Exercise - Getting Started with Data Analysis

IST 5520 - Fall 2022, Chen

In this exercise, we'll try some data management and visualization methods in pandas and seaborn packages. Please complete the programming tasks and submit your jupyter notebook with answers to Canvas.

1. Download data file “ToyotaCorolla_FullData.csv”. Import the data file as a pandas dataframe.
    a. How many observations are in the dataset?
    b. How many variables are in the dataset?
    c. Calculate the range (i.e., minimum and maximum) of price, KM, doors, and cylinders.
2. Explore the manufacturing year of used corolla.
    a. How many unique manufacturing years are in the dataset?
    b. Count the number of observations per manufacturing year.
    c. How many observations of cars that were manufactured in year 2000?
    d. Draw a barchart to show the number of observations across manufacturing years.
3. Explore price.
    a. Draw a distribution plot (histogram or/and density plot) of the price column.
    b. Does the price follow a normal distribution?
    c. Draw a barchart to show the number of observations across different fuel types.
    d. Draw box plots of price for each fuel type.
    e. Calculate the average price of cars of each fuel type.
4. Explore the relationship between price and age of used corolla.
    a. Draw a scatterplot to show the relationship between price and age.
    b. What is the relationship between price and age? Does the relationship change within cars of each fuel type?
5. Explore the relationship between price and mileage of used corolla.
    a. Draw a scatterplot to show the relationship between price and mileage.
    b. What is the relationship between price and mileage? Does the relationship change within cars of each fuel type?