Matt Fertakos – worked alone

Q1: The plot types that show every data point include scatterplots, Cleveland dotplots, and coplots.

Q2: The plot types that show aggregated or summarized data are histograms, boxplots, and QQ plots.

Q3: A conditional variable is a third variable applied to a plot of two variables to show how the conditional variable interacts with the plotted variables. An example would be plotting subsets of the data in slices on a range of the elevation they were collected at against temperature and windspeed. The conditional variable would be elevation, and then you can compare the relationship between temperature and wind speed at different elevations by looking at the varying plots.

Q4: Three of the common measure of spread or dispersion that were mentioned in the reading include: interquartile range, (absolute) range, variance, and standard deviation.

Q5:
1) Range: Measure the range of values from absolute minimum to absolute maximum. Represents all of the data points.
2)Interquartile range: Measures the range between the 25^{th} and 75^{th} quantiles of the data. Interquartile range differs from absolute range in that the IQR shows the middle 50% of the data when the number of data points is separated into four parts (quarters) with 3 quartiles. The IQR can be used to identify outliers.

Q6: Currently I am working with a dataset of historical nursery catalogs that show the location, species being sold, and date from hundreds of years ago until today. One reason data exploration is important in this dataset is because it allows me to see the spread of the data and how evenly it is distributed across years. This information will inform me if I need to supplement this data with data from other sources to make sure there is coverage for all years leading up to today. For this exploration, I would use a scatter plot of the total range of years represented in the data vs the number of records available for each year. Another reason data exploration is important in this dataset is the number of catalogs available for each species in the dataset. This information is important because there needs to be enough data per species to perform adequate future analyses. To explore this avenue I would use a histogram plotting the number of catalogs for each species on the x axis and the number of species on the y axis. This will show the amount of species with a varying amount of catalog presences. The results of this histogram will tell me if I need to supplement the data I have with more historical data to have a more robust representation of different species, and therefore a more well rounded study.