**Name: Syeda Zainab**

**Roll no: 24 MS DS020**

**Sub: Data Science Tool & Techniques**

**Assignment: EDA Exercise**

**Submitted to: Ma’am Dr.Sanam Narejo**

## **Exercise**

For this assignment, you'll be using the [seeds](https://archive.ics.uci.edu/ml/datasets/seeds) data set, generated by recording X-Ray measurements of various wheat kernels.



### **Histograms**

Question 1: Write python code that

1. Loads the seeds dataset into a dataframe.

**df=pd.read\_csv("wheat.csv",index\_col=7)**

**print(df.columns)**

1. Creates a slice of your dataframe that only includes the **area** and **perimeter** features **my\_dataframe1=df[['area','perimeter']]**
2. Creates another slice that only includes the **groove** and **asymmetry** features

**my\_dataframe2=df[['grove','asymmetry']]**

1. Creates a histogram for the 'area and perimeter' slice, and another histogram for the 'groove and asymmetry' slice. Set the optional display parameter: **alpha**=0.75

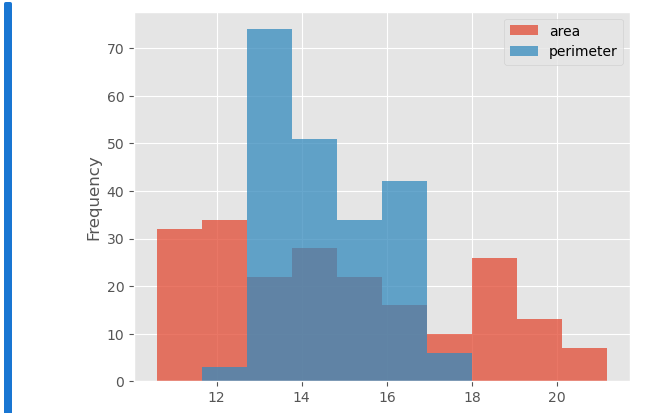
**my\_dataframe1.plot.hist(alpha=0.75)**

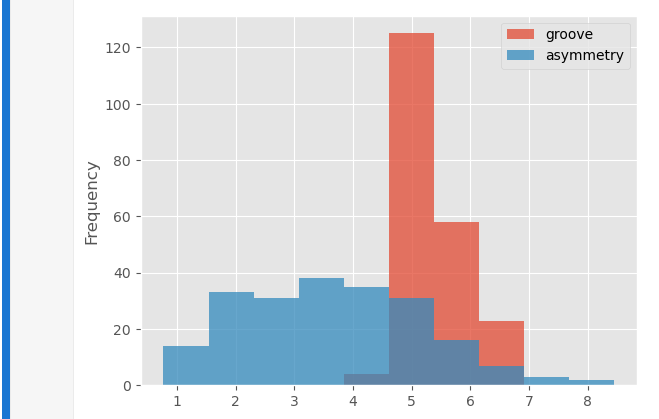
**plt.title="The Relation of Area and Perimeter of wheat"**

**plt.show()**

**my\_dataframe2.plot.hist(alpha=0.75)**

**plt.show()**

****

****

Once you're done, run your code and then answer the following questions about your work:

1. Looking at your first plot, the histograms of area and perimeter, which feature do you believe more closely resembles a Gaussian / normal distribution?

**Ans:** In first histogram the perimeter feature more closely resembles a normal distribution and in second char the feature groove is more closely resembles.

1. In your second plot, does the groove or asymmetry feature have more variance?

**Ans:** Yes as per above mentioned histogram there is more variance between groove and asymmetry.

### **2D Scatter Plots**

Question 2: Write python code that

**Import libraries**

import pandas as pd

import matplotlib

import matplotlib.pyplot as plt

matplotlib.style.use('ggplot')

1. Loads up the seeds dataset into a dataframe

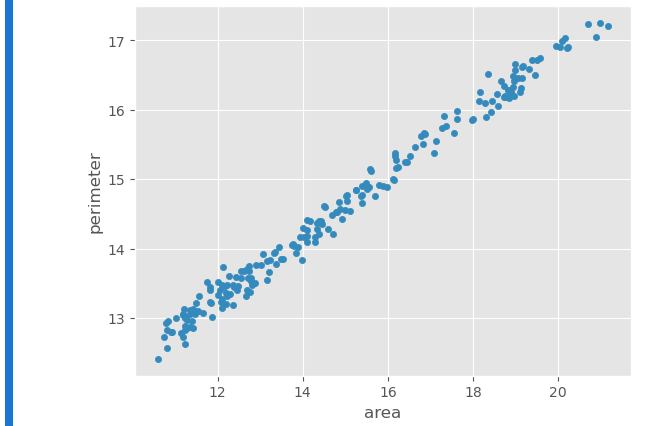
**df=pd.read\_csv("wheat.csv",index\_col=7)**

