Data Visualization - Homework 4

Matt Isaac

December 8, 2017

i

```
# Required packages

library(imager)
library(ggplot2)
library(dplyr)
```

a.

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.

The following rules have been followed, resulting in a bad graphic:

1. *Rule 5: Graph data out of context.* (Note: the reasoning below could also be applied to Rule 1: show as little data as possible.)
   There is a lot of data here that isn't represented, but could have been. The price has been shown for one book in various stores. This is only one book for a particular subject, out of hundreds of books on biology and various other subjects. The trends that we can see for this specific book may not be representative of the prices of all text books across the different stores. Books of a certain subject might be cheaper online than books of a different subject. Because the purpose of this graph (as stated on the graphic itself) is to compare "Buying at a book store vs. online", it would be helpful to use data from more than one book. This could have been accomplished in several different ways. First, the creator of the graph could have taken the average of the prices of several comparable biology textbooks sold from each vendor. Or, the creator of the graph could have added the price of a math textbook from each vendor, the price of a psychology textbook from each vendor, etc.

2. *Rule 4: Ignore the visual metaphor.*
   This rule was followed when the creator of the graph decided to begin the vertical axis scale at 60. The visual metaphor of the area of the bars representing the price is lost. This effect could become even worse if the viewer sees the bars as a stack of books (three-dimensional bar). Then there is even a larger misinterpretation if the viewer interprets the volume of the stack of books as being proportional to the price.

3. *Rule 9: Alabama first!*
   The categories in this plot are somewhat sorted: there two stores are first, then the online stores are plotted next. However, within the group of online stores, there is no particular order - not even alphabetical. This aspect of the plot should be changed to order the categories by price.

4. *Rule 10: Label: (b) incompletely.* This rule was followed by not labeling which vendors are traditional stores and which vendors are online stores.

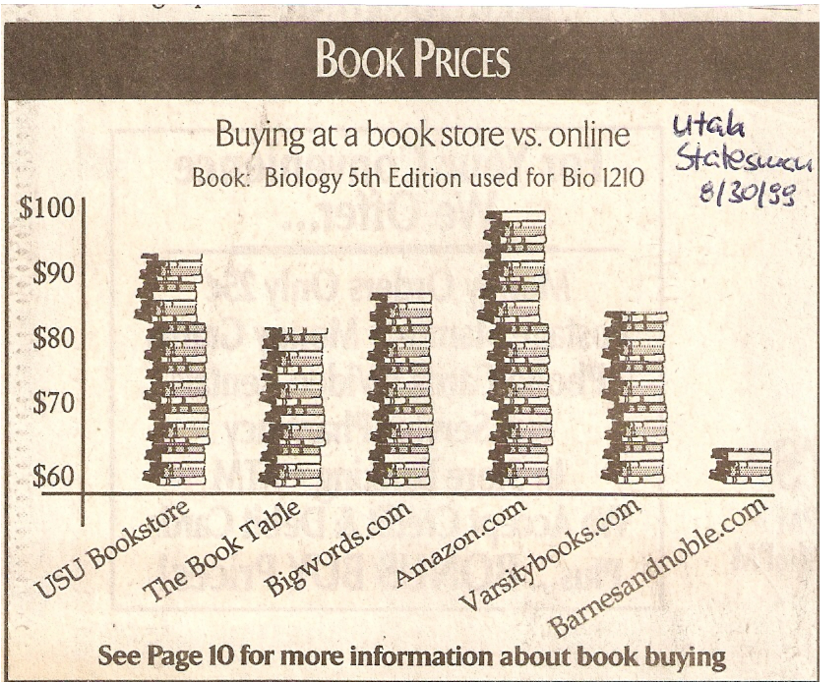
5. *Rule 11: More is murkier.*
   The bars have been displayed as stacks of books. This could be interpreted as an added "dimension" in two ways. First, it is simply added detail that distracts from the interpretation of the plot. Secondly, the stacks of books appear to have three dimensions (shadow of left side). This follows the rule of adding superfluous elements that make the plot "murkier".

b.

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.

