Assignment 3

CP468 – Artificial Intelligence

Dr. Sumeet Sehra

July 10th, 2023

Steven Tohme

Natalie Song

Question 1

a. This scenario would be a regression problem since our dependent variable (CEO salary) is continuous. We are interested in inference because we want to know how the factors affect CEO salary. Our n, number of observations, the value would be 500 (500 CEOs considered), and our p, number of features, the value would be 3 (profit, number of employees, and industry).

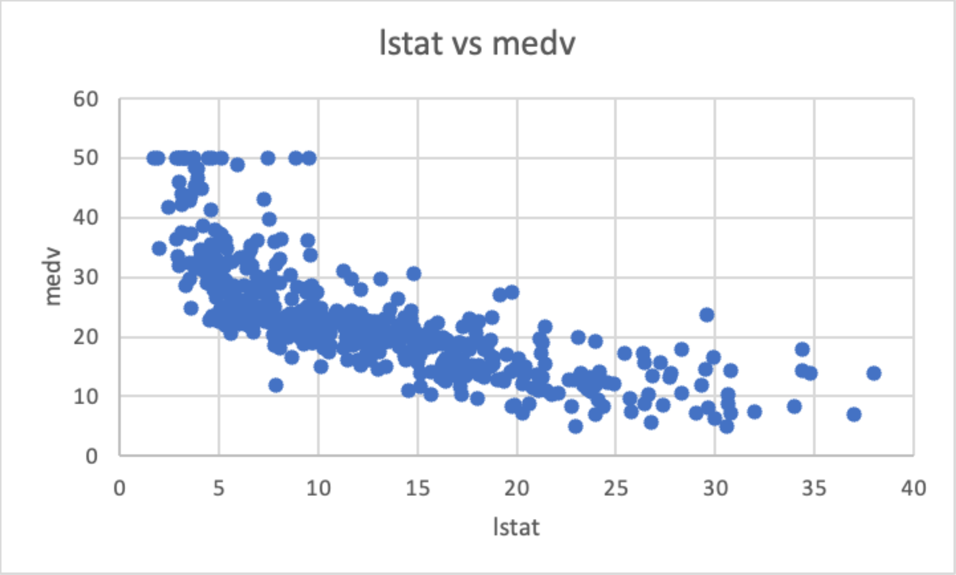
b. This scenario would be a classification problem since our variable (Success or Failure of a product) is not continuous. This is also a prediction problem because we only care about success or failure, not the underlying effect of the relationship between factors. Our n value is 20 (20 similar products launched), and our p value is 13 (price charged for the product, marketing budget, competition price, and ten other variables).

c. This scenario would be a regression problem because our dependent variable (% change in the USD/Euro exchange rate) is quantitative. This is also a prediction problem because we only care about the % change of the exchange rate, not how the relationship between variables changes it. Our n value is 52 (weekly data for all of 2012), and our p value is 3 (the % change in the US market, the % change in the British market, and the % change in the German market).

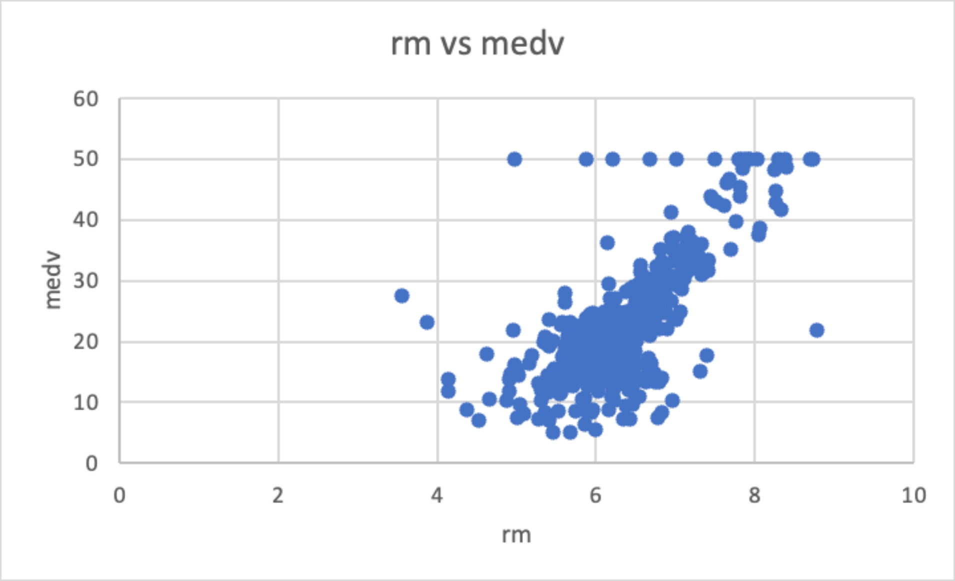
Question 2:

a. It has 506 rows and 14 columns. The rows represent the U.S Census Tracts in the Boston area and the columns represent the measures of the Census variables.