**print(df.columns)**

1. Create a 2d scatter plot that graphs the **area** and **perimeter** features

**df.plot.scatter(x='area',y='perimeter')**

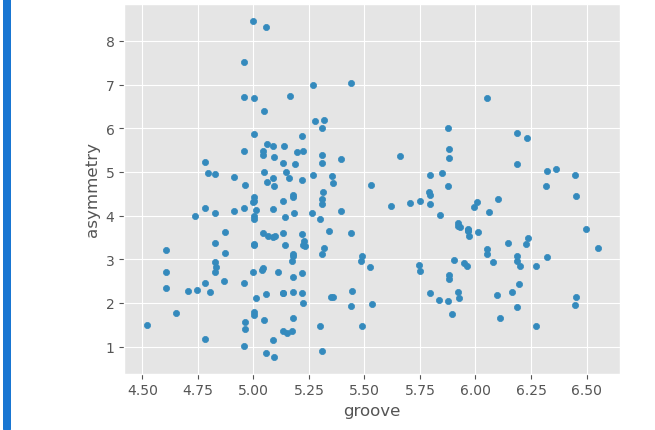
**plt.show()**



1. Create a 2d scatter plot that graphs the **groove** and **asymmetry** features

**df.plot.scatter(x='groove',y='asymmetry')**

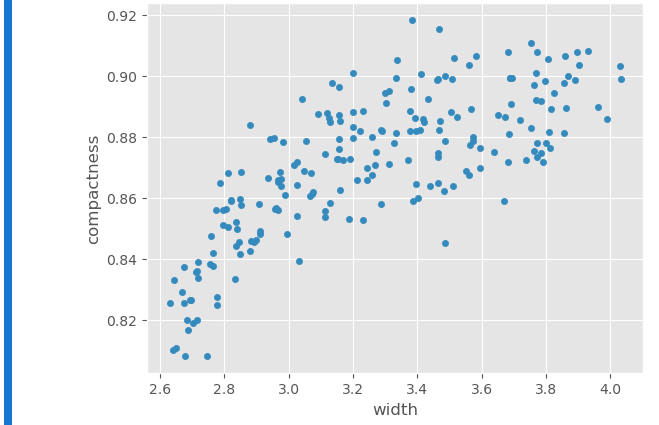
**plt.show()**



1. Create a 2d scatter plot that graphs the **compactness** and **width** features

**df.plot.scatter(x='width',y='compactness')**

**plt.show()**



Once you're done, answer the following questions about your work:

1. Which of the three plots seems to totally be lacking any correlation?

**Ans:** According to above visualized graphs the is no correlation or we can say that it is totally lacking correlation.

1. Which of the three plots has the most correlation?

**Ans:** According above data visualization the scatter graph 1 visualize good correlation or we can say that strong correlation between perimeter and area feature, and in third scatter chart of compactness and width visualizes weak correlation.

### **3D Scatter Plots**

Question 3: Write python code that

Import libraries

import pandas as pd

import matplotlib

import matplotlib.pyplot as plt

from mpl\_toolkits.mplot3d import Axes3D

matplotlib.style.use('ggplot')

1. Loads up the seeds dataset into a dataframe. You should be very good at doing this by now.

**df=pd.read\_csv("wheat.csv",index\_col=7)**

**print(df.columns)**

1. Graph a 3D scatter plot using the **area**, **perimeter**, and **asymmetry** features. Be sure to label your axes, and use the optional display parameter **c**='red'.

**fig=plt.figure()**

**ax=fig.add\_subplot(111,projection='3d')**

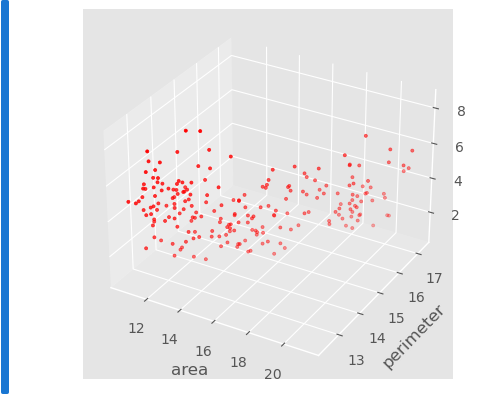
**ax.set\_xlabel('area')**

**ax.set\_ylabel('perimeter')**

**ax.set\_zlabel('asymmetry')**

**ax.scatter(df.area,df.perimeter,df.asymmetry,c='red',marker='.')**

**plt.show()**



1. Graph a 3D scatter plot using the **width**, **groove**, and **length** features. Be sure to label your axes, and use the optional display parameter **c**='green'.