The original plot is shown below.

```
im1<-load.image("Fig1_books.jpg")
plot(im1, axes = FALSE, xlab = "original graphic")
```



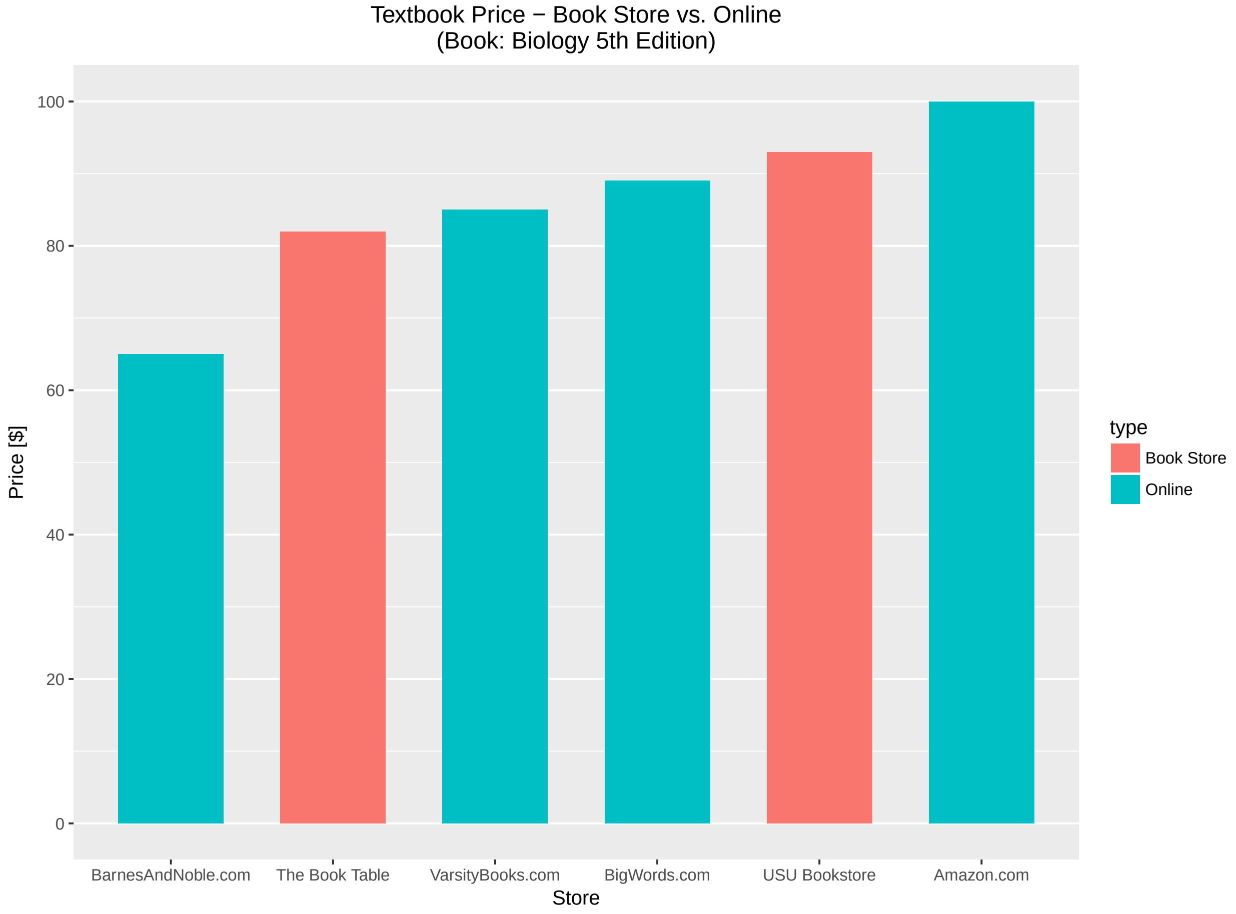
The improved plot is shown below.

```
price <- c(93, 82, 89, 100, 85, 65)
price <- sort(price)
store <- c("BarnesAndNoble.com", "The Book Table", "VarsityBooks.com", "BigWords.com", "USU Bookstore",
type <- c("Online", "Book Store", "Online", "Online", "Book Store", "Online")
books_df <- data.frame(store, price, type)

books_df$store <- factor(books_df$store, levels = c("BarnesAndNoble.com",
                                                    "The Book Table",
                                                    "VarsityBooks.com",
                                                    "BigWords.com",
                                                    "USU Bookstore",
                                                    "Amazon.com"))

ggplot(data = books_df, aes(x = store, y = price, fill = type)) +
  ylim(0, 100) +
  xlab("Store") +
  ylab("Price [$]") +
  theme(panel.grid.major.x = element_blank(), plot.title = element_text(hjust = 0.5)) +
  ggtitle("Textbook Price - Book Store vs. Online\n(Book: Biology 5th Edition)") +
  scale_y_continuous(breaks = seq(0, 100, by = 20)) +
```

```r
  geom_bar(stat = "identity", width = 0.65)
```



