Assignment 01

Introduction to R, RStudio, dplyr, and ggplot2

This assignment is intended to give you experience working with the R program. Submit your responses to each of the questions below in a printed document and label the sections as indicated below within your 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. Any mathematics/equations also need to be appropriately typeset within the document. This assignment is worth 9 points.

Preparation: Install Packages

Open RStudio and install the following packages, if you have not already installed them:

- dplyr
- ggplot2
- sm

Once these have been installed successfully, you should not need to install them again.

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.

Part I

In 2013, Andy read 40 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	1453
February	422
March	848
April	1679
May	1655
June	1630
July	710
August	557
September	978
October	920
November	647
December	2698

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 2013. Report this value.
2. Use the `sm.density()` function from the **sm** package to create a density plot of the marginal distribution of `pages`. 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 neccesary.
3. Use the `mean()` function to compute the mean number of pages Andy read per month in 2013. Report this value.
4. Use the `sd()` function to compute the standard deviation of number of pages Andy read per month in 2013. Report this value.

Part II

Use RStudio to open the *goodreads.csv* dataset and assign it into an object called `read`. This file contains data from Andy's GoodReads entries. The data consists of 12 variables and 291 observations. The variables are:

- `title`: Book title
- `author`: Primary author of the book
- `my_rating`: Andy's GoodReads rating (on a 5-pt scale)
- `avg_rating`: Average GoodReads rating (on a 5-pt scale)
- `publisher`: Publishing company
- `binding`: Book binding (Harcover or Paperback)
- `pages`: Length of the book (in pages)
- `year_published`: Year the book was published
- `month_read`: Month Andy finished reading the book
- `year_read`: Year Andy finished reading the book
- `bookshelf`: GoodReads bookshelf (to-read; currently-reading; read; quit-reading)

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.

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 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 in the article: Snedker, K. A., Herting, J. R., & Watson, E. (2009). Contextual effects and adolescent substance use: Exploring the role of neighborhoods. *Social Science Quarterly, 90*(5), 1272–1296.
    - 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. Re-create the formatting of Table 1 as closely as you can. Finally, make sure the table you create also has an appropriate caption.

Part III

Use RStudio to open the *beauty.csv* dataset and assign it into an object called `beauty`. Use the data to answer the following questions.

7. Use `ggplot()` to create a scatterplot of the relationship between professors' beauty ratings (`btystdave`) and their average course evaluation rating (`avgeval`). (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 neccesary.

Part IV

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

8. The variable `female` in the data set is a dummy variable indicating the gender of the professoer; 0 = male and 1 = female. Use **dplyr** syntax to create a new variable in the dataset called `sex` that has the levels `Male` and `Female` rather than 0 and 1. (There are many ways to do this. For example, see http://www.theanalysisfactor.com/r-tutorial-recoding-values/.) After you do this, copy-and paste the output from `head(beauty)` into your word-processed document. Change the font of this output to a mono-spaced font. (Here is a list of mono-spaced fonts.)

9. 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 sex Change the point colors to some non-default palette of your choice. Also, facet the plot using sex Finally, add a regression line to the faceted plot by including the layer `geom_smooth(method = "lm", se = FALSE)`. Be sure the plot has appropriate labels (on both axes, and on the legend if you include it), 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 neccesary.