DFSC 5340.02 Assignment 6

Due: Tuesday Nov 24@11:59PM
Total Points: 120

1. (20 pts) The Firearms data file given by
http://users.stat.ufl.edu/~aa/smss/data/Firearms.dat
shows U.S. statewide data on x = percentage of people who re-port owning a gun and y = firearm death rate (annual number of deaths per 100,000 population).

(a) Find the prediction equation and interpret.

(b) The correlation is 0.70. Identify an outlier and show that r = 0.78 when you remove this state from the data file.

1 (a) The mean for X is 33.092 [X = ownership]
" " for y " 11.514 [y = death rate]

Sum for values of X = 1654.6
" " " " y = 575.7

Sum of square of X = $\Sigma(x_i - M_x)^2$
= 8976.3168

$\Sigma(x_i - M_x)(y_i - M_y)$
= 1914.4456

Prediction equation $\hat{y} = a + bx$ (we know)

$$b = \frac{\Sigma(x_i - M_x)(y_i - M_y)}{\Sigma(x_i - M_x)^2}$$

= 0.21328

$M_y - bM_x = 11.514 - (.21328 \times 33.092)$
= 4.45622

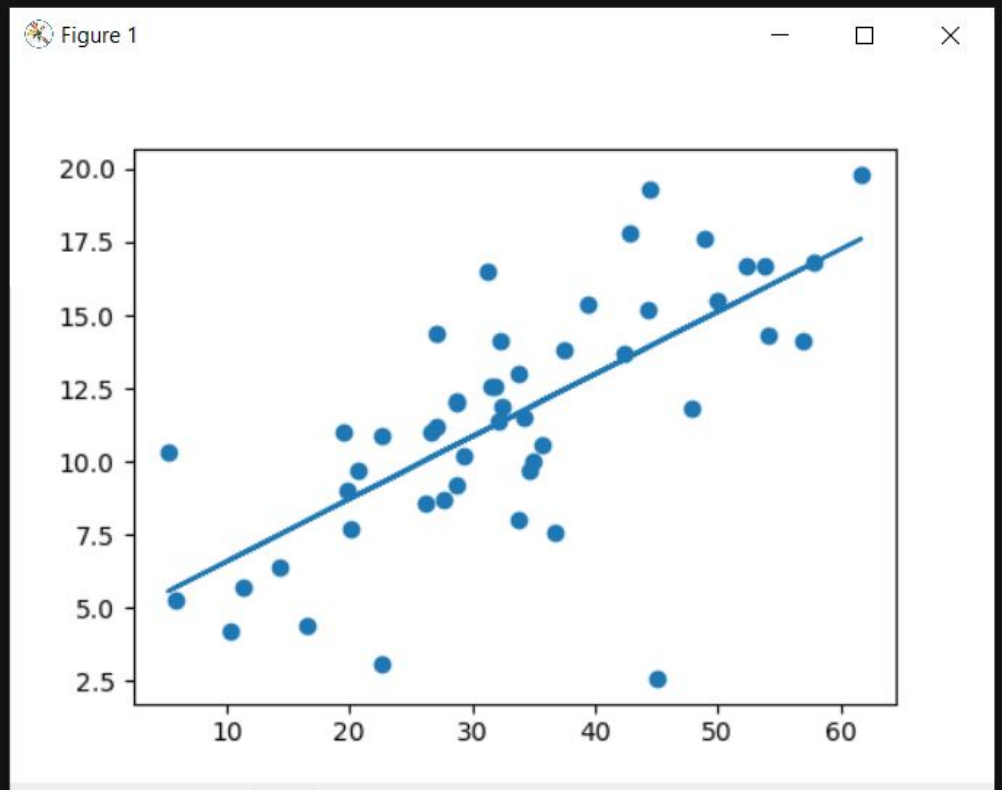
$\hat{y} = 4.45622 + 0.21328x$

```python
import pandas as pd
import matplotlib.pyplot as plt
from pandas import *
from sklearn.linear_model import LinearRegression
import numpy as np
import math

data = pd.read_csv('C:/hmm/courses/3rd semester/Special topics/Assignment/6/New assignment/Qu1/csv1.csv', index_col=0)
```

