Problem # A

Consider the data given in CSV file **Consumer Data** obtained from a public repository. Consider the following data description:

Table 1: Data Description

Field	Description
Channel	The mechanism through which the goods were consumed. Contains two values: Hotels or Retail.
City Town	The City/Town from which the data is collected.
Fresh	Annual spending (in SAR) on the fresh products.
Milk	Annual consumption (in liter) of milk products.
Grocery	Annual spending (in SAR) on the grocery products.
Frozen	Annual spending (in SAR) on the frozen products.
Detergents_Paper	Annual spending (in SAR) on the detergents and paper products.
Delicassen	Annual spending (in SAR) on the delicatessen products.

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

A-1: **Given Data.** Read and display the data. Identify the fields of the data and count the number of rows and columns.

A-2: **Type Consistency.** Identify the type for each field based on value. Report any inconsistency.

A-3: **Filter noise.** Any record whose channel value is neither "Retail" nor "Hotels" should be removed.

A-4: **Data Wrangling/Munging.** Columns "Fresh", "Milk", "Grocery", "Frozen", "Detergents_Paper" and "Delicassen" should be numeric. Resolve the inconsistency. The value in the "Milk" column is in liters and should be converted to SAR. The conversion can be done by multiplying values in the "Milk" column by 9.8. Round the values in the milk column to the nearest integer.

A-5: **Handling NaN values.** All missing values in "Channel" field corresponds to "Hotels". Similarly, all the missing values in "City Town" fields corresponds to "Khobar". The missing values in "Detergents Paper" field should be replaced by mean, rounded to the nearest integer.

A-6: **Encoding.** Pick field "Channel", and relabel "Hotels" as 1, and "Retail" as 0.

A-7: **Feature Generation.** Create a new filed, called "Region". The values in region should be as follows:

Region value for "Riyadh", "Qaseem" or "Hail" should be "Central".
Region value for "Tabuk", "Makkah" or "Madinah" should be "Western".
Region value for "Khobar", "Dammam" or "Dhahran" should be "Eastern".

A-8: **One-Hot-Encoding.** Do One-Hot-Encoding for "Region" column. Do not delete the original column.

A-9: **Standardization.** For all the following columns, do standard scalarization, such that the mean value is 0, and the standard deviation is 1: "Fresh", "Milk", "Grocery", "Frozen", "Detergents_Paper" and "Delicassen"

A-10: **Clean & Prepared Data.** Display the modified data and the default summary statistics for all the columns.

Problem #B

Consider data given in CSV file **Student Data** obtained from a public repository[2]. Consider the following data description:

Table 2: Data Description

Field	Description
gender	Gender of the student
race/ethnicity	The ethnicity the student belongs to (masked)
parental level of education	The highest level of education of either of the parent of the student
lunch	Whether the student pays for lunch at standard or free/reduced rate
test preparation course	Student took and completed the preparation course before the test or not
math score	Score student received in the math assessment (out of 30)
reading score	Score student received in the reading assessment (out of 25)
writing score	Score student received in the writing assessment (out of 40)
gk score	Score student received in the general knowledge (GK) assessment (out of 100)

Do the following tasks using the data given in **Student Data** 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 description of both numeric and non-numeric columns.

B-2: **Type Consistency.** For each column, identify type of each field and verify that each column in Python is identified correctly.

B-3: **Inconsistent Data.** Looking at the data, two types of inconsistencies were discovered, some of the scores were entered with negative sign (by mistake). Or in some cases, we found larger values than the possible maximum score. For all such entries, assume the score is out of 100, and scaleit out of corresponding maximum score.

B-4: **Handling NaN values.** For any missing data in the score columns, take the average score of the student from other exam scores and replace the NaN value.

B-5: **Handling NaN values.** For any missing data in the non-numeric columns, or replace the NaN value as follows: NaN value is to be replaced by the mode based on the race/ethnicity.

B-6: **Label Encoding.** Convert "parental level of education" and lunch using label encoder.

B-7: **One-Hot-Encoding**. For the 'test preparation course', convert it using one-hot-encoding.

B-8: **Normalization.** To be able to compare scores, scale all the scores between [0,1].

B-9: **Clean & Prepared Data.** Display the modified data, and the summary statistics for all columns