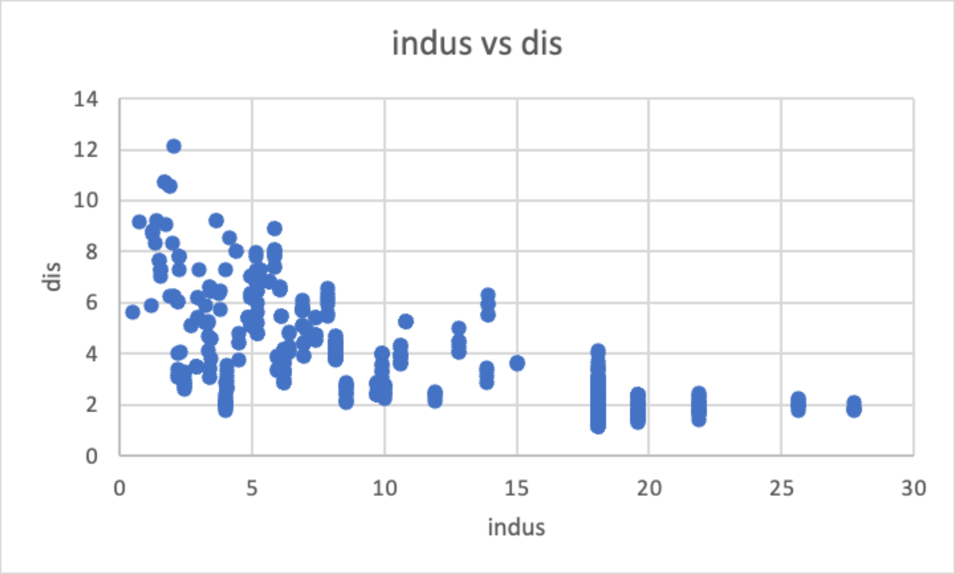
b.



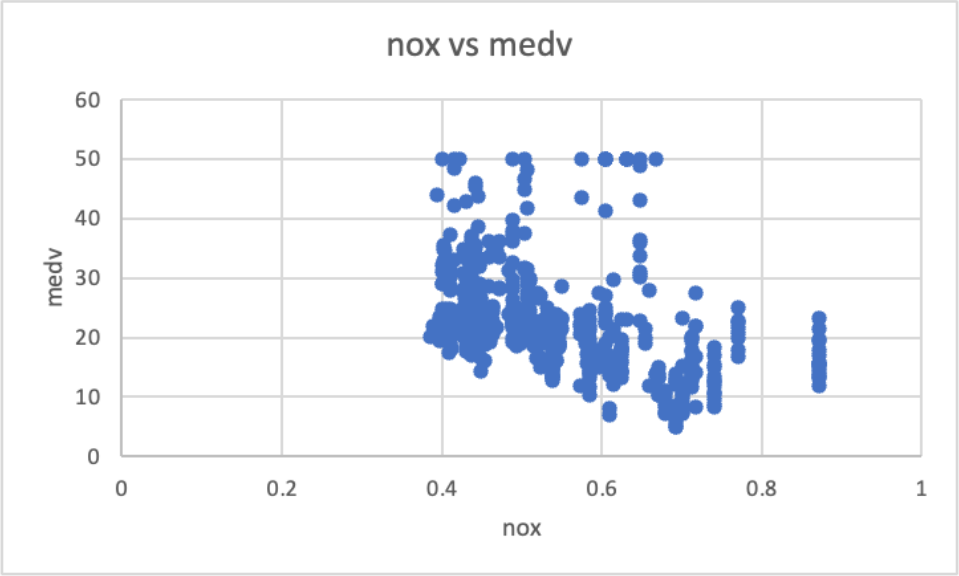
There is a seemingly strong inverse relationship between the lower status of the population and the median value of owner-occupied homes, so the greater the lstat, the lower the medv.



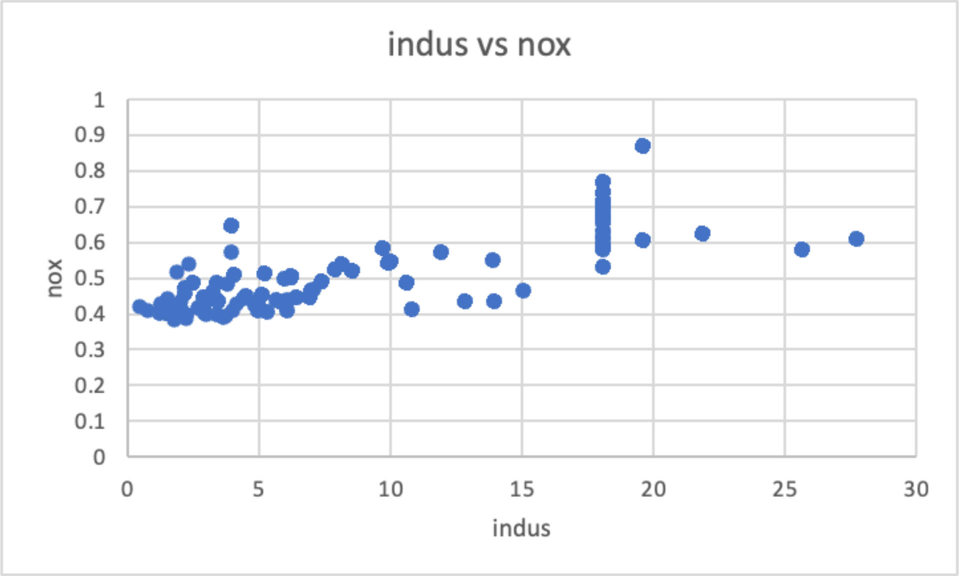
There is a seemingly strong positive correlation between the average number of rooms per dwelling and the median value of owner-occupied homes, meaning as the number of rooms increase, the value also increases.