c.

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.

I believe that the modified chart displays these data in a more simple and straightforward way that cannot be easily misinterpreted. The first thing that I did was change the title from “Book Prices” to “Book Price” There is only one book being evaluated in these data, and I felt that the title “Book Prices” gave the impression that prices from more than one book were compared. I next made sure that the vertical axis extended all the way down to zero. Because we are using a bar plot, the area of the bars is interpreted as being proportional to the numeric value associated with it. Thus, if the scale doesn’t start at zero, the interpretation of areas will be incorrect. For example, the book as purchased from Amazon.com appears to be about 7 times more expensive than if it is purchased from BarnesAndNoble.com. However, if we look at the actual values of the prices we can see that the book from Amazon is not even twice as expensive as the book purchased from BarnesandNoble.com. Correcting the y-axis as described will eliminate this incorrect interpretation. The labeling of the bars was also incomplete and needed to be addressed. The plot title stated that the purpose of the plot was to compare the prices of text books purchased online to text books purchased in a book store. However the bars associated with online prices and the bars associated with book store prices were not readily distinguishable from each other. In order to remedy this, I color coded the bars by store type and provided a legend. This makes the graph easier to read because the viewer can see the desired comparison at a glance instead of having to read the horizontal axis labels and mentally keep track of which category a particular bar is in. Lastly, I changed the order of the stores along the horizontal axis so they were sorted in a meaningful way. I ordered these by price value so the viewer can easily compare which vendors have higher prices. The color coding helps the viewer understand how store type may affect price. It is easy to see that online stores are not always cheaper and not always more expensive than traditional stores.

ii

a.

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.

The following rules have been followed, resulting in a bad graphic:

1. *Rule 1: Show as little data as possible (minimize the data density).*
   For as much space is used for the graphic, there are only ten data points on the plot. The graphic is amazingly unhelpful, especially for how much space it is taking up.

2. *Rule 3: Ignore the visual metaphor altogether.*
   The creator of this plot did not make it clear what was to be interpreted as having a meaningful data. When I first looked at this plot, I thought that the length of the line was the visual representation of the data. When I realized that that wasn't the case, I thought it might be the slope of the line - that wasn't it either. Finally, I realized that the vertical positioning of the point on the page was representative of the percentage associated with it.

3. *Rule 2: Hide what data you do show.*
   The data from this plot is incredibly hard to understand visually. The numbers on the page are more useful than the points in understanding the data. The lines that are attempting to label the data look too much like part of the data itself. These lines are extremely distracting and good at hiding the trends in the points.

4. *Rule 7: Emphasize the trivial (ignore the important).*
   The creator of this plot listed the exact value percent decrease directly on the plot. In this case, viewers will be more interested in comparing the approximate percent decrease from company to company instead of knowing the exact percent decrease. For example, viewers will be more interested in knowing that Volkswagen had the least percent decrease, and that General Motors and Ford both had approximately the same (30%) percent decrease than knowing that Volkswagen had a percent decrease of exactly 14%, and that General Motors and Ford had 31% and 32% decreases respectively.

