Candidate: Andrés Felipe Bolaños Acosta.
Position: code reviewer.
E-mail: afb.acosta@gmail.com.
Mobile phone: +5511973662765.
Portfolio: https://github.com/AFBA1993/.
Linkedin: https://www.linkedin.com/in/andr%C3%A9s-felipe-bola%C3%B1os-acosta-b14b691a4/.
Google Scholar: https://scholar.google.com.br/citations?user=GLKkmsoAAAAJ&hl=pt-BR.

Yandex Tasks' Answers

Task 3

How would you answer the student's question below? Your task is to get your message across in such a way that a beginner can understand your explanation. You can do this any way you want (pictures, GIFs, metaphors, anything) so long as it makes your explanation clear. Answer the question: "What is the difference between DataFrame and Series?" Indicate how much time you spent completing this task.

Answer:

The difference between a DataFrame and Series is that a DataFrame is a collection of Series. A practical analogy is depicted by Fig. 1. This figure illustrates a train where a simple wagon represent a series and the entire train denotes a dataframe. I would also highlight that the best analogy are intercity and load trains, because the space inside their wagons have labels (seats or space allocated for load). Consequently, those labels behave as indexes either in dataframes or series. It is also important to mention that the capacity of wagons in the previously mentioned type of trains is similar, for instance, the wagons A and B of a certain train have 26 seats in each of them. Likewise, the series that compose a dataframe must have the same size. (2 hours and 15 minutes were spent for solving this task)

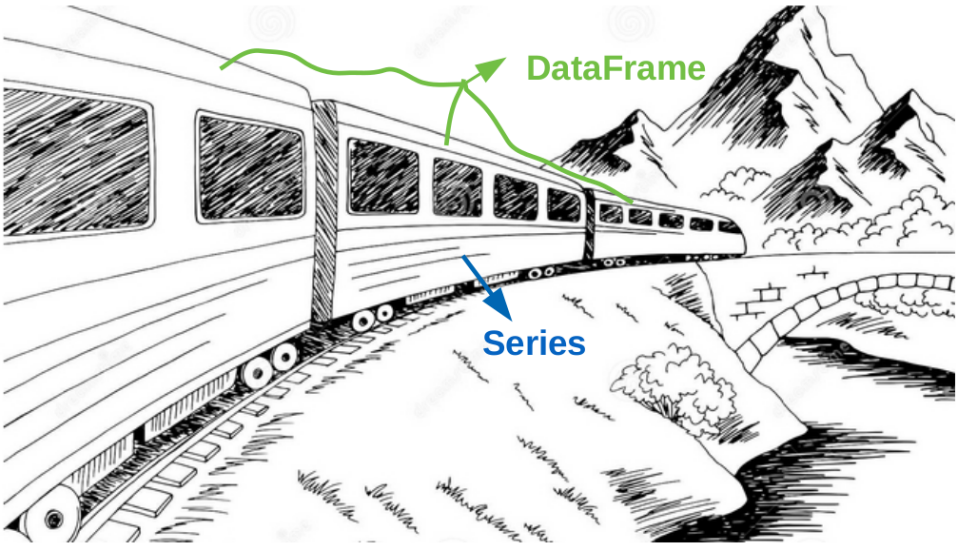

Figure 1: Train analogy: deference between Series and DataFrame.
Font: https://www.dreamstime.com/train-travel. (Adapted)

Task 4

You are given two random variables X and Y.

$$\begin{aligned} E(X) = 0.5,\ Var(X) = 2 \\ E(Y) = 7,\ Var(Y) = 3.5 \\ cov(X, Y) = -0.8 \end{aligned} \tag{1}$$

Find the variance of the random variable $Z = 2X - 3Y$.

Answer:

The covariance of X and Y is different from zero, which means that X and Y are dependent. Therefore:

$$Var(aX - bY) = a^2 Var(X) + b^2 Var(Y) - 2ab\, cov(X, Y) \tag{2}$$

Further, Eq.(3) results from applying Eq.(2) to the random variable Z.

$$Var(2X - 3Y) = 4Var(X) + 9Var(Y) - 12ab\, cov(X, Y) \tag{3}$$

Finally, the problem conditions defined by Eq.(1) are substituted into Eq.(3)

$$Var(Z) = Var(2X - 3Y) = 49.1 \tag{4}$$

Task 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.