There is a negative correlation between the proportion of non-retail business acres per town and weighted mean of distances to five Boston employment centres, so as the indus gets larger, the dis gets smaller.

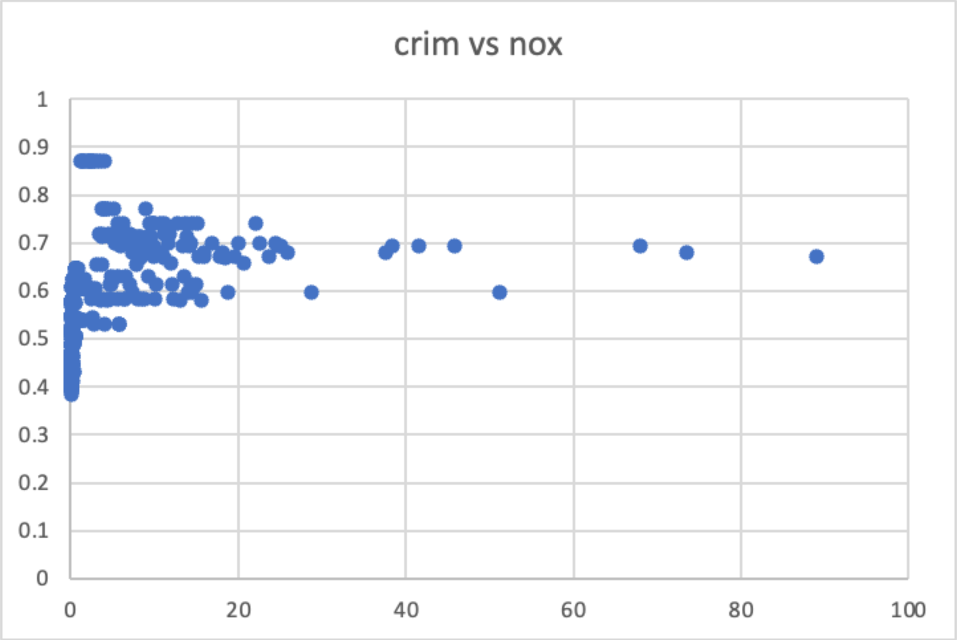


There is a negative correlation the nitrogen oxide concentration and the median value of owner-occupied homes, so as the nitrogen oxide levels increase, the value of homes decreases.

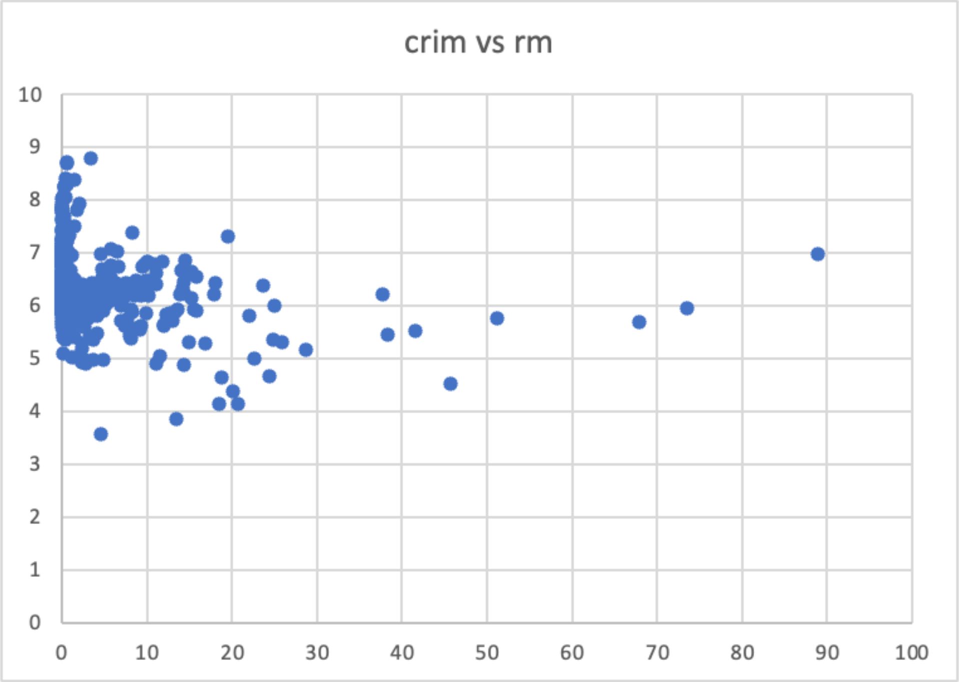


There is a slight positive correlation between indus and nox, so as the proportion of non-retail business acres per town, the higher the nitrogen oxides concentration.

