CE888 Lab 2

Damian Machlanski

January 2019

1 Introduction

This lab session consisted of two main parts: data visualisation and the bootstrap technique. The implementation incorporates NumPy for numerical operations, Pandas for loading CSV data and Seaborn for producing plots.

2 Data Visualisation

The aim of this section was to explore the data provided (vehicles.csv) by plotting scaterplots and histograms. In addition, some summary statistics were calculated to get more insight about the data. This section's code can be found in vehicles.py file.

The data contained two columns of integer values, which indicate Miles Per Gallon (MPG) achieved by each car with respect to current and new fleet. Figure 1 presents the scaterplot of vehicles.csv data:

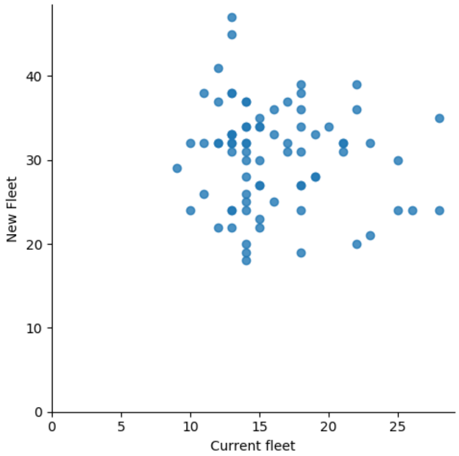
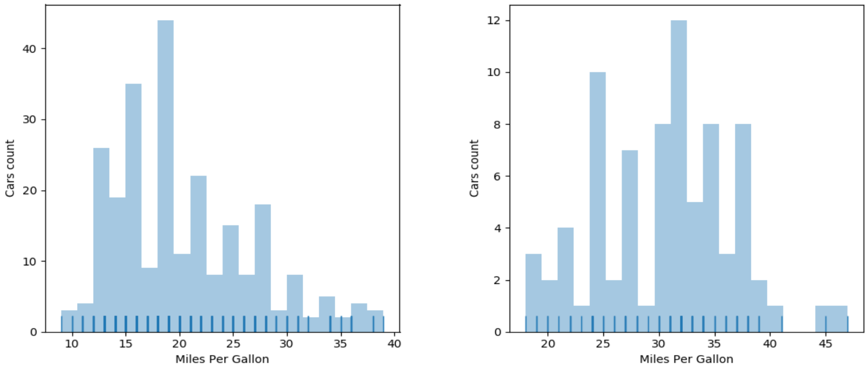


Figure 1: Two fleets compared by Miles Per Gallon (MPG) achieved.

The current fleet's MPG values range from 10 to 25, whereas the new fleet varies from 20 to 40. This suggests that the new group of cars might perform better on average as the more miles per gallon achieved the better. Figure 2 shows vehicles count with respect to the MPG achieved.



(a) Old fleet

(b) New fleet

Figure 2: Number of vehicles in each group and their MPG achieved.

It is worth noticing that the old group seems to be centered around 18 MPG, while the new one around value 32. This again suggests better performance of the new fleet. Finally, the following summary statistics were obtained for both groups of vehicles:

Measure	Current fleet	New fleet
Mean	20.144578	30.481013
Median	19.000000	32.000000
Var	40.983113	36.831918
Std	6.401805	6.068931
MAD	4.000000	4.000000

Table 1: Summary statistics of both fleets.

3 The Bootstrap

Second part aimed at implementing the bootstrap function (bootstrap.py) and then using it to analyse data sets with Confidence Interval (CI) set to 95%. Figure 3 shows bootstrapping applied to salaries data (salaries.csv), which, more or less, looks similar to the chart presented during the lecture. Thus, the implemented bootstrap function probably works as expected. Also, as anticipated, all three measures, namely mean and its upper and lower bounds, look to be more stable as the number of bootstrap iterations increases.

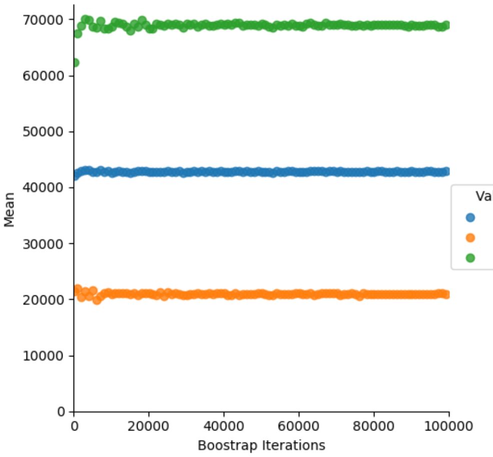
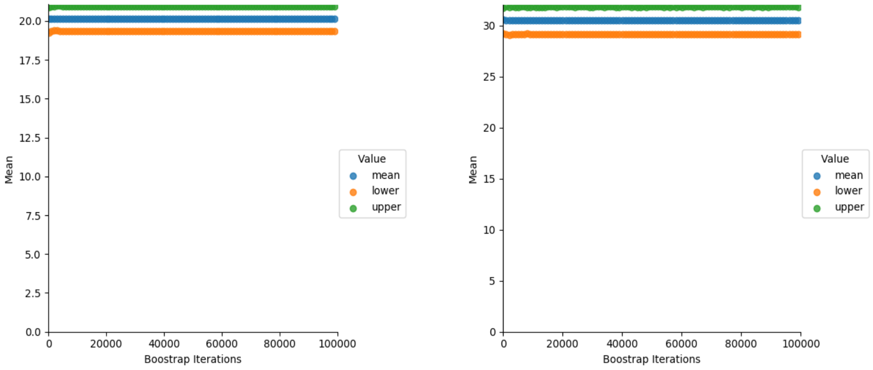


Figure 3: Bootstrapping applied to salaries data set.

Next step involved applying the same technique to vehicles data in order to find upper and lower bounds of the mean of both fleets. Figure 4 demonstrates results achieved for current and new fleets. The current fleet's mean settled around 20 MPG, whereas the new group reported a value approx. 30 MPG. Both upper and lower bounds were not so distant from their means (more or less than 1). Therefore, the new fleet showed to be clearly better than the current fleet, in terms of miles per gallon achieved.



(a) Old fleet

(b) New fleet

Figure 4: Bootstrapping applied to vehicles data set.