Answer:

Eq.(5) determines the MSE regarding Omer's model.

$$MSE_{\text{Omer}} = \frac{1}{500} \sum_{i=1}^{500} (y_i - \hat{y}_i)^2 \tag{5}$$

In Eq.(5) the terms y as well as $\hat{y}$ refer to the predicted value and the real one, respectively. In addition, the subscript i represents an object within the 500 data. The summation term in In Eq. (5) can be separated in two parts, from the firt object up to 400 objects, and from 401 to 500 objects, which yields:

$$MSE_{\text{Omer}} = \frac{1}{500} \left[\sum_{i=1}^{400} (y_i - \hat{y}_i)^2 + \sum_{i=401}^{500} (y_i - \hat{y}_i)^2 \right] \tag{6}$$

Consequently, Eq.(7) results from substituting into Eq.(6) the difference between Omer's model predicted values and the real ones in both parts of the data.

$$MSE_{\text{Omer}} = \frac{1}{500} \left[\sum_{i=1}^{400} (0.5)^2 + \sum_{i=401}^{500} (-0.7)^2 \right] \tag{7}$$

$$MSE_{\text{Omer}} = \frac{1}{500} \left[400(0.5)^2 + 100(-0.7)^2 \right] \tag{8}$$

Finally, the MSE for Omer's model is calculated by Eq. (8) and it is equal to 0.298.

$$MSE_{\text{Omer}} = 0.298 \tag{9}$$

With respect to the MSE for Limor's model, it is obtained by adding the constant k to Eq.(7) as defined by Eq.(10).

$$MSE_{\text{Limor}} = \frac{1}{500} \left[\sum_{i=1}^{400} (0.5 + k)^2 + \sum_{i=401}^{500} (k - 0.7)^2 \right] \tag{10}$$

$$MSE_{\text{Limor}} = \frac{1}{500} \left[400(0.5 + k)^2 + 100(k - 0.7)^2 \right] \tag{11}$$

Therefore, the MSE for Limor's model becomes a function depending on the k term as defined by Eq.(11). In order to calculate the minimum of MSE_{Limor}, Eq.(11) was implemented in a Python program and the function 'minimize' from 'scipy.optimize' packages was used to find the minimum of Eq.(11) , as displayed by Fig. 2.

$$k = -0.26 \tag{12}$$

$$MSE_{\text{Limor}} = 0.2304 \tag{13}$$

The value for Limor's MSE occurs when $k = -0.26$ and it's value corresponds to 0.2304. This values could be also attained from an analytical solution, which consists of finding the root of the first derivative equals zero of Eq.(11). After some mathematical manipulation Eq.(11) becomes:

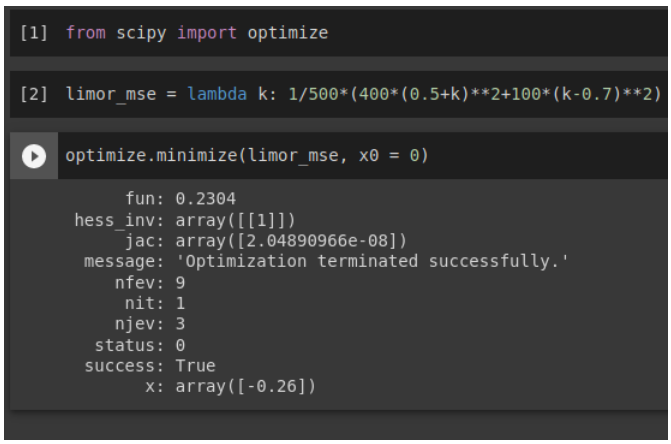

Figure 2: Python code for computing the minimum of Eq.(11).

$$MSE_{\text{Limor}} = 0.298 + 0.52k + k^2 \tag{14}$$

$$\frac{d(MSE_{\text{Limor}})}{dk} = 0.52 + 2k = 0 \tag{15}$$

The root of Eq.(15) is then determined by Eq.(12) and consequently Limor's MSE corresponds to 0.2304.