Dimas Rinarso Putro / drp354@nyu.edu

Applied Data Science

GX5004: HW 1

PROBLEMS:

To the extent that it helps you to learn, you may work with fellow students on this assignment. R and Python have extensive libraries online that can guide you on this assignment.

Total Number of Points: 20 points

1. Go to http://www.random.org/integers/ and generate two series of 10 random integers with values between 0 and 9. Call them Y and X. Using R or Python: [7 points]

ANSWER:

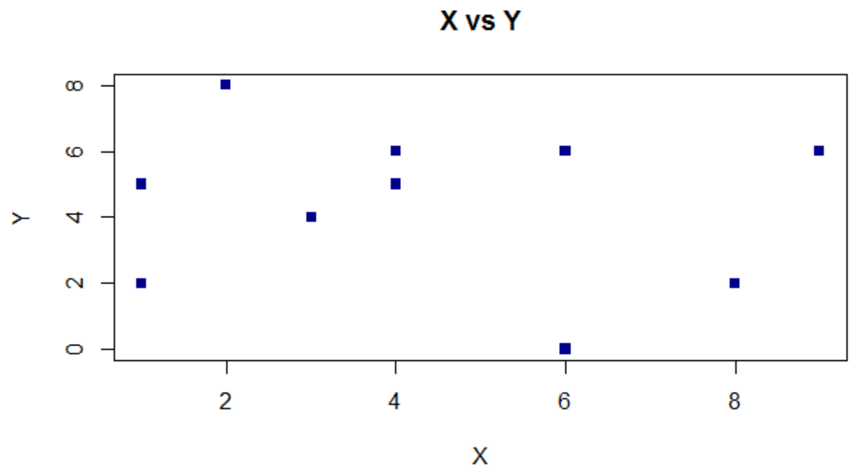
The random integer pairs obtained from the website:

```
Y <- c(2,6,4,6,8,2,0,5,5,6)
X <- c(1,4,3,6,2,8,6,4,1,9)
```

a. Generate a scatter plot of (Y, X).

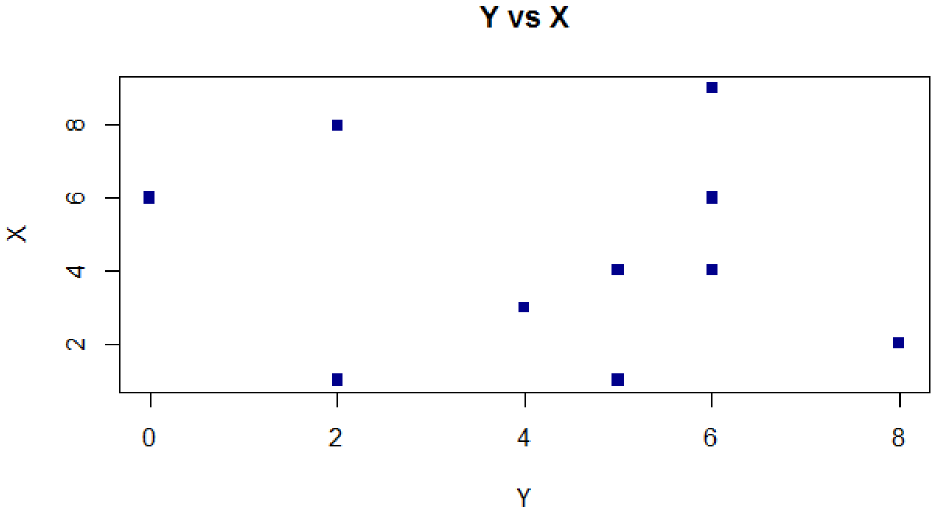
Input:

```
plot(X,Y,
     pch = 15,
     col = "darkblue",
     main = "X vs Y",
     xlab = "X",
     ylab = "Y")
```



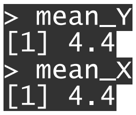
And not the plot of “Y,X” as mentioned in the problem:

```
plot(Y,X,
    pch = 15,
    col = "darkblue",
    main = "Y vs X",
    xlab = "Y",
    ylab = "X")
```



b. Calculate the means of Y and X.