Question number 1 | Question number 1 | Question number 1

```
C:\Users\Abantika\AppData\Local\Programs\Python\Python37\python.exe "C:/hmm/courses/3rd semester/Special topics/Assignment/6/New
```



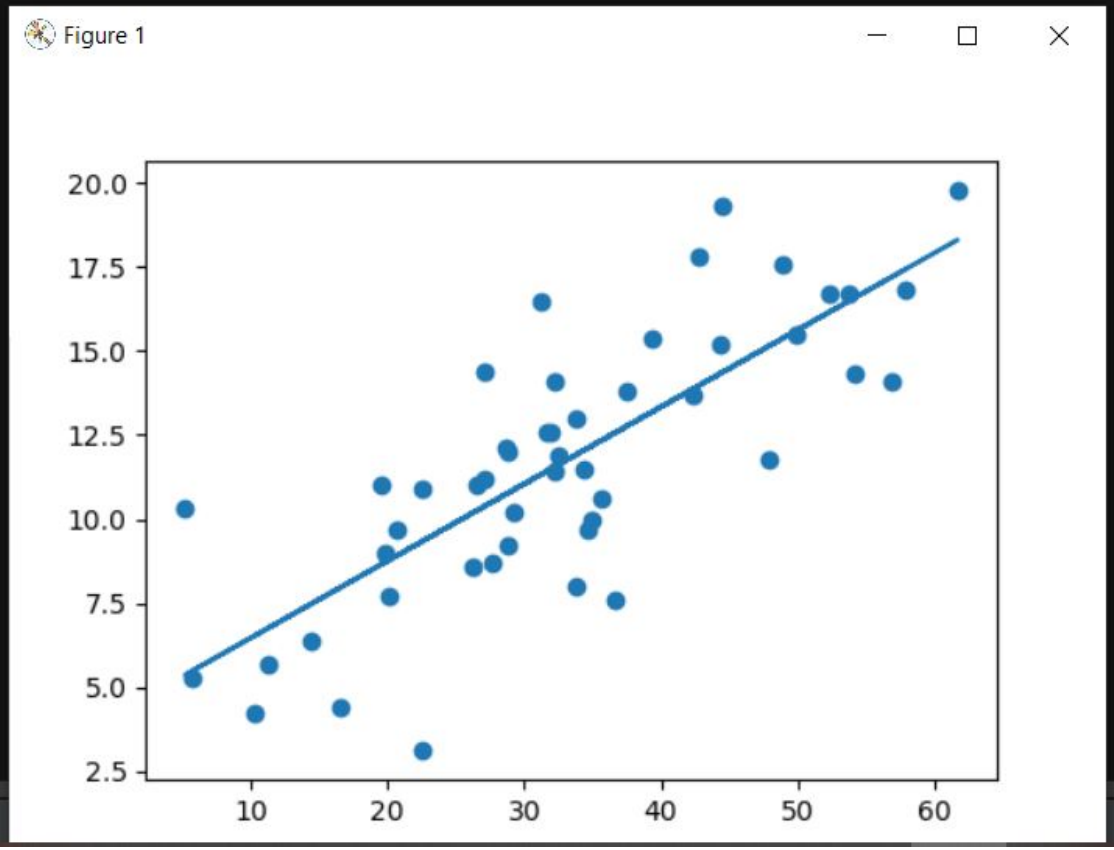
1(b) We removed the state from the data file. Here is the output after that:

Question number 1.py | Question number 2.py

```
18  r_square = model.score(x, y)
19  rvalue = math.sqrt(r_square)
20  print(rvalue)
21
22  #show the graph
23
24  plt.plot(x,model.coef_*x + model.intercept_)
25  plt.scatter(x,y)
26  plt.show()
```

Run: Question number 2 | Question number 1 | Question number 2

```
C:\Users\Abantika\AppData\Local\Programs\Python\Python37\python.exe "C:/hmm/courses/3rd semester/Special topics/Assignm
0.7822901765090373
```



PEP 8: block comment should start with '# '

2. (40 pts) Access the following Crime2 data.
http://users.stat.ufl.edu/~aa/smss/data/Crime2.dat
Let y = violent crime rate and x = poverty rate.

