Stat 5810, Section 003
Statistical Visualization I
Fall 2018
Homework 3

ShaunMicheal Bartschi

A01975136

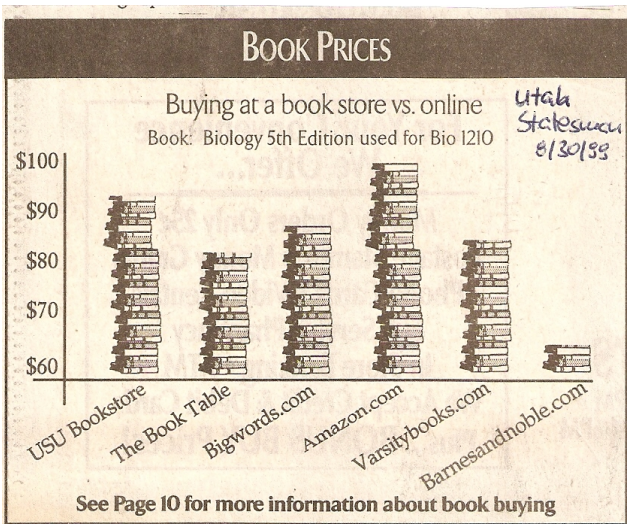
December 9, 2018

Stat 5810 Statistical Visualization I	Fall 2018

Homework Assignment 3 (11/21/2018)

40 Points — Due Sunday 12/9/2018 (via Canvas by 11:59pm)

(i) (10 Points) Carefully look at the graph below:



(a) (3 Points) Explain which rule(s) (how to construct a bad graphic) from our lecture notes the graph designer has followed, i.e., list the rule(s) (number and name) and explain why it has been followed. Do not blindly list rules (numbers and names) as you will lose points if you incorrectly quote a rule. Be sure that you understand the difference between rules 3 and 4.

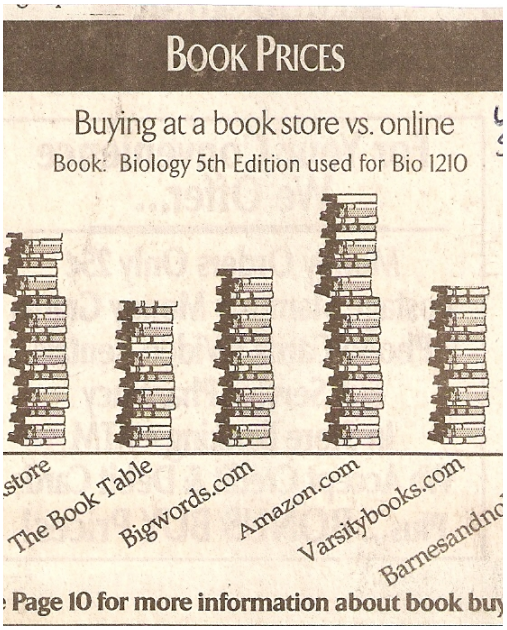
(b) (4 Points) Demonstrate how this poor graph might be improved. Using the data from the graph (or your best approximation if necessary), construct a superior representation of the same information, using R, similar to the improvements from Section 6.2 of our lecture notes. You can use any R package of your choice to create the improved version. Include a scan or a photo of the bad graph in your answer, next to your improved version. Also include your R code.

