**Data Screening Assignment:**

Download the data screening file to analyze. You would be analyzing this data to see if there is an interaction between gender and other information provided on several variables. You want to screen the whole dataset at once to look for issues, since all the variables will be used in several different hypotheses.

**Study:**

The participant was asked to imagine they were interviewing a person for a job. Participant gender was recorded, and then each participant was randomly assigned to an “information” group, where different résumés were given to each group (see below). Participants were then allowed to talk to the job candidate (a researcher in disguise) for five minutes, and finally, completed several questionnaires. Here are the variable descriptions:

Subject number – a number assigned to each subject, only for data entry use.

Gender – each participant’s gender.

Information – which category each subject was assigned. Information poor participants were only given a few pieces of information about a person (short résumé), while information rich participants were given more information about a person (long résumé).

Likability – each participant rated how likeable a person was given some information and a short time talking to them (scale is in percentages).

Externalatt – Explicit rating of a participant’s attitude assessed by asking the participant how they felt about the person on a 9 point Likert scale.

Internalatt – Implicit rating of a participant’s attitude about a person assessed by the implicit attitudes test (9 point Likert scale).

Ovcompt – a rating scale about how much a participant thought the participant was overcompensating for a flaw on their résumé (0-20 scale).

Slfestm – a rating of each participant’s self esteem (0-10 scale).

Negmood – a rating of how negative/positive a participant was during the fake interview (lower numbers are more positive, runs 0-10 scale).

However, before you start this analysis, you want to screen the data for any entry errors, missing data, and violations of assumptions. You want to check the data for the following:

1. Accuracy
   1. Check the data for out of range scores. Include a SPSS box showing you do/do not have out of range scores.
   2. If necessary, fix the out of range scores.
      1. Describe how you fixed them.
      2. Include a SPSS box showing that you fixed the accuracy issues.
2. Missing data
   1. Include a SPSS box that shows that there are / not missing data.
   2. What type of missing data do you appear to have?
   3. If necessary, “fix” the missing data (remember there are several options).
      1. Describe what you did to the missing data.
      2. Include a SPSS box showing that you fixed the missing data (you may repeat a box you had earlier).

Note: on this section, you will walk through both univariate and multivariate screening options because it helps to practice. We will continue to use data screening on the next assignments, but you’ll just do the multivariate parts like we talked about in class (since they tend to match).

1. Outliers
   1. Univariate:
      1. Create Z-scores for the continuous variables.
      2. What’s the cut off z score for univariate outliers?
      3. Indicate if you have any univariate outliers based on those z-scores by listing the variable and the z-score for the participant. (i.e. Participant 4, likeability z = #).
   2. Multivariate:
      1. Calculate Mahalanobis distance scores. Include the top 5 Mahalanobis scores.
      2. What is your *df* for the cut off score?
      3. What is the cut off score?
      4. How many outliers did you have?
         1. Are those outliers the same people who have crazy z-scores in the univariate section?
      5. If necessary, “fix” the outliers.
         1. What did you do with those outliers (remember there are several options)?
2. Multicollinearity
   1. Include a correlation table of the continuous variables.
   2. Are any of the variables too highly correlated?
3. Normality
   1. Univariate
      1. Include a box with skew and kurtosis values.
      2. What is the cut off score for these values to be non-normal?
      3. Do any of your variables appear to be problematic for normality?
   2. Multivariate
      1. Include the multivariate normality histogram.
      2. Interpret the graph. Does it indicate multivariate normality?
4. Linearity
   1. Univariate:
      1. Include a scattermatrix (look at the graph notes) of the continuous variables.
      2. Do any of the pairwise (two variables at a time) combinations appear to be curvilinear?
   2. Multivariate:
      1. Include the multivariate PP plot.
      2. Interpret the graph. Does it indicate multivariate linearity?
5. Homogeneity/Homoscedasticity
   1. Multivariate:
      1. Include the multivariate residuals plot.
      2. Interpret the graph. Does it indicate homogeneity?
      3. Does it indicate homoscedasticity?

Write up an analysis of what you find in this data, including all the information you answered above. This write up should include the following for credit:

1. APA style for results sections
   1. Double space
   2. Times New Roman 12 point
   3. Two decimals
   4. Centered, bolded Results
2. Short description of the study/variables.
3. Accuracy – did you have problems? What did you do to fix it?
4. Missing data - did you have problems? What did you do to fix it?
5. Outliers - did you have problems? What did you do to fix it?
6. Multicollinearity – did you have any issues?
7. Assumptions:
   1. Normality
   2. Linearity
   3. Homogeneity
   4. Homoscedasticity
   5. On each, did you meet the assumption / have problems?