Data Analysis With Python – Module 6 – Case Studies

Case Study #1: Linear Regression of Height & Weight

This case study will use data science and machine learning to predict mass when height is known. The dataset is a sample of women aged 30-39, derived from here: https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_example

Please perform the following steps to complete this case study:

1. Create a new empty Jupyter Notebook.
2. Import all the modules required for:
    - numpy
    - pandas
    - matplotlib
    - seaborn
    - LinearRegression
3. Read the height_mass.csv file into a Pandas DataSet called: people
    - Use the pandas read_csv method.
4. Use a Seaborn distplot to show the distribution for Mass.
    - https://seaborn.pydata.org/examples/distplot_options.html
    - Use bins=50
    - What does the plot tell you about the data?
    - Insert a markdown cell and note your observations.
5. Use a Seaborn distplot to show the distribution for Height.
    - Use bins=50
    - What does the plot tell you about the data?
    - Insert a markdown cell and note your observations.
6. Use a Seaborn jointplot to plot x=Height, y=Mass
    - Does this plot confirm what the distplot showed?
    - Insert a markdown cell and note your observations.
7. Split the data into training and testing data.
    - Prepare your x and y:
        - x: Drop the Mass column.
        - y: Specify the Mass column.
    - Use sklearn train_test_split to split the data.
8. Create the model and fit it to the training data.
    - Create a sklearn LinearRegression model.
    - Use the fit method to fit it to the training data.
9. Predict values based on testing data.
    - Use the predict method to predict values with the x testing data.
10. Print out error metrics:
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
11. Predict some specific mass. Choose any height within the range of the data, and see whether the prediction is close to reality.

- Use the predict method and feed it a 2d array like: [[1.70]]
- Add a markdown cell and explain how well the prediction matched reality, with specific attention to the RMSE error.

12. Use seaborn to display an lmplot with the linear regression line shown (fit_reg=True).