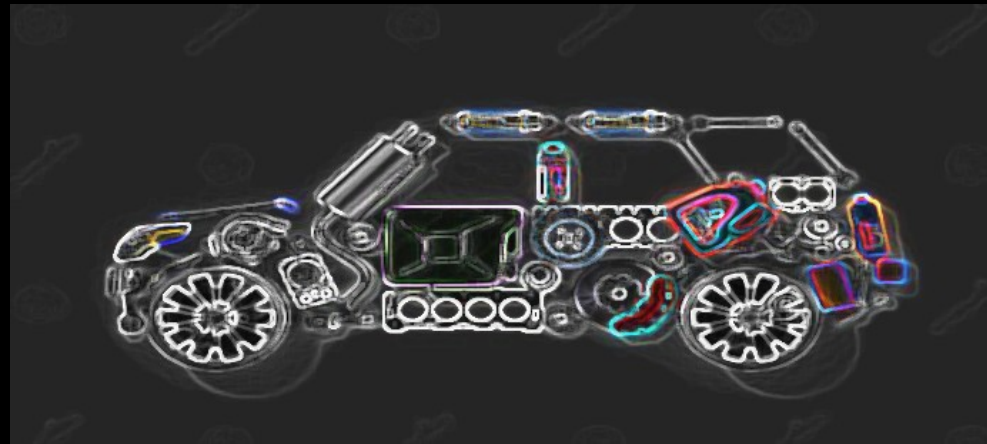


1985 AUTO IMPORTS - CAR SPECIFICATION DATA



Name: Anusree R
Course: DA & DS (Cohort A)
EDA Project: Milestone 2
Date: 20/12/2024

Project Overview

Data set consists of three types of entities:

- (a) the specification of an auto in terms of various characteristics,
- (b) its assigned insurance risk rating
- (c) its normalized losses in use as compared to other cars.

OBJECTIVE:

The primary objective of analyzing the **1985 Automobile Dataset** is to explore and understand the relationships between various attributes of automobiles, such as price, engine specifications, fuel efficiency, and design characteristics, to gain insights into factors influencing automobile performance and pricing segmentation.



Data Understanding

Shape of the Dataset: 205 observations & 26 attributes/features

Datatypes: float64(5), int64(5), object(16)

List of attributes in the dataset:

- Symboling
- normalized-losses
- Make
- fuel-type
- Aspiration
- body-style
- drive-wheels
- engine-location
- wheel-base
- Length
- Width
- Height
- price
- Curb-weight
- Engine-type
- Num-of-cylinders
- engine-size
- fuel-system
- Bore
- Stroke
- compression-ratio
- Horsepower
- peak-rpm
- city-mpg
- highway-mpg
- Number of doors

Missing / Invalid Values Handling

Data Quality Issues

There were no column names, created column names manually & updated the dataset

Using Replace method, '?' converted into NaN for better handling & to detect missing values.

Used "pd.to_numeric" type conversion in order to ensure if any value contains non-numeric string type, it will convert into NaN.

After handling '?' with NaN and convert numeric columns to appropriate types, Using "df.isnull().sum()", able to identify Missing values in:

- I. normalized-losses : 41
- II. Num-of-doors : 2
- III. Bore: 4
- IV. Stroke: 4
- V. Horsepower: 2
- VI. Peak-rpm: 2
- VII. Price: 4



Techniques Used for Handling

Used **KDEPlot & df['normalized-losses'].skew()** to analyze the distribution of the Data points & its skewness of every Attribute consisting missing values.

Imputation Based on Distribution:

- For numerical data, replace missing values with the mean or median depending on the KDE plot's peak.
- For categorical data, used mode or the most frequent category

Numerical Assessment of Skewness:

- Skewness > 0 : Right-skewed.
- Skewness < 0 : Left-skewed.



Techniques Used for Handling

Mean Replacement: Numerical data with a normal distribution (symmetrical, no skew).

- Advantages: Preserves the overall mean of the data.
- **Bore - Symmetrical not skewed**
- **Peak-rpm - Symmetrical not skewed**

Median Replacement: Numerical data with a skewed distribution (right or left skew) & Data containing outliers.

- Advantages: Robust to outliers and better represents the central tendency for skewed data.
- **Normalized loss - Positively/Right skewed**
- **Stroke - Negatively/Left skewed**
- **Horsepower - Positively/Right skewed**
- **Price - Positively/Right skewed**

Mode Replacement: Categorical data (e.g., fuel-type, body-style). Numerical data with many repeated values.

