CSE 536 Fall 2020: Takehome 3

Instructions:

- Due date: Tuesday, December 8. Send an email from your UofL account by noon (EST) with your answers as an attachment.
- The following questions ask you to write some R code. Return an *R script* with all your answers. An R script is a plain text file (not Word) containing only R code; anything else should be commented out. This file can be produced by using RStudio, typing your R code in the Console window, and then
  - saying "yes" when you quit and the system asks you whether you want to save your workspace image;
  - using 'Save As' in the File menu to save what you did; or
  - using cut and paste to put your commands on a plain text file.

  When giving your file a name, use your username and extension '.R'. For instance, if your username is Jjones01, call the file 'Jjones01.R'. NOTE: if you are unsure of whether you produced an R script, you should be able to open your file in RStudio (menu File → Open File) and then press the 'Run' button. If your file is a correct R script, it should run without problems.
- Use only the R commands that we have seen in class. Do not write your own functions or use libraries we have not shown in class (none of that should be necessary at all!).
- The exam has two sets of questions. The second one is more open-ended; it may ask you to solve some issues without specifying how. You can use both R and command line operations, as you see fit; but *document for me everything you do* so I can repeat it in my computer.
- The points for each set of questions is indicated. OPTIONAL questions are worth half the penalty of regular ones if answered incorrectly, and twice the points if answered correctly.

If you have any doubts or questions, please do not hesitate to contact me!

1. (70 points) Install the package "nycflights13". This will give you the data frame called "flights" which contains the NY city 2013 flights data.

   Use R to express the following questions. Note that you are not asked to provide an answer, only how you would write the question in R.

   (a) Are there missing values in `tailnum`? If so, how many?

   (b) We know the origin airport is one of LGA, JFK, EWR. Find out what percentage of all flights originates on each airport.

   (c) What is the average delay for each origin airport? (OPTIONAL: run an ANOVA test to see if the differences between LGA and JFK are significant).

   (d) Among all flights originating at EWR, what percentage go to IAH?

   (e) On each row, attribute/variable `dep_delay` is supposed to be the difference between attributes/variables `dep_time` and `sched_dep_time`. Check that this is indeed the case.

   (f) Find the carriers that do not fly out of one or more of the origin airports (HINT: create a contingency table).

   (g) Is there a relation between the origin airport and the departure delay? Visualize the relation between the two to answer this question (hint: you can make the origin airport a factor).

   (h) Create a plot to check if some airlines are better than others at timely departures.

(i) (OPTIONAL) Try a linear regression to predict arrival delay; use at most 3 factors. How good are your results? NOTE: you can include factors among your independent variables/predictors by creating dummy variables.

2. (30 points) Download the dataset "animals.csv" from Blackboard into your computer. For this part, you can use R or the command line; if you use the command line, just comment out all parts that are not in R so I can see what you did and still run R on RStudio.

   (a) Load the data into a data frame. Show the command that you use.

   (b) Take a look at the data. Note that there are missing values. However, missing values may appear differently in different attributes, depending on how you loaded the data (hint: take a look at non-numeric attributes like `sex` and numeric attributes like `hindfoot_length` or `weight`). Fix this and any other attributes so that all missing values are explicitly marked (NOTE: you can manipulate the data file on the command line and reload the file; just document what you did in a comment).

   (c) How many species (`species_id`) are there in this dataset? How are they distributed? You can plot or use any other approach to describe the attribute.

   (d) Make `sex`, `genus` and `species` be factors. Use `addNA()` function if necessary to make sure NAs are their own level. Determine which one of the three has more influence on weight.

   (e) Determine whether there is some correlation between when hind foot length and weight. You can do this visually or numerically but give a written assessment.