Stat 5550, Section 001
Statistical Visualization I
Fall 2020

Chance Gunter

A02232323

Homework 3

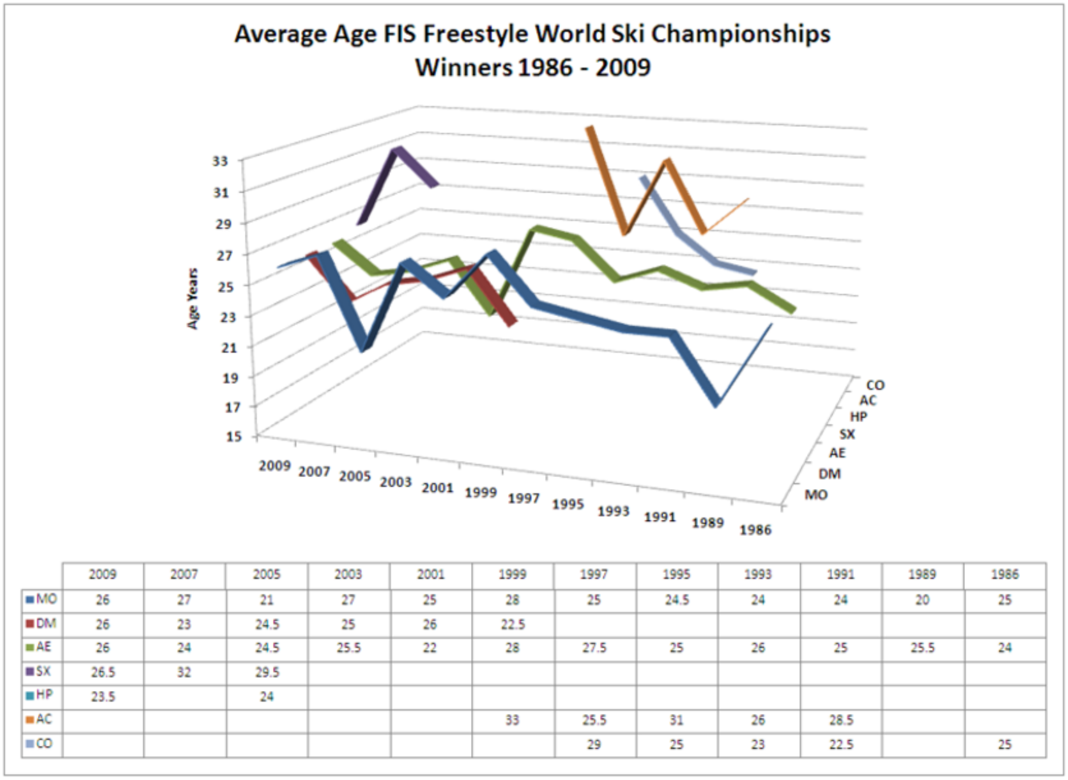
December 11, 2023

Stat 5550 Statistical Visualization I Fall 2023

Homework Assignment 3 (11/13/2023)

40 Points — Due Monday 11/13/2023 (Claim of Bad Graph) &
Thursday 11/16/2023 (In-Class Presentation) &
Monday 12/11/2023 (Written HW via Canvas by 11:59pm)

(i) (10 Points) **Fix a Bad Graph, Number 1:**
Carefully look at the graph in Figure 1.



	2009	2007	2005	2003	2001	1999	1997	1995	1993	1991	1989	1986
■ MO	26	27	21	27	25	28	25	24.5	24	24	20	25
■ DM	26	23	24.5	25	26	22.5						
■ AE	26	24	24.5	25.5	22	28	27.5	25	26	25	25.5	24
■ SX	26.5	32	29.5									
■ HP	23.5		24									
■ AC						33	25.5	31	26	28.5		
■ CO							29	25	23	22.5		25

Figure 1: Original: Obtained from `https://wiki.fis-ski.com/index.php/Image:Average_age_winner_FS_WSC_1986-2009.PNG` in Fall 2021. The graph is no longer available at this URL at this time.

(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.

<u>Answer:</u>

Figure 1 follows several “Bad Graphics” rules:

- Rule 9: Alabama First- The X axis and the z axis are improperly sorted. The X axis visually is moving right to left and not left to right for the dates. The z axis is not sorted alphabetically.
- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously- The data is highly illegible and ambiguous. This is due to the amount of Nan’s recorded and the use of a third axis. To make matters worse the labels in the table bellow the graph do not depict the same colors as those in the graph above (at least by Shade). Lastly we view date as oldest to Newest moving left to right and the graph has the opposite of this of Newest to Oldest.
- Rule 11: More is murkier: (a) more decimal places and (b) more dimensions- This data has the added dimension of Category or Sport. As a result the graph has more dimensions than necessary. Due to this it is harder to understand the impacts the Null data is causing. Another thing affected by this, the third axis makes it more difficult to gauge grid lines regarding the Ages on the Y axis.

(b) (4 Points) Demonstrate how this poor graph might be improved. Using the data from the graph (or better, from the table from underneath the graph), 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.

