IMT 573: Problem Set 5

Descriptive Data Analysis

Jenny Skytta

Due: May 01, 2022

Collaborators: Indepedent work

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Download the `05_ps_descdataanalysis.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `05_ps_descdataanalysis.Rmd`.
2. Replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps4_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup: In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(readr)
```

```
library(lubridate)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Problem 1: Comic Books Are Still Made By Men, For Men And About Men In this problem set, we will recreate some of the data analysis, specifically data visualizations used to support the article https://fivethirtyeight.com/features/women-in-comic-books/.

First, use the following code to load the data.

```
# Load the data
urlfile = "https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/dc-wikia-data
dc_comics <- read_csv(url(urlfile))

urlfile = "https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/marvel-wikia-
marvel_comics <- read_csv(url(urlfile))
marvel_comics$YEAR <- marvel_comics$Year
```

(a) Visualization Recreation Next, choose one of the data visualizations in the article to recreate. Be clear which you are aiming to reproduce.

```
#adding column with universe to each df
dc_comics$universe <-  "DC"
marvel_comics$universe <-  "MARV"

#selecting columns to allow for rbind
sex_marv <- marvel_comics %>%
  select(SEX, universe, YEAR)

#selecting columns to allow for rbind
sex_dc <- dc_comics %>%
  select(SEX, universe, YEAR)

#combining selected tibbles
new_fem_char_comb <- rbind(sex_marv, sex_dc) %>%
  filter(YEAR > 1980) %>% #filtering over 1980
  group_by(SEX, YEAR, universe) %>% #grouping
  summarise("sex_per_group" = n()) %>% #total of sex
  ungroup() %>%
  mutate("sex_total" = sum(sex_per_group)) %>% #new column
  mutate("sex_proportion" = (sex_per_group / sex_total) * 100) %>%
  filter(SEX == "Female Characters") %>% #filtering for females
    group_by(YEAR) #grouping by year
```

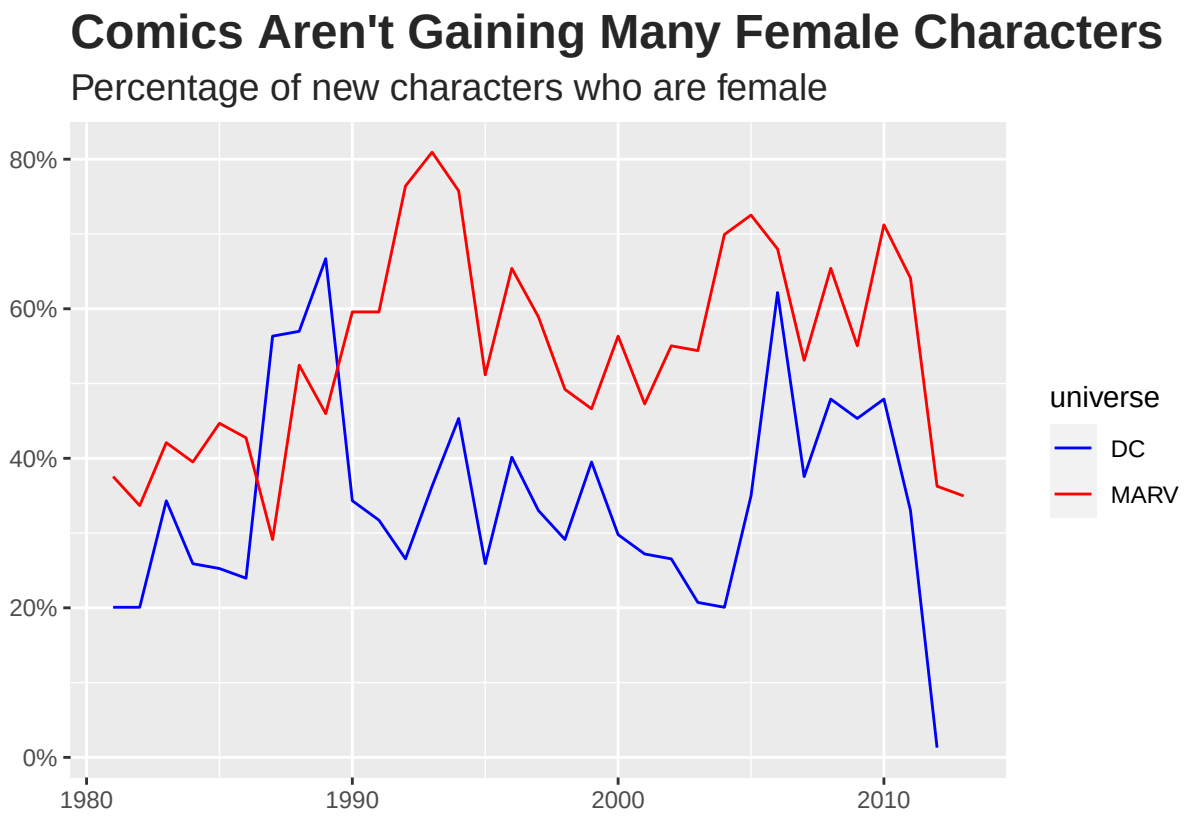
```
## `summarise()` has grouped output by 'SEX', 'YEAR'. You can override using the `.groups` argument.
```

```
# wrapping title
my_title_manual2 <- "Percentage of new characters who are female"

#plot of female characters
ggplot(data = new_fem_char_comb) +
  geom_line(mapping = aes(x = YEAR, y = sex_proportion, fill = universe, color = universe)) +
```

```
    scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
    scale_color_manual(values=c('Blue','Red')) +
    labs(title = "Comics Aren't Gaining Many Female Characters",
       subtitle = my_title_manual2,
       y = "",
       x = "" ) +
  theme(
  plot.title = element_text(color = "grey15", size = 18, face = "bold"),
  plot.subtitle = element_text(color = "grey15", size = 14))
```

```
## Warning: Ignoring unknown aesthetics: fill
```



