Stat 5810 Statistical Visualization I Fall 2017

Homework Assignment 3 (11/16/2017)

67 Points — Due Thursday 11/30/2017 (via Canvas by 11:59pm)

(i) (19 Points) **Slot Machines:** In this question, you have to work with the *vlt* data set from the *DAAG* package. This data set was collected from three windows of a video lottery terminal playing the game "Double Diamond". There are seven possible symbols that may appear in each window. See the *vlt* help page and `https://ww2.amstat.org/publications/jse/v3n2/datasets.braun.html` for further details.

(a) (1 Point) Load all required R packages to answer this question. Show your R code.

(b) (3 Points) You likely have to do some considerable data processing to get the data into the right format for your graphs. Do all data manipulations in R. You cannot manipulate your data outside of R, e.g., in Excel. The choice of your approach/package for the visualization likely determines which data format you need. It may be a good idea that you first get the data into the right format and then replace the numeric values with the matching names from the help file. Otherwise, you may easily mess up with the factor levels as they may not be in the same order for the three windows. Show your R code and the head of your final data structure.

The following R code may serve as a template how to replace the numeric values in the data file with the matching names from the help file. No need for if–statements or any plyr functions. Check your results afterwards and make sure that the numeric values indeed are replaced correctly! A few simple summary statistics should help.

```
> myColors <- c(3, 0, 0, 1, 3, 3)
> myColors
[1] 3 0 0 1 3 3
> myColorNames <- c("Black", "Red", NA, "Green")
> myColorNames
[1] "Black" "Red"   NA      "Green"
```

```
> myColorNames[myColors + 1]

[1] "Green" "Black" "Black" "Red"   "Green" "Green"

> as.factor(myColorNames[myColors + 1])

[1] Green Black Black Red   Green Green
Levels: Black Green Red

> # possible way how to check for correct tansformation
> summary(as.factor(myColors))

0 1 3
2 1 3

> summary(as.factor(myColorNames[myColors + 1]))

Black Green   Red
    2     3     1
```

(c) (5 Points) Draw equally scaled bar charts to see if the distributions of frequencies are the same for each window. Describe important features. Try at least four different layouts of your bar charts. No need to further refine all of these. Just choose the layout that best answers this question and refine this bar chart by adjusting scales, axis labels, title, etc. Include your R code and the resulting graphs (4 initial layouts and the final version).

(d) (5 Points) Treat prize as a categorical variable with 5 outcomes. Draw at least three different graphs that show the distribution of prizes. No need to further refine all of these. Just choose the graph that best summarizes this distribution and refine this graph as needed. The challenge here is that a large number of the plays did not result in any win. Breaking up the axis with the counts and leaving out the "middle part" is cheating with graphs and will result in 0 points for this question part. Include your R code and the resulting graphs (3 initial graphs and the final version).

(e) (5 Points) Draw graphs of your choice to answer the question whether the variables prize and night are independent. Based on your graph(s), do these two variables seem to be independent? No need for any formal statistical test. Explain your answer. Try at least three different graphs. No need to further refine all of these. Just choose the graph that best answers this question and refine this graph as needed. Include your R code and the resulting graphs (3 initial graphs and the final version).

(ii) (16 Points) **Hair and Eye Colors:** In this question, you have to work with the *HairEyeColor* data set (from baseR). It shows the distribution of hair color, eye color, and sex in 592 statistics students. See the *HairEyeColor* help page and any of the cited references for further details.

(a) (6 Point) Create six different mosaic plots that show all possible layouts for the three variables, using baseR. Also show the standardized residuals, based on the assumption that all three variables are independent. Include your figures and your R code.

(b) (5 Point) Overall, each of your mosaic plots above should show seven colored areas. These may relate to hair color, eye color, or sex. Six of the seven shaded areas come in pairs (i.e., three pairs of two related areas each) and one is a unique combination of the three variables. Optimize (i.e., add labels, etc.) the mosaic plot that best displays the pairs and the unique combination. Pairs should be located next to each other and not in different regions of the plot. There are two (of the six) mosaic plots that meet this condition and could be optimized. You only have to optimize one. Include your resulting figure and your R code.

