Student name: David Murillo

I worked alone

September 25, 2022

Q1 (1 pt.): Which of the plot types show every data point?

Answer: Scatter plot and coplot

Q2 (1 pt.): Which of the plot types show aggregated or summarized data?

Answer: Histogram and Boxplot

Q3 (3 pts.): Explain what a conditional variable means in the context of graphical data exploration.

Answer: Is a variable that may obscured the relationship between a dependent and independent variables, the conditional variable could be examine on a scatterplot as a *z*, and, *x* and *y* will represent the independent and dependent variables respectively. For example, in the relation of length and weight of x species, this relationship could be obscured by the sex od the individuals, age, site of sample, etc.

Q4 (1 pt.): List *at least three* of the common measures of spread or dispersion that were mentioned in the readings.

Answer: Variance, Standard deviation, Coefficient of variation, Median absolute deviation, range, interquartile range.

Q5 (2 pts.): Choose *two of the measures* in your list and explain how they capture different aspects of the concept of spread.

Answer: Standard deviation: is the root mean squared deviation from the mean or expected value. The standard deviation has special meaning for normally distributed variables, because the mean ± 1 standard deviation captures approximately 68% of the values, ± 2 standard deviations captures approximately 95% of the values, and ± 3 standard deviation captures more than 99% of the values. Also, the standard deviation is in the same units as the measurement variable, this allows us to interpretable the values on an easier way.

Coefficient of variation: defined as the standard deviation divided by the mean, often multiplied by 100 to express as a percentage, allows us to compare the spread for variables measured on different scales, also is very easy way to see spread.

Q6 (5 pts.): 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.

Answer: 1) Data exploration could reveal aspect of the data that may help me construct a more appropriate environmental model to answer my original question. For example, I am study richness of bird's species on different coffee crop systems, I could made a QQplot to see if the richness of birds has a normal distribution, also if my data have extreme values using a boxplot, I could made a histogram to see what proportion of my data fall in different categories. With the QQplot result, I could implement a lineal model if the data looks a normal distribution, or use other model as a generalized lineal model with Poisson distribution if the data does not look as a normal distribution, this could be support by the histogram, if this shows and decreases of the amount of observation started in the cero category. And the boxplot result I could decide if I eliminate the extreme values to perform my analysis, this will depend if this observation was for error of the type the data or was a real observation.

2) Data exploration allow me to identify conditional variable that could affect my principal variable to study, for example, with the data of richness of birds, I could make a coplot of richness and elevation, with the system as conditional variable. The coplot will help me to identify is the elevation may have a relation with richness, if is the relationship could be different between systems. My research question is to know if the richness of birds differs depending the coffee crop system, but, if the elevation make a effect on the richness of bird?, I have to be sure to have good representation of the elevation for different coffee crop systems.