University of Bristol
EFIMM0095 (Data Science for Economics)
Professor Vincent Han

Coursework 1

Due: 25 February at 11:00am

- This coursework is a team project. The output will be evaluated as a team, that is, all team members will receive the same score. Therefore, cooperation among team members is important. If there is a free-rider problem, try to figure out how to resolve it. Please show work ethic to your team members.
- All the questions should be answered using Python. Report all the codes used in answering the questions.

1. Consider the data set "Auto.csv," which contains a sample of automobiles and their characteristics. We want to explore this sample by producing summary statistics and visualising the data.
   (a) Browse the data set by creating a DataFrame with the first 20 rows.
   (b) Find the sample mean, sample standard deviation, minimum and maximum (but not other statistics) of all the variables (except *name*). Report the results in a table (DataFrame) where the index is the name of the variables.
   (c) Find the sample mean, sample standard deviation, minimum and maximum (but not other statistics) of all the variables (except *cylinders* and *name*) for cars with 4 cylinders, and find them for cars with 8 cylinders. Report the results in two separate tables, using the same table format as (b).
   (d) Discuss the findings in (c) for *weight* by comparing the means and standard deviations between the two types of cars. To support your discussions, draw box plots that depict the relationship between *cylinders* and *weight*.
   (e) Draw histograms of *weight* for cars with 4 cylinders and then with 8 cylinders, both in one figure. Also draw the empirical cumulative distribution functions of *weight* for them in one figure.
   (f) Discuss the findings in (e). In particular, focus on aspects that were not reflected in the means and standard deviations that you compared in (c).
   (g) Answer (c) and (e) with *mpg* above and below 25 (instead of 4 vs. 8 cylinders).
   (h) Discuss the findings in (g) focusing on *weight*. To support your discussions, draw scatter plots that depict the relationship between *mpg* and *weight*. Try to draw the two scatter plots in one figure to facilitate the comparison.
   (i) How many Ford Mavericks are there in the dataset? Do all Ford Mavericks have the same specs? If not, summarize the difference.