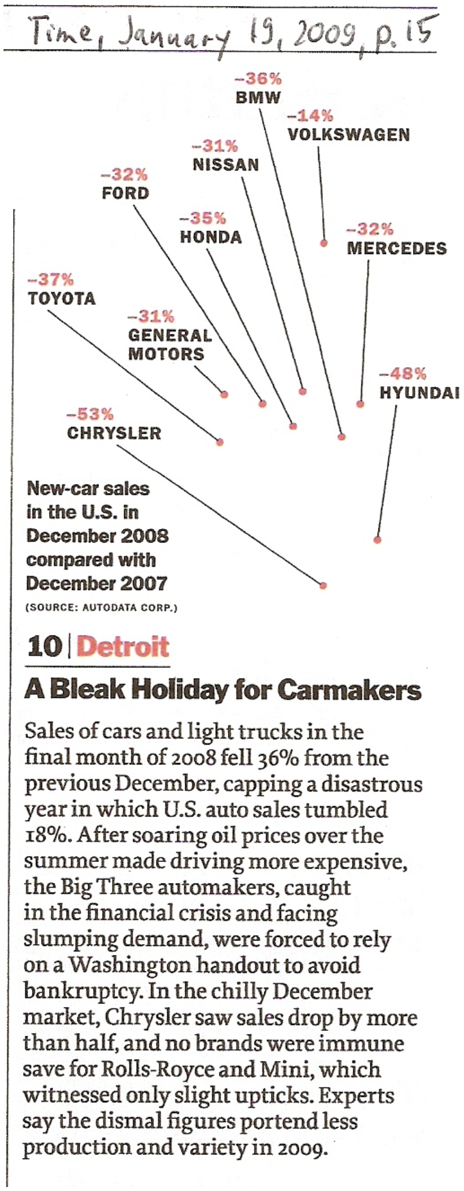
5. *Rule 9: Alamaba first!*
   There no meaningful order to the data - not even alphabetical. Even after studying this plot for quite some time, I could not understand any rhyme or reason as to why the creator of this graph ordered the data like he or she did. This is worse than alphabetic order - at least alphabetic order is some kind of order.

6. *Rule 10: Label (a) illegibly.*
   I know that I have already complained about the lines used to label the data, but I feel like it was one of the worst qualities of this graph. I had very difficult time making sense of this system of labeling and trying to understand what the data was indicating. I think that this is one of the most illegible ways of labeling these data.

b.

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.

```r
im2 <- load.image("Fig2_cars.jpg")
plot(im2, axes = FALSE)
```



Time, January 19, 2009, p. 15

-36%
BMW

-14%
VOLKSWAGEN

-31%
NISSAN

-32%
FORD

-35%
HONDA

-32%
MERCEDES

-37%
TOYOTA

-31%
GENERAL
MOTORS

-48%
HYUNDAI

-53%
CHRYSLER

New-car sales
in the U.S. in
December 2008
compared with
December 2007

(SOURCE: AUTODATA CORP.)