(a) Write a Python, Matlab, R or SPSS/Stata script, and show that the prediction equation, and interpret the y-intercept and the slope.

(b) Find the predicted violent crime rate and the residual for Massachusetts, which had x = 10.7 and y = 805. Interpret.

(c) Two states differ by 10.0 in their poverty rates. Find the difference in their predicted violent crime rates.

(c) From the prediction equation, can you tell the sign of the correlation between these variables? How?

2(a) From the CSV file we get the mean of x which is poverty rate is 14.2588

The mean of y (crime) is 612.84

Sum of x = 727.2

Sum of y = 312.55

$\Sigma (x_i - M_x)^2 = 1050.76$

$\Sigma (x_i - M_x)(y_i - M_y) = 51514.0706$

$\hat{y} = a + bx$ (we know)

Here, $b = \frac{51514.0706}{1050.76} = 49.0253$

$M_y - bM_x = -86.20026$

$\hat{y} =$ **$-86.20026 + 49.02537\,x$**

2(b) $x = 10.7$, $y = 805$

Residual $= (y - \hat{y})$

$\hat{y}$ at $x = 10.7 = -86.2003 + 49.0253(10.7)$

$= 438.3704$

Residual $= 805 - 438.3704$

$=$ **366.629**

2(c) Difference $= 10 \times 49.0253 =$ **490.2537**

2(d) $\Sigma (x_i - M_x)^2 = 1050.764$, $\Sigma (y_i - M_y)^2 = 9728474.7$

$\Sigma (x_i - M_x)(y_i - M_y) = 51514.071$

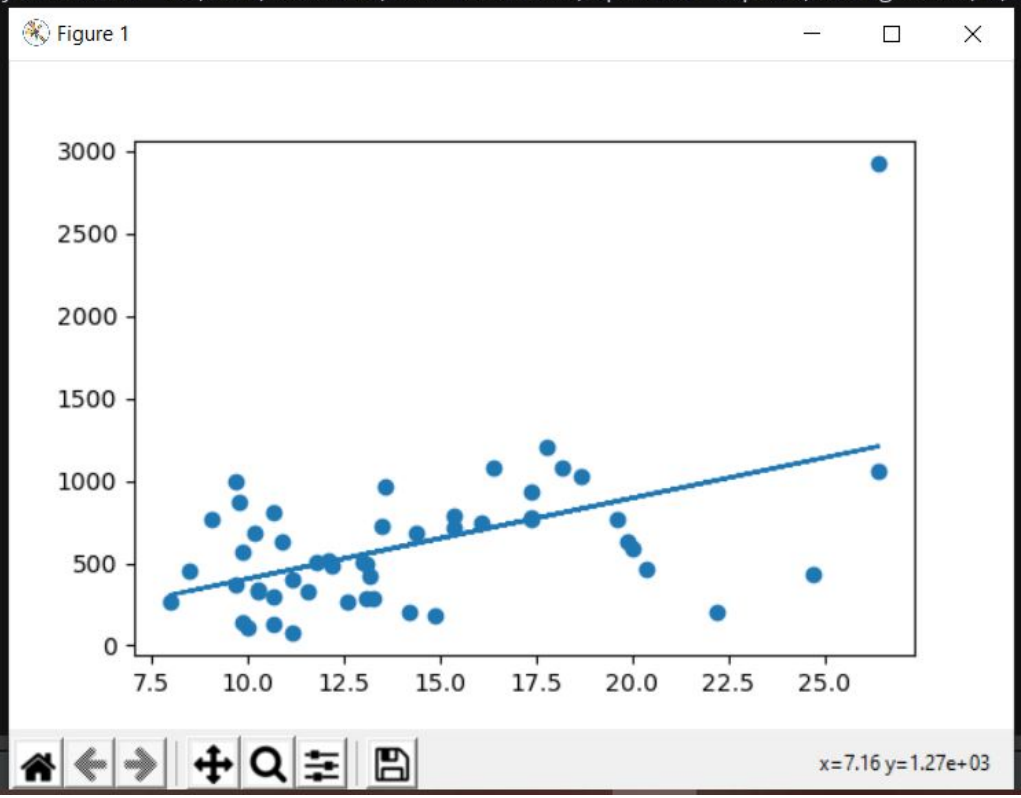
$r = \frac{\Sigma (x_i - M_x)(y_i - M_y)}{\sqrt{\Sigma (x_i - M_x)^2 * \Sigma (y_i - M_y)^2}} =$ **0.5095**

