## Assignment 1 - Data Preprocessing

The Health Facts database (Cerner Corporation, Kansas City, MO), includes a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. Health Facts is a voluntary program offered to organizations which use the Cerner Electronic Health Record System. The database contains data systematically collected from participating institutions electronic medical records.  All data were deidentified in compliance with the Health Insurance Portability and Accountability Act of 1996 before being provided to the investigators. Continuity of patient encounters within the same health system (HER system) is preserved. The Health Facts data for this assignment was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States. The dataset for the assignment includes encounters that satisfy the following criteria:

(1) It is an inpatient encounter (a hospital admission).

(2) It is a “diabetic” encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.

(3) The length of stay was at least 1 day and at most 14 days.

(4) Laboratory tests were performed during the encounter.

(5) Medications were administered during the encounter.

Your job is to clean the data and make it ready for developing predictive analytics. Look for missing values, outliers and recode the categorical data to make it ready for analysis. Follow the instructions below and report every step that you take in details and explain your results.

**Instructions:**

1. Download the data and meta data attached.
2. Declare (?) as missing values for all attributes.
3. Eliminate the attributes that include extensive number of missing values.
4. Eliminate the examples (rows) that include missing values.
5. Normalize the data for “num\_lab\_procedures”, “num\_medications”, and “num\_procedures”.
6. Filter examples and only keep the rows where the normalized “num\_medications” values equal 3 or less.
7. Recode A1CResult into Dummy variables.
8. Report descriptive statistics for age, gender, race, length-of-stay and number-inpatient. Make sure to use appropriate measures for qualitative and quantitative variables.
9. Create a Bar chart that represent number of patients with A1CResuts>8 by race.
10. Create Pie chart that presents the prevalence of readmission by gender.
11. Create a histogram of length-of-stay. Limit the number of bins to 7.

**Deliverables:**

1. Your submission should include a MS Word file that includes detailed explanation of all the ten steps listed above that you took to clean the data and detailed explanation of the results. Upload your report in Canvas.
2. Upload your**RapidMiner Processes** or**R Code**for all steps.