Note that the original graph uses abbreviations for the following terms: Mogul (MO), Dual Mogul (DM), Aerial (AE), Ski Cross (SX), Half Pipe (HP), Acroski (AC), and Slopestyle (CO).

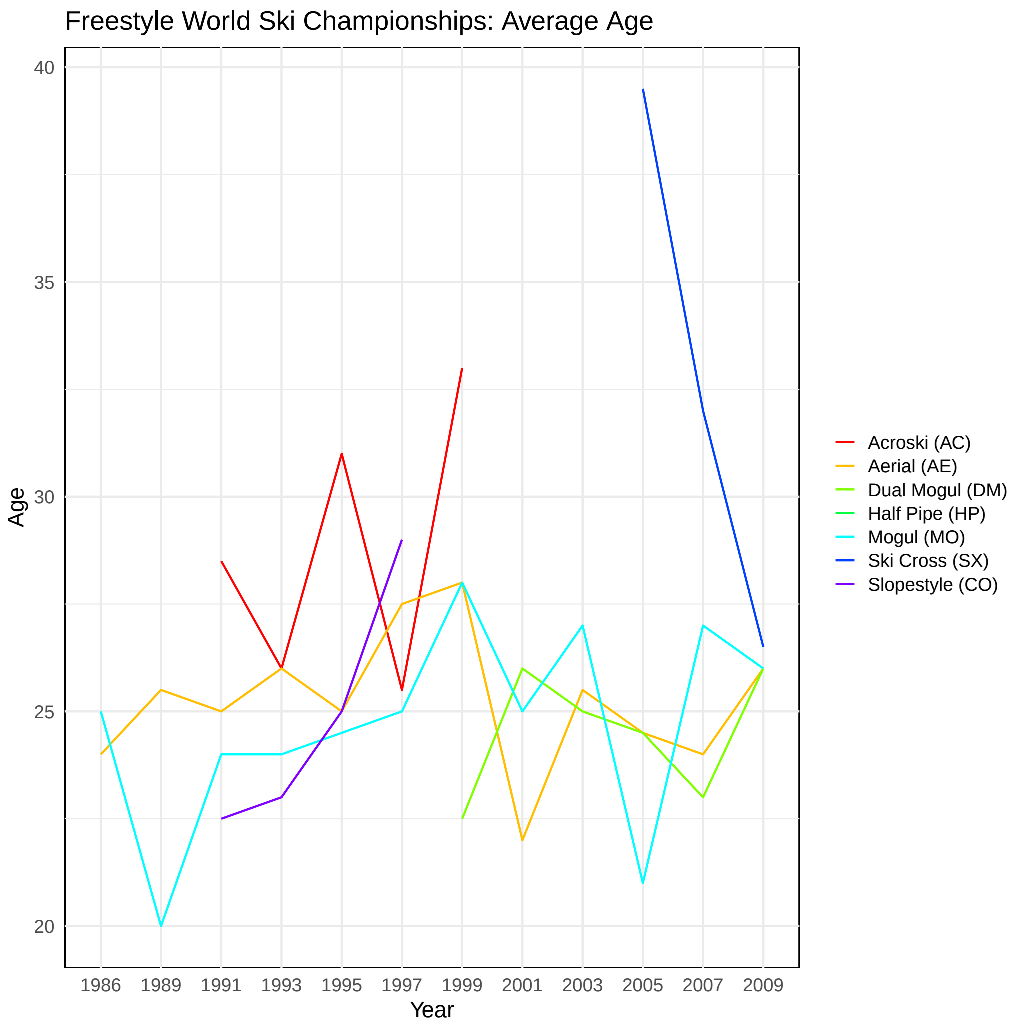
In your improved version of this graph, work with the numbers from the table below the graph and not the graph itself. It appears that the designer of the original graph messed up the graph...

Include a scan or a photo of the bad graph in your answer, next to your improved version. Also include your R code.