- Advantages: Preserves the most frequent category or value, maintaining data consistency.
- **Num of doors - Does not contribute much to my analysis, Dropped this column**

Outlier Handling

Identifying Outliers:

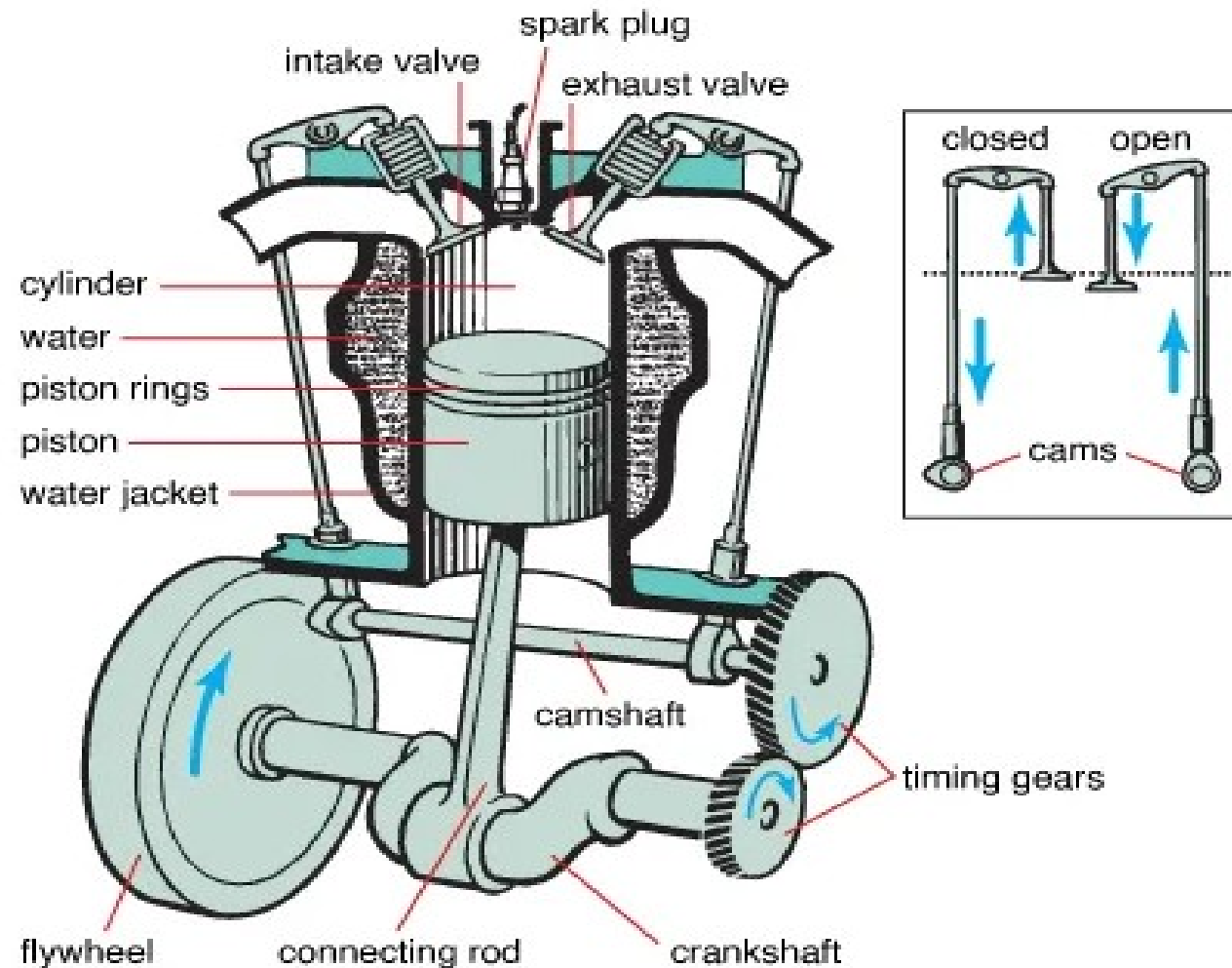
- Visually, using Box plot Identified outliers for all the attributes of the existing dataset
- Used **Z-score method** to detect and removes outliers based on Z-scores for numeric columns:
- **'symboling', 'normalized-losses', 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg', 'highway-mpg', 'price'**
- `cleaned_df = df[(z_scores < threshold).all(axis=1)]` - Original Data Frame altered to include only the rows where all Z-scores are below the threshold(3)
- Dataset shape after removing outliers via Z-score method: (182, 25)
- Using **IQR method** to detect and remove outliers in **'engine-size'**:
- `outliers = (df['engine-size'] < (Q1 - 1.5 * IQR)) | (df['engine-size'] > (Q3 + 1.5 * IQR))`
- Dataset shape after removing outliers via IQR method: (195, 25)

Derived Metrics

- **Power-to-Weight Ratio** → Derived from “horsepower” and “curb-weight”
- **Fuel Efficiency Ratio** → Derived from “city-mpg” and “highway-mpg”
- **Price per Horsepower** → Derived from “price” and “horsepower”
- **Weight-to-Engine Size Ratio** → Derived from “curb-weight” and “engine-size”

Overall understanding of car mechanism with the Attributes

Working parts of an engine



Aspiration

Fuel System

Cylinders

Bore and Stroke

Horsepower

Peak RPM

Compression Ratio

Different Body STYLES



Sedan



Convertible



Hardtop



Hatchback



Wagon

Data Visualization

Univariate Analysis:

1. Visualize the distribution of price.
2. Analyze the count of fuel-type.
3. Make: Distribution of car makes

Insights:

1. Car prices which are ranged from 5000 to 10000, are most purchased. Car price ranging from 30,000 to 35,000 are least purchased.
2. Gas based cars are more preferred than Diesel
3. Toyota, Nissan & Mazda are most manufactured during the year 1985 compared to other cars.