The sign is positive. The slope is also same as the sign.

```
19
20      r = math.sqrt(r_square)
21      print("r:", round(r,2))
22      print("The prediction equation:",model.intercept_,"+",model.coef_,"X")
23
24
25      plt.plot(x,model.coef_*x + model.intercept_,)
26      plt.scatter(x,y)
27      plt.show()
```

Run: Question number 2_1 | Question number 2_1 | Question number 2

```
C:\Users\Abantika\AppData\Local\Programs\Python\Python37\python.exe "C:/hmm/courses/3rd semester/Special topics/Assignment/6/New assignm
The coefficient of determination: 0.2595983878254692
r: 0.51
The prediction equation: [-86.200959] + [[49.02536979]] X
```



Packages installed successfully: Installed packages: 'sklearn' (today 2:05 AM)

3. (30 pts) Access the Houses data file
http://users.stat.ufl.edu/~aa/smss/data/Houses.dat
at the text website. For the response variable taxes and explanatory variable size, write a Python, Matlab, R or SPSS/Stata script, complete the following tasks.

(a) Graphically portray the association and describe it.

(b) Find and interpret the prediction equation.

(d) Find and interpret the correlation and r2.

code1.py [C:\Users\Abantika\AppData\Local\Temp\code1.py] - C:\hmm\courses\3rd semester\Special topics\Assignment\6\New assignment\Question number 3\3_1.py - PyCharm

File Edit View Navigate Code Refactor Run Tools VCS Window Help

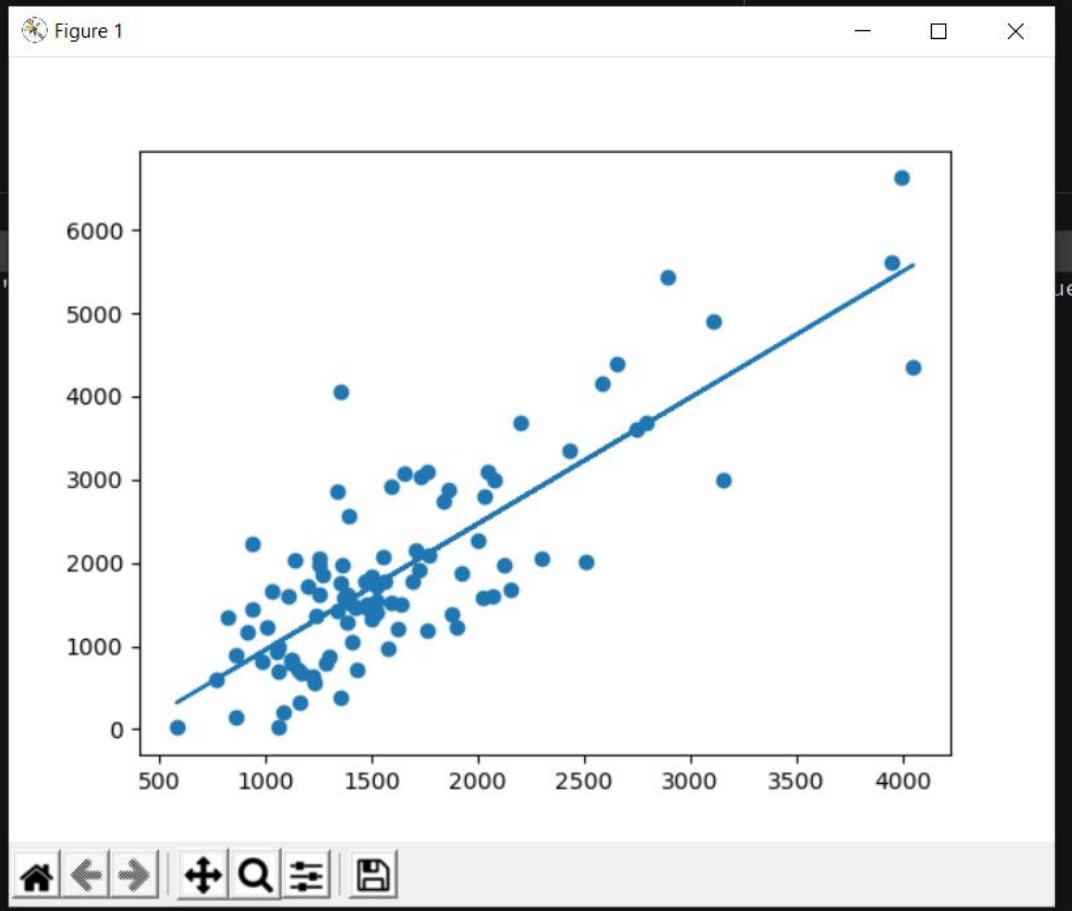
C: hmm courses 3rd semester Special topics Assignment 6 New assignment Question number 3 3_1.py