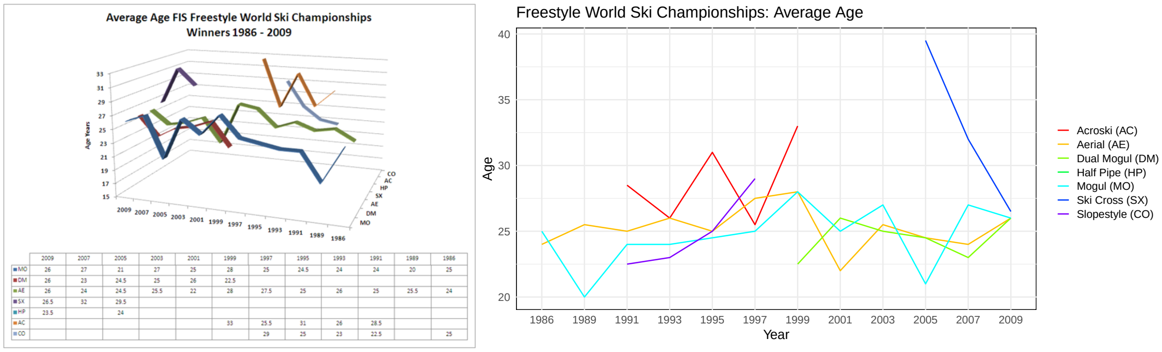
<u>Answer:</u>

```r
# Start with these vectors and further transform them for your needs

year <- c(seq(2009, 1989, -2), 1986)
mo <- c(26, 27, 21, 27, 25, 28, 25, 24.5, 24, 24, 20, 25)
dm <- c(26, 23, 24.5, 25, 26, 22.5, NA, NA, NA, NA, NA, NA)
ae <- c(26, 24, 24.5, 25.5, 22, 28, 27.5, 25, 26, 25, 25.5, 24)
sx <- c(26.5, 32, 39.5, NA, NA, NA, NA, NA, NA, NA, NA, NA)
hp <- c(23.5, NA, 24, NA, NA, NA, NA, NA, NA, NA, NA, NA)
ac <- c(NA, NA, NA, NA, NA, 33, 25.5, 31, 26, 28.5, NA, NA)
co <- c(NA, NA, NA, NA, NA, NA, 29, 25, 23, 22.5, NA, 25)

# Combine data into a data frame
df <- data.frame(year, mo, dm, ae, sx, hp, ac, co)

# Reshape the data into long format for ggplot
library(tidyr)
library(ggplot2)

# Rename columns
new_names <- c("Years", "Mogul (MO)", "Dual Mogul (DM)", "Aerial (AE)",
               "Ski Cross (SX)", "Half Pipe (HP)", "Acroski (AC)",
               "Slopestyle (CO)")

# Rename the columns of the data frame
colnames(df) <- new_names

df_long <- pivot_longer(df, cols = -Years, names_to = "Category",
                        values_to = "Age")


# Create a grouped bar plot using ggplot
```

```
ggplot(df_long, aes(x = factor(Years), y = Age, color = Category,
                    group = Category)) +
  geom_line() +
  labs(title = "Freestyle World Ski Championships: Average Age",
       x = "Year", y = "Age") +
  scale_color_manual(values = rainbow(ncol(df))) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "white"),
        legend.position = 'right', legend.direction = "horizontal",
        legend.title = element_blank()) +
  guides(color = guide_legend(title.position = "top", title.hjust = 0.5,
                             keywidth = 0.8, keyheight = 0.8,
                             nrow = 7))
```



```
# Save the plot
ggsave("hw03_sol_Fixed_Freestyle.pdf", width = 8, height = 4)
```



	2009	2007	2005	2003	2001	1999	1997	1995	1993	1991	1989	1986
■ MO	26	27	21	27	25	28	25	24.5	24	24	20	25
■ DM	26	23	24.5	25	26	22.5						
■ AE	26	24	24.5	25.5	22	28	27.5	25	26	25	25.5	24
■ SX	26.5	32	29.5									
■ HP	23.5		24									
■ AC						33	25.5	31	26	28.5		
■ CO							29	25	23	22.5		25

Figure 2: Original (left) and revised (right) Version: Average Age of Freestyle World Ski Championships Winners.

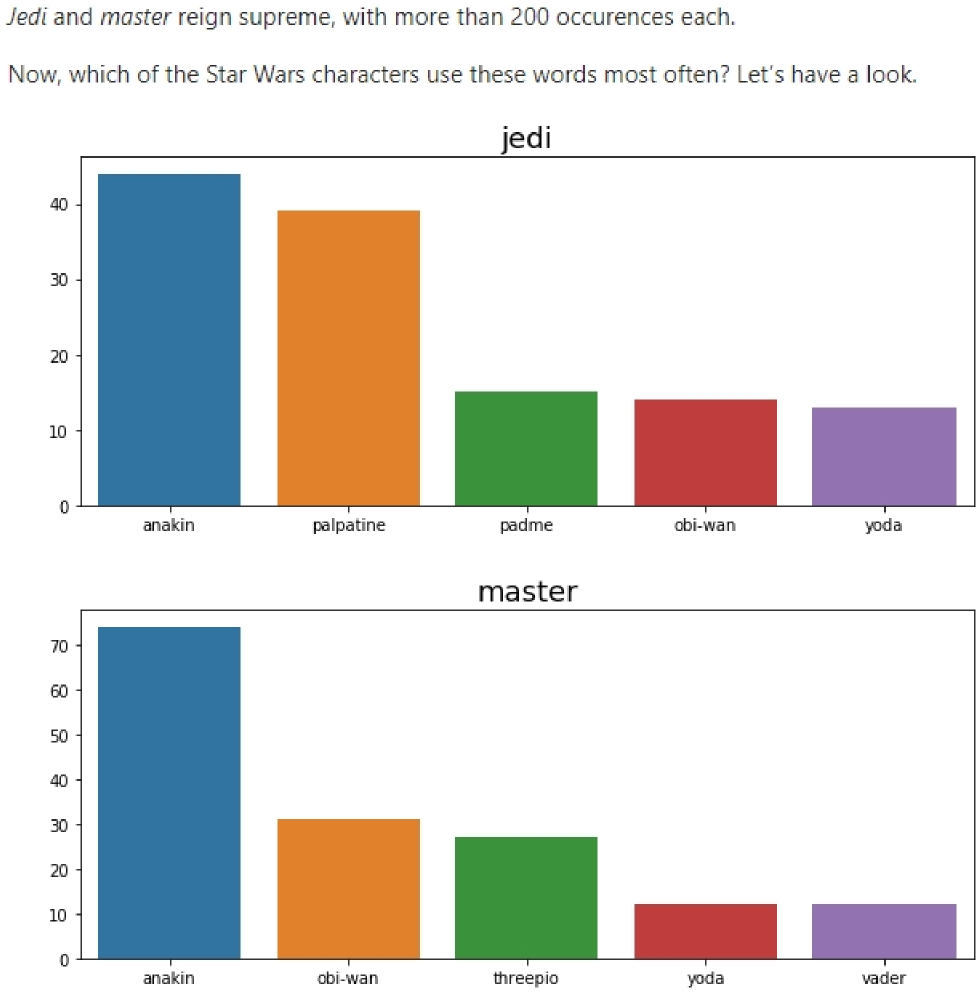
(c) (3 Points) Include a short write–up (about half a page) as to how you believe your version improves on the poor original. More specifically, indicate what you have modified and why this improves the representation of the underlying data.

