COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Machine Learning, I

CS 5805

Homework # 3

In this LAB, you will practice data transformation and dissimilarities using the Minkowski distance. Object orient programing will also be exercised in this LAB. All numbers must be displayed with 2-digit decimal precision.

Set the numpy random seeds to 5808 for all questions.

1. Write a simple python **class** named "data preprocessing "with the one attribute and the following methods: [15 pts]
    a. Normalized
    b. Standardized
    c. IQR
    d. Show_original
    e. Show_normalized
    f. Show_standardized
    g. Show_IQR

where:

Normalized: normalized all features inside the dataset.

Standardized: standardized all features inside the dataset.

Show originals: Plot all the raw features

Show_normalized: Plot the normalized features.

Show_ standardized: Plot the standardized features.

Show_ IQR: Plot the scaled features using median and quantiles.

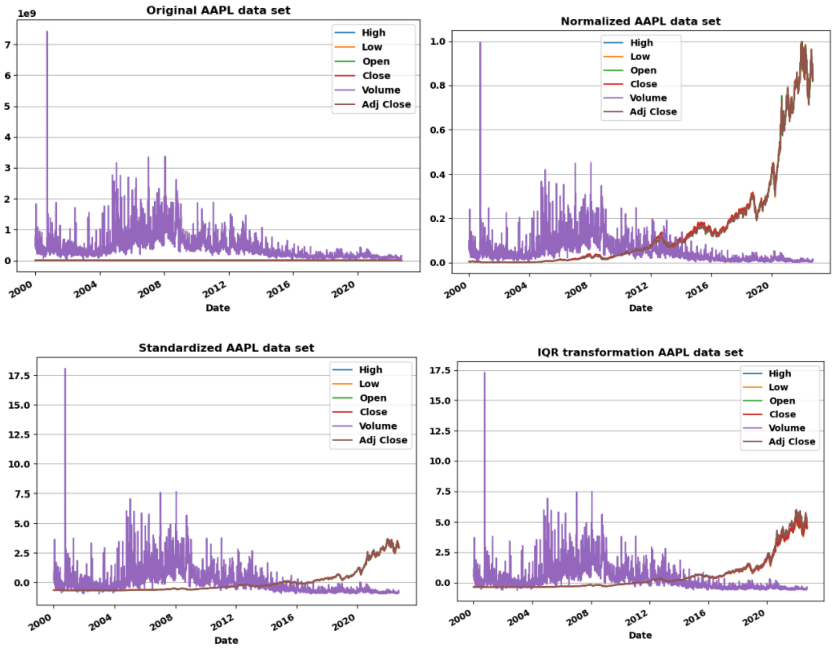
Hint: The python program for these questions should be saved under different file name. The rest of the problems should be saved under another file name (i.e., LAB_answer.py).

2. Develop another python file (i.e., LAB_answer.py), that imports the created object in the previous problem. Then read 'Apple' stock using the pandas_datareader package inside the Yahoo API. Pick the start and end date as [10 pts]

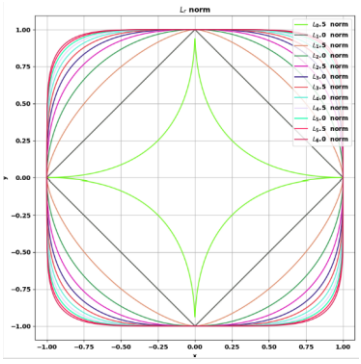
```
from pandas_datareader import data
import yfinance as yf
yf.pdr_override()
df = data.get_data_yahoo('AAPL', start='', end='')
```

```
start="2000-01-01", end="2022-09-25"
```

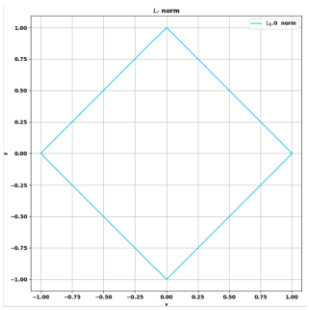
Create an instance with the Apple dataset [with all features] and plot the original, normalized, standardized and IQR transformed features using the created methods in the previous question.