Bivariate Analysis:


Explore relationships between:

1. Price and Engine size
2. Average Price vs Car make
3. Symboling vs Risk
4. Drive-wheels vs Fuel Efficiency ratio
5. Peak-rpm vs Horsepower
6. Fuel-system vs body-style
7. Fuel-system vs Car Make



Insights

1. Price of the car Vs Engine hold positive correlation & directly proportional to each other
2. Mercedes-Benz and Porsche → Highly priced (Luxury & Performance Focus, Brand Prestige, Higher Production Costs & Target Audience. Whereas, Plymouth, Dodge, and Chevrolet → Low priced (Economy and Accessibility, Utility and Practicality, Market Competition & Cost-Effective Production
3. Higher Symboling values usually indicate vehicles with higher risk levels
4. Drive wheels:
 - FWD → More fuel efficient: 62 %
 - RWD → Less fuel efficient than FWD: 33.1%
 - 4WD / All WD → Least fuel efficient: 4.8%

- 
5. Higher peak RPMs are often associated with higher horsepower, especially in engines designed for performance, whereas Economy-focused engines deliver peak horsepower at lower RPMs to optimize fuel efficiency.
 6. Simpler Systems (1BBL, SPFI, 2BBL): Align with practical and economical body styles like Hatchbacks, Sedans, and Wagons.
Performance-Oriented Systems (4BBL, MFI, SPDI): Pair with sporty and premium body styles like Convertibles, Hardtops, and high-performance Sedans.
Diesel Systems (IDI): Suit utility-focused body styles like Wagons and durable Sedans.
 7. MPFI preferred in all kinds of car makes, coz of more fueling & performance factors. SPFI fuel system is seen only in isuzu, as its less manufactured in the year 1985. This fuel system involved newer technology, high production costs & new for consumer practice.

Multivariate Analysis:

1. Explore relationships between price, engine size, horsepower, and fuel type using pair plots.
2. How City-mpg vs Highway-mpg vary based on the fuel-system?
3. Distribution of city-mpg and curb-weight for each body-style
4. Comparing horsepower, weight, engine size across price segments



Insights

1.

Metric	Petrol	Diesel
Price	Budget to luxury, varies widely	Higher initial cost, better fuel efficiency
Engine Size	Smaller engines for economy, larger for power	Larger engines prioritize torque
Horse Power	High horsepower for performance	Moderate horsepower, torque-focused

3.

Highest City MPG rating seen in : Hatchback & then in Sedan type of Body Style.

Lowest City MPG rating seen in: Convertible type of Body Style

Highest Curb Weight rating seen in: Wagon & then in Sedan type of Body Style.

Lowest Curb Weight rating seen in: Convertible type of Body Style

4.

Luxury Cars Like **Porche, Mercedes, BMW, Audi & Volvo** had higher curb weight, Horsepower & Engine size comparing to high priced & mid-ranged vehicles like **Alfa-Romero, Mercury, Peugeot, Saab, & Mazda, Nissan, Honda, Isuzu, Dodge etc.**

Horsepower & Engine Size are seen to be Directly proportional to each other.

2.

Fuel System	City-mpg	Highway-mpg	Variation Gap	Remarks
1BBL	Low	Moderate	Large	Inefficient in city; steady speeds improve performance.
2BBL	Slightly better than 1BBL	Moderate to good	Moderate	Improved airflow compared to 1BBL, better highway mpg.
MPFI	High	Excellent	Small	Optimized fuel delivery ensures efficiency in all conditions.
SPFI	Moderate	Good	Moderate	Better than carburetors but less precise than MPFI.
4BBL	Low	Moderate to good	Large	Designed for power, not efficiency.
IDI	Low	Better	Moderate	Energy loss in pre-combustion chambers affects efficiency.
MFI	Moderate	Good	Moderate to large	Mechanically controlled, less precise than MPFI.
SPDI	High	Excellent	Small	Real-time optimization makes it highly efficient.

Hypothesis Testing: T-Test

Independent two sample T-test done to check if there is a significant difference in horsepower between high-risk and low-risk cars.

Results:

- $\mu_{\text{high-risk}} = \mu_{\text{low-risk}}$
- $p\text{-value} > \text{significance level (e.g., 0.05)}$, indicating insufficient evidence to reject H_0
- The conclusion is correct and statistically valid: there is no evidence to support a difference in horsepower between the two groups.

Conclusion



The 1985 automobile market reflected a balance between performance and cost, with advanced technologies emerging to improve fuel efficiency. While larger, more powerful cars dominated the high-price segments, there was a growing demand for more efficient and economical vehicles in the mid-range and low-price categories.

Target audience insights suggest that high-risk vehicles cater to performance enthusiasts, while low-risk vehicles are designed for cost-conscious, safety-focused consumers.



THANK
YOU

