* The assignment must be individual work and must not be copied or shared. Any tendency of cheating/copying evidence will lead to a 0 mark for the assignment.
* There are two types of problems – conceptual paper-based and application notebook-based. The conceptual paper-based problems need to be worked on paper and submitted in a single pdf file. Please name the file as *Name\_HW#*.
* The application problems must be worked on in the notebook using python. If the problem does not mention the libraries/packages to use, students must only use pandas, numpy, and spacy. All problems must be in a single notebook file as *Name\_HW#.*
* All files must be compressed into a single zip file as *Name\_HW#.*

**Question 1. Gaussian Distribution [15 pts]**

Verify that the univariate Gaussian distribution given by

satisfies

Then, show that the Gaussian also satisfies

**Question 2. Probability Theory [20 pts]**

Consider *k-*many fair coins flip where each coin has probability of . Suppose you are flipping coins *N* many times and you increase the number of flips (*t*) by intervals of 50 when *N* increases by one, see the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 |  | 200 |
|  | 50 | 100 | 150 |  | 10000 |

1. (10 pts) Implement a function, “coin\_flip\_prop”, that calculates the probability of having heads using *NumPy.random.binomial(k,*,*t)=h*, where *k* is the number of coins, is the probability of each coin, is the number of total flips, and *h* is head. To confirm the coin\_flip\_prop function applicability, show that the probability reaches 0.5 as *N* increases from 1 to 200 when *k=1*. Make a visualization.
2. (5 pts) Numerically calculate the probability of having 2 heads when 5 coins are flipped.
3. (5 pts) Run the coin\_flip\_prop function above for . Report the average probability.

**Question 3. KMeans [25 pts]**

1. (5 pts) The data set, HW1\_Q3.csv, has unlabeled data. Visualize the data set and make an educated guess of what each column is. Do you find any relationship or trend between columns? Explain.
2. (10 pts) Implement a function KMeans using NumPy to cluster the data set for
3. (7 pts) Using myKMeans, report the centroid and covariance of each cluster.
4. (3 pts) Use *sklearn.KMeans* to find the appropriate *k* value.

**Question 4. Exploratory Data Analysis [25 pts]**

There are two data files provided (d1: GDP.csv, d2: Country.csv). The task of this problem is to practice data mining and exploratory data analysis (EDA) to learn about the data.

1. (5 pts) Load two data files, d1 and d2, and merge two files using the country code. How many countries were merged?
2. (5 pts) Visualize the number of countries in each region and report the region having the most and the least number of countries.
3. (5 pts) Compare the GDP distribution of each region and determine the regions having the similar GDP distribution using KS-test. Report the regions.
4. (5 pts) Normalize the GDP distribution for each region and compare. Did the KS-test result change? Explain why or why not.

**Question 5. Principal Component Analysis [10 pts]**

Consider the data table below.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1.23 | 1.57 | 1.69 | 1.5 | 1.83 | 1.57 | 1.01 | 1.61 | 1.02 | 1.32 | 1.79 |
|  | 1.51 | 2.46 | 2.86 | 2.25 | 3.35 | 2.46 | 1.02 | 2.59 | 1.04 | 1.74 | 3.2 |
|  | 2.67 | 3.25 | 3.53 | 3.1 | 3.93 | 3.25 | 2.48 | 3.34 | 2.49 | 2.79 | 3.81 |

Suppose we are going to make a model where is the new data points transformed from PCA. Use *sklearn.decomposition* to do the problem.

1. (5 pts) Find the unit vector of the principal component. Then transform the data.
2. (5 pts) Visualize that t is linear with respect to