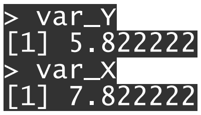
```
mean_Y = mean(Y, na.rm = FALSE)
mean_X = mean(X, na.rm = FALSE)
mean_Y
mean_X
```



```
> mean_Y
[1] 4.4
> mean_X
[1] 4.4
```

c. Calculate the variances of Y and X.

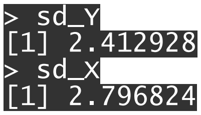
```
mean_Y = mean(Y, na.rm = FALSE)
mean_X = mean(X, na.rm = FALSE)
mean_Y
mean_X
```



```
> var_Y
[1] 5.822222
> var_X
[1] 7.822222
```

d. Calculate the standard deviations of Y and X.

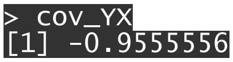
```
sd_Y = sd(Y, na.rm = FALSE)
sd_X = sd(X, na.rm = FALSE)
sd_Y
sd_X
```



```
> sd_Y
[1] 2.412928
> sd_X
[1] 2.796824
```

e. Calculate the covariance of Y and X.

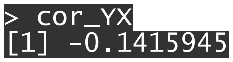
```
cov_YX = cov(Y, X)
cov_YX
```



```
> cov_YX
[1] -0.9555556
```

f. Calculate the correlation of Y and X.

```
cor_YX = cor(Y, X)
cor_YX
```



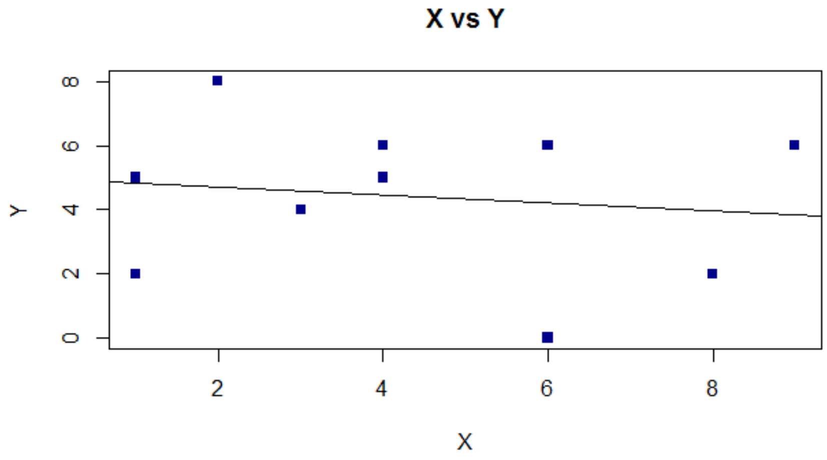
```
> cor_YX
[1] -0.1415945
```

g. Given the results above, suppose I ask you to predict the value of Y if I give you a value of X = 13. How would you respond?

The approach is to go find linear bivariate regression technique: $y = \beta_0 + \beta_1 x$. In R language it can be approached by:

Note that in this number X will be represented in X-axis and Y in Y-axis respectively.

```
lm.yr = lm(Y ~ X)
abline(lm.yr)
```



```
summary(lm.yr)
```

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2045 -1.6129  0.3679  1.7344  3.3068 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   4.9375     1.5515   3.183   0.0129 *
X            -0.1222     0.3020  -0.405   0.6964  
---
Signif. codes:  0 ·**·0.001 ·*·0.01 · ·0.05 · ·0.1 · ·1

Residual standard error: 2.534 on 8 degrees of freedom
Multiple R-squared:  0.02005,	Adjusted R-squared:  -0.1024 
F-statistic: 0.1637 on 1 and 8 DF,  p-value: 0.6964
```

We know from the summary above the coefficient of regression line, which is

X = -0.1222 and Intercept = 4.9375

Hence,

Y = -0.1222(X) + 4.9375.

For X=13, we could get:

