# EDA_Automobile

This dataset consist of data From 1985 Ward's Automotive Yearbook. Here are the sources:

1. 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
2. Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038
3. Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

Tools & Libraries • Python • Jupyter Notebook • Pandas • Numpy • Seaborn • Matplotlib

Data Description

Attribute Information:

Attribute: Attribute Range

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.

18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

Data Cleaning

I made the following changes and created the following variables: • I first checked head and tail. • I get to know shape and info. Dataset has 205 row and 26 columns. • I came to know all data types. • Then I handled missing and duplicate values. • Normalized losses had most missing values "?". It is replaced with Nan values. Then filled with mean of the rest of the column. And other columns missing values are dropped as they are less in no. • No duplicate values are present. • Changed data type of price. And renamed column names.

EDA

I looked at the different-different trends of the data and Below is a few highlights of the analysis. • Number of doors in cars • Fuel usage by cars • Body styles by each car • Which drive_wheels, engine_location, engine-type, num-of-cylinders, fuel-system is used • Heatmap showing the correlation between numeric type data.