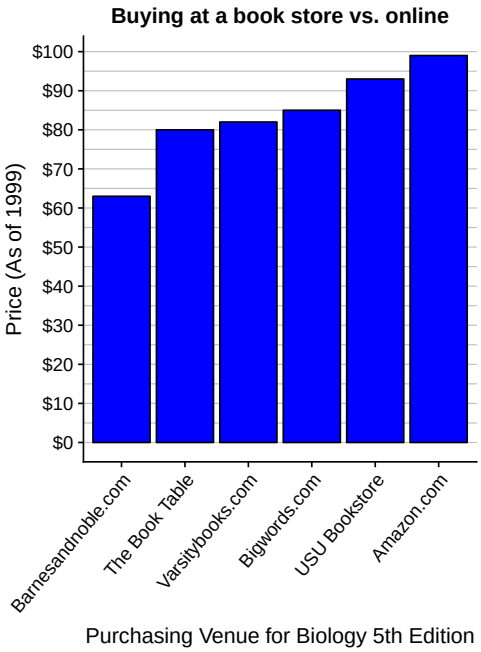
(c) (3 Points) Include a short write–up (about half a page) as to how you believe your version improves on the poor original. More specifically, indicate what you have modified and why this improves the representation of the underlying data.

Rules that it breaks:

- Rule 4 (Only Order Matters) - Choosing to represent the bars with offcentered books leads to the impression that the area may be changing in density.
- Rule 9 (Alabama First) - It appears that the graph is order from brick and mortar stores (further arranged by distance from campus), and then online book retailers (which appear to be unorganized after this point).

```
> setwd("C:/Users/Shaun/Desktop/StatVis/HW3")
> library(ggplot2)
> library(ggthemes)
> library(grid)
> library(gridExtra)
> library(extracat)
> library(lvplot)
> library(lattice)
> library(scales)
> library(jpeg)
> library(cowplot)
> library(magick)
> book <- data.frame("store" = c("USU Bookstore","The Book Table",
+                                "Bigwords.com","Amazon.com",
+                                "Varsitybooks.com", "Barnesandnoble.com"),
+                    "price" = c(93,80,85,99,82,63))
> book$store <- factor(book$store,levels = book$store[order(book$price)])
> original <- ggdraw() + draw_image("hw03_q01_Fig1_books.jpg",scale=1.5)
> improved <- ggplot(book, aes(x=store,y=price))+geom_col(fill = 'blue',
+                                                         color = 'black') +
+   theme(axis.text.x = element_text(angle = 50, hjust = 1),
+         panel.grid.major.y = element_line(color = 'grey'),
+         panel.grid.minor.y = element_line(color = 'grey')) +
+   xlab("Purchasing Venue for Biology 5th Edition") +
+   ylab("Price (As of 1999)") +
+   ggtitle("Buying at a book store vs. online") +
+   scale_y_continuous(labels=dollar_format(), breaks =
+                        c(0,10,20,30,40,50,60,70,80,90,100))
> grid.arrange(original,improved, nrow=1)
```



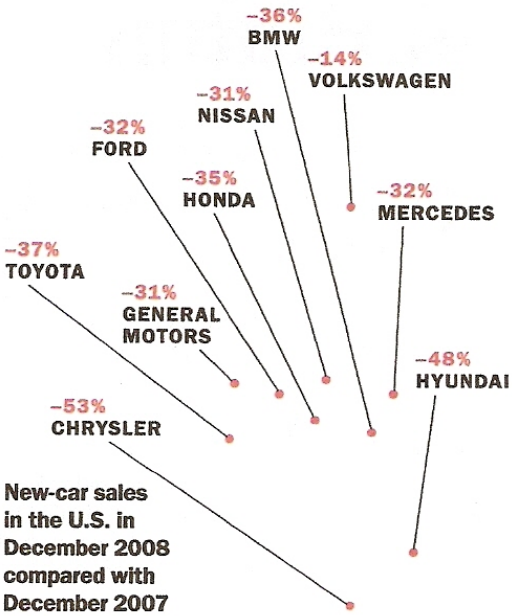


By rescaling the graph and by placing the sellers in accending order, I believe that I have been able to improve upon th original graph.

Now, it is more clear how the various retailers compare in selling prices to one another, the area under the bars is now accurately representative of price, and it is easier to find the relavite price of each of the respective books.

