MTH 552: Problem Set *#1*

Problem #1: Consider the "World economic development" dataset (source: International Monetary Fund) (eco_dev_data.xlsx) containing data on the following economic development indicators of 121 countries.

Economic Development Indicators

Indicator and abbreviation	**Aspect of Economic development**
GNP per capita at PPP (GNPPER)	Income level
GDP growth rate (GDPGR)	Growth of economy
Gross domestic investment as percentage of GDP (DOMINV)	Level of investment
GDP deflator (GDPDFL)	Inflation
Agriculture value added as percentage of GDP (AGRVLAD)	Structure of output
Industry value added as percentage of GDP (INDVLAD)	Structure of output
Export of goods and services as percentage of GDP (EXP)	Openness of economy
General government consumption as percentage of GDP (GOVCON)	Role of government
Resource balance as percentage of GDP (RESBL)	Net borrowing/lending on account of merchandise trade
Domestic credit provided by the banking sector as percentage of GDP (DOMCRDT)	Private sector financing
Ratio of gross international reserve to imports (GRIIMP)	Strength of foreign exchange reserve
Number of months of import cover (IMPCOV)	Strength of foreign exchange reserve
Interest spread (INTSPRD)	Efficiency of financial market

(a) Detect multidimensional outliers using a principal component projection of the data.

(b) Detect rough clusters of world economies from the PCA projection and compare the clusters with clusters formed using a k-means clustering method with the same number of clusters.

(c) Obtain, if possible, a ranking of the world economies in terms of state of economic development using PCA.

(d) Obtain variable clustering using PCA.

Problem #2: The “Car dataset” (cars_data.xlsx) gives data of 32 cars on 11 variables which comprise of important features of the automobile. Variable codes given in the dataset and their description is given below:

Variable code	**Variable description**
Mpg	Miles/(US) gallon
Cyl	Number of cylinders
Disp	Displacement (cu.in.)
Hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears

(a) Obtain principal component projection of the data.

(b) How much proportion of total variation is explained by the first 3 principal components collectively?

(c) Obtain clusters from the PCA projection. Can you identify important characteristics of the identified clusters?

(d) Can you identify any outliers in the data?

(e) Obtain clustering of the automobiles, using k-means and hierarchical clustering and compare the clusters with the clusters obtained through PCA

Problem # 3: Consider the “Bank financial ratios data set’ (Source: Reserve Bank of India) (Public_bank_fin_ratio.xlsx) containing 19 important financial ratios of Indian public sector banks during the period from financial year 1996-1997 to financial year 1999-2000.

Ratio	**Abbreviation**
Return on Equity	ROE
Return on asset	ROA
Cost of deposit	COD
Cost of borrowing	COBR
Return on advances	ROAD
Return on Investment	ROI
Operating profit to total asset	OPTAST
Interest income to total income	INTINTOT
Other non-interest income to total income	OTHINTOT
Commission etc. to total income	COMINTOT
Net Interest income to total asset	NIITAST
Spread=Return on (advance+investment)-cost of deposits	SPRD
Staff expenses to total expenses	STEXTEX
Provisions and contingencies to total asset	PRVTAST
Intermediation cost (other operating expenses) to total asset	INTRMTAST
Net NPA to net advances	NET_NPA
Capital adequacy ratio	CAR
Business per employee	BUSEMP
Profit per employee	PFTEMP

(a) Using PCA obtain an appropriate projection of the financial ratios data and trace the trajectory movement of the banks with significant movements on the projected plane over the years.

(b) Rank the financial institutions for all the 4 years under study. Explain whether such a ranking based on PCA is justified. In case the ranking scheme is justified, discuss the major changes over the years in the rank table.

(c) Apply clustering techniques to find clusters for the 4 years and find out significant inter-cluster movements, if any, over the years.

Problem #4: The "Wine dataset" (source: UCI Machine Learning Repository) (Wine_data.xlsx) gives data that are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different wineries. The analysis determined the quantities of 13 constituents found in each of the three types of wines, these are 1) Alcohol, 2) Malic acid, 3) Ash, 4)Alcalinity of ash, 5) Magnesium, 6) Total phenols, 7) Flavanoids, 8) Nonflavanoid phenols, 9)Proanthocyanins, 10)Color intensity, 11)Hue, 12)OD280/OD315 of diluted wines and 13)Proline.

(i) Ignoring the Type variable, representing the winery, obtain the clustering of the wine samples based on the features using;

(a) Hierarchical clustering with complete and single linkage,

(b) K-means method with K=3.

(ii) Calculate total within cluster sum of squares around respective cluster centroids and other measures for comparing the clusters detected in (a), (b) .

(iii) Suggest appropriate number of clusters under (a) and (b).

(iv) Validate the results of your clustering exercise with the "Type" variable after you have obtained the clusters.

IMPORTANT NOTE:

- You may do coding in R/MATLAB/PYTHON/SAS.
- Data sets are attached with mail.
- Answer to questions along with the codes/plots/justifications have to be presented.
- Problems would be assigned on "first come first served" basis (after I receive solution through email).