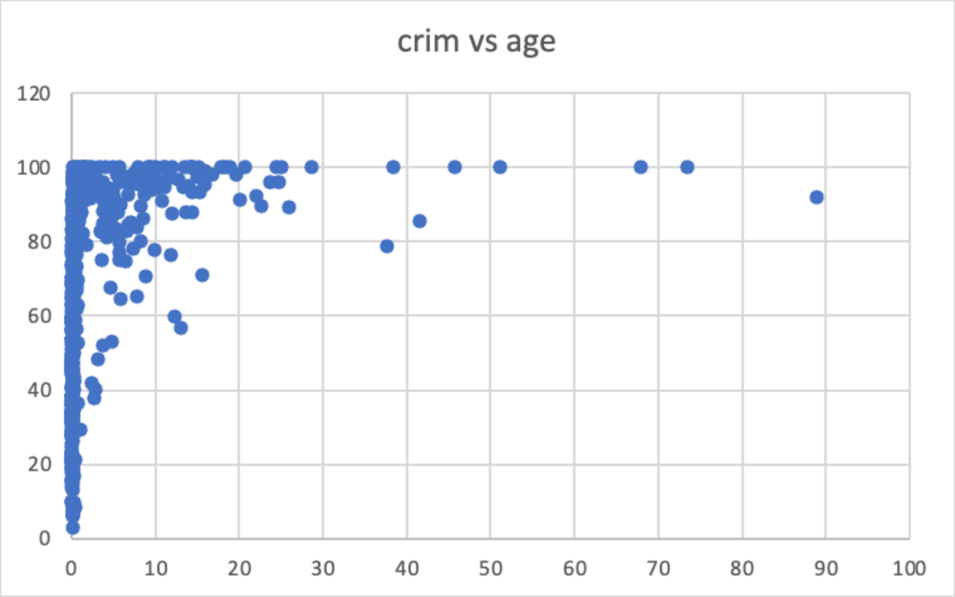
c. **crim is the x-axis on each scatter plot



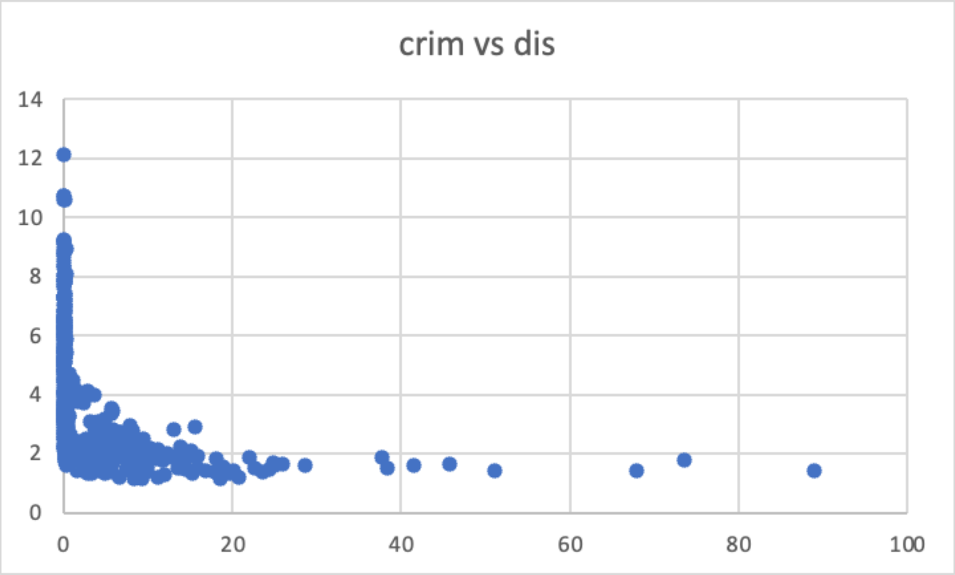
There is a positive correlation between crime rate and nitrogen oxide levels, so as the crime rate increases, so does the nitrogen oxide levels, however, it levels out at about a crime rate of 20.



