Grace Okamoto
Data 115: Introduction to Data Analytics
Week #2 Assignment
Fall 2021

1.

a. One advantage .csv files have is that their file size is much smaller than those of .xlsx files.

b. Three approaches for addressing missing data are inspecting the rest of the data to see if two or more entries can be combined, to extrapolate the missing numbers based off the averages of other similar entries, or to simply remove the entry if it's unimportant.

c. According to the lecture slides, a matrix usually only contains integers or floats, doesn't contain labeling information, and can be easily converted into a vector. Conversely, a spreadsheet can contain a wider variety of data, usually contains labels, and is much more difficult to convert into vector form.

2.

	Name	Age	Birthdate	Height	Favorite Snack Food
Data Type	String	Numeric (Int)	Date	Numeric (Int)	String

3.

Name, Age, Number of pets, Number of Siblings, Smoker,
Dave, 31, 0, 3, False
Olaf, 76, NaN, 6, True
Carol, 44, 2, 1, False
Alisa, NaN, 1, NaN, True

4. MCAR, or Missing Completely At Random, is when data is missing without correlation to any of the other variables being studied. MAR, or Missing at Random, is when data is missing with correlation to another variable, and MNAR, or Missing Not at Random, is when missing data is directly caused by a variable.

5. In row 43, the information for Democratic and Republican gubernatorial votes was missing, and I didn't feel comfortable averaging the data from the rest of the sheet as there was a great deal of variety in the numbers provided. Thus, I handled it just as I handle all my life's problems by removing it and pretending it doesn't exist. In rows 48 and 52, I removed the negative numbers, as they appeared to be a data entry error. I verified this by ensuring that the new number of votes for the Democratic and Republican governors added up to the number of voters. After creating the formulas at the bottom of the spreadsheet, I despaired because my numbers were off and I found it necessary to reevaluate the way the Democratic and Republican voter percentages were calculated. I also decided to remove any entries with missing voter data. For row 15, I replaced the percentage with the value greater than 1 with the data that was already on the spreadsheet. I then added the Democratic and Republican percentages together in a new column to ensure that all values totaled 100%, replacing some values with the values provided in the sheet when necessary to meet this criteria. After being satisfied with the numbers provided by

this, I then concluded that 49.93 percent of all voters supported the Democratic gubernatorial candidate.