```
> (-0.1222*13)+4.9375
[1] 3.3489
```

So expected Y value for X=13 is **3.3489**.

2. Recall we discussed the role that random variables play in applied data science. We also discussed the distinction between discrete and continuous random variables. For the random variables below, indicate the more appropriate random variable (discrete or continuous) and why you believe this to be the case. [7 Points]
    a. The number of taxi rides taken in a month by a NYC resident.
    b. The speed of a bicyclist on Jay Street.
    c. The luminosity of light emitted by street lamps in Brooklyn.
    d. The income of bankers working on Wall Street.
    e. The number of hotels in Manhattan.
    f. The ambient sound generated by trash trucks picking up trash at midnight.

ANSWER:

a. The number of taxi rides taken in a month by a NYC resident should be treated as a **discrete** data because the number of taxi rides per month countable.

b. The speed of a bicyclist on Jay Street should be treated as a **continuous** data because the data number is large and speed can have any kind of value.

c. The luminosity of light emitted by street lamps in Brooklyn should be treated as **continuous** data because of its physical element (for instance energy, projection of lightwave) that can be projected in time.

d. The income of bankers working on Wall Street should be treated as **continuous** data because at any given time the wage could change because of the performance of its trading performance which also related to time

e. The number of hotels in Manhattan should be treated as **discrete** data because it can plot by arranging for instance number of hotel per area at single point of time.

f. The ambient sound generated by trash trucks picking up trash at midnight should be treated as **continuous** data because of its physical element (frequency, projection of sound wave) that can be projected in time.

g. The quality of coffee served in the student lounge should be treated depends on how we quantify the “quality of coffee” into data. For instance, if we set a category of rating, let’s say 1-5 and have customers filling up questionnaire, for instance, then it’s a **continuous** data.

3. Consider the salaries of bankers on Wall Street. One argument that could be advanced is that Wall Street bankers have high salaries because they have attained high levels of education. Another argument could be advanced that Wall Street banker have high salaries because they have high Intelligence Quotients (IQ). If you had a dataset that provided you with the salaries of a sample of Wall Street bankers, together with their education levels and IQs, discuss how you might explore these arguments. Are there other methods beyond relying on this sample of bankers that could allow you to explore these arguments? (As with most applied data analytics, there is no right or wrong answer to this question. You may find it helpful to consider what we discussed in class to address it.) [6 points]

ANSWER:

Below are some basic steps might be useful to explore the arguments above:

1. Defining the problems and measurement framework

   First, we need to define the dataset and have a basic understanding of the problems, which is "which of the factors (IQ or Education level) plays important role in affecting the salary". It seems that the data provided only for the relations to salary, IQ and education level. IQ and education are the variables that might affect salary.

2. Obtaining the data

   We need to plot the dataset into these columns:

   a. Get the data of bankers VS education level
   b. Get the data of bankers VS IQ
   c. Get the data of bankers VS salaries

3. Data cleaning

   We need to perform data cleaning of those 3 datasets so salary, IQ and education level are respective to same bankers. So there are no bankers that do not have all three data so that the data can be confirmed appropriately and do not contain bias arguments.

4. Find correlation

   a. Sort those 3 datasets based on the value of salaries (ascending).
   b. Find correlation between:
      i. Salaries VS education level
      ii. Salaries VS IQ
   c. Find which one is more positively correlated.
   d. The bigger positive correlation value meaning that factor plays more important role or give more impacts to the high salaries.

5. Break down into more detail, explore more data if possible

   Although these measurements and data analysis could give you a glimpse of the relation of IQ and Education level to the bankers salary, there are other data which are not provided which possibly have higher impact to the level of salary. For instance, from the data that were sorted out in cleaning process in number 3, we might have groups of data which shows that neither IQ nor Education Level determine the salary level. Bankers might have more bonus because at the point of sampling A banker is handling stocks that is showing good performance and cannot be compared to people of same IQ (let's say banker B) that is not on the same project. However, the data we have cannot justify this because:

a. Data that were excluded from the data cleaning process above might only have IQ **or** Education level data, not **both**.
b. We do not have a dataset of the performance of projects or stocks related to each individual

Therefore, for more objective data analysis, it might be a good thing to include some other factors, for instance, such as project/stock trading performance per time (e.g. derived from stock values over time) that are handled each individual, to predict the "bonus" aspect of their salary.

6. Do some "street-level" data confirmation and feedbacks

   In determining what other dataset that might useful for the data analysis, we could also make human communication approach by sending surveys or interviewing their managers to get a clearer image of what parameters the use when justifying the performance of their employee and how they put that as a factor in their wages, to avoid bias in data and find clues to other dataset that we might need to support more objective measurement.