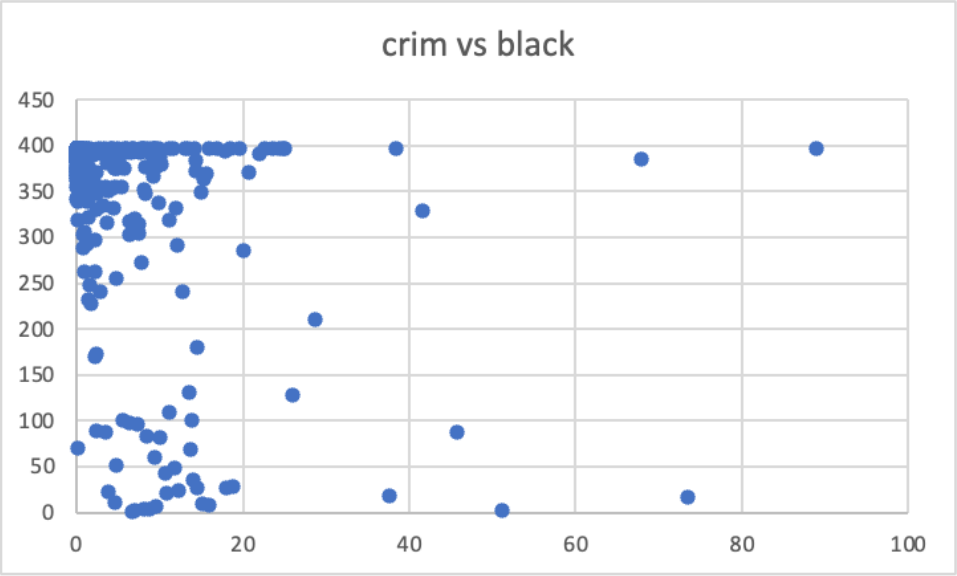
There is a negative correlation between crime rate and number of rooms per dwelling, so as the crime rate increases, the number of rooms per welling decreases.



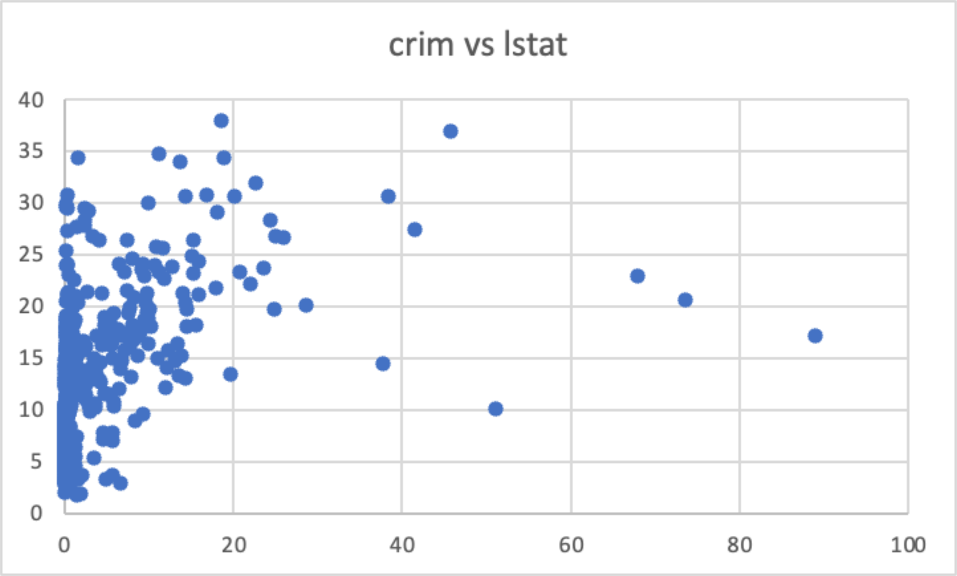
There is a slight positive correlation between crime rate and age, so as the age gets higher, the crime rate do as well.



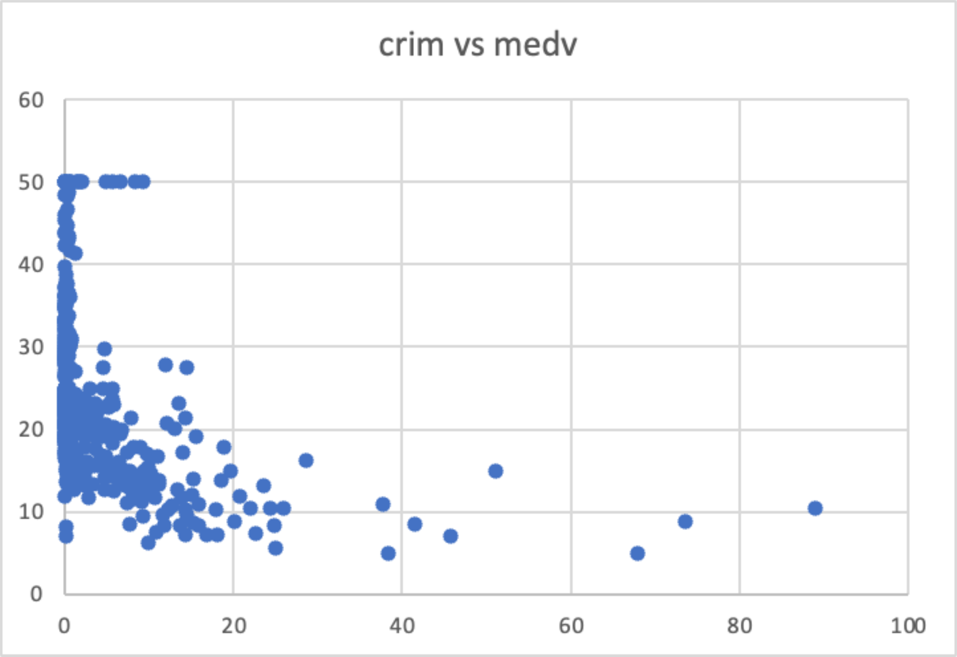
There is a negative correlation between crime rates and distances to five employment centres, meaning that as the crime rate increases, the distance to employment centres gets smaller.