Answer:

I created Figure 2 (right) to rectify the issues described in part (a) above.

The first thing that I did was remove the third dimension of category. Following this I orientated the years to read from left to right. The ending result is a line plot with a clearly labeled legend (Full names of Categories) and axis with appropriate labels. The original graphic had too much going on with the multi-dimensionality, the dates facing the wrong way reduced readability. Fixing these issues allowed for a clear representation of Freestyle Ski Ages amongst different sports.

(ii) (10 Points) **Fix a Bad Graph, Number 2:**
Carefully look at the graph in Figure 5.



Jedi and *master* reign supreme, with more than 200 occurences each.

Now, which of the Star Wars characters use these words most often? Let's have a look.

Figure 3: Original: Obtained from `https://www.quora.com/What-is-the-single-most-said-word-in-the-first-6-Star-Wars-movies`, Fall 2018, accessed 11/13/2023.

Here some background about this graph. Someone asked this question on Quora: "What is the single most said word in the first 6 Star Wars movies?" The graph in Figure 5 was designed by Hakon Hapnes Strand to answer this question. The graph shows the top-5 characters respectively that speak jedi and master most often. Visit the web page for further details or to see other (bad) Star Wars related graphs.

Repeat parts (a) through (c) from the previous question for this graph.

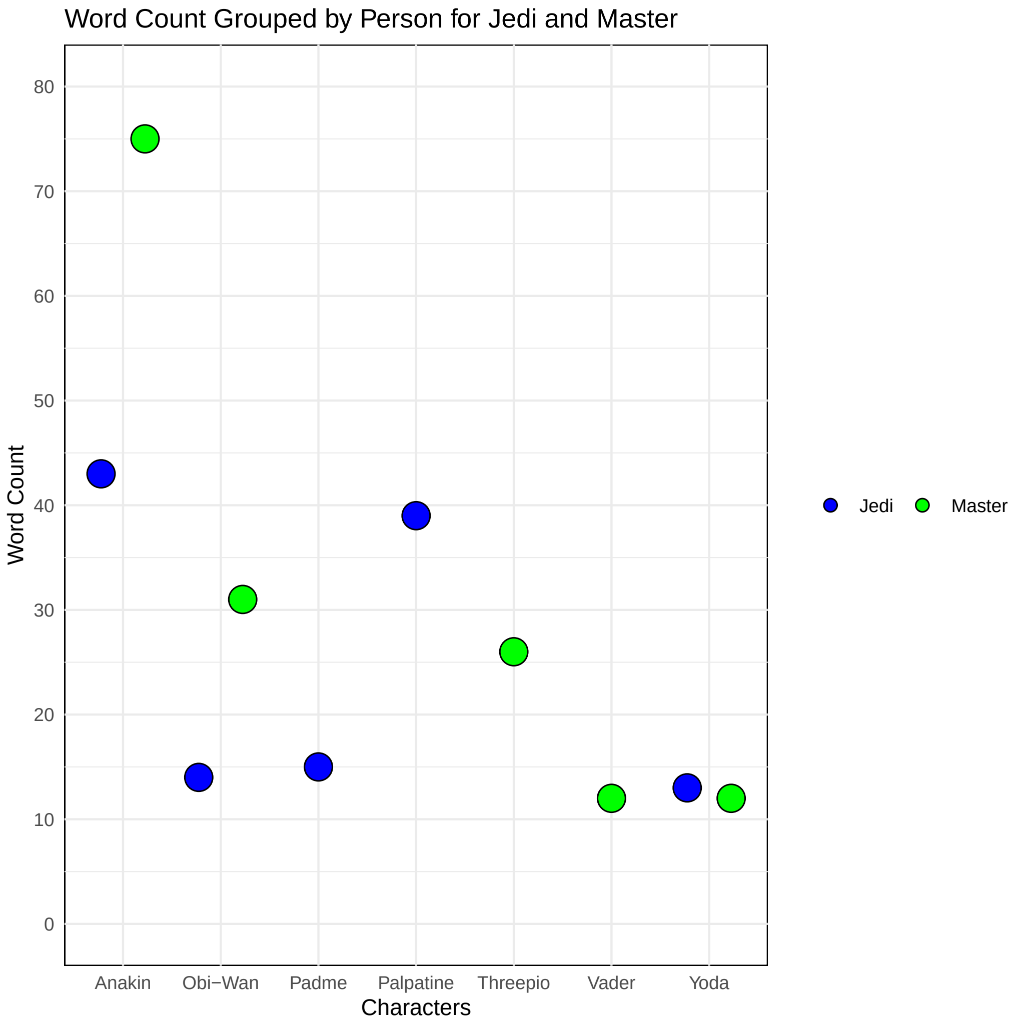
(a) Answer:

