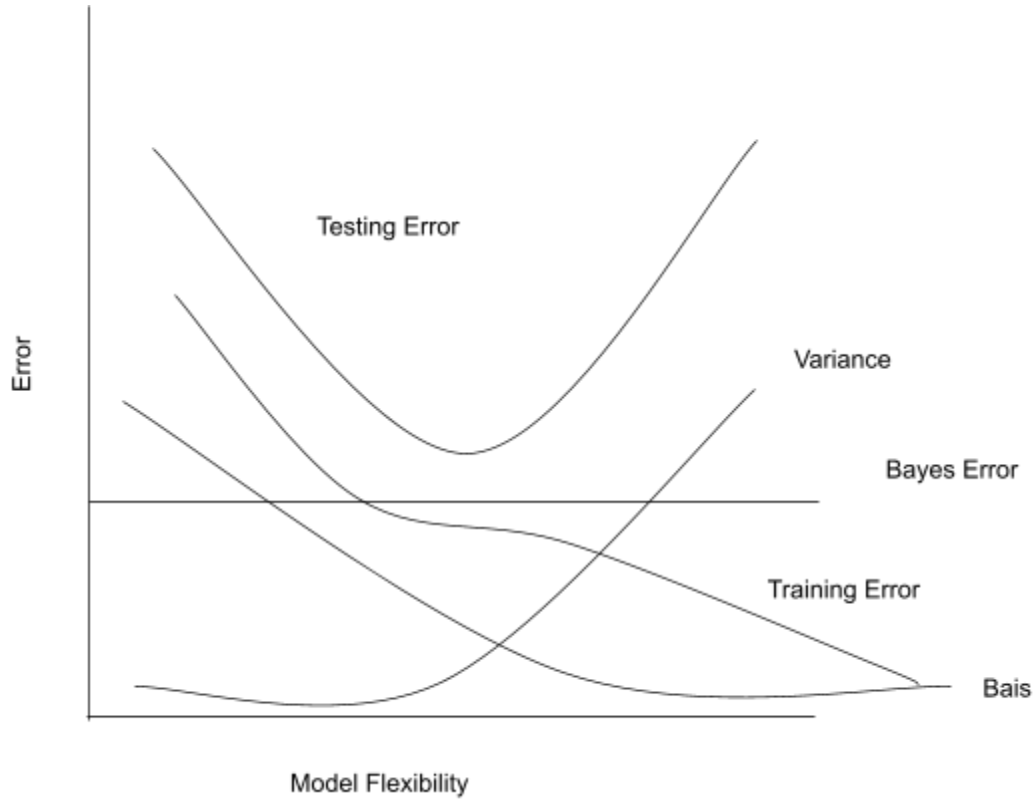
Theoretical Questions

1.
    a. Flexible will be better because of its improved performance with large data samples
    b. Flexible will perform worse because with minimal observations it will overfit the data
    c. Flexible will perform better for highly non-linear data
    d. Flexible would perform worse in this situation because it would adapt more to the bad data
2.
    a. Regression and Inference: n = 500 p = 3
    b. Classification and Prediction: n = 20 p = 14
    c. Classification and Prediction: n = 52 p = 4
3.
    a.



    b. Bais - As the model increases the flexibility the bias will naturally decrease as it fits the data
    Variance - As the model become better fitting to the sample it becomes worse at predicting the overall data
    Bayes - Uncontrollable error
    Testing Error - The sum of the other graph lines
    Training Error - With a more overfitting model the training error will decrease
4.

a. Image recognition, Email spam detection, and Test grading
    i. All of which have distinct categories that they can be filtered into
b. Predicting stock prices, car prices, and house values
    i. Depending on certain stats or characteristics of these items you can create a prediction of how the price will trend or what the price should be based on other examples.
c. Setting school boundaries, flower classification, marketing data
    i. Clustering can put objects into groups based on location/data which allows it to be classified into groups that can be used.

