DATAFRAME VS SERIES

3. Answer the question: "What is the difference between DataFrame and Series?" Indicate how much time you spent completing this task.

A DataFrame is a two-dimensional data structure with table headings and indexes similar to your tables in Microsoft Excel or Google Spreadsheets, while a Series is a one-dimensional data structure with indexes but without table headings. In addition, A Series could be created from a DataFrame but not vice-versa.

A DataFrame

	user_id	event_date	event_type
0	c40e6a	2019-07-29 00:02:15	registration
1	a2b682	2019-07-29 00:04:46	registration
2	9ac888	2019-07-29 00:13:22	registration
3	93ff22	2019-07-29 00:16:47	registration
4	65ef85	2019-07-29 00:19:23	registration
5	90852e	2019-07-29 00:21:16	registration
6	357151	2019-07-29 00:25:53	registration
7	71ac11	2019-07-29 00:28:51	registration
8	af679d	2019-07-29 00:30:46	registration
9	a48f29	2019-07-29 00:41:54	registration
10	b65930	2019-07-29 00:47:22	registration
11	956ad6	2019-07-29 00:54:13	registration
12	8aa5b4	2019-07-29 00:58:22	registration
13	5fb555	2019-07-29 01:07:24	registration
14	37fa41	2019-07-29 01:10:32	registration
15	b5787e	2019-07-29 01:11:22	registration
16	b2e16e	2019-07-29 01:14:16	registration
17	ca3c58	2019-07-29 01:18:51	registration
18	bea18b	2019-07-29 01:41:59	registration
19	48cac1	2019-07-29 01:42:46	registration

A Series

```
0       c40e6a
1       a2b682
2       9ac888
3       93ff22
4       65ef85
5       90852e
6       357151
7       71ac11
8       af679d
9       a48f29
10      b65930
11      956ad6
12      8aa5b4
13      5fb555
14      37fa41
15      b5787e
16      b2e16e
17      ca3c58
18      bea18b
19      48cac1
20      5290a3
```

4. You are given two random variables X and Y.
   $E(X) = 0.5$, $Var(X) = 2$
   $E(Y) = 7$, $Var(Y) = 3.5$
   $cov(X, Y) = -0.8$
   Find the variance of the random variable $Z = 2X - 3Y$

Solution

General formula for random variable:

$V(X + Y) = V(X) + V(Y) + 2*cv*(SD_x)*(SD_y)$,

$V(aB) = a^2*V(B)$ where a is a real number

$SD_x = (V(X))^{1/2}$

$SD_y = (V(Y))^{1/2}$

Therefore,

$V(2X-3Y) = V(2X) + V(-3Y) + 2 * cv * (SD_{2x})*(SD_{-3y})$

$V(2X) = 2^2*2 = 8$, $V(-3Y) = -3^2*3.5 = 31.5$, $SD_{2x} = (8)^{1/2}$, $SD_{-3y} = (31.5)^{1/2}$, Cv = -0.8

$V(2X-3Y) = 8 + 31.5 + 2 * -0.8 * (8)^{1/2} * (31.5)^{1/2} = 14.10$

5. Omer trained a linear regression model and tested its performance on a test sample of 500 objects. On 400 of those, the model returned a prediction higher than expected by 0.5, and on the remaining 100, the model returned a prediction lower than expected by 0.7. What is the MSE for his model? Limor claims that the linear regression model wasn't trained correctly, and we can do improve it by changing all the answers by a constant value. What will be her MSE? You can assume that Limor found the smallest error under her constraints. Return two values - Omer's and Limor's MSE.

Defining variables;

actual_value = y, predicted_value = y_pred, error = abs(y – y_pred)

Estimating MSE for Omer's model

y_pred – y = 0.5 #for first 400 test samples

y – y_pred = 0.7 # for remaining 100 test samples

$MSE = (0.5^2 * 400 + 0.7^2 * 100) / 500 = 0.298$

Estimating MSE for Limor's model

There's not sufficient information to determine MSE for Limor's model. A lot of parameters affect model performance during Hyper-parameter tuning, therefore more information are required to estimate Limor's MSE.