There is a very slight negative correlation between crim and black, however, there are many points with the same crime rate but vastly different proportions of certain population by town.



The higher the lower status of the population is, the higher the crime rate.



The higher the median value of owner-occupied homes, the lower the crime rate.

d)
Crime rate:

Min	1st Quart	Median	Mean	3rd Quart	Max
0.00632	0.082045	0.25651	3.61352356	3.6770825	88.9762

Since the median and max are around 0.26% and 89%, the crime rates are high. Also, the max is much higher than the 3rd quartile indicating high crime rates.

Tax Rates:

Min	1st Quart	Median	Mean	3rd Quart	Max
187	279	330	403.2	666	711

Based off the quartiles, 75% pay under $666 while 25% pay over $666. Since the max is $711, most of the tracts aren't paying a high price, so the tax rates are relatively low.

Ptratio:

Min	1st Quart	Median	Mean	3rd Quart	Max
12.6	17.4	19.05	18.46	20.2	22

Based off the quartiles with mean 18.46 and median 19.05, most have a pt ratio close to the 3rd quartile, so they have a high pt ratio when compared to the max.

e) Num of 1 = 35, so there are 35 census tracts that bound the Charles River

f) Median = 19.05

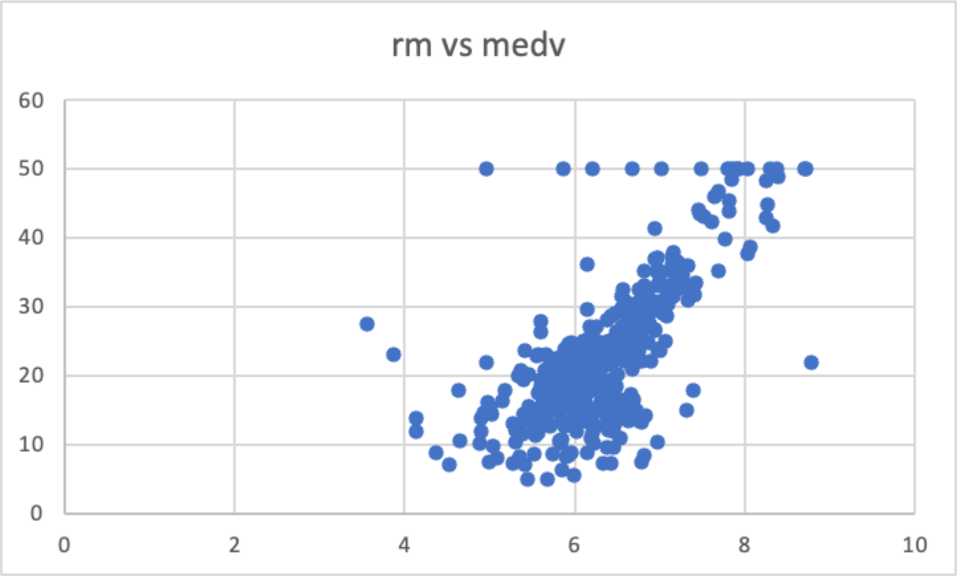
g) Census tract on row 400 has the lowest median value of owner-occupied homes.

400:

Crim	Zn	Indus	Chas	Nox	Rm	Age	Dis	Rad	Tax	Ptratio	Black	Lstat	medv
38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.9	30.59	5

h) More than 7 rooms per dwelling = 64

More than 8 rooms per dwelling = 13



We can see from the graph that as the number of rooms increases, the price also increases, but not always and even an outlier exists of very lower price than houses with less rooms.

Question 3:

Entropy of target attribute: Repeat Customer

Entropy(S) = -p(YES)log2p(YES) - p(NO)log2p(NO)

YES =6 and NO=4

Entropy(S) = -(6/10)log2(6/10)-(4/10)log2(4/10) = 0.97

Entropy of Attribute: Age

Age has 4 values: 20..30, 31..40, 41..50, 51..60

20…30

YES = 3 NO = 2

Entropy(S|Age=20…30) = -(3/5)log2(3/5)-(2/5)log2(3/5)=0.74

31…40

YES = 1 NO = 0

Entropy(S|Age=31…40) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

41…50

YES = 2 NO = 0

Entropy(S|Age=41…50) = -(2/2)log2(2/2)-(0/2)log2(0/2) = 0.00

51…60

YES = 0 NO = 2

Entropy(S|Age=51…60) = -(0/2)log2(0/2)-(2/2)log2(2/2) = 0.00

Gain(S|Age) = Entropy(S) – 5/10(Entropy(S|Age=20…30))-1/10(Entropy(S|Age=31…40))-2/10(Entropy(S|Age=41…50))-2/10(Entropy(S|Age=51…60))