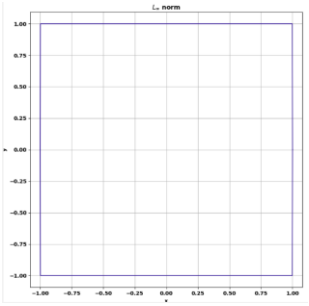
3. Develop a python program that implements Minkowski distance equation for various L-norms r = 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6. Plot all norms in one graph. The final graph should be similar to the folowing. Add appropriate x-label, y-label, title, legend and grid to your graph. [10 pts]



4. [No need to use Python] Write a <u>proof</u> that shows the diamond shape shown below corresponds to L-1 Minkowski distance with unity distance. Hint: you need to start with the boundary line equtions and proves $|x|+|y|=1$. [10 pts]



5. [No need to use Python] Write a <u>proof</u> that shows the L_∞ Minkowski distance become a square box in 2 dimensions. Consider the Minkowski distance $L_\infty$to be equal to 1. [10 pts]



6. Develop a python program that generate normally distributed random variable $x \sim N(\mu = 1, \sigma^2 = 2)$ and sample size =1000. Then generate a new random variable y such that [35 pts]

$$y = x + \epsilon$$

$$\epsilon \sim N(\mu = 2, \sigma^2 = 3)$$

   a. Construct the estimated covariance matrix (use the definition for the constriction of covariance and don't <u>not</u> use the built-in .conv() function in python) $\widehat{\sum}$ of feature matrix X = [x,y]. Display the covariance matrix through a table [Hint: Use <u>PrettyTable</u> or Tabulate package] with the title of 'Estimated covariance Matrix". Justify why the diagonal elements of calculated covariance matrix $\widehat{\sum}$ is correct.

   b. Calculate the eigenvalue and eigen vector of the covariance matrix $\widehat{\sum}$. Display the eigenvalues and eigenvectors through a table with an appropriate title.

Eigenvalue & Eigenvector	
$\lambda_1 =$	$\lambda_2 =$
Eigenvector related to λ_1	Eigenvector related to λ_2

   c. Graph the scatter plot between x and y. On the same graph plot the eigenvectors with different colors. What is the meaning of eigenvectors corresponding to the maximum and minimum eigenvalues? If you were to drop one feature (x or y) which attribute, do you drop and why? Add an appropriate x-label, y-label, title, legend, and grid to the plot.

d. Calculate the singular values of the feature matrix $\tilde{X}$ [centered X] and display the results through a table on the console with an appropriate title [use np.linalg.svd]. What is the relationship between the singular values of feature matrix $\tilde{X}$ and the e-values of the $\tilde{X}^T\tilde{X}$?

e. Using the DataFrame.corr() calculate the correlation matrix for X = [x,y]. What is the sample Pearson correlation coefficient between x and y? What is the relationship between the correlation coefficient matrix and the estimated covariance matrix. Justify your answer using the formula equation for the covariance and correlation coefficient.

7. Develop a python function program that performs 1st, 2nd and 3rd order differencing to time series dataset in a general case. Then test your program with the following time series input dataset $x(t)$: -4 to 4 with step size = 1 and $y(t) = x(t)^3$. Perform 1st, 2nd and 3 order differencing on the $y(t)$ and display the original dataset, 1st, 2nd and 3rd order differenced dataset through a table as shown below. Plot the original dataset, versus the 1st, 2nd and 3rd order differencing in one graph versus x(t). Justify why the graph makes sense using the definition of differencing and the equation for $y(t)$. Add appropriate title, legend, x-label, y-label, and grind to the plot. [10 pts]

x(t)	y(t)	$\Delta y(t)$	$\Delta^2 y(t)$	$\Delta^3 y(t)$
-4				
-3				
-2				
-1				
0				
1				
2				
3				
4				