(ii) (10 Points) Carefully look at the graph and text below:

Time, January 19, 2009, p. 15



10 | Detroit

A Bleak Holiday for Carmakers

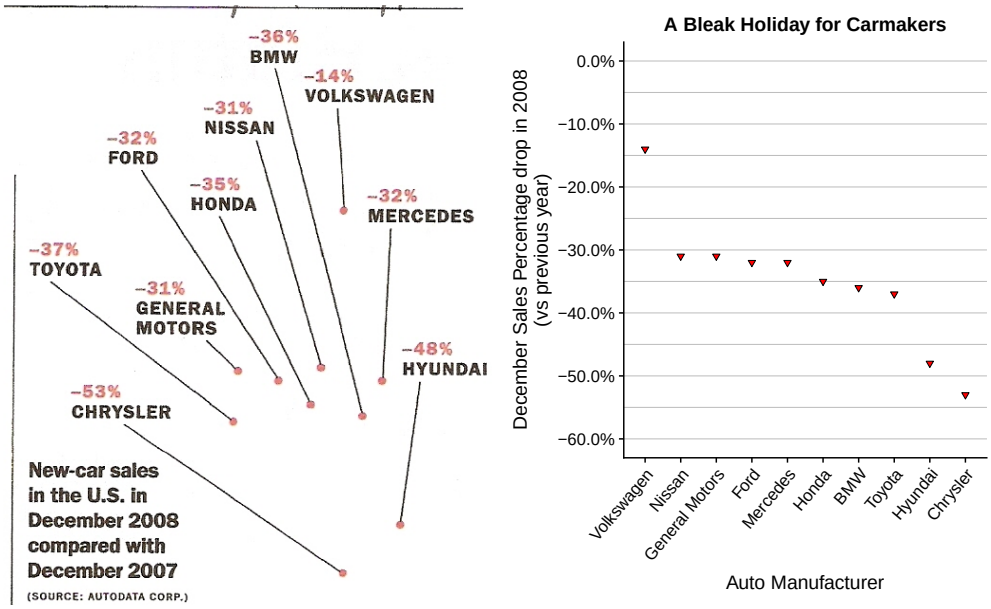
Sales of cars and light trucks in the final month of 2008 fell 36% from the previous December, capping a disastrous year in which U.S. auto sales tumbled 18%. After soaring oil prices over the summer made driving more expensive, the Big Three automakers, caught in the financial crisis and facing slumping demand, were forced to rely on a Washington handout to avoid bankruptcy. In the chilly December market, Chrysler saw sales drop by more than half, and no brands were immune save for Rolls-Royce and Mini, which witnessed only slight upticks. Experts say the dismal figures portend less production and variety in 2009.

Repeat parts (a) through (c) from the previous question for this graph. When constructing your improved version, keep in mind that these numbers are decreases (and not increases)!

Rules that it breaks:

- Rule 5 (Graph Data Out of Context) -
- Rule 9 (Alabama First) -
- Rule 10 (Missing Labels) -
- Rule 12 (Think of a New way to do it) -

```
> original <- ggdraw() + draw_image("p2_clipped.png")
> auto <- data.frame(maker = c("BMW","Volkswagen","Nissan","Ford",
+                              "Honda","Mercedes","Toyota","General Motors",
+                              "Hyundai","Chrysler"),
+                    loss = c(-0.36,-0.14,-0.31,-0.32,-0.35,-0.32,
+                             -0.37,-0.31,-0.48,-0.53))
> auto$maker <- factor(auto$maker,levels = auto$maker[order(-auto$loss)])
> improved <- ggplot(auto, aes(x=maker,y=loss)) +
+   geom_point(color = 'black',fill = 'red', shape=25) +
+   theme(axis.text.x = element_text(angle = 50, hjust = 1),
+         panel.grid.major.y = element_line(color = 'grey'),
+         panel.grid.minor.y = element_line(color = 'grey')) +
+   scale_y_continuous(labels = percent, breaks =
+                        c(0,-0.1,-0.2,-0.3,-0.4,-0.5,-0.6),
+                      limits = c(-0.6,0)) +
+   xlab("Auto Manufacturer") +
+   ylab("December Sales Percentage drop in 2008\n(vs previous year)") +
+   ggtitle("A Bleak Holiday for Carmakers")
> grid.arrange(original,improved, nrow=1)
```



By redoing this graph, is has now been done in such a way that graphical interpretation can actually be done on it. Prior to reploting the data, it was completely unclear what the individual points were meant to represent.