Gain(S|Age) = 0.97-(5/10)*0.74 – (1/10)*0 – (2/10)*0 – (2/10)*0 = 0.6

Entropy of Attribute: City

City has 3 values: NY, LA, SF

NY

YES = 5 NO = 2

Entropy(S|City=NY) = -(3/5)log2(3/5)-(2/5)log2(3/5) = 0.74

LA

YES = 1 NO = 1

Entropy(S|City=LA) = -(1/2)log2(1/2)-(1/2)log2(1/2) = 1

SF

YES = 0 NO = 1

Entropy(S|City=SF) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0

Gain(S|City) = 0.97-(7/10)*0.74 – (2/10)*1 – (1/10)*0 = 0.97-0.518-0.2-0 = 0.252

Entropy of Attribute: Gender

Gender has 2 values: M, F

M

YES = 1 NO = 2

Entropy(S|Gender=M) = -(1/3)log2(2/3)-(1/3)log2(2/3) = 0.39

F

YES = 5 NO = 2

Entropy(S|Gender=F) = -(5/7)log2(5/7)-(2/7)log2(2/7) = 0.86

Gain(S|Gender) = 0.97-(3/10)*0.39 – (7/10)*0.86 = 0.97-0.117-0.602 = 0.251

Entropy of Attribute: Education

Education has 3 values: High School, College, Graduate

High School

YES = 0 NO = 2

Entropy(S|Education=High School) = -(0/2)log2(0/2)-(2/2)log2(2/2) = 0.00

College

YES = 5 NO = 2

Entropy(S|Education=College) = -(5/7)log2(5/7)-(2/7)log2(2/7) = 0.86

Graduate

YES = 1 NO = 0

Entropy(S|Education=Graduate) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

Gain(S|Education) = 0.97-(2/10)*0 – (7/10)*0.86 – (1/10)*0 = 0.97-0-0.602-0 = 0.368

Largest Information Gain: Age

Age	City	Gender	Education	Repeat Customer
20…30	NY	F	College	YES
20…30	LA	M	High School	NO
20…30	LA	M	College	YES
20…30	NY	F	High School	NO
20…30	NY	F	College	YES

Age	City	Gender	Education	Repeat Customer
31…40	NY	F	College	YES

Age	City	Gender	Education	Repeat Customer
41…50	NY	F	College	YES
41…50	NY	F	Graduate	YES
Age	City	Gender	Education	Repeat Customer
51…60	SF	M	College	NO

Age 20…30

Entropy of Attribute: City

City has 2 values: NY, LA

NY

YES = 2 NO = 1

Entropy(Age=20…30|City=NY) = -(2/3)log2(2/3)-(1/3)log2(1/3) = 0.92

LA

YES = 1 NO = 1

Entropy(Age=20…30|City=LA) = -(1/2)log2(1/2)-(1/2)log2(1/2) = 1

Gain(20…30|City) = 0.74-(3/5)*0.92-(2/5)*1 = 0.74-0.552-0.4 = -0.212

Entropy of Attribute: Gender

Gender has 2 values: M, F

M

YES = 1 NO = 1

Entropy(Age=20…30|Gender=M) = -(1/2)log2(1/2)-(1/2)log2(1/2) = 1

F

YES = 2 NO = 1

Entropy(Age=20…30|Gender=F) = -(2/3)log2(2/3)-(1/3)log2(1/3) = 0.92

Gain(20…30|Gender) = 0.74-(2/5)*1-(3/5)*0.92 = 0.74-0.4-0.552 = -0.212

Entropy of Attribute: Education

Education has 2 values: High School, College

High School

YES = 0 NO = 2

Entropy(Age=20…30|Education) = -(0/2)log2(0/2)-(2/2)log2(2/2) = 0.00

College

YES = 3 NO = 0

Entropy(S|Education=College) = -(3/3)log2(3/3)-(0/3)log2(0/3) = 0.00

Gain(20…30|Education) = 0.74-(2/5)*0-(3/5)*0 = 0.74

Age 31…40

Entropy of Attribute: City

City has 1 value: NY

NY

YES = 1 NO = 0

Entropy(Age=31…40|City=NY) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

Entropy of Attribute: Gender

Gender has 1 value: F

F

YES = 1 NO = 0

Entropy(Age=31…40|Gender=F) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

Entropy of Attribute: Education

Education has 1 value: College

College

YES = 1 NO = 0

Entropy(Age=31…40|Education=College) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

Age 41…50

Entropy of Attribute: City

City has 1 value: NY

NY

YES = 2 NO = 0

Entropy(Age=41…50|City=NY) = -(2/2)log2(2/2)-(0/2)log2(0/2) = 0.00

Entropy of Attribute: Gender

Gender has 1 values: F

F

YES = 2 NO = 0

Entropy(Age=41…50|City=NY) = -(2/2)log2(2/2)-(0/2)log2(0/2) = 0.00

Entropy of Attribute: Education

Education has 2 values: College, Graduate

College

YES = 1 NO = 0

Entropy(Age=41…50|Education=College) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

Graduate

YES = 1 NO = 0

Entropy(Age=41…50|Education=Graduate) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

Age 51…60

Entropy of Attribute: City

City has 1 value: SF

SF

YES = 0 NO = 1

Entropy(Age=51…60|City=SF) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

Entropy of Attribute: Gender

Gender has 1 value: M

M

YES = 0 NO = 1

Entropy(Age=51…60|Gender=M) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

Entropy of Attribute: Education

Education has 1 value: College

College

YES = 0 NO = 1

Entropy(Age=51…60|Education=College) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

