**Description of Dataset and Problem Statement**

This dataset contains data of several patients in the US refilling different medications from pharmacies. Each row in data is a patient who refills a particular medicine. The description of different columns is detailed below.

**SEX**: Contains the gender of patient- 1 means male and 2 means female.

**Agegrp**: contains the age-group to which the patient belongs. 0-17 means age-group 1 and there are a total of 5 different age-groups in increasing age range from 1 to 5.

**REGION**:The region contains the area where the patient belongs. For example, 3 depicts the South region. There are a total of 5 region codes.

**Refill\_count**: depicts the count of the medicine the patient bought (This attribute can be used as dependent variable to predict using regression).

**ADMTYP**: contains the category of a patient, whether the patient belongs to the surgery department, maternity department, etc. These departments have been coded.

Then, there are several columns related to diagnostic and procedure codes. So, if a patient underwent a particular procedure, then there is value 1 in the cell otherwise it is 0 in the cell.

**Medicine**: column depicts 3 medicines. Patient can consume medicine 1, 2, or 3.

--------------------------------------------------------------------------------------------------------------------

Perform the data cleaning if required. Perform the descriptive analytics to understand the data and infer from the data.

Perform the predictive analytics (regressive analysis) on predicting the **Refill\_count** using different regression techniques and compare the results. Perform the data pre-processing (normalization, standardization, correlation analysis & feature section, dimension reduction using PCA) and compare the results of regression with unprocessed data.

Consider 70% of data for training and remaining 30% of data for testing.

Infer the results obtained from descriptive and predictive analytics.