Figure 5 follows several “Bad Graphics” rules:

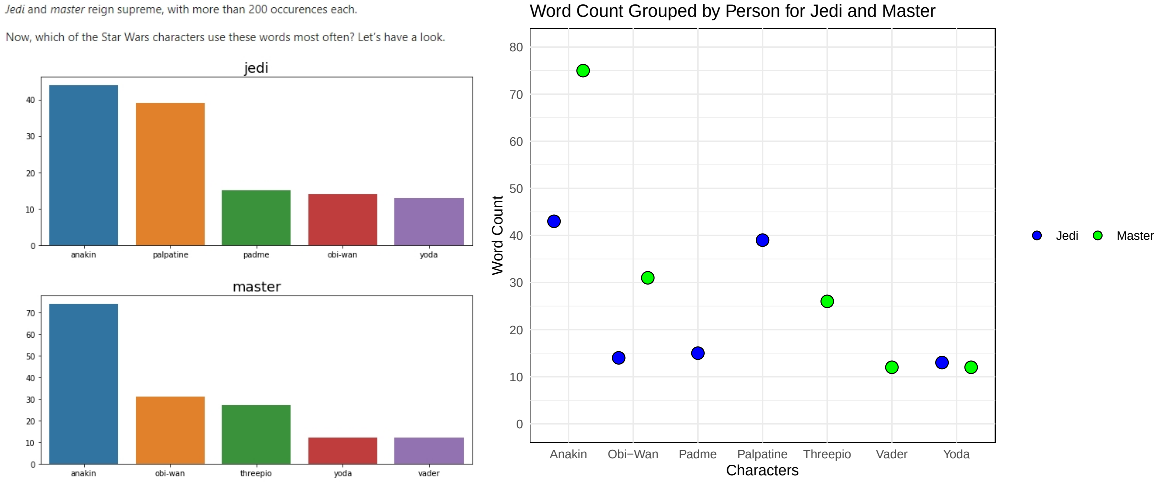
- Rule 1: Show as little data as possible (minimize the data density)- There are only 5 data points on each of the plots, but the figure is considerably filled with ink used for the bars..
- Rule 10: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously- The title and additional subtext do not provide an adequate description of the graph. For example, there is no indicator that this is over the first 6 movies, and there is no description as to why these characters were chosen. In movies 4-6, Darth Vader uses the word Jedi extensively but is not featured in the Jedi plot.
- Rule 3: Ignore the visual metaphor altogether- The plot depicting Jedi and the plot depicting Master are the same height, but are of a different scale. As a result, the amount in which Anakin says Jedi (42) looks the same as the amount in which Anakin says Master (72). The axes of the two graphs should be the same if they are comparing the same y-value of word count.

(b) Answer:

```r
# Start with these vectors and further transform them for your needs

characters <- c(
  "Anakin", "Palpatine", "Padme", "Obi-Wan", "Yoda",
  "Anakin", "Obi-Wan", "Threepio", "Yoda", "Vader"
)
word_count <- c(
  43, 39, 15, 14, 13,
  75, 31, 26, 12, 12
)

Word <- c("Jedi", "Jedi", "Jedi", "Jedi", "Jedi",
         "Master", "Master", "Master", "Master", "Master")

star_wars_df <- data.frame(characters, word_count, Word)

ggplot(star_wars_df, aes(x = characters, y = word_count, fill = Word)) +
  geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 1,
               position = "dodge") +
  labs(title = "Word Count Grouped by Person for Jedi and Master",
       x = "Characters", y = "Word Count") +
  scale_fill_manual(values = c("Jedi" = "blue", "Master" = "green")) +
  scale_y_continuous(breaks = seq(0, 80, by = 10),
                     labels = seq(0, 80, by = 10),
                     limits = c(0, 80)) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "white"),
        legend.position = ’right’, legend.direction = "horizontal",
        legend.title = element_blank()) +
  guides(color = guide_legend(title.position = "top", title.hjust = 0.5,
                              keywidth = 0.8, keyheight = 0.8,
                              nrow = 7))

## Bin width defaults to 1/30 of the range of the data.  Pick better value with
## ‘binwidth‘.
```



```
# Save the plot
ggsave("hw03_sol_Fixed_StarWars.pdf", width = 7, height = 5)

## Bin width defaults to 1/30 of the range of the data.  Pick better value with
## `binwidth`.
```



Jedi and *master* reign supreme, with more than 200 occurences each.