Age 20…30 needs further splitting:

Education has the largest gain, so it is the next node:

Age	City	Gender	Education	Repeat Customer
20…30	NY	F	College	YES

20…30	LA	M	High School	NO
20…30	LA	M	College	YES
20…30	NY	F	High School	NO
20…30	NY	F	College	YES

Age	Education	City	Gender	Repeat Customer
20…30	High School	LA	M	NO
20…30	High School	NY	F	NO

Age	Education	City	Gender	Repeat Customer
20…30	College	NY	F	YES
20…30	College	LA	M	YES
20…30	College	NY	F	YES

20…30|High School

Entropy of Attribute: City

20…30|High School has 2 values: LA, NY

LA

YES = 0 NO = 1

Entropy(Age=20…30|Education=High School|City=LA) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

NY

YES = 0 NO = 1

Entropy(Age=20…30|Education=High School|City=NY) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

Entropy of Attribute: Gender

20…30|High School has 2 values: M, F

M

YES = 0 NO = 1

Entropy(Age=20…30|Education=High School|Gender=M) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

F

YES = 0 NO = 1

Entropy(Age=20…30|Education=High School|Gender=F) = -(0/1)log2(0/1)-(1/1)log2(1/1) = 0.00

20…30|College

Entropy of Attribute: City

20…30|College has 2 values: LA, NY

LA

YES = 1 NO = 0

Entropy(Age=20…30|Education=College |City=LA) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

NY

YES = 2 NO = 0

Entropy(Age=20…30|Education=College |City=NY) = -(2/2)log2(2/2)-(0/2)log2(0/2) = 0.00

Entropy of Attribute: Gender

20…30|High School has 2 values: M, F

M

YES = 1 NO = 0

Entropy(Age=20…30|Education=College |Gender=M) = -(1/1)log2(1/1)-(0/1)log2(0/1) = 0.00

F

YES = 2 NO = 0

Entropy(Age=20…30|Education=College |Gender=F) = -(2/2)log2(2/2)-(0/2)log2(0/2) = 0.00

Question 4

a. $Salary = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
$Salary = 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01(x_1 \cdot x_2) - 10(x_1 \cdot x_3)$

Let a be a constant GPA and b be a constant IQ

$MaleSalary = 50 + 20a + 0.07b + 35(0) + 0.01(ab) - 10(a \cdot 0)$

$MaleSalary = 50 + 20a + 0.07b + 0.01ab$

$FemaleSalary = 50 + 20a + 0.07b + 35(1) + 0.01(ab) - 10(a \cdot 1)$

$FemaleSalary = 50 + 20a + 0.07b + 35 + 0.01ab - 10a$

$MaleSalary = FemaleSalary$

$$50 + 20a + 0.07b + 0.01ab = \ 50 + 20a + 0.07b + 35 + 0.01ab - 10a$$

$$0 = 35 - 10a$$

$$When\ a\ \geq 3.5,\ MaleSalary \geq FemaleSalary$$

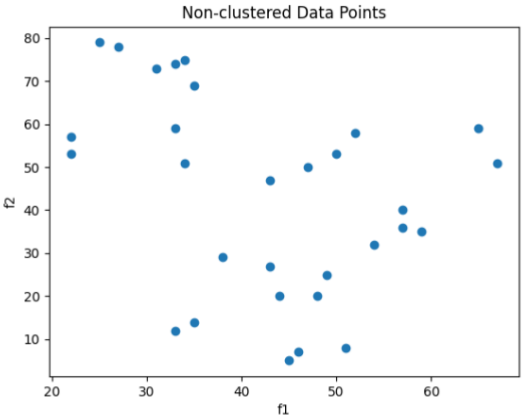
∴ iii. is correct. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

b. $Salary = 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_4 - 10x_5$
$Salary = 50 + 20(4) + 0.07(110) + 35(1) + 0.01(4 \cdot 110) - 10(4 \cdot 1)$
$Salary = \$137,100$

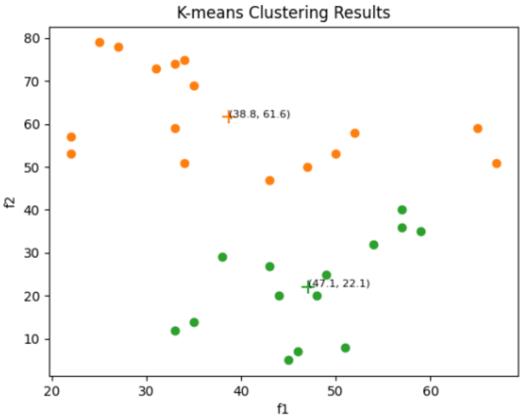
c. False. The statement is not necessarily true. The size of a coefficient alone does not determine the evidence of an interaction effect.

Question 5

a.



c.



d. Cluster 1 (in green) is of size 14, Cluster 2 is of size 16.
e. Centroid 1 at (47.1, 22.1), Centroid 2 at (38.8, 61.6).