Name: Jiahao Yu
Email: jiahao.yu.2003@outlook.com
Country: United Kingdom
College/Company: University of Bristol
Specialization: Data Science

Data understanding:

1. What type of data you have got for analysis:
   An excel file consisting of strings (object) and integers (int64).

2. What are the problems in the data:
   a. NA values:
      i. Most columns of data are complete.
      ii. For Race, Ethnicity, Region, there are slight missing of data.
      iii. For Ntm_Speciality, there are moderate missing of data.
      iv. For Ntm_Speciality_Bucket, Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score, Change_Risk_Segment, there are significant missing of data.
   b. Outliers:
      i. Dexa_Freq_During_Rx and Count_Of_Risks exist outliers.
   c. Skewness:
      i. Dexa_Freq_During_Rx is skewed to the right.
      ii. Count_Of_Risks is skewed to the right.

3. What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why:
   a. NA values:
      Drop the NA values, because if we substitute them to other values, our final model accuracy may be affected.
   b. Outliers:
      We can detect outliers based on Chebyshev's theorem. Drop outliers as they may affect the accuracy of the model.
   c. Skewness:
      i. Approach 1: Duplicate the data to double the size of data, then perform training. Since all the data are valid. If we duplicate the data, it may subsequently make the model more accurate.
      ii. Approach 2: Drop some data to eliminate skewness, but this approach may result in inaccurate predictions.