Lab Session 2: Resampling Statistics

Simon Harris
CE888: Data Science and Decision Making
University of Essex

January 26, 2019

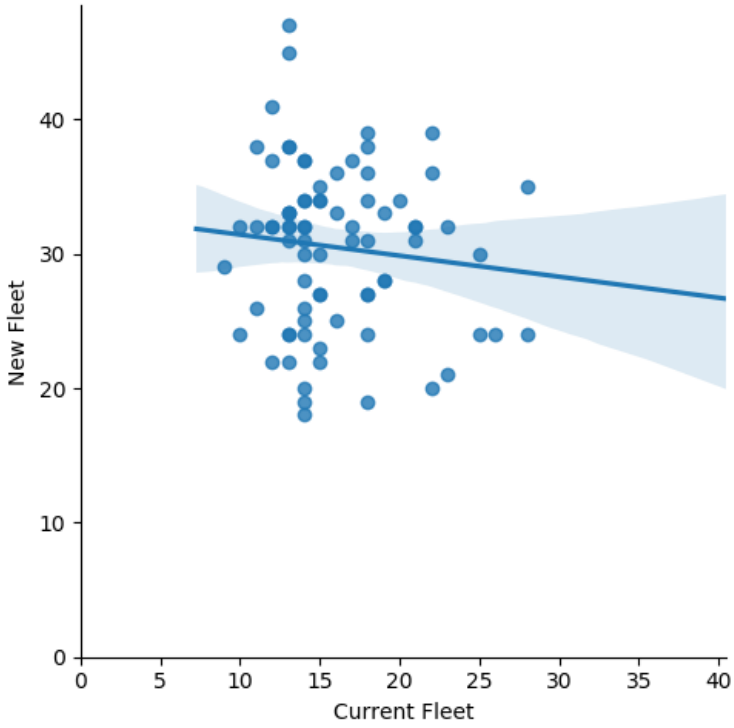
A Brief writeup on Lab 2 and getting stuck in with some of Python's Data Science toolkits.

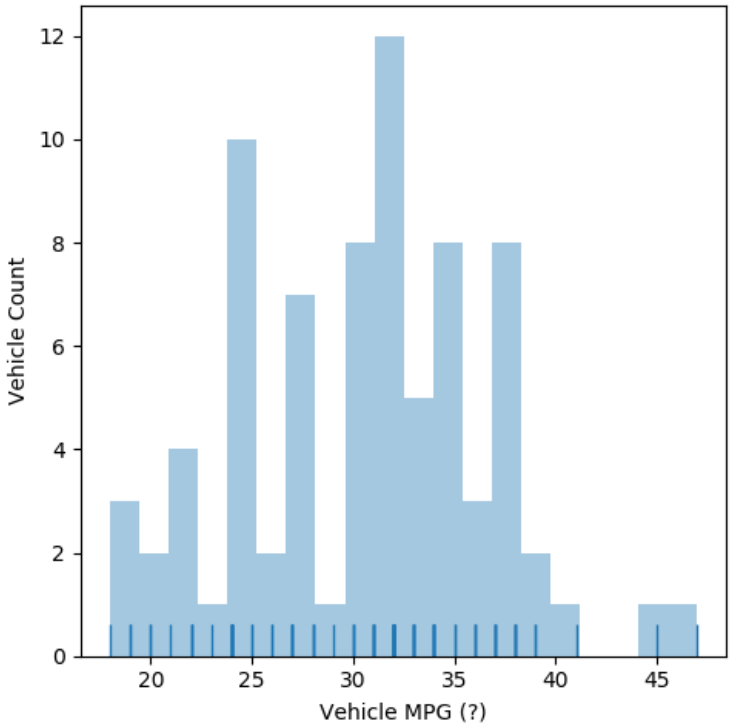
1 Setting Up

1. **Did not** set up an Overleaf account as I'm familiar with LaTeXand have tools to work with it
2. **Did** send an email to Ana with my GitHub username
3. **Did** download the slides and initial code, and upload to GitHub
4. **Did** download PyCharm, but am reluctant to use it to a great extent as I have tools I prefer

2 Data Visualisation

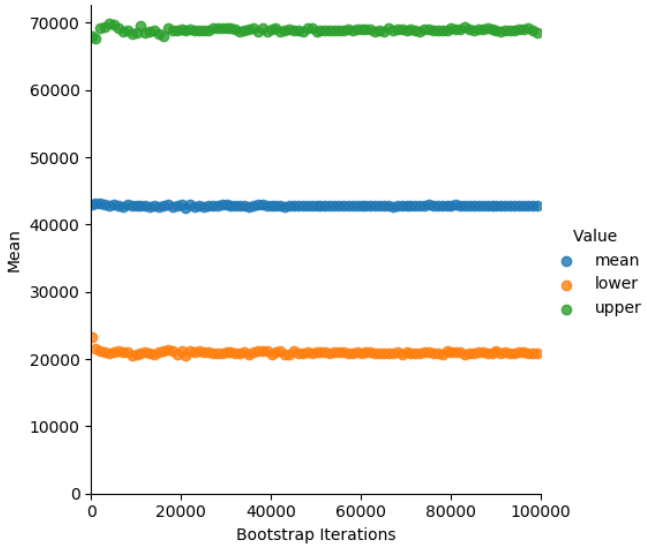
Here I was pleased to be able to create a working `vehicles.py` and generate the scatter plot and histogram. This was admittedly mainly a case of copying and trimming the `salaries.py` code, but I made sure to take the time to understand it and learn about the libraries involved.


New Fleet
Current Fleet
0
10
20
30
40
0
5
10
15
20
25
30
35
40


Vehicle Count
Vehicle MPG (?)
0
2
4
6
8
10
12
20
25
30
35
40
45

3 Exercise: Bootstrap 1

This was somewhat more challenging, and it took me several hours to write very few lines of code! However, I was struck by how powerful those few lines are, and how much more complicated the solution would have been without libraries, in particular NumPy. In addition, I invested time in reviewing the lecture slides in order to understand the resampling algorithm, and also the libraries involved, which was a worthwhile endeavour.



4 Exercise: Bootstrap 2

This was mainly just a case of calling my `bootstrap()` function from the previous exercise. It took me a while to find an overall mean for the dataset, but finally discovered that `df.mean().mean()` provides that.

I wasn't sure how many iterations and how large a sample size to use, so I chose arbitrary numbers, both fairly high but not high enough to take forever to run. Output was as follows:

```
Overall mean: 25.31279548574043
Old Lower: 19.313253012048193
Old Upper: 20.883734939759034
New Lower: 29.189873417721518
New Upper: 31.78481012658228
```

I'm not sure how to decide whether these are "comparable". Given a 95% confidence interval, the numbers for the old fleet are consistently below the overall mean, while the new fleet provides numbers consistently above it. Assuming that high numbers are good, one could conclude the new fleet to be better.