Data Screening Check List
(Multivariate Analysis)

Name of Data file: World Health Report 2017

1. Accuracy of the Data file – Check the data from Kaggle against original resource ✓

2. Missing Data – Check for missing values and see if they are significant number of missing values and investigate what to do about them. ✓

3. Outliers - Check for both univariate (outliers on one variable alone) and multivariate (outliers on a combination of variables) outliers.
   i. Check if missing value codes are being treated as data entries and outliers. ✓
   ii. If the outlier was not part of the population, remove it.
   iii. The outlier is part of the population you wanted but, in the distribution, it is seen as an extreme case. In this case you have three choices 1) delete the extreme cases or 2) change the outliers' scores so that they are still extreme but they fit within a normal distribution (*for example: make it a unit larger or smaller than last case that fits in the distribution*) 3) if the outliers seem to be part of an overall non-normal distribution than a transformation can be done but first check for normality.

   b. Detecting Outliers
      i. List any dichotomous variables with uneven splits (delete any more than a 90-10 split)
         1. No dichotomous variables ✓

      ii. Among continuous variables –
         1. Univariate outliers

Outlier	How is handled

         2. Multivariate Outliers

Outlier	How is handled

4. Normality –Check for normality of the variables. There are two aspects to normality of a distribution, skewness.
    a. Shapiro test ✓
    b. Fixing Non-Normality – If a variable is not distributed normally then a transformation can be done.

5. Homoscedasticity, Homogeneity of Variance and Homogeneity of Variance-Covariance Matrices
    a. Homoscedasticity – Check for homoscedasticity of normal variables by performing a bivariate scatterplot.
       Heteroscedasticity is corrected by transformations.

Pair of variables to be checked.	Pass / Fail
Economy and freedom	
Generosity and freedom	
Economy and Generosity	

    b. Homogeneity of Variance – This assumption is important when you have grouped data.
        i. Levene's test as a measure of homogeneity of variance. Reaching a significant value ($p < .05$) on the Levene's test means that you have heterogeneity of variance, but the test is very conservative.
    c. Homogeneity of Variance-Covariance Matrices – This is a multivariate assumption that is similar to homogeneity of variance. It roughly states that an entry in a variance-covariance matrix using one DV should be similar to the same entry in a matrix using another DV.

6. Common Data Transformations –1) square root, used when there is moderate skewness/ deviation, 2) logarithm, used when their substantial skewness/ deviation and 3) inverse, used when there is extreme skewness/ deviation. The best approach is to try square root first to see if that fixes your problem, if not then move to logarithm and then to inverse until you have satisfactory normality/ homoscedasticity.