Fourth Industrial Revolution (4IR) Summer School
Module 2 – Day 3

Part 2 - Data Preprocessing

Task1:

Utilizing the functionalities provided by the Pandas and NumPy libraries, write a Python program to accomplish the following tasks on the given "MobileApps" dataset:

A. Manage the missing values as follows:
   o   Show a summary of the missing values in each column.
   o   Remove any record that has at least 3 missing attributes.
   o   Fix the values in the "Installs" column to display only the number, removing the "+" sign.
   o   Fill in the remaining missing values such that the categorical variables are filled with the mode of their respective columns, and the numerical values are filled with the mean of their respective columns.

Task2:

Utilizing the functionalities provided by the Pandas and NumPy libraries, write a Python program to accomplish the following tasks on the given "concrete" dataset:

A. Use the Use IQR (Interquartile range) to detect the outliers of the dataset.
   o   Show the count of outliers detected in each column.
   o   Replace the outliers with the means of the corresponding columns.
B. Use the Z-score to normalize the whole dataset.
C. Save the output of the different items in one Excel file using appropriate sheet titles.