(b) Reflection After producing your own visualization, comment on your ability to recreate the visual from the article exactly. Are there places where you see discrepancies? Why might this be the case? Do you need any additional information not present in the article to be able to do this?

```
#assessing unique observations in variables
#unique(lgbt$GSM)
#unique(lgbt$SEX)

#adding column with universe to each df
dc_comics$universe <-  "DC"
marvel_comics$universe <-  "MARV"

#selecting columns to allow for rbind
lgbt_marv <- marvel_comics %>%
  select(SEX, GSM, universe, YEAR)
```

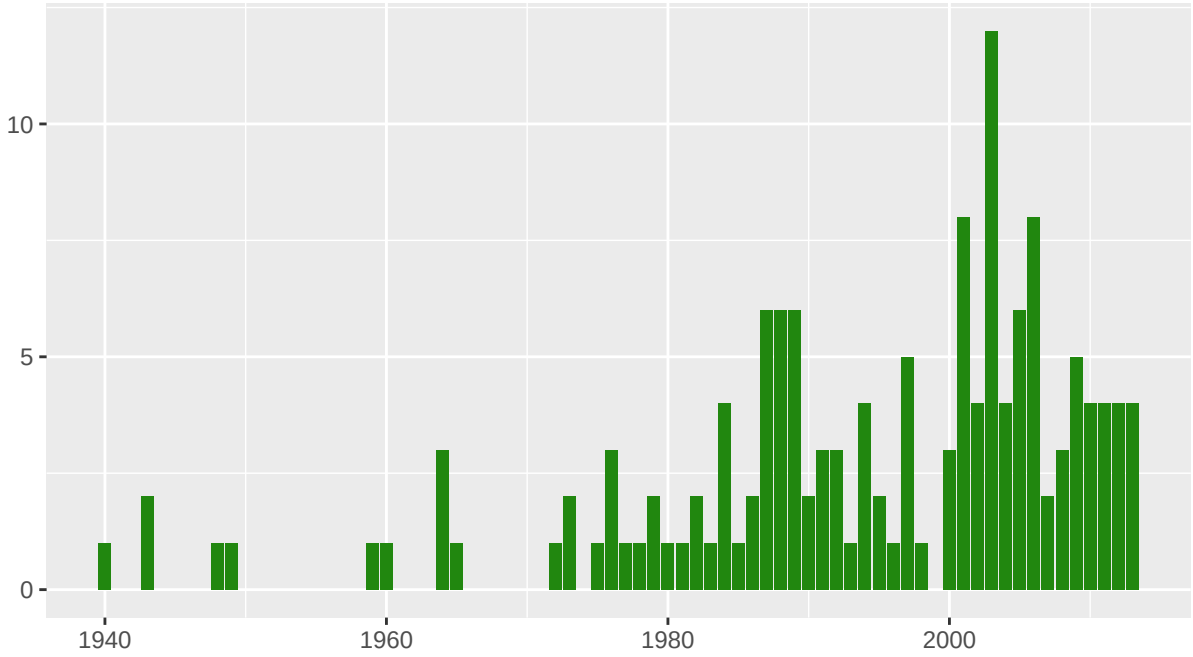
```r
#selecting columns to allow for rbind
lgbt_dc <- dc_comics %>%
  select(SEX, GSM, universe, YEAR)

#combining the data using rbind and filtering for parameters
lgbt <- rbind(lgbt_dc, lgbt_marv) %>%
  filter(SEX == "Transgender Characters" |
           SEX == "Genderfluid Characters" |
           #SEX =="Genderless Characters" | # removed
           #SEX == "Agender Characters" |  # removed
           GSM == "Homosexual Characters" |
           GSM == "Bisexual Characters")

# wrapping title
my_title_manual <- "LGBT characters introduced into DC and Marvel comics per year,\n including retroacti

#creating barchart
ggplot(data = lgbt) +
  geom_bar(mapping = aes(x = YEAR, fill = GSM), position = "dodge", fill="#21870f", show.legend = FALSE)
  scale_y_continuous(breaks=seq(0, 20, 5)) +
  labs(title = "Comics Are Gaining A Few LGBT Characters",
       subtitle = my_title_manual,
       y = "",
       x = "" ) +
  theme(
  plot.title = element_text(color = "grey15", size = 18, face = "bold"),
  plot.subtitle = element_text(color = "grey15", size = 14))
```

```
## Warning: Removed 6 rows containing non-finite values (stat_count).
```



At first I tried to make the line charts of percentages by sex and I could not get the information to work; likely because I over-wrangled. I reassessed and focused on the lgbt data which made a decent looking barchart. Other than not getting the y axis and x axis labels to work correctly, it looks mostly identical but I do see a few measurements that aren't clear post 2000. It would be helpful to have specifics of what was being measured. For the most part, I was guessing at the specifics and included genderfluid, transgender, homosexual and bisexual characters but I didn't see lesbian characters represented in the datasets when I used the unique function.

Citations

GGPlot Title, Subtitle and Caption https://www.datanovia.com/en/blog/ggplot-title-subtitle-and-caption/

R: ggplot2, can I set the plot title to wrap around and shrink the text to fit the plot? https://localcoder.org/r-ggplot2-can-i-set-the-plot-title-to-wrap-around-and-shrink-the-text-to-fit-t

Code written above is from the previous course IMT 511 which used the below text to support class scripts. https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover —