Problem #A

Consider data given in file **StudentScoresDataA** . Consider the following data description:

Table 1: Data Description

Field	Description
Gender	Gender of the student
Location	Home City of the student
Quiz-1	Score of the student in Quiz-1
Quiz-2	Score of the student in Quiz-2
Quiz-3	Score of the student in Quiz-3
Quiz-4	Score of the student in Quiz-4
Major-1	Score of the student in Major-1
Major-2	Score of the student in Major-2
Major-3	Score of the student in Major-3
Final	Score of the student in the final exam.

Do the following tasks using data given in **StudentScoresDataA** and Table-1:

B-1: **Given Data.** Read and display the data. Identify the number of rows and columns. Does any column have missing data? Display the description of both numeric and non-numeric columns.

B-2: **Type Consistency.** For each column in **StudentScoresDataA**, identify type of each field and verify that each column in Python is identified correctly. If there is any inconsistency, then resolve it.

B-3: **Normalization.** For each score column in **StudentScoresDataA**, apply the standard scaler, such that the mean is zero and standard deviation is one.

B-4: **Visualization.** Draw pairwise scatter plots each pair of score columns in **StudentScoresDataA**. Also, for each score column draw KDE (Kernel Density Estimation). Differentiate the pairwise plots and KDE plot by 'Gender' column.

B-5: **Correlation Analysis.** Do the following:

Calculate the correlation between all the score columns of **StudentScoresDataA**.

Identify top 3 variables that are highly correlated with 'Final' score column.

Which pair of score columns are strongly correlated?

B-6: **PCA.** Do the following:

Get first two principal components of the data without considering 'Gender', 'Location' and 'Final' columns.

Add the two principal components to the data frame and rename the components 'PC1' and 'PC2' respectively.

Construct a scatter plot using the first two principal components of the data. Can the principal components separate 'Final' variable?

Differentiate the above plot using 'Gender'. In a separate plot, differentiate the above plot using 'Location'.

How much variation do each principal component capture?

What are the coefficients (the *u* vector) of the linear combination of input variables for the first PC?

Problem #B

Consider data given in CSV file **StudentScoresDataB** . Consider the following data description:

Table 2: Data Description

Field	Description
Quiz-1	Score of the student in Quiz-1
Quiz-2	Score of the student in Quiz-2
Quiz-3	Score of the student in Quiz-3
Quiz-4	Score of the student in Quiz-4
Major-1	Score of the student in Major-1
Major-2	Score of the student in Major-2
Major-3	Score of the student in Major-3
Final	Score of the student in the final exam.

Do the following tasks using data given in **StudentScoresDataB** and Table-2:

B-1: **Given Data.** Read and display the data. Identify the number of rows and columns. Does any column have missing data? Display the statistical summaries of all the columns.

B-2: **Type Consistency.** For each column in **StudentScoresDataB**, identify the type for each field based on value. Also, identify the datatypes in Python. Report and resolve any inconsistency.

B-3: **Normalization.** For each column in **StudentScoresDataB**, apply the standard scaler, such that the mean is zero and standard deviation is one. Display the summaries of all the columns.

B-4: **Cross Normalization.** For each column in **StudentScoresDataC**, apply the standard scaler fitted (learned) from **StudentScoresDataB** data. Display the summaries of all the columns in **StudentScoresDataC** data.

B-5: **OLS Regression.** The hypothesis is that the quiz and major exam scores are linearly related to final exam score. Use the following formula to calculate the OLS coefficient estimates of all **StudentScoresDataB** data. Take column 'Final' as the output column, and all other columns as input column.

$$\boldsymbol{\theta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

B-6: **OLS Regression.** The hypothesis is that the quiz and major exam scores are linearly related tofinal exam score. Do the following:

Use the sklearn library to calculate the OLS coefficient estimates of all **StudentScoresDataB** data. Take column 'Final' asthe output column, and all other columns as input column.

Compare the coefficients obtained in Part *B*-5 with the above coefficients. Report any differences in between the coefficients from Parts *B*-5 and *B*-6.

Using the above OLS coefficient estimates, calculate the MSE for data given in **StudentScoresDataC**.

B-7: **Ridge Regression.** It may be possible that the quiz and major exam scores are not really independent. Thus, the coefficients needs regularization (penalization). Do the following:

Do the ridge analysis, taking all **StudentScoresDataB** data as the training data. Use 10-fold cross validation, and pick thebest value of alpha from $10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}$

Using the above coefficient estimates, calculate the MSE for data given in **StudentScoresDataC**.

B-8: **Lasso Regression.** It may be possible that not all the quiz and major exam scores are helpful in predicting final score. Thus, the coefficients needs selection (penalization). Do the following:

Do the lasso analysis, taking all **StudentScoresDataB** data as the training data. Use 10-fold cross validation, and pick thebest value of alpha from $10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}$

Using the above coefficient estimates, calculate the MSE for data given in **StudentScoresDataC**.

B-9: **Regression Analysis.** Compare and contrast the coefficient estimates obtained from Parts *B*-6, *B*-7, *B*-8 and *B*-9.