10 | Detroit

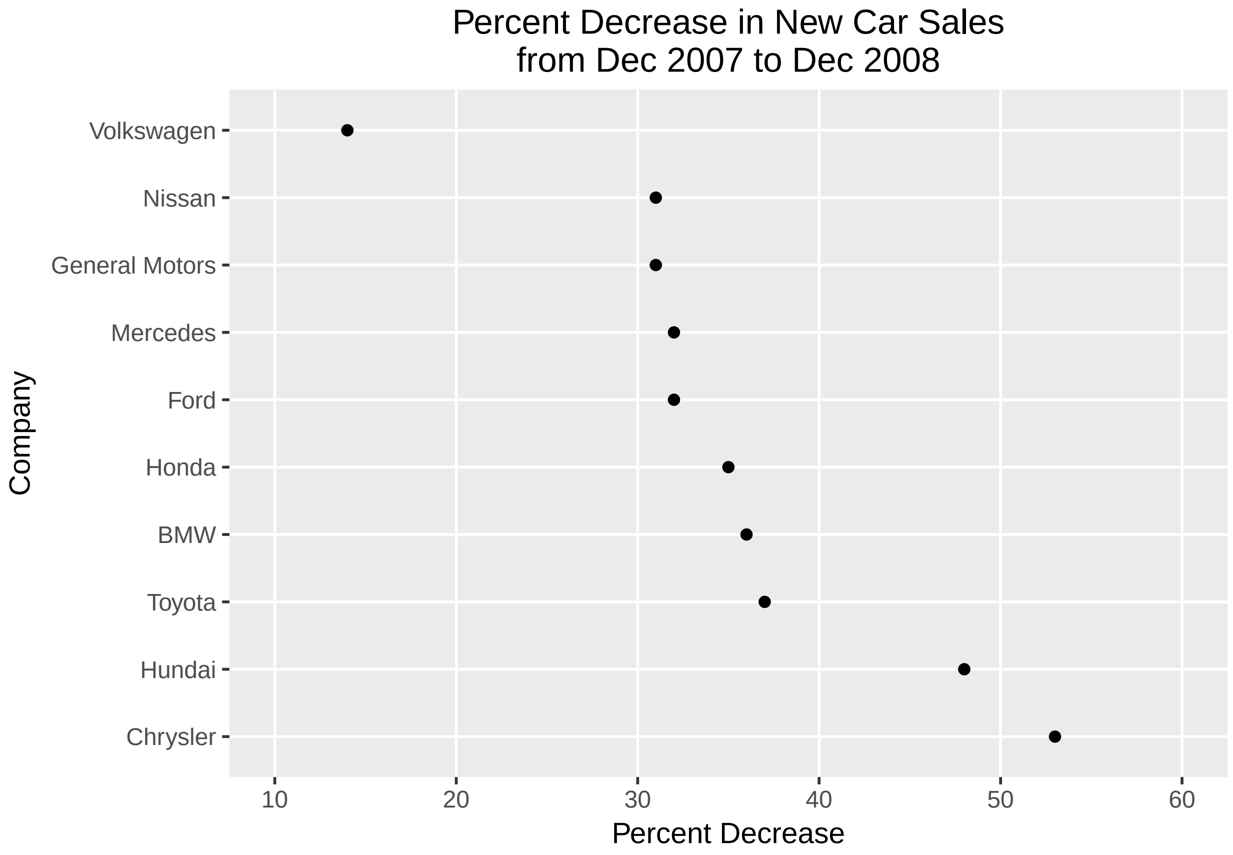
A Bleak Holiday for Carmakers

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.

```
company <- c("Chrysler", "Toyota", "General Motors",
            "Ford", "Honda", "Nissan", "BMW",
            "Volkswagen", "Mercedes", "Hundai")
p_decr <- c(-53, -37, -31, -32, -35, -31, -36, -14, -32, -48)
carSales <- data.frame(company, p_decr)
```

```r
carSales <- mutate(carSales, perc_of_ly = 100 + p_decr)
carSales <- mutate(carSales, p_decr_pos = p_decr * -1)

carSales <- carSales[order(carSales$p_decr),]
carSales$company = factor(carSales$company, levels = c("Chrysler", "Hundai", "Toyota",
                                                       "BMW", "Honda", "Ford", "Mercedes",
                                                       "General Motors", "Nissan",
                                                       "Volkswagen"))
ggplot(data = carSales, aes(x = p_decr_pos, y = company)) +
  xlab("Percent Decrease") +
  ylab("Company") +
  scale_x_continuous(breaks = c(0, 10, 20, 30, 40, 50, 60, 70), limits = c(10, 60)) +
  ggtitle("Percent Decrease in New Car Sales\nfrom Dec 2007 to Dec 2008") +
  theme(panel.grid.minor.x = element_blank(), plot.title = element_text(hjust = 0.5)) +
  geom_point()
```



c.

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.

