Problem Set 6

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

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

(a) What character separates distinct fields in this file?

(b) Print the code you use to read in the unaltered file using `read.csv` to the rendered `.pdf`.

(c) How many rows does this dataset contain? Print the code you use determine this to the rendered `.pdf`.

(d) How many columns does this contain? Print the code you use determine this to the rendered `.pdf`.

(e) What variables are contained in this dataset? Print the code you use determine this to the rendered `.pdf`.

(f) What fruit has the highest retail price? Print the code you use determine this to the rendered `.pdf`.

(g) What fruit has the highest price, normalized for quantity (this is the `CupEquivalentPrice` variable)? Print the code you use determine this to the rendered `.pdf`.

(h) How many fresh fruits are described in this dataset? Print the code you use determine this to the rendered `.pdf`.

(i) What happens if you remove all quotes from the file and then read in the file using `read.csv`. Explain in at most one sentence.

2. On my teaching page, you can find a dataset summarizing some Massachusetts employment statistics called `CESReport.csv`. It was downloaded from here: https://lmi.dua.eol.mass.gov/LMI/CurrentEmploymentStatistics

(a) Read this data into `R` using `read.csv`. You want to obtain a data frame with 172 observations and 15 variables. Print the code you use determine this to the rendered `.pdf`.

(b) What are the modes of the variables in this data?

(c) What happens if you apply `as.numeric` to the `Year` variable? Explain in at most sentence.

(d) What happens if you apply `as.numeric` to the `January` variable? Explain in at most sentence.

3. For this problem, we'll keep working with the baseball database that we have used in class.

(a) Create a table with 1 rows and 3 columns, where each column corresponds to one of the following datasets, "Salaries", "Master", "Batting", and each element describes the number of observations in the corresponding dataset using `kable`. Make sure that the table is self contained and readable. Print the code you use obtain the number of observations per dataset to the rendered `.pdf`.

(b) Using `dbGetQuery`, read in all observations for all variables in the "Batting" data. Print the code you use to the rendered `.pdf`.

(c) Using `dbGetQuery`, read in all observations for all variables in the "Batting" data. Print the code you use to the rendered `.pdf`.

(d) Using `dbGetQuery`, read in all observations for the `playerID`, `yearID`, and `HR` variables in the "Batting" data. Print the code you use to the rendered `.pdf`.

(e) Now we are going to start conducting some timing comparisons. Recall that in `arcovariance.R` file on my teaching page we used the `Sys.time()` function to time different operations in `R`. We will use it again here. Record the time it takes to perform the tasks in (b) and (c). Print the code you use to the rendered `.pdf`.

(f) Provide your results from (d) and explain why they do or do not make sense, in at most one sentence.

(g) Using `dbGetQuery`, read in all observations for the `playerID`, `yearID`, and `HR` variables in the "Batting" data. **Then** subset the data to just observations with year 2000 or later. Print the code you use to the rendered `.pdf`.

(h) Using `dbGetQuery`, read in observations during and after 2000 for the `playerID`, `yearID`, and `HR` variables in the "Batting" data. Print the code you use to the rendered `.pdf`.

(i) Record the time it takes to perform the tasks in (f) and (g). Print the code you use to the rendered `.pdf`.

(j) Provide your results from (h) and explain why they do or do not make sense, in at most one sentence.