5. Pros: Can represent very complex data
   Cons: Can overfit the data and find trends that aren't really present
   A more flexible approach is better when there is lots of data with low standard deviation.
   A less flexible approach is better for more linear data that might have lots of variance.
6. Parametric algorithms use a mathematical equation to describe the relationship between a set of inputs and outputs whereas non parametric algorithms use only the data to inform the decisions. It is much less restrictive but requires a large amount of data to make good predictions whereas parametric can be useful in situations where there may not be a ton of data or there is a clear correlation between the inputs and outputs.
7.
    a.

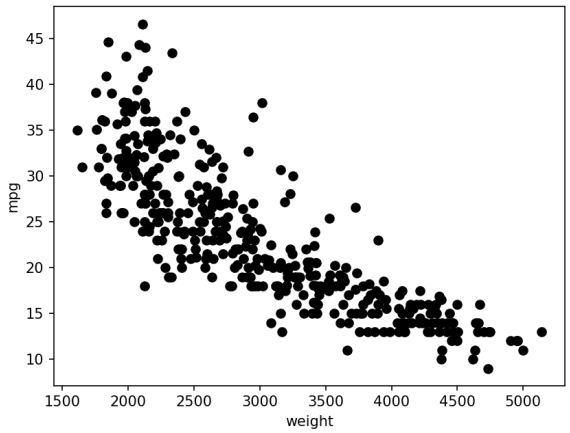
1	2	3	4	5	6
3	2	sqrt(10)	sqrt(5)	sqrt(2)	sqrt(3)

    b. We would predict K = 1 to be green based on point 5 which is the closest
    c. We would predict K = 3 to be red based on point 1 which is also 3
    d. Small because we require a highly flexible model to fit the non-linear data

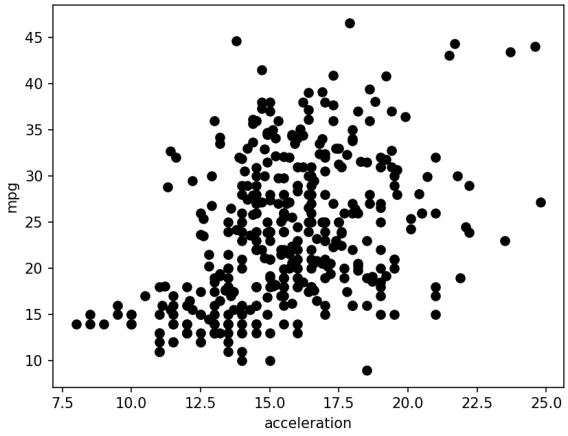
Applied Questions

1.
    a. Qualitative: Name
       Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin
    b. MPG:37, cylinders: 5, displacement: 387, horsepower: 184, weight: 3527, acceleration: 16, year: 12, origin:2
    c. MPG: 9.0-46.6, cylinders: 3-8, displacement: 68-455, horsepower: 46-230, weight: 1613-5140, acceleration: 8-24.8, year: 70-82, origin: 1-3
    d. MPG: range: 35, mean: 24.34, std: 7.89, cylinders: range: 5, mean: 5.39, std: 1.66, displacement: range: 387, mean: 188.39, std: 100.42, horsepower: range: 184, mean: 101.24, std: 36.19, weight: range: 3348, mean: 2936.81, std: 810.99, acceleration: range: 16, mean:15.7, std: 2.71, year: range: 12, mean: 77.13, std:3.13, origin: range: 2, mean: 1.60, std: 0.82

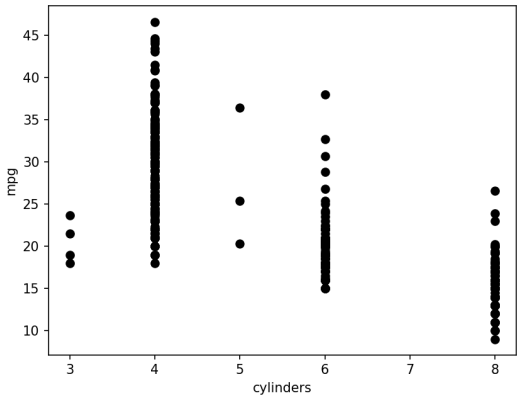
e.