**fig=plt.figure()**

**ax=fig.add\_subplot(111,projection='3d')**

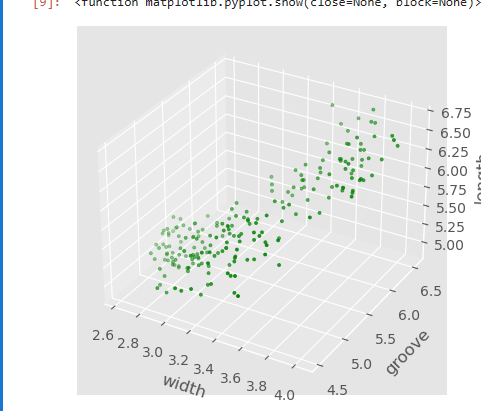
**ax.set\_xlabel('width')**

**ax.set\_ylabel('groove')**

**ax.set\_zlabel('length')**

**ax.scatter(df.width,df.groove,df.length,c='green',marker='.')**

**plt.show**



Once you're done, answer the following questions about your work.

1. Which of the plots seems more compact / less spread out?
2. Which of the plots were you able to visibly identify two outliers within, that stuck out from the samples?

### **Parallel Coordinates**

Question 4: Write python code that

import pandas as pd

import matplotlib.pyplot as plt

1. Loads up the seeds dataset into a dataframe

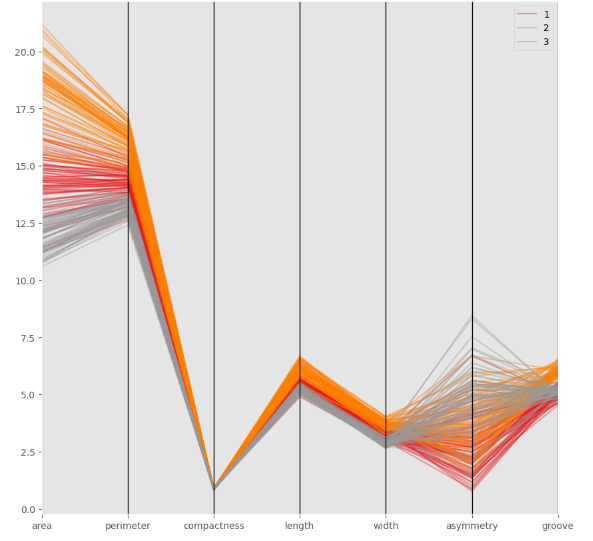
**df = pd.read\_csv('wheat.csv')**

1. Drop the **area**, and **perimeter** features from your dataset. Use .drop method on data frame to drop specified columns
2. Plot a parallel coordinates chart, grouped by the **wheat\_type** feature. Be sure to set the optional display parameter **alpha** to 0.4

**plt.figure(figsize=(10, 10))**

**pd.plotting.parallel\_coordinates(df, 'wheat\_type', colormap='Set1', alpha=0.4)**

**plt.show()**



Once you're done, answer the following questions about your work.

1. Which class of wheat do the two outliers you found previously belong to?

**Ans**: In class 3 of Wheat type having tow outliers according to above visualization.

1. Which feature has the largest spread of values across all three types of wheat?

**Ans:** Area has largest spread values and in second number the asymmetry feature has largest spread value across all three types of wheat.

### **Andrew’s Plot**

Question 5: Write python code that

import pandas as pd

import matplotlib.pyplot as plt

1. Loads up the seeds dataset into a dataframe

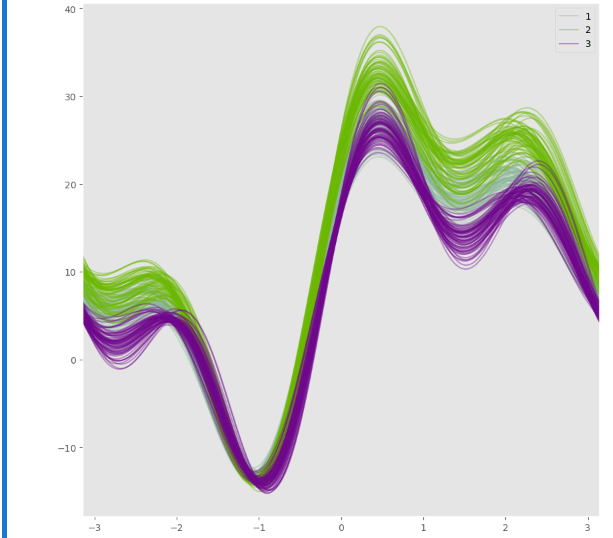
**df = pd.read\_csv('wheat.csv')**

1. Plot anandrew’s curve chart, grouped by the **wheat\_type** feature. Be sure to set the optional display parameter **alpha** to 0.4

**plt.figure(figsize=(10, 10))**

**pd.plotting.andrews\_curves(df, 'wheat\_type', alpha=0.4)**

**plt.show()**



Once you're done, answer the following questions about your work.

1. Are your outlier samples still easily identifiable in the plot?

**Ans:** outliers are not properly easily identifiable

### **IMSHOW**

Question 6: Write python code that

import pandas as pd

import matplotlib.pyplot as plt

import matplotlib

import matplotlib.image as mpimg

1. Loads up any image of your choice, into a dataframe.
2. Print shape and type of the object holding image.
3. Plot image using **imshow.**

**img=mpimg.imread("flower.jpg")**

**print(img.shape)**

**plt.imshow(img)**

****