Course Project

Exploratory Data Analysis for Machine Learning

Naphon Santisukwongchot

Data Description

It contains data about various car models, with a focus on different performance metrics of the cars. The dataset was originally extracted from the 1974 Motor Trend US magazine, which provides specifications and performance measurements for 32 different car models.

The dataset consists of various car attributes such as miles per gallon (mpg), engine displacement, number of cylinders, weight, and more.

Summary Attribute

mpg: Miles per gallon (fuel efficiency), a continuous variable representing the number of miles the car can travel per gallon of fuel.
cyl: Number of cylinders in the car's engine. This is a categorical variable, representing the type of engine.
disp: Displacement (in cubic inches) of the car's engine. A continuous variable that measures the engine's volume.
hp: Horsepower of the car's engine. A continuous variable indicating the engine's power.
drat: Rear axle ratio. This continuous variable represents the ratio of the car's rear axle gears.
wt: Weight of the car (in 1000 lbs). A continuous variable that shows the car's weight.
qsec: 1/4 mile time. This continuous variable shows the time it takes the car to complete a quarter-mile drag race.
vs: Engine shape (0 = V-shaped, 1 = straight). A categorical variable that indicates the engine's configuration.
am: Transmission type (0 = automatic, 1 = manual). This categorical variable represents whether the car has an automatic or manual transmission.
gear: Number of forward gears. A categorical variable indicating how many gears the car has.
carb: Number of carburetors. A categorical variable representing how many carburetors the car uses.

Initial Plan for Data Exploration

Step 1

Load and Inspect the Data

Import pandas.
Import seaborn.
mtcars.head()
mtcars.info()
mtcars. describe()

Step 2

Check for Missing Data

mtcars.isnull().sum()

Step 3

Examine Data Distributions

Plot histograms or density plots for continuous variables.

Use boxplots to identify potential outliers in numerical features.

Initial Plan for Data Exploration

Step 4

Explore Categorical Variables

Mtcars[ ].value_counts()

Create bar plots to visualize the distribution of categorical data

Step 5

Visualize Relationships Between Variables

Understand relationships between features (both continuous and categorical)

Create scatter plots or pair plots

Step 6

Feature Engineering

Create new features or transform existing ones to improve model performance.

scaler = StandardScaler()

Initial Plan for Data Exploration

Step 7

Detect Outliers

Visualize potential outliers using box plots or scatter plots.

Apply statistical methods to identify and handle outliers.

Step 8

Check Multicollinearity

Identify highly correlated features that might cause multicollinearity issues in regression models.

Correlation matrix
Variance Inflation Factor

Step 9

Summarize Findings

Summarize key insights and prepare the data for modeling or further analysis.

Data Cleaning

Handle Missing Data : mtcars.isnull().sum()

Addressing Duplicate Data : mtcars.duplicated().sum()

Detect and Handle Outliers : sns.boxplot(x='mpg', data=mtcars)

Convert Data Types : mtcars.dtypes

Feature Engineering

```
from sklearn.preprocessing import StandardScaler

#Feature Scaling
scaler = StandardScaler()
mtcars[['mpg', 'hp', 'wt']] = scaler.fit_transform(mtcars[['mpg', 'hp', 'wt']])

#Getdummies
Import pandas as pd
pd.get_dummies(mtcars, columns=['cyl', 'vs', 'am', 'gear', 'carb'], drop_first=True)
```

Key Insight

General Overview of the Data

The dataset contains 32 observations and 11 columns. The 11 columns include both numerical and categorical variables.

- Numerical variables include mpg, hp, wt, and qsec (quarter-mile time).
- Categorical variables include cyl, vs (engine type), am, gear, and carb (number of carburetors).

Key Insight

Descriptive Statistics

Central Tendencies:

- The average mpg (miles per gallon) is approximately 20.1, with the values ranging from 10.4 to 33.9, suggesting variability in fuel efficiency across different cars.
- The average wt (weight) is about 3.3 tons, with values ranging from 1.5 to 5.4, indicating substantial differences in the size and weight of the cars.

Variability:

- The standard deviation of mpg and hp is significant, indicating a wide range of performance (fuel economy and horsepower) across the cars.

Key Insight

Conclusion

- Key variables such as weight, horsepower, and transmission type are important determinants of fuel efficiency (mpg).
- The relationship between weight and horsepower is critical, with heavier and more powerful cars being less fuel-efficient.
- Manual transmission is associated with higher fuel efficiency, making it an important feature to consider.
- Feature engineering such as log transformations and interaction terms could improve model performance.

Hypothesis 1

Hypothesis: Weight of the Car (wt) Is Inversely Related to Fuel Efficiency (mpg)

- Null Hypothesis (H_0): There is no significant relationship between the weight of the car (wt) and its fuel efficiency (mpg).
- Alternative Hypothesis (H_1): Heavier cars have lower miles per gallon (mpg) compared to lighter cars.

Hypothesis 1

Hypothesis: Weight of the Car (wt) Is Inversely Related to Fuel Efficiency (mpg)

```
import pandas as pd
import scipy.stats as stats

correlation, p_value = stats.pearsonr(mtcars['weight'], mtcars['mpg'])

print("Pearson Correlation Coefficient:", correlation)
print("p-value:", p_value)

alpha = 0.05

if p_value < alpha:
    print("Reject the null hypothesis: There is a significant negative correlation between the weight of the car and fuel efficie
else:
    print("Fail to reject the null hypothesis: There is no significant correlation between the weight of the car and fuel efficie
```

```
Pearson Correlation Coefficient: -0.831740933244335
p-value: 2.9727995640500577e-103
Reject the null hypothesis: There is a significant negative correlation between the weight of the car and fuel efficiency.
```

Hypothesis 2*

Hypothesis: Cars with More Cylinders (cyl) Have Higher Horsepower (hp)

- Null Hypothesis (H_0): There is no significant relationship between the number of cylinders (cyl) and horsepower (hp).
- Alternative Hypothesis (H_1): Cars with more cylinders have significantly higher horsepower.

Hypothesis 3*

Hypothesis: Transmission Type (am) Affects Fuel Efficiency (mpg)

- Null Hypothesis (H_0): There is no significant difference in the miles per gallon (mpg) between cars with manual and automatic transmissions.
- Alternative Hypothesis (H_1): Cars with manual transmissions have significantly higher miles per gallon (mpg) than cars with automatic transmissions.

Summary

The mtcars dataset provides a useful and relatively clean collection of vehicle specifications, including key variables such as miles per gallon (mpg), weight (wt), horsepower (hp), and number of cylinders (cyl), among others. It is small, consisting of only 32 observations, which limits its ability to fully represent the variability found in the broader automotive market. The data appears to be well-structured, with no missing values, but its limited size and the potential for outliers or anomalies may affect the generalizability of any findings. Additionally, while it contains valuable information for understanding fuel efficiency and car characteristics, more data points covering a broader range of car models, years, and features would help improve the robustness of analyses and provide a more comprehensive understanding of the relationships between variables. To enhance the quality of analysis, acquiring data from a larger sample size, including different car manufacturers, newer models, and additional features like car price, fuel type, or engine efficiency, would be beneficial.