Project Work 1

Data exploration

27.10.2024

In this project, we will explore the "mpg" dataset, which is readily available upon loading the **ggplot2** package in R. This dataset comprises 234 observations of various car models, each accompanied by multiple attributes, including fuel economy, engine type, and drive type. Each row corresponds to a unique car model, and the dataset features 11 variables: 5 numeric variables—**displ** (engine displacement), **year**, **cyl** (number of cylinders), **cty** (city mpg), and **hwy** (highway mpg)—and 6 categorical variables—**manufacturer**, **model**, **trans** (type of transmission), **drv** (type of drive train), **fl** (fuel type), and **class** (car type).

The primary objective of this analysis is to uncover insights into the factors that influence fuel efficiency. To get started, you should first install and load the **ggplot2** package. Next, load the dataset with the command data(mpg) and examine its structure using str(mpg). For additional information about the dataset, you can use ?mpg.

Tasks

1. Visualize the distribution of car classes (class variable) with a bar plot and determine the proportion of each car type. Identify the most common car type.
2. **Engine Size and Fuel Efficiency**
    - Plot a scatterplot of engine displacement (displ) vs. highway fuel efficiency (hwy).
    - Describe any observable trend. Does a larger engine size generally correlate with lower fuel efficiency?
3. **Analyzing Drive Type**
    - Create subsets for each drive type (drv), i.e., f (front-wheel), r (rear-wheel), and 4 (4wd).
    - Compare the average highway fuel efficiency (hwy) across these subsets using mean() and sd().
    - Identify which drive type tends to have the highest and lowest fuel efficiency.
4. **Exploring Fuel Economy by Cylinder Count**
    - Visualize the distribution of city fuel economy (cty) across cars with different numbers of cylinders (cyl) using a boxplot.
    - Determine if there’s a noticeable trend in fuel efficiency with increasing cylinder count.

5. **Factors Influencing Fuel Efficiency**

    - Investigate which variables might influence fuel efficiency (cty and hwy) the most.
    - Use appropriate visualizations (scatterplots, boxplots) to examine relationships between class, displ, cyl, and fuel efficiency.
    - Summarize your findings and suggest possible reasons for the relationships observed.