Assignment 01

Introduction to Statistical Computing

The goal of this assignment is to give you experience working with the R statistical computing environment. Submit your responses to each of the questions below in a printed document. All graphics should be resized so that they do not take up more room than necessary and also should have an appropriate caption. This assignment is worth 10 points.

Preparation: Install Packages

If you have not already installed them, you will need to install the **dplyr** and **ggplot2** packages to complete this assignment. Once these have been installed successfully, you should not need to install them again. Remember: Install once; load every session.

Preparation: Script File

Open a new script file. Save the script file as `Assignment-01.R`. Save all of the R syntax you use to answer the questions on this assignment in this script file.

Denote each question in the script file using comments. For example,

```
################################################
### Question 1
################################################

<< syntax >>
```

Add comments throughout your syntax as liberally as you feel is necessary to help you recall what the syntax does in the future. Although you do *not need to submit this in with your assignment*, it will be useful for building good coding habits and potentially for future assignments.

Part I

In 2019, Andy read 96 books. The number of pages Andy read each month is reported in Table 1.

Table 1
Number of Pages Read per Month

Month	Pages
January	2819
February	1737
March	2989
April	2026
May	2707
June	2487
July	2927
August	2459
September	2058
October	2573
November	2429
December	4655

Use Excel (or some other program) to enter these data into a spreadsheet. The first column you should name `month` and the second should be named `pages`. The data entered should have 13 rows (including the variable names) and two columns. Save this spreadsheet as a CSV file. Then import the data into RStudio into an object called `reading`.

1. Use the `sum()` function to find the total number of pages Andy read in 2019. Report this value.
2. Use the `geom_density()` function from the **ggplot2** package to create a density plot of the marginal distribution of `pages`. You can see many examples of how to do this here. Be sure the plot has appropriate labels and has a caption. Include this plot in a word-processed document. Resize the plot so it does not take up any more space than necessary.
3. Use the `mean()` function to compute the mean number of pages Andy read per month in 2019. Report this value.
4. Use the `sd()` function to compute the standard deviation of number of pages Andy read in 2019. Report this value.

Part II

Use RStudio to open the *goodreads.csv* dataset and assign it into an object called `andy_books`. This file contains data on all the books Andy has read since late 2010 (see the data codebook). Use the data to answer the following questions.

5. Use **dplyr** to select only the books on the "read" bookshelf; these are the books that Andy actually finished reading. Assign these books into a new object and count the number of rows in this object. Report this value along with the dplyr syntax you used to obtain this value. Change the font of your syntax to a mono-spaced font. (Here is a list of mono-spaced fonts.)

6. Using the data frame object that only includes the books Andy finished reading, compute the following three summaries: (a) the total number of pages read each month; (b) the average number of pages read each month; and (c) the standard deviation of the number of pages read for each month. (Hint: Group by month and then use summarize to make your computations.) Report these values in a word-processed table.

    - *To format this table:* Examine the structure and formatting of Table 1 at http://zief0002.github.io/epsy-8251/misc/creating-tables/creating-tables.html. Notice that variables are presented in rows and summary statistics are presented in columns. Mimic the format and structure of this table to create a table to present the numerical summary information asked for in this question. Mimic the format and structure of this table to create a table to present the summary information. Finally, make sure the table you create also has an appropriate caption.

7. Which two months did Andy read the fewest pages? Explain why this might be the case. (Hint: Consider the academic schedule.)

Part III

Use RStudio to open the *evaluations.csv* dataset and assign it into an object called `evaluations`. Read through the data codebook so you understand the variables included in the data. Use the data to answer the following questions.

8. Use `ggplot()` to create a scatterplot of the relationship between professors' beauty ratings and their average course evaluation rating. (Put the beauty ratings on the x-axis.) Change the axis labels so that both the x- and y-axis have labels that suitably describe the variables being plotted. (For help on this, read the *Axes* page of the Cookbook for R website.) Finally, add a figure caption that adequately explains your figure (e.g., see the *APA Format: Using Tables and Figures* section at http://www.svsu.edu/writingcenter/apa/). Include this plot in a word-processed document. Resize the plot so it does not take up any more space than necessary.

Part IV

In this section, you will again, work with the data in the `evaluations` object you created in Part III.

9. The variable `tenured` in the data set is a dummy variable indicating whether the professor is tenured or not. Use **dplyr** syntax to create a new variable in the `evaluations` dataset called `tenure_status` that has the levels `Not Tenured` and `Tenured` rather than 0 and 1. (There are many ways to do this. For example, see https://www.gerkelab.com/blog/2018/08/recode-if/.) After you do this, copy-and paste the output from `head(evaluations)` into your word-processed document. Change the font of this output to a mono-spaced font.

10. Use `ggplot()` to again create a scatterplot of the relationship between professors' beauty ratings and their average course evaluation rating. This time, color the observations by tenure status. Change the point colors to some non-default palette of your choice. Also, facet the plot using tenure status. Be sure the plot has appropriate labels (on both axes, and on any legend included), and has a caption. Include this plot in a word-processed document. Resize the plot so it does not take up any more space than necessary.