There appears to be a negative correlation between mpg and weight.



There is either a small or no correlation between acceleration and mpg



It appears there may be a negative correlation between cylinders and mpg

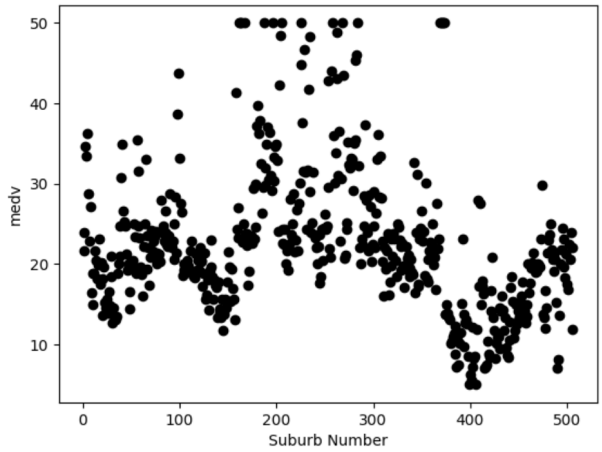
f. Yes, my plot between weight and mpg seems to show a negative correlation between the two. It is not completely clear but the plot between cylinders and mpg may also show a negative correlation. Other variables that I did not plot may have a correlation as well and could be used to predict the mpg of a vehicle.

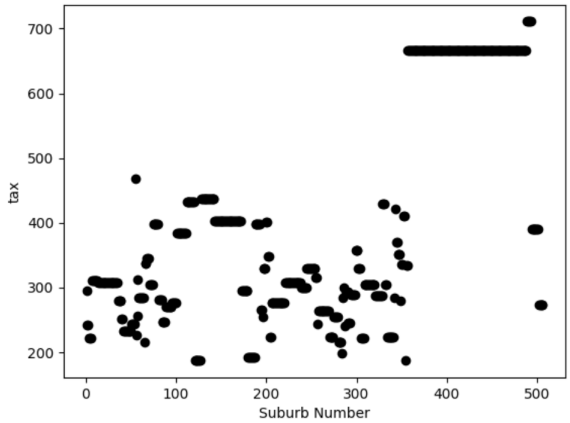
g.

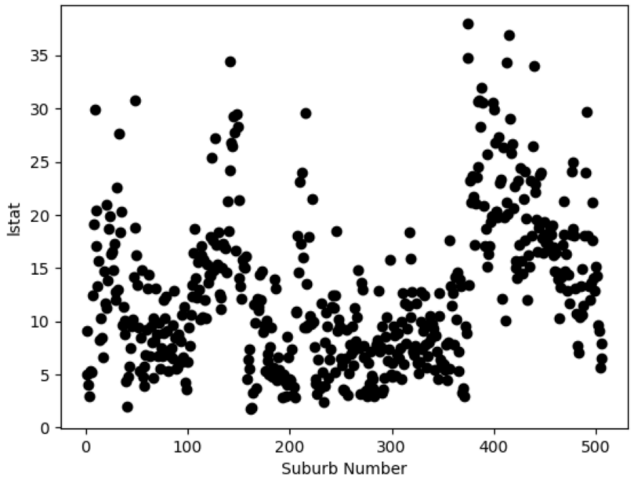
2.

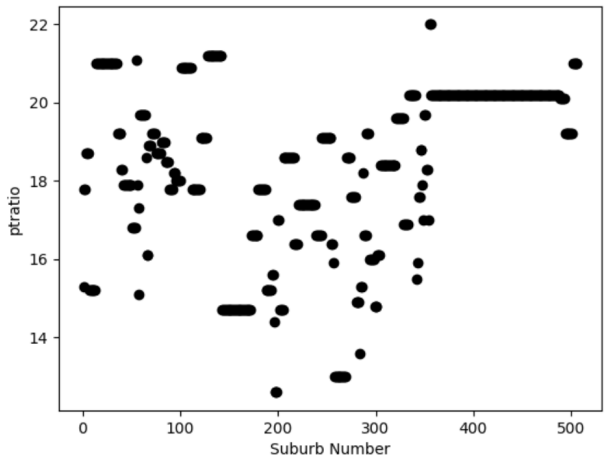
a. Rows: 506 Columns: 13
Rows represent data from specific suburbs
Columns represent different criteria relating to quality of life