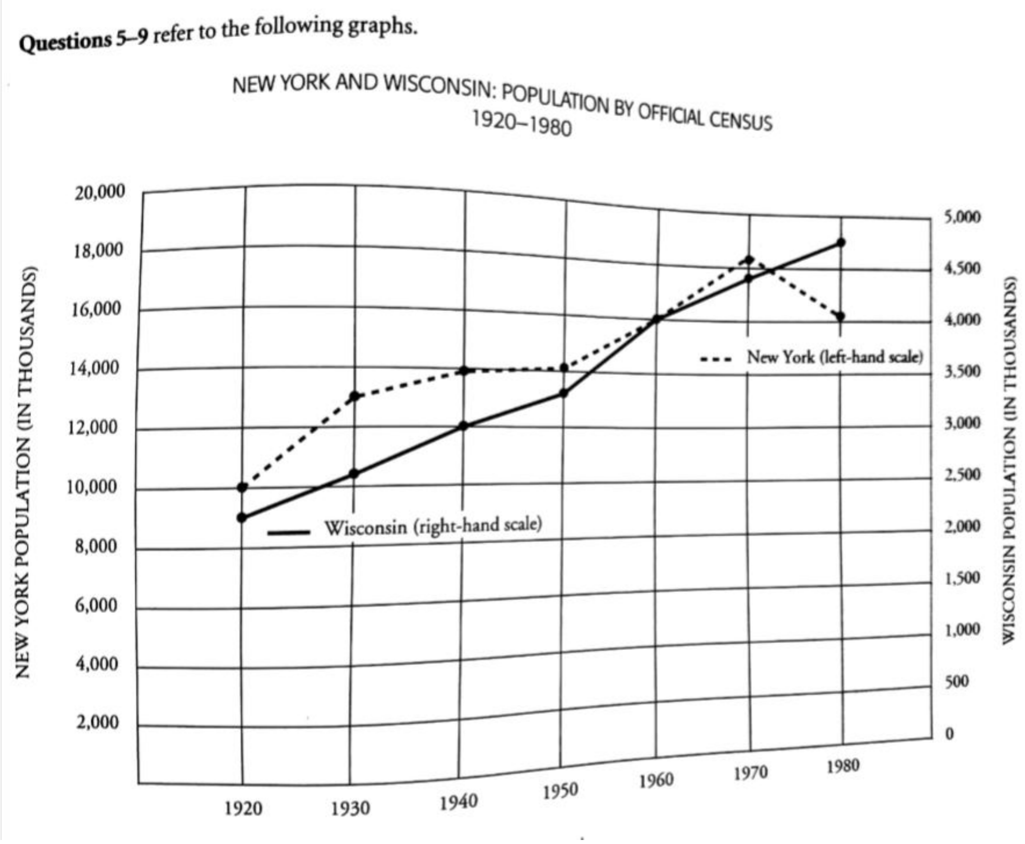
I don't know if this bad graphic uses a type of graph that falls in any specific category of graph. The closest thing that I think that it would be is some sort of dot plot. For this reason, I decided to create a dot plot with these data that would be much easier to understand. I was unable to add any more data, as I don't know where I would get that data. If I were able to, however, it would helpful to add sales for the other eleven months between December 2007 and December 2008. In this case, a line chart could be used and we would have a lot more information about the decrease in car sales. The dot plot mainly changes the labeling system from the original plot to something much more legible and easy to understand. The labels cannot be confused as part of the data itself. Next I sorted these data from highest percent decrease to lowest percent decrease. This is an extremely beneficial change from the original plot. In order to compare sales from different companies on the previous plot, one must find the name and the listed percentage for the companies in question, and then compare these numbers. However, I found this more difficult than it sounds. The lines on the plot were so distracting and the labeling confusing, I had a hard time keeping track of the points that I wanted to compare. On the revised plot, it is easy to find the companies in question on the vertical axis. Because the companies are sorted by percent decrease, one can simply look at the positions of the companies in question on the vertical axis and automatically know which had a greater percent decrease. I decided to include both the horizontal and vertical grid lines. The horizontal grid lines help the viewer quickly match the point with the corresponding company. The vertical grid lines help the viewer to quickly estimate the percent decrease.

iii

a.

I claimed the following graphic as my "bad graph".

```r
im3 <- load.image("Bad_Graph.jpg")
plot(im3, axes = FALSE)
```



Questions 5–9 refer to the following graphs.

b.

Present bad graph: (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 1 min!

c.

Repeat part a. through c. from the two previous questions for your bad graph:

a.

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.