Question number 2_1.py | 3_1.py | Question number 1.py

```
18  print("The r2:", r_square)
19
20  r=math.sqrt(r_square)
21  print("The r:", round(r,2))
22  print('Slope:', model.coef_)
23  print('The y-intercept:', model.intercept_)
24  print("Prediction equation= ",model.intercept_,"+",model.coef_,"X")
25
26
27  plt.plot(x,model.coef_*x + model.intercept_)
28  plt.scatter(x,y)
29  plt.show()
```

Run: Question number 2_1 | 3_1 | Question number 2

```
C:\Users\Abantika\AppData\Local\Programs\Python\Python37\python.exe
The r2: 0.6704265211319951
The r: 0.82
Slope: [[1.51720732]]
The y-intercept: [-563.56553956]
Prediction equation=  [-563.56553956] + [[1.51720732]] X
```



4. (30 pts) A study was conducted using 49 Catholic female undergraduates at Texas A&M University. The variables measured refer to the parents of these students. The response variable is the number of children that the parents have. One of the explanatory variables is the mother's educational level, measured as the number of years of formal education. For these data, $\bar{x} = 9.88$, $s_x = 3.77$, $\bar{y} = 3.35$, $s_y = 2.19$, the prediction equation is $\hat{y} = 5.40 - 0.207x$, the standard error of the slope estimate is 0.079, and SSE = 201.95.

   (a) Find the correlation and interpret its value.

   (b) Test the null hypothesis that mean number of children is independent of mother's educational level, and report and interpret the P-value.

   (c) Sketch a potential scatterplot such that the analyses you conducted in (a) and (b) would be inappropriate.

4 (a) $\bar{x} = 9.88$

$S_x = 3.77$

$\bar{y} = 3.35$

$S_y = 2.19$

$\hat{y} = 5.40 - 0.207x$

(a) $r = \left(\frac{S_x}{S_y}\right) b$ (We Know)

$= -0.356$

The correlation is negative.

4 (b) Hypothesis H_0 = is independent

" H_a = not "

test static $= r / \sqrt{\frac{(1-r^2)}{(n-2)}}$ [Here, $n = 49$, $r = -0.356$]

|test static| = 2.62

The number of children is not independent of mothers educational level, here the test static is 2.62 which is smaller than 3.79.

P-value is $0.050004 > 0.05$

Question number 2_1.py | 3_1.py | 4_4.py

```
4   import numpy as np
5   import math
6   import seaborn as sns
7   import matplotlib.pyplot as plt
8   from sklearn.linear_model import LinearRegression
9
10  data=pd.read_csv('C:/hmm/courses/3rd semester/Special topics/Ass
11
12  col1=["Mother_education"]
13  x=data[col1]
```

Run: 4_4 | 4_4 | Question number 2

```
C:\Users\Abantika\AppData\Local\Programs\Python\Python37\python
The r2: 1.0
The r: 1.0
The y-intercept: [5.4]
The slope: [[-0.207]]
The prediction equation ŷ =  [5.4] + [[-0.207]] X
```

Figure 1