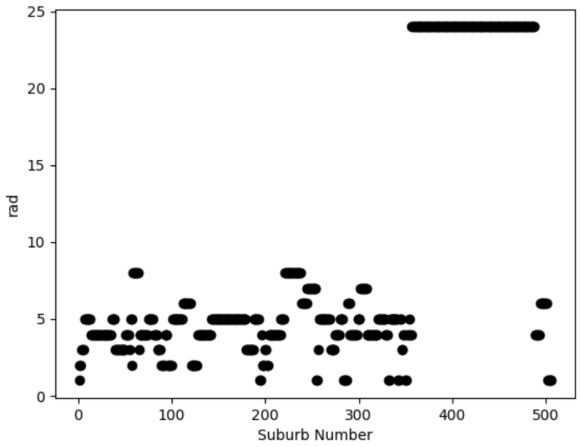
b.

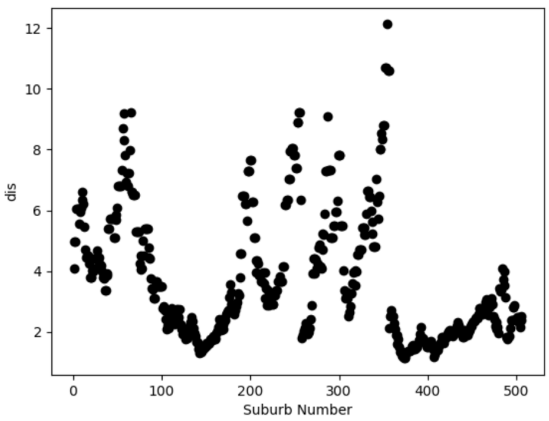


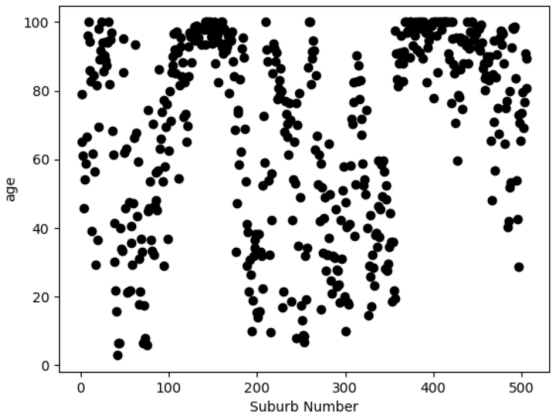


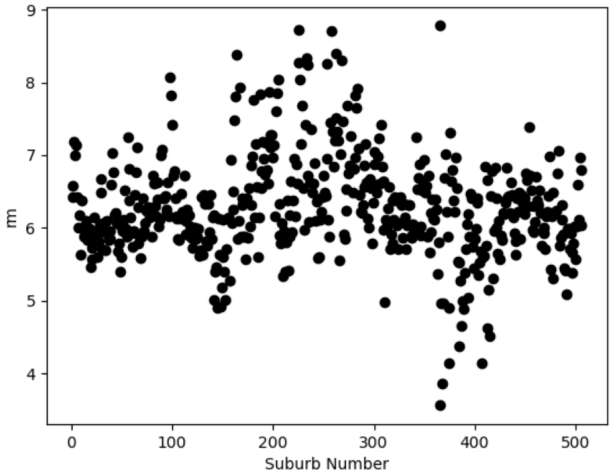


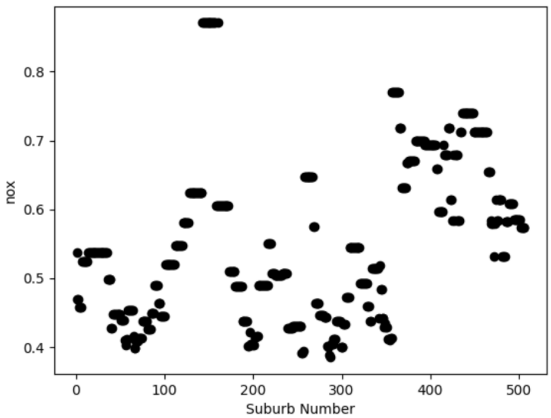

ptratio
Suburb Number

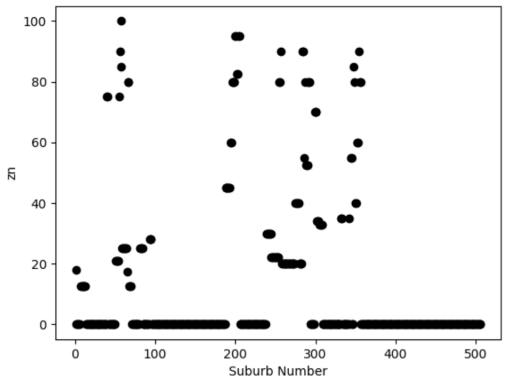

rad
Suburb Number

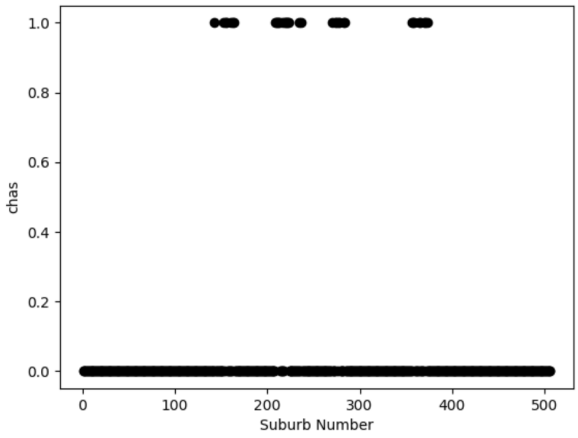

dis
Suburb Number

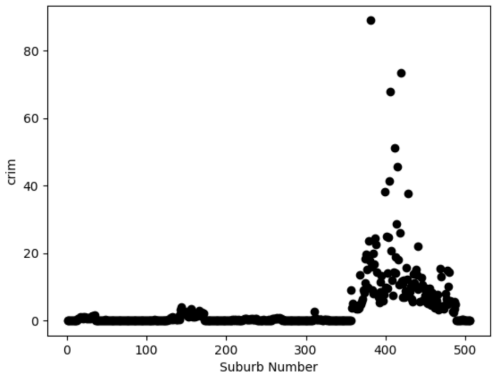

age
Suburb Number


rm
Suburb Number


nox
Suburb Number







c. rad (Index of accessibility to radial highways)
   tax (full-value property-tax rate per $10,000)
   ptratio (pupil-teacher ratio by town)
   All of the scatter plots of these graphs listed above show a positive correlation with the increase in per capita crime rate by suburb
d. Ranges:
   crim - 88.96988 rad - 23 tex - 524 ptratio - 9.4
   Proportionally all of the ranges seem to have very big jumps between the suburbs numbered below 350 and above 350. The only set that shows the

correlation without a large jump in the data would be ptratio which has a high of 22 and a low of 12.6.

e. 35
f. 19.05
g. Suburbs 399 and 406 share the lowest median value of owner-occupied homes in $1000s with a score of 5. With the exception of crime they find themselves within the general area of the other suburbs that they fall near numerically. Within this group you can observe an above average age, indus, rad, tax, and lstat.
h. 64 suburbs average more than 7 rooms per dwelling and 13 average more than 8. In the suburbs that average more than 8 rooms per dwelling there are some observable trends such as low crim, high age, low zn, and low dis. It seems like there are other factors that have a stronger influence on the other data points than rm as we observe large ranges across the suburbs with a higher average than 8.