Now, which of the Star Wars characters use these words most often? Let's have a look.

Figure 4: Original (left) and revised (right) Version: Most Frequent Speakers of the Words Jedi and Master in the first 6 Star Wars Movies.

(c) Answer:

I created Figure 4 (right) to rectify the issues described in part (a) above.

The starwars graphs tried to convey to much with a small amount of data. To rectify this I combined the previous bargraphs into one dotplot. The charts shared several of the same charaters and were scalled different so in doing so it allowed for a better prespective of the wordcount data as a whole. In addition to this there should be more description regarding why they chose this data, for example in movies 4-6 Vader says Jedi a lot however it was not represented. As a result some of the characters only have 1 data point while others have 2.

(iii) (20 Points) **Find your own bad graph (FYOBG)!:**
Now, you have to find your own bad graph and fix it.

(a) (5 Points) Find a bad graph from the past 5 years (since 2018) 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. Also include a photo, scan, or screenshot of your bad graph.

Also, summarize what makes this a bad graph. You do not have to cite the exact bad-graphs rules from the lecture notes at this time, but it must be clear that you found an F-grade graph (a D- still would be a pass for a graph) and recognized its worst problems.

Web pages, journal articles, newspapers, magazines, videos, 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.

If you find a bad graph in the public, take a photo and describe where it was located, e.g., in a poster you noticed on a certain floor in a certain USU building. In this case, also include a photo of the entire source and not only the bad graph. Other suitable sources might be ads, signs, things shown on TV, etc. For these, also submit a photo of the bad graph, if possible a photo of the entire source, and describe where you found it.

Your bad graph must meet at least one of the rules for bad graphs from Chapter 6 in our lecture notes.

Each student must claim a bad graph by Monday, 11/13/2023, 11:59pm. Each student must claim a different graph. If you claim a graph that was previously claimed by someone else (or that was used in a previous semester), 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.

Here is my collection of bad Covid graphs (plus a few good ones):
https://math.usu.edu/~symanzik/talks/2021_SouthwestMichiganChapter.pdf

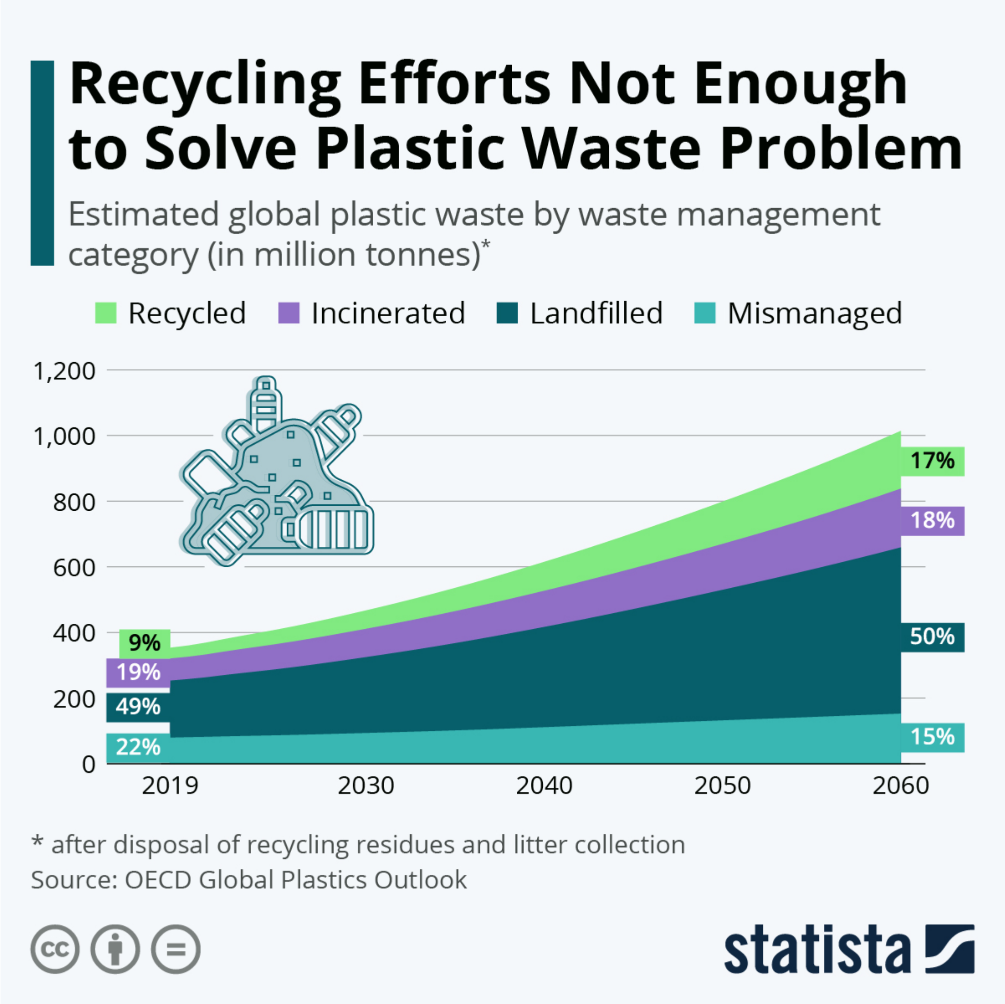
You cannot use any bad graph that is posted on these or any other bad-graph-collection web sites, books on bad graphs, or other similar publications. The goal of this HW is NOT to reuse a bad graph that has already been marked as bad by someone else, but rather to identify a (new) bad graph when we see it!

