Stat 311 Homework 1

This assignment has a couple of problems related to the material in Lesson 1 and will also introduce you to `rmarkdown`, a package that allows you to embed R code and writing into a document. We will use a specific format created by OpenIntro in conjunction with `rmarkdown` to create HTML file output that you will upload to Canvas.

Since this will be your first assignment that requires R/RStudio, you will need to install some packages before you begin the assignment. Please see the GettingStartedWithR video that is linked as part of the homework assignment.

The main steps that are outlined in the video include:

1. Create a folder for HW1.
2. Download the HW1Template.Rmd file and place a copy in your HW1 folder.
3. Open RStudio (be sure to open R Studio and not R).
4. Set your working directory to the HW1 folder.
5. Install the rmarkdown package (see the video for how to do this)[1].
6. Install the tidyverse, openintro and ggplot2 packages as well following the same process.
7. Open the HW1Template.Rmd file and do a Save as something like HW1_FirstNameLastName.Rmd. We ask you to append your name to any .Rmd files so if you email them to us for help debugging, it is easy for us to associate files with students.
8. You are now ready to begin the assignment. Watch the video to see how to attach packages to a current R session, and how to insert both text and code into the assignment.
9. When you are done with your assignment, run the spell checker in R (**go to Edit > Check Spelling…**) and then knit to HTML (see the video for instructions on knitting and what your output will look like).

[1] Note: you only install packages one time unless you reinstall R. Thus, steps 5 and 6 above will only need to be done for HW1. Attaching packages with the `library( )` function, however, must be done every time you open a new R session.

Complete the problems below. Problems 1 – 5 do not require any code. Simply type your answers into the .Rmd file. You can use #### a), etc. to add sublevels within a problem. Problems 6 – 7 require some very basic R to get you started and familiar with how to add R code within a markdown file.

To reinforce the concepts in the Lesson 1 lectures and to practice a few R commands, I recommend that you try some of the OpenIntro tutorials that are listed on on page 59 of the IMS text (Section 3.2 in the book). I have included the links here as well. The first four links have more lecture content with just a bit of R. The last link is oriented towards R.

https://openintro.shinyapps.io/ims-01-data-01

https://openintro.shinyapps.io/ims-01-data-02

https://openintro.shinyapps.io/ims-01-data-03

https://openintro.shinyapps.io/ims-01-data-04

https://www.openintro.org/go?id=ims-r-lab-intro-to-r --the end of this lab has links to three cheatsheets that you may find helpful with working in R/RStudio.

Stat 311 Homework 1

Stat 311 Homework 1

1. The following table summarizes data from the JSR Launch Vehicle Database, 2019 Feb 10 Edition, and shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success of failure).

	1957 - 1999		**2000 - 2018**	
	Failure	**Success**	**Failure**	**Success**
Private	13	295	10	562
State	281	3751	33	711
Startup	0	0	5	65

   a) What variables were collected on each launch to create the summary table shown above?
   b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.
   c) Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

2. In a study of three nationally representative large-scale datasets from Ireland, the United States, and the United Kingdom (n = 17,247), teenagers between the ages of 12 to 15 were asked to keep a diary of their screen time and answer questions about how they felt or acted. The answers to these questions were then used to compute a psychological well-being score. Additional data were collected and included in the analysis, such as each child's sex and age, and on the mother's education, ethnicity, psychological distress, and employment. The study concluded that there is little clear-cut evidence that screen time decreases adolescent well-being (Orben and Baukney-Przybylski, 2018).

   a) What type of study is this? Briefly explain.
   b) Identify the explanatory variables.
   c) Identify the response variable.
   d) Comment on whether the results of the study can be generalized to the population, and why.
   e) Comment on whether the results of the study can be used to establish causal relationships.

3. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context.

   a) Randomly sample 200 households from the city.
   b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.
   c) Divide the city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.
   d) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.
   e) Sample the 200 households closest to the city council offices.

Stat 311 Homework 1

4. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group (Bucciol and Piovesan, 2011).

   a) Identify the population of interest and the sample in this study.
   b) Comment on whether the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

5. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

   a) Percentage of all videos on YouTube that are cat videos.
   b) 2%.
   c) A video in your sample.
   d) Whether a video is a cat video.

6. **Using R as a calculator**

   a) Use the $ notation shown in the video to write out the equations for the circumference and area of a circle. Remember that $C = 2\pi r$ and $A = \pi r^2$. To get Greek letters to show in an equation, use `\pi` for example. To get a multiplication sign in the formula use `\times`.

   b) Write code to calculate the circumferences and areas of circles with radii of 0.5 and 2 cm. If you want, you can create a variable such as `rad <- c(0.5,2)` to save both radii in a single object. Then you can use rad in your calculations to get two circumferences and two areas in one line of code each. [e.g, `C <- 2 * pi * rad` will produce the object C that has two values, one for rad = 0.5 and one for rad = 2]

   c) Summarize the results of part (b). Remember to use complete sentences and to **include units when appropriate**. **Also, round your answers—do not just copy/paste numeric output given in R. This applies to all problems that have computer output.**

Stat 311 Homework 1

7. **Reading spreadsheet data into R**

   Data files we use in this class will be provided as .csv files. We have provided example code that uses the **read_csv( )** function to read the patient data (Patient_Data.csv) into R, creating a tibble. We also included code to turn the categorical variable Sex into a factor variable (factor variables will be important when we build models), count the number of observations for each Sex and to create a new subset of data that only contains the observations for females. Lastly, the glimpse function is used to display information about the subset.

   Modify the code we provided to do the following:

   a) Use the `count` function to count the number of observations in each of the MartitalStat categories. What percentage of patients are widowed? [Use R as a calculator to get the percentage]
   b) Create a subset of data that includes all observations for patients that are married. Include a call to `glimpse` to show the summary of the data.
   c) Of the married patients, what percentage are male? [Add a line of code to get this number]