Reading Questions- Week 3

Mercy Melo (No other collaborators)

**Question 1:** Which of the plot types show every data point?

Scatterplots, Cleveland dotplots, and QQ plots show every data point on the main plot area. Coplots will show all data points in the lower panels of the figure and will show summarized data in the upper panel (conditional data).

**Question 2:** Which of the plot types show aggregated or summarized data?

Histograms and boxplots show only summarized data. Coplots will show summarized data in the upper panel, but will also have raw data in the lower panels.

**Question 3:** Conditional plot, conditioning variable, and related terms occurred throughout the Zuur and McGarigal readings. Explain what a conditional variable means in the context of graphical data exploration.

Conditional variables can be used to specify a secondary interaction within the data for plotting purposes. In a conditional histogram, for instance, a conditional variable can be used to subset the data to create a second (or more) histograms representing data differentially based on its value of the conditional variable. For instance, we can use conditional histograms of a survey to separate out male and female respondents. By using sex as the conditional variable, all survey data from males can be represented in one histogram while survey data from females can be represented in a separate histogram.

Similarly, conditional variables are an integral part of coplots. In coplots, two variables are plotted against each other in scatterplot-type figures that are organized by their value of the conditional variable. Unlike with conditional histograms, the conditional variable for a coplot can be categorical or continuous.

**Question 4:** List *at least three* of the common measures of spread or dispersion that were mentioned in the readings.

Variance, standard deviation, coefficient of variation, and range are all measures of spread.

**Question 5:** Choose *two of the measures* in your list and explain how they capture different aspects of the concept of spread.

Standard deviation signifies spread by measuring the distance of different datapoints from the mean. In normally-distributed data, one standard deviation will encompass about 68% of all datapoints to a specific side of the mean. A second standard deviation (in conjunction with the first) will encompass around 95% of the datapoints on that side of the mean. In this way, a larger standard deviation would indicate that a larger proportion of datapoints are far away from the mean than in a dataset with a lower standard deviation.

Range directly measures the spread of the data by reporting the minimum and maximum values. Although this measure lacks detail on the distribution of data within these limits, it still provides important information about where the outer limits lie.

**Question 6:** List two of the important reasons to perform data exploration (numerical and/or graphical). For each of the two reasons you identify, describe the quantities or plots you would use and the insight you would gain.

To identify bounds of data and outliers, you could use histograms, boxplots, or even scatterplots to gain insight into how to refine measurement techniques in the future. For example, if you found that your highest (in elevation) point that you found a species of songbird at was 400m, you could better cater your next survey to limit data collection to a reasonable cutoff around this value. Identifying the range of data and outliers can also be helpful in setting up/designing further statistical analysis.

To identify any need to transform data, you could use a QQ plot to see how the data strays from a normal distribution. If the data doesn’t fit the normality line very well, you could transform the data and plot it again to see if it fits the normality line better post-transformation. Knowing whether you need to transform data is helpful in preparing data for further analysis.