(b) (5 Points) I will combine 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 in-person lecture on Thursday 11/16/2023**: (i) Mention the source of the graph; (ii) indicate the rule(s) [from the 12 bad graph rules in Section 6.1 of the lecture notes] 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. Always refer to the number and name of these rules. You should practice in advance that you don't speak longer than 2 min!

Note: If you are unable to attend our last in-person lecture on Thursday 11/16/2023, you have to do the following: Send me a 2 min recording that shows your name, the bad graph, and the URL on the screen. Then provide the information for (i) to (iii) listed in (b) above. When completed, send me your recording via `https://bft.usu.edu/` **or provide me with a link to Box, Google Drive, etc. that allows me to access your recording. This recording is due on Wednesday 11/15/2023 by 12noon (so I can check it in advance). I will play your recording in our last in-person lecture on Thursday 11/16/2023 as part of the presentations.**

***** In any case, you must contact me in advance via e-mail if you are not able to attend our last in-person lecture on Thursday 11/16/2023. *****

(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.


Recycling Efforts Not Enough to Solve Plastic Waste Problem

Estimated global plastic waste by waste management category (in million tonnes)*

* after disposal of recycling residues and litter collection

Source: OECD Global Plastics Outlook

Figure 5: Original: Obtained from `https://www.statista.com/chart/27756/global-waste-management-projections/`, Fall 2023, accessed 11/13/2023.

Answer:

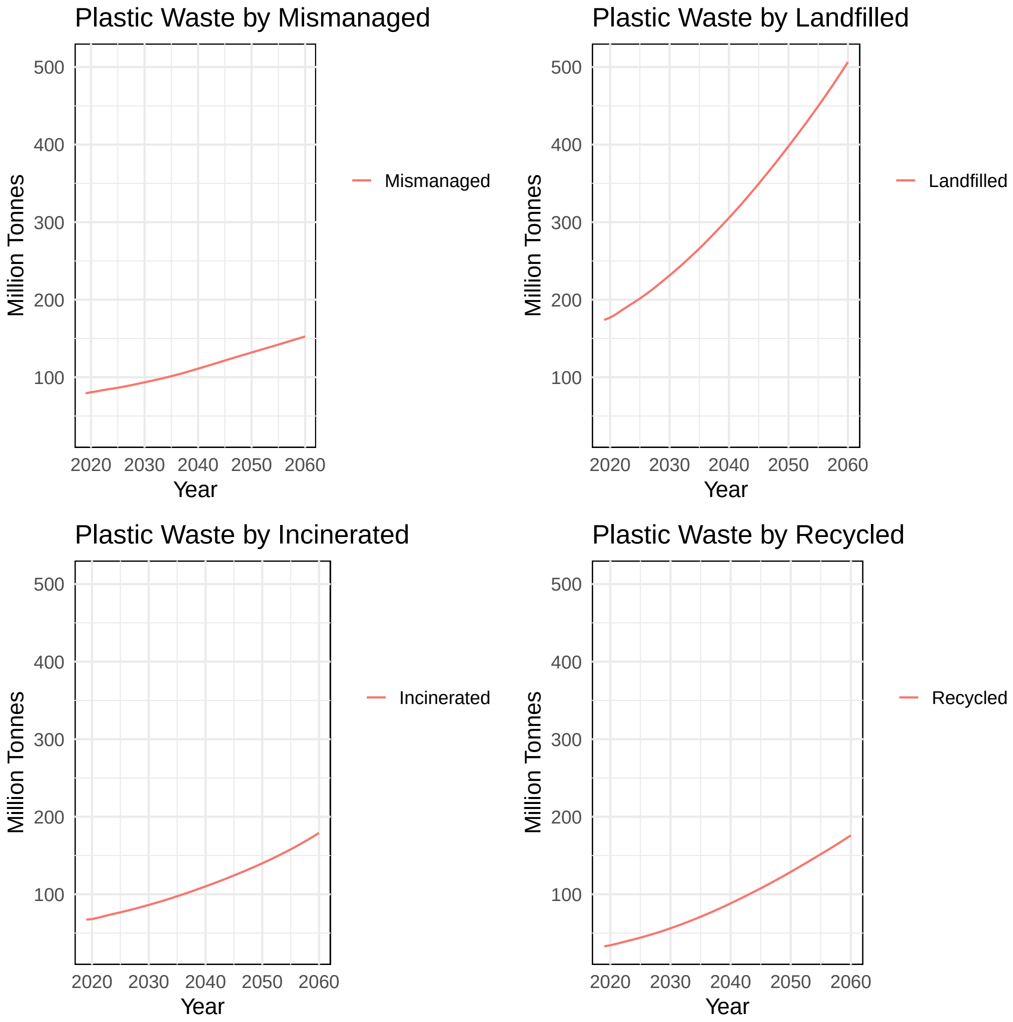
(a) My selected bad graph follows several "Bad Graphics" rules:

