Problem Set 7

Keep your rendered `.pdf` to no more than 5 pages. Only provide code in the rendered `.pdf` when it is specifically requested.

1. Revisit the data on fruit prices from the USDA: https://www.ers.usda.gov/data-products/fruit-and-vegetable-prices. Again, download the file linked below "ALL FRUITS – Average prices (CSV format)."

(a) Create a new variable called `FruitType` that just describes the specific type of fruit e.g. "Apples," not how it's prepared. Print the code you use to do this to the rendered `.pdf`.

(b) Create a new variable called `Preparation` that describes how the corresponding fruit prepared. Note, there will be some fruits for which this variable is blank. Print the code you use to do this to the rendered `.pdf`.

(c) Which fruit type is is prepared in the most different ways? Print the code you use to do this to the rendered `.pdf`.

(d) Which fruit type is is prepared in the most different ways? Print the code you use to determine this to the rendered `.pdf`.

(e) Create a table with 7 rows and 1 column, where each row corresponds to a different type of preparation (including no information on preparation) and each column corresponds to the number of fruits prepared in that way using `kable`. Make sure that the table is self contained and readable. Print the code you use to create this table to the rendered `.pdf`.

2. Again, find the dataset summarizing some Massachusetts employment statistics called `CESReport.csv`. It was downloaded from here: https://lmi.dua.eol.mass.gov/LMI/CurrentEmploymentStatistics

Start by reading this data into `R` using `read.csv`, skipping the first six lines. I want you to obtain a data frame with 172 observations and 15 variables (I know that that includes some stuff we might want to get rid of later, that will be part of this exercise).

(a) Using functions we've learned about for manipulating strings, modify the `January` variable values to remove commas and asterisks. Print the code you use to do this to the rendered `.pdf`.

(b) Expanding on your code in (a), write a for loop over the columns that correspond to names of months that removes all commas and asterisks and then converts the values to numeric.

(c) Summarize the month variable values in a table with 12 rows, one for each month, and 4 columns, one for the number of non-missing observations, one for the mean of the non-missing observations, and one for the standard deviation of the non-missing observations. Create this table using `kable`. Make sure that the table is self contained and readable. Print the code you use obtain the table to the rendered `.pdf`.

(d) Your table in (c) will show that different months are missing different amounts of data. In at most two sentences and based on examining which rows of the data are missing, explain the missing data.

3. Once more, find the dataset summarizing some Massachusetts employment statistics called `CESReport.csv`. It was downloaded from here: https://lmi.dua.eol.mass.gov/LMI/CurrentEmploymentStatistics

This time, read in this data into `R` using `read.csv`, skipping the first six lines and setting `stringsAsFactors = TRUE`. Again, obtain a data frame with 172 observations and 15 variables.

In at most two sentences, explain what is returned if you convert the `Year` variable to numeric after reading the data in in this way. What types of values does it take on? Are they what you would expect? If they're not what you'd expect, explain the discrepancy.