Now, by ordering them in decending order according to losses (as well as using red down facing triangle to represent the loss), the car manufactures losses are now comparable, and the axis are approprately labeled.

(iii) (20 Points) Now you need to find your own bad graph!

(a) (5 Points) Find a bad graph you want to discuss and improve for this homework question. "Claim" your bad graph via a personal announcement to me via e-mail or in Canvas. Be specific which graph you want to improve, in particular if there is more than one bad graph shown. Include the URL for a graph found on the web, the full reference with page number and figure number for a bad graph from a publication, or all necessary details for a bad graph found in a newspaper, in class materials, etc.

Web pages, journal articles, newspapers, magazines, and scholarly books are all appropriate sources. Good sources for bad graphs are CNN, Time magazine, the Utah Statesman, Wikipedia, and many other online sites, but also textbooks and journal papers.

Your bad graph must meet at least one of the rules for bad graphs from Chapter 6 in our lecture notes. I will award extra credit points to all those students who found graphs with the maximum number of "approved" rules for bad graphs (i.e., if 3 students find bad graphs with 5 rules followed, all 3 students will get the same extra credit points, but if 1 student finds a bad graph with 6 rules followed, only that student will get the extra credit points). Note that listing all 12 rules won't help, as I may only "approve" 4 of the rules for your graph (and you may actually lose points for listing too many unapproved rules).

Each student must claim a graph by Tuesday 11/27/2018, 11:59pm. Each student must claim a different graph. If you claim a graph that was previously claimed by someone else, you must find and claim a different graph.

Note that numerous web sites with overviews of bad graphs exist. Some examples are
`https://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6`,
`https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/`,
`https://www.buzzfeednews.com/article/katienotopoulos/graphs-that-lied-to-us`,
and several more. **You can't use any graph that is posted on these or any other bad-graph-collection web sites, books, or other publications. The goal of this HW is NOT to reuse a bad graph that has already been marked as bad by someone else, but rather to identify a (new) bad graph when we see it!**

(b) (5 Points) I will collect all proposed bad graphs into a PowerPoint presentation.

Each bad graph is shown on a single page, including the name of the student who claimed it and the source. **Each student will have a maximum of 2 min to introduce the bad graph in our last lecture on Thursday 11/29/2018**: (i) Mention the source of the graph; (ii) indicate the rule(s) that have been followed to make it a bad graph; and (iii) briefly outline how you are going to improve this graph, e.g., whether you change the type of the graph, modify the layout, etc. You should practice in advance that you don't speak longer than 2 min!

Note: If you are unable to attend class on Thursday 11/29/2018, you have two options: (1) Claim your bad graph by Sunday 11/25/ 2018 and present it in class on Tuesday 11/27/2018. All other requirements remain the same; or (2) Call my office number at 435 797 0696 by Tuesday 11/27/2018. Leave a voice mail, starting with your name, and then provide the information for (i) to (iii) listed in (b) above. I will then play your voice mail in class on Thursday 11/29/2018 as part of the presentations.