(c) (5 Point) Describe and explain your mosaic plot from (b) above. What can be seen? Assume that a reader is not familiar with mosaic plots, so you have to start with the basic layout you used. How can we best interpret the three pairs and the unique combination? Isn't there an important lurking variable that is missing from this data set, but that would help to even better explain the observed pattern? Which variable is this — and how could it be used to explain the pattern?

(iii) (10 Points) **Barley Data:** Reconstruct and optimize the final version of the *barley* data dot plot from Section 5.7 (Dot Charts for Univariate Data) in our lecture notes, using *ggplot2*. Make sure that you use the same sorting (of the varieties and of the sites) and colors (for the years) as in our version of this plot that was created via the *lattice* dotplot function. Include your final figure and your R code.

(iv) (22 Points) **Movies Data:** In this question, you have to work with the *movies* data set from the *ggplot2movies* package. We are only interested in the length of the movies. Ignore all other variables. See the *ggplot2movies* help page for further details.

(a) (2 Point) Load all required R packages to answer this question. Show your R code. Do not blindly trust the information on the help page! How many movies are really included in this data set?

(b) (3 Point) Create a default histogram via *ggplot2*. Do not optimize this histogram. Interpret this histogram. Do you believe the information shown in the histogram? Yes or no? How can you check? Explain your answer. Include your figure and your R code.

(c) (4 Points) Create a histogram via *ggplot2* of those movies that are 180 min (3 hours) or less in length. Use a bin width of 1 min, centered at 0 min. Describe and interpret this histogram. What happened? Look carefully – the pattern is not random (*plotly* may help if you don't see it immediately)! Include your figure and your R code.

(d) (4 Points) Overlay a density plot on your previous histogram from (c). Make sure that you do not oversmooth and also not undersmooth (you clearly want to get rid of the pattern described in (c)). Optimize this figure. Include your final figure and your R code.

(e) (9 Points) Create a regular box plot, violin plot, and letter value boxplot of those movies that are 180 min (3 hours) or less in length. Optimize each of these. Choose an R package of your choice for each of these three graphs. Include your final figures and your R code. Does any of these three graphs "match" your previous histogram, i.e., lead to a similar interpretation of the data? Explain your answer (thus, you have to carefully explain/interpret what is visible in each of these three graphs).

General Instructions

(i) Create a single html or pdf document, using R Markdown, Sweave, or knitr. You only have to submit this one document.

(ii) Include a title page that contains your name, your A-number, the number of the assignment, the submission date, and any other relevant information.

(iii) Start your answers to each main question on a new page (continuing with the next part of a question on the same page is fine). Clearly label each question and question part.

(iv) Before you submit your homework, check that you follow all recommendations from Google's R Style Guide (see `https://google.github.io/styleguide/Rguide.xml`). Moreover, make sure that your R code is consistent, i.e., that you use the same type of assignments and the same type of quotes throughout your entire homework.

(v) Give credit to external sources, such as stackoverflow or help pages. Be specific and include the full URL where you found the help (or from which help page you got the information). Consider R code from such sources as "legacy code or third-party code" that does not have to be adjusted to Google's R Style (even though it would be nice, in particular if you only used a brief code segment).

(vi) **Not following the general instructions outlined above will result in point deductions!**

(vii) For general questions related to this homework, please use the corresponding discussion board in Canvas! I will try to reply as quickly as possible. Moreover, if one of you knows an answer, please post it. It is fine to refer to web pages and R commands, but do not provide the exact R command with all required arguments or which of the suggestions from a stackoverflow web page eventually worked for you! This will be the task for each individual student!

(viii) Submit your single html or pdf file via Canvas by the submission deadline. Late submissions will result in point deductions as outlined on the syllabus.