- Rule 8: Jigging the Baseline- Due to the stacked nature of the this plot it is hard to tell the growth of the individual categories. For Example 49-50 percent we see a change in total of 1 percent while the y axis is in units leaving the viewer to preform som ration analysis to find an estimate of the final unit value. This is further complicated by the missing 400 unit grid line.

- Rule 1: Show as Little Data as Possible- Similar to the U.S. vs Japan labor statistic this is a comparison of percentages while the scale is based on units. Another thing to note is that during the duration from the beginning to the end is smoothed and not accurately representing the value at that given point in time.
- Rule 3: Ignore the visual metaphor altogether- The highlighted percentages are misleading, the graph indicates for incinerated, landfill, and mismanaged waste categories that there is a change from 2019-2060. However, That is not reflected in the visual representation where the largest reflections of this are 49-50 percent and the 22-15 percent.

(b) An improved version can be created as follows:

```
data_df <- read.csv("data.csv")

library(gridExtra)

# Assuming you have a data frame named ’data’ with columns: Year,
# Category, Value

categories <- unique(data_df$Category[data_df$Category != ""])
print(categories)

## [1] "Mismanaged"  "Landfilled"  "Incinerated" "Recycled"

y_range <- range(data_df$Value, na.rm = TRUE)

# Create a list to store the plots
plots_list <- list()

for (category in categories) {
  # Subset the data for the current category
  subset_data <- data_df[data_df$Category == category, ]

  # Create the plot
  p <- ggplot(subset_data, aes(x = Year, y = Value, color = Category)) +
    geom_line() +
    labs(title = paste("Plastic Waste by", category),
         x = "Year", y = "Million Tonnes") +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "white"),
        legend.position = ’right’, legend.direction = "horizontal",
        legend.title = element_blank()) +
  guides(color = guide_legend(title.position = "top", title.hjust = 0.5,
                             keywidth = 0.8, keyheight = 0.8,
                             nrow = 7)) +
    ylim(y_range)  # Set y-axis limits

  # Add the plot to the list
  plots_list[[category]] <- p
}
# Arrange the plots in a 2x2 matrix
arranged_plots <- grid.arrange(grobs = plots_list, ncol = 2)
```



```
# Save the arranged plots to a PDF file
ggsave("hw03_sol_Fixed_Recycling.pdf", arranged_plots, width = 7, height = 5)
```

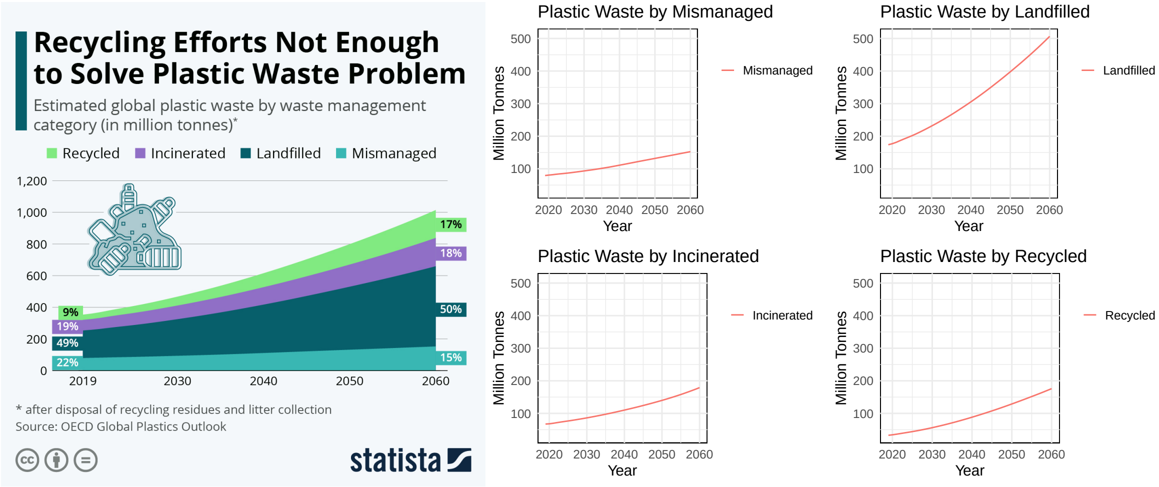


Figure 6: Original (left) and revised (right) Version: Statista Recycling Image and the Revised Subplots

(c) Improvements of the bad graph:
To fix these Rule issues I created several subplots. These subplots are split out by subcategory of waste management type. All the axises are