***** In any case, you must contact me in advance via e-mail if you are not able to attend class on Thursday 11/29/2018. *****

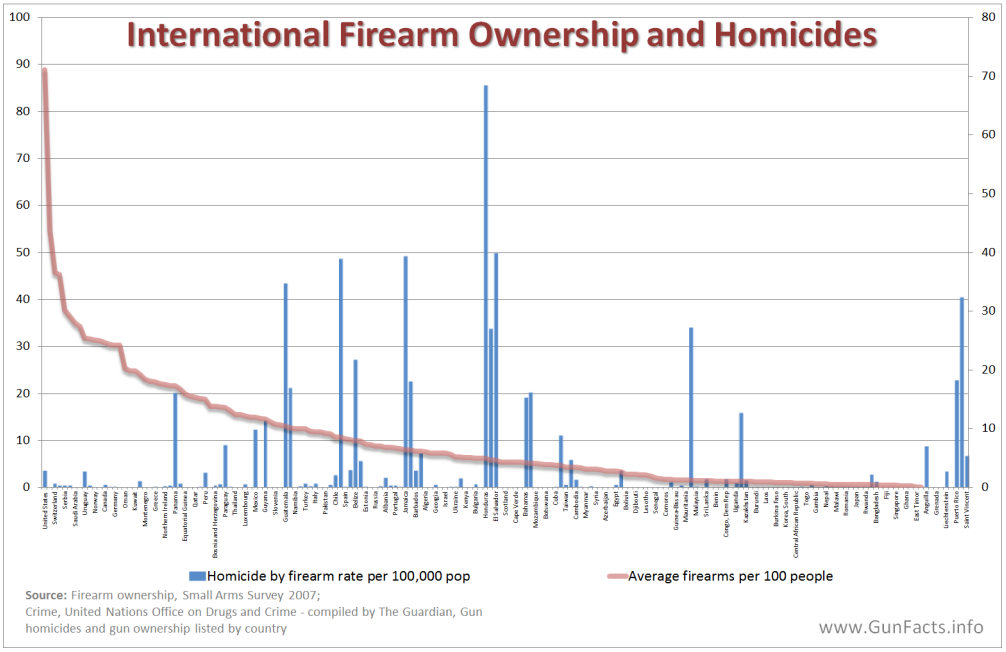
(c) (10 Points) Repeat parts (a) through (c) from the two previous questions for your bad graph. In your discussion, include the exact source of your bad graph, i.e., the information you initially sent to me when you claimed your graph.

Bad Graph Coutesy of `http://www.gunfacts.info/wp-content/uploads/2013/09/GUNS-IN-OTHER-COUNTRIES-Firearm-Ownership-and-Homicides-Rates-per-Country.png`

Data Courtesy of the Guardian

`https://docs.google.com/spreadsheets/d/1chqUZHuY6cXYrRYkuE0uwXisGaYvr7durZHJhpLGycs/edit`

Original Graph:



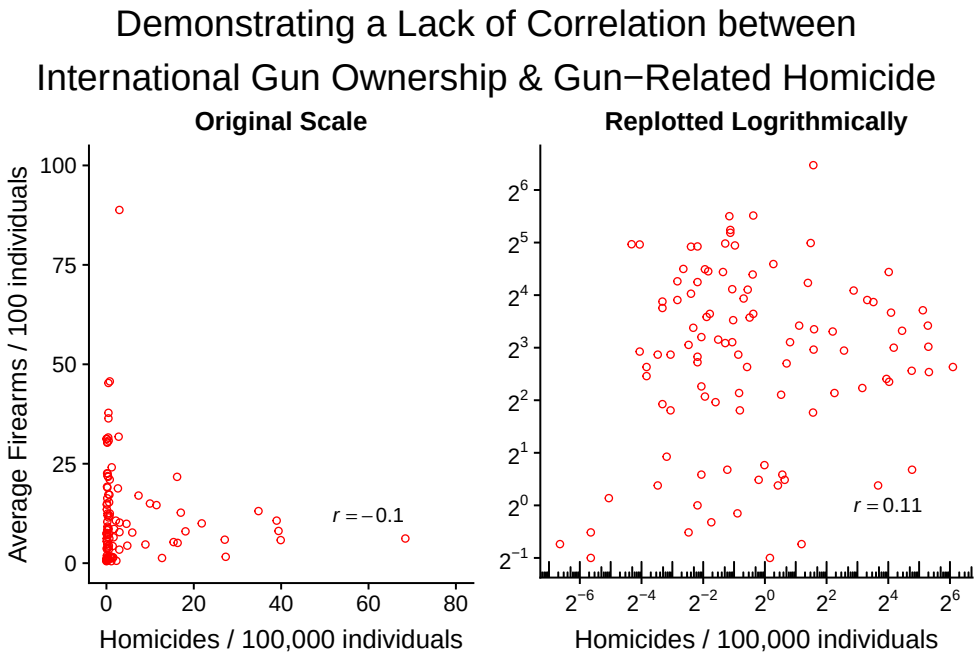
Rules that it breaks:

- Rule 7 (Emphasize the trivial) -
- Rule 10 (Label Incompletely) -
- Rule 11 (More is Murkier) -

Proposed Improved Graph:

```
> dat = read.csv("GunViolence_GunOwnership.csv", header = TRUE, sep=",")
> dat <- dat[!(is.na(dat$Homicide.by.firearm.rate.per.100.000.pop) |
+                is.na(dat$Average.firearms.per.100.people) |
+                dat$Average.firearms.per.100.people == "0" |
+                dat$Homicide.by.firearm.rate.per.100.000.pop == "0"),]
> ownership <- dat$Average.firearms.per.100.people
> homicide <- dat$Homicide.by.firearm.rate.per.100.000.pop
> corr_eqn <- function(x,y, digits = 2) {
+   corr_coef <- round(cor(x, y), digits = digits)
+   paste("italic(r) == ", corr_coef)
+ }
> labels = data.frame(x = 60, y = 12, label = corr_eqn(homicide, ownership))
> loglabels = data.frame(x = 16, y = 1, label = corr_eqn(log(homicide),
```

```
+                                                          log(ownership)))
> p1 <- ggplot(dat, aes(x=homicide, y=ownership)) +
+   geom_point(color = 'red',shape = 1) +
+   geom_text(data = labels, aes(x = x, y = y,
+                          label = label), parse = TRUE) +
+   xlab("Homicides / 100,000 individuals") +
+   ylab("Average Firearms / 100 individuals") +
+   ggtitle("Original Scale") +
+   theme(plot.title = element_text(hjust = 0.5)) +
+   xlim(0,80) + ylim(0,100)
> p2 <- ggplot(dat, aes(x=homicide, y=ownership)) +
+   geom_point(color = 'red', shape = 1) +
+   scale_x_continuous(trans='log2',
+     breaks = c(1/64,1/16,1/4,1,4,16,64),
+     labels = trans_format("log2", math_format(2^.x))) +
+   scale_y_continuous(trans='log2',
+     breaks = c(1/2,1,2,4,8,16,32,64,128),
+     labels = trans_format("log2", math_format(2^.x))) +
+   annotation_logticks(sides="b") +
+   geom_text(data = loglabels, aes(x = x, y = y,
+                          label = label), parse = TRUE) +
+   xlab("Homicides / 100,000 individuals") +
+   ggtitle("Replotted Logrithmically") +
+   theme(plot.title = element_text(hjust = 0.5),axis.title.y =
+           element_blank())
> grid.arrange(p1, p2, nrow = 1,
+              top=textGrob("Demonstrating a Lack of Correlation between\nInternat
+                           gp=gpar(fontsize=20,font=8)))
```



By showing the data as a scatterplot instead (and excluding the Country the data is describing), we are able to more clearly illustrate that there is no correlation between Gun Homicide and Average Gun ownership. And by plotting the data on two different scales, we see that even logrithmically, there is no meaningful relationship between the two variables on an international scale.

Also, this plot is much easier to interprete that meaning. Including the correlation coefficient for each of them helps to illustrate the statistically insignificant relationship.

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.