Rules Followed:

1. *Rule 1: Show as little data as possible.*
   Since this is a dot plot, there is no reason to show below 8,000 on the left hand scale (or 2,000 on the right hand scale). This space could be used more effectively for something else - to provide more space for text in a journal article, if nothing else.
2. *Rule 3: Ignore the visual metaphor altogether.*
   This rule has been broken by the way the lines have been positioned on the graph. The visual assumption of this graph is that if a line is higher on the graph then it has a larger value. However, this isn't the case for these plots. The last data point (1980) shows the Wisconsin line above the New York line. However, if we look at the scales, Wisconsin population is only at about 4700, while New York population is at about 16,000 (in thousands). This is not even half as large of a population as New York, but the line ends higher on the graph. Although the lines lie nearly directly on top of each other, they should not be directly compared.
3. *Rule 5: Graph data out of context.*
   Our data are graphed out of context because New York was founded in 1788, and Wisconsin was founded in 1848. There is a lot of data that is not included on the plot for the years that have passed since the founding of these states.
4. *Rule 8: Jiggle the baseline*
   The baseline was "jiggled" by using different vertical scales on the left and right sides. The left hand vertical axis displays the data (in thousands) from 2,000 to 20,000 in steps of 2,000. The right hand vertical axis displays the data (in thousands) from 0 to 5,000, in steps of 500. Although both scales start at zero, the first tick mark on the left side is 2000, while the first tick mark on the right side is 500. As noted in the discussion of Rule 3, this makes it so viewers of the plot cannot directly compare the lines to each other, although they do lie directly on top of each other. It is interesting to note that the top tick mark of the right-hand scale (5,000) barely reaches past the second tick mark on the left-hand scale (4,000). Dual scales are especially appropriate when both scales are plotting a similar unit. In this case, both scales are showing population levels.
5. *Rule 10: Label (a) illegibly.*
   The labeling of the lines is one of the more confusing parts of this graph. The label for the solid line says "Wisconsin (right-hand scale)". However this label is on the *left* side of the graphic. Likewise, the label for the dashed line indicates "New York (left-hand scale)", but is placed on the *right* side of the graphic - right next to the right hand scale. The first time I tried to understand this plot, I found myself wanting to use the scale that was nearest to the label instead of the scale on the opposite side of the plot. A small improvement to this plot would be putting the labels nearer to the scale that should be used.
6. *Rule 11: More is murkier.*
   The problem of two axes has already been discussed above. However, I felt that the extra axis made this plot much "murkier". Thus, this rule has also been followed.

b.

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.

