Assignment 2 - Data Visualization. Due September 22, 11:59pm 2019

EPIB607 - Inferential Statistics[a]

[a]Fall 2019, McGill University

This version was compiled on September 3, 2020

The first step in understanding data is to hear what the data say, to "let the data speak for themselves". Numbers speak clearly only when we help them speak by organizing, displaying, and summarizing. In this assignment you will explore how to visualize your data and critique figures from published papers. These papers are from the 'Terms and Concepts' in-class exercise. All questions are to be answered in an R Markdown document using the provided template. You are free to choose any function from any package to complete the assignment. Concise answers will be rewarded. Be brief and to the point. Each question is worth 25 points. There will be a bonus of 5 points if your assignment is reproducible. Label your graphs appropriately with proper titles and axis labels. Please submit both the compiled HTML report and the source file (.Rmd) to myCourses by September 22, 2019, 11:59pm. Both HTML and .Rmd files should be saved as 'IDnumber_LastName_FirstName_EPIB607_A2'.

Histograms | Bar plots | Line plots | ggformula package | mosaic package

Template

The `.Rmd` template for Assignment 2 is available here

The `mosaic` package (optional)

The `mosaic` package provides a consistent and user-friendly interface for descriptive statistics, plots and inference. You may find it useful to complete an interactive tutorial on its plotting functions. (note: this is optional and will not be counted for any marks). First install the following packages:

```
install.packages(c("pacman", "mosaic"), dependencies = TRUE)
pacman::p_install_gh("rstudio/learnr")
```

Then, from RStudio, run the following command which will open a new page in your web broswer:

```
learnr::run_tutorial("introduction", package = "ggformula")
```

An advanced tutorial on customizing plots is available also:

```
learnr::run_tutorial("refining", package = "ggformula")
```

1. Indoor Tanning Among US High School Students

Read the paper by Qin et al. (2018) *State Indoor Tanning Laws and Prevalence of Indoor Tanning Among US High School Students,2009–2015* available here

a) (8 points) Consider Figure 1: What visual cues (or aesthetics) are being used? Briefly describe the main takeaways from Plot (b).
b) (5 points) Do you think the two graphs are clear? Is there anything you would have done differently?
c) (12 points) Consider Table 1: produce a plot that displays the indoor tanning prevalence and confidence intervals among female students by restriction type (no restriction, age restriction and parental permission) between 2009 and 2015 (you have to manually enter the data).

2. Folate Nutrition Status in Mothers of the Boston Birth Cohort, Sample of a US Urban Low-Income Population

a) (5 points) Table 2 shows the folate intake frequncy across different pregnancy stages. Here, we will visualize a part of information shown in this Table. For the `Total Sample (n=7612)` data only, create a summary-level `data.frame` that has 3 columns: (pregnancy) **period**, (folate intake) **frequency** (including 'missing' field), and **count** (of total sample). Alternatively, you can create an individual-level data frame that has 2 columns: (pregnancy) **period** and (folate intake) **frequency**, repeating each combination of the two variables by their corresponding sample counts.

b) (8 points) Display the data by a grouped barplot, where the folate intake frequency is grouped by pregnancy period, and the height of each bar is the corresponding sample counts. Don't forget to add an appropriate title and caption to the plot. What can you interpret from this plot?

c) (12 points) Another way to draw a grouped barplot is to group the pregnancy period by folate intake frequency. Draw this plot and compare it with last plot. What can you interpret from this plot? Briefly describe the main takeaways from each of these two plots.

3. Flint Blood Lead Levels

Lead in the environment is persistent, bio-accumulative, and toxic. Chronic exposure to lead in children is associated with many negative health outcomes even when the Blood Lead Levels (BLLs) are measured as low as 1.0-10.0 µg/dL. An analysis of childhood exposure to lead is described in the article *Blood Lead Levels of Children in Flint, Michigan: 2006-2016*.

a) (2 points) As presented in, Figure 3, is BLL on a continuous or discrete scale? Is Frequency on a continuous or discrete scale?

b) (3 points) Briefly comment on a strength of Figure 3. i.e.: what information jumps out at you when you look at this figure?

c) (3 points) Briefly comment on a weakness of Figure 3. i.e.: what information is presented in the figure, but is more difficult to interpret/see?

d) (17 points: 5 for data entry, 6 for visualization, 3 for advantage, and 3 for disadvantage) Figure 3 data in tabular form is show below. Create a `data.frame` of the data and a new visualization using the data. Explain why your data visualization is better and/or worse than Figure 3. The visualization does not need to be radically different than Figure 3, but it can be if you are feeling especially rad today. The aim is to make a visualization that is *not* a bar chart and to explain at least one advantage and one disadvantage of your new and exciting visualization with respect to Figure 3. You may facet the visualization (e.g. https://cran.r-project.org/web/packages/ggformula/vignettes/ggformula-blog.html#facets) as long as you justify your choice. In the data shown below, the `bll_cat` column is the category of BLL and `cat_freq` is the frequency percentage of the particular BLL category in a given year. All values are taken from Figure 3 and the values for `bll_cat` are approximate.

```
#     year bll_cat cat_freq
#  1  2012    <0.5      2.0
#  2  2013    <0.5      4.5
#  3  2014    <0.5      6.0
#  4  2015    <0.5      4.5
#  5  2016    <0.5      6.0
#  6  2012 0.5-1.0     28.5
#  7  2013 0.5-1.0     37.0
```

```
#  8  2014 0.5-1.0     35.0
#  9  2015 0.5-1.0     35.0
#  10 2016 0.5-1.0     44.5
#  11 2012 1.1-2.0     45.0
#  12 2013 1.1-2.0     38.0
#  13 2014 1.1-2.0     33.0
#  14 2015 1.1-2.0     42.0
#  15 2016 1.1-2.0     36.0
#  16 2012 2.1-3.0     14.0
#  17 2013 2.1-3.0     12.0
#  18 2014 2.1-3.0      9.0
#  19 2015 2.1-3.0     12.5
#  20 2016 2.1-3.0      7.0
#  21 2012 3.1-4.0      5.0
#  22 2013 3.1-4.0      4.5
#  23 2014 3.1-4.0      3.0
#  24 2015 3.1-4.0      3.0
#  25 2016 3.1-4.0      2.5
#  26 2012 4.1-5.0      2.5
#  27 2013 4.1-5.0      2.5
#  28 2014 4.1-5.0      2.0
#  29 2015 4.1-5.0      2.0
#  30 2016 4.1-5.0      1.0
#  31 2012    >5.0      3.0
#  32 2013    >5.0      2.0
#  33 2014    >5.0      3.0
#  34 2015    >5.0      3.5
#  35 2016    >5.0      2.5
```

4. Physicochemical properties and phenolic content of honey from different floral origins and from rural versus urban landscapes

a) (5 points) Comment on figure 2a and explain what is wrong with it.

b) (10 points) Which visual cues are being used in Figure 2b? Re-plot Figure 2b. You can make changes to the aesthetics or reproduce it. You do not need to reproduce the error bars or the sample size (n). Be sure to include labels and a title.

c) (5 points) Which visual cues are being used in Figure 4? Is this an effective visualization? Is there a better way to represent the data? (you may take inspiration from https://serialmentor.com/dataviz/directory-of-visualizations.html or http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#3.%20Ranking)

d) (5 points) List errors in Table 1 and Figure 4 and what could have been improved.