**Data Screening Assignment:**

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. Include the appropriate output into this document while answering the questions, and include your Excel file in your blackboard submission.

**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:

* 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 on a 9 point Likert scale.
* Ovcompt – a rating scale about how much a participant thought the participant was overcompensating for a flaw on their résumé on a 0-20 scale.
* Selfesteem – an average rating of each participant’s self esteem on a 0-10 scale.
* Negmood – a rating of how negative/positive a participant was during the fake interview where lower numbers are more positive on a 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:

**Accuracy:**

1. Check the data for out of range scores.
   1. Include a summary showing you do/do not have out of range scores. You can use the data analysis table that you create in Excel.
   2. If necessary, fix the out of range scores.
      1. Indicate what the problems were in the dataset.
      2. Make all out of range values NA/missing (do not list them as NA, rather just delete the wrong information from the cell in Excel).
      3. Include a summary showing that you fixed the accuracy issues. You can rerun the data analysis table to show that the scores have been fixed.

**Missing data:**

1. What type of missing data do you appear to have?
   1. Since this data set is small, we can view most of it at once in Excel. Does it appear to be random? Or did everyone skip the same question?
2. Include the percentages of complete data by participant (going across).
   1. Remember, you can use the formula in Excel =counta(highlight cells) / total number of columns \* 100
   2. This dataset is too small to properly estimate for each participant (i.e. any missing data point puts them over the 5% rule). However, for practice, we are going to estimate any participants who have less than 20% missing data.
3. Include the percentages of complete data by column, after excluding participants with too much missing data.
   1. Formula: =counta(highlight cells) / total number of rows \* 100
4. Estimate the missing continuous data.
   1. Include the new percentages of missing data by column.

**Outliers:**

1. Calculate *z* scores for the six categorical variables.
   1. Formula: (click cell – average(highlight column)) / stdev(highlight column). Remember you can drag the formula across and down to get all of the columns and rows.
   2. What is the cut off score for an outlier?
   3. How many outliers did you have?
   4. Exclude participants who have *z* scores over the cut off.

**Additivity:**

1. Include a correlation table of the continuous variables.
2. Are any of the variables too highly correlated?

**For this section, in the video, we created one overall score to compare against our random variable. Given the smaller size of this dataset, try comparing your random variable (Y) to all the continuous variable (X) columns.**

**Normality:**

1. Include the multivariate normality histogram.
   1. I cover in a later video that it might be easier to first create bins for the histogram.
   2. Cut and paste the following into Excel:

|  |
| --- |
| Bins |
| -2 |
| -1.5 |
| -1 |
| -0.5 |
| 0 |
| 0.5 |
| 1 |
| 1.5 |
| 2 |

* 1. Click data > data analysis > histogram.
  2. Highlight the standardized residuals for the input range.
  3. Highlight the new bins for the bin range.
  4. Click chart output and ok.
  5. This procedure will make it easier to interpret because it forces the X-axis to have the -2 to 2 range we would normally expect.

1. Interpret the graph. Does it indicate multivariate normality?

**Linearity:**

1. Include the normality PP plot.
2. Interpret the graph. Does it indicate multivariate linearity?

**Homogeneity/Homoscedasticity:**

1. Include the multivariate residuals plot.
   1. By using multiple variables, you will get multiple graphs. To create one overall graph, highlight the Predicted Y and Standardized Residuals column. It may be easier to cut and paste just those two columns into a separate sheet first.
   2. Click insert > scatter plot.
2. Interpret the graph. Does it indicate homogeneity?
3. Interpret the graph. Does it indicate homoscedasticity?

**Write Up:**

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

1. Result section style (APA and AMA):
   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. Additivity – 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?