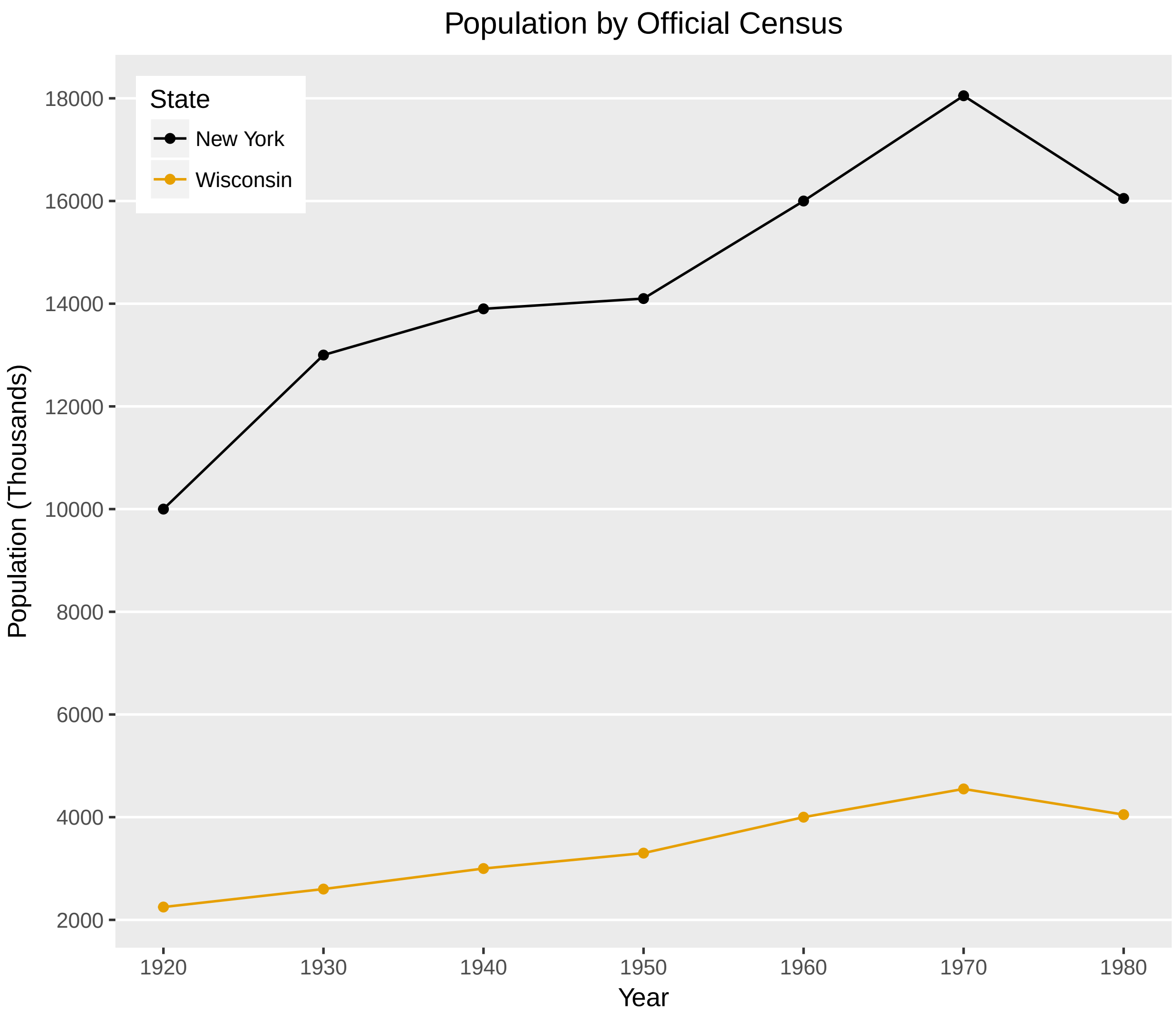
```r
library(ggplot2)
library(graphics)
library(ggthemes)
#colorblind_pal()

dataf <- read.csv("population_data.csv", header = FALSE)

year <- dataf[,1]
pop <- dataf[,2]
State <- dataf[,3]


dataf <- data.frame(year, pop, State)
```

```r
ggplot(dataf, aes(x=year, y=pop, group = State, color = State)) +
  xlab("Year") +
  ylab("Population (Thousands)") +
  ggtitle("Population by Official Census") +
  ylim(0, 20000) +
  scale_y_continuous(breaks = seq(2000, 20000, by = 2000)) +
  scale_x_continuous(breaks = c(1920, 1930, 1940, 1950, 1960, 1970, 1980)) +
  theme(plot.title = element_text(hjust = 0.5), legend.position = c(0.1, 0.9),
        panel.grid.minor.x = element_blank(), panel.grid.major.x = element_blank(),
        panel.grid.minor.y = element_blank()) +
  scale_color_colorblind() +
  geom_line() +
  geom_point()
```



c.

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.

The first change (as well as the most important change) that I made was to plot both lines on the same scale. Plotting them on the same scale was the aspect of the original graph that led to the most confusion. With the lines plotted on the same scale, viewers no longer are confused by the Wisconsin line ending higher than the New York line. Where the old lines crossed over each other several times in the original plot, the revised plot makes it clear that New York always has a larger population than Wisconsin. Removing the second scale also made the labeling more clear. I color coded the lines on the plot and included a legend. I also made sure that the legend listed New York and Wisconsin in the same vertical order as they appear on the graph. This makes reading the graph even easier for people unfamiliar with these data. Lastly, I began the vertical axis at 2000 instead of at 0. With a dot plot, it is unnecessary to include space below the points included on the plot. I also removed all of the horizontal axis grid lines and the minor vertical axis grid lines. These were unnecessary to include in the plot and ony added "noise" to this graphic. Lastly, I included tick marks and labels on the vertical and horizontal axes that provided enough resolution to understand these data adequately.