**Examining Your Data.**

As previously discussed in the previous chapter there are several types of data and different steps are taken before successfully employing Multivariate Analysis. However, before running we must learn how to walk. In this case, walking refers to learning what our date is trying to tell us. Probably the easiest and most visually best way to start doing so is with a graphical representation of the data. From there the next step is to handle missing data, handling outliers, and finally, transform data adequately.

**Visualizing Data**

Data visualization is a key player in giving meaning to the data obtained. As humans, we are visual creatures and it serves us greatly by grasping the layout of the data. A way to get started is by simply observing the distribution of a category. Using a Histogram to get a deep understanding of the frequency an event is present in a certain category. Another, way to visualize data is to look at variables and their relationship with one another. This type of visualization is best to describe is by using Scatterplot. Interestingly enough, a Scatterplot Matrix could potentially be deployed in a matter that would represent all combinations of variables. Then, there are grouping differences in which categories are compared with another to dissect outliers and changes within the comparison of said groups. Whilst working with these types of visualizations a the very least a deeper understanding of the data should come through and with these knowledge outliers and missing, data should be discovered.

**Missing Data**

Missing data is a natural process in any analytical review. Data points could be missing from an array of reason, ranging bad data entry to corrupted files. Regardless of the reasons an Analyst must understand the processes leading to the missing data in order to select an appropriate course of action. Fortunately, a four-step verification process has been developed.

**Step 1. The type of Missing Data**

Ignorable Missing Data, this type of data can sometimes be forecasted, and a level of uncertainty can be tolerated. For example, if you are driving to a friend’s house with the aid of a GPS system and for some reason, you lose signal momentarily, that missing piece of data is not going to intervene with e the overall route. It is going to be an inconvenience, but not world breaking. Then there is Not Ignorable missing data, this type of missing data could potentially affect the overall layout of the analysis. The missing data could be coming from a failure in procedural factors, bad data entry, and disclosure restrictions. However, there needs to be a deep understanding as to how this data is missing; randomly? Categorically? Etc…

Step 2. Determine the Extent of Missing Data

As mentioned in step one, there needs to be a deep understanding as to how the data is missing; if the cause of the missing can be fixed then it should be addressed. To understand the level of missing data the best way is to assess the extend and patters of the missing data. Typically, by tabulating the percentage of variables missing with each case and the number of cases with missing data for each variable will paint a clear picture. However, let’s keep the following rules of thumb. **Missing data under 10% for an individual case or observation can generally be ignored, except when the missing data occurs in a specific nonrandom fashion. Then, the number of cases with no missing data must be sufficient for the selected analysis technique if replacement values will not be substituted for the missing data**. Lastly, there is an option to ultimately delete the cases/variables and work with a smaller dataset.

**Step 3. Diagnose the Randomness of the Missing Data Proceses**

To diagnose the levels of randomness we must looks at the statistical significance between two groups. If the difference is significant enough that could indicate the data missing is nonrandom. The second approach to diagnose the missing data is by MAR or MCAR. meaning if one category is affected for a variable or the multiple categories are affected by a variable. This could potentially identify isolated incidents.

**Step 4. Select the Imputation Method**

After evaluating the missing data, the Analyst must take action. One decision an analyst could take is Imputation. This process is to estimate the missing value based on valid values of other variables. However, be warned this could be dangerous. Then there is a range approach, a process where a correlation is attributed, and value is giving based on that range. In short, there is a platitude of way to replace the missing values, but as argued before caution is key.

**Outliers**

This section is self-explanatory. Outliers are data points that do not fall within the range of other variables within the category. These could be either extremely low or high points. Because they are crucial in the analysis of data there are several methods one utilizes in order to detect them. The univariate Method examines all metrics variables to identify unique or extreme observations. The Bivariate Methods focus their use on a specific variable relationship such as the independent vs dependent variables. The multivatire methods are best suited for examining a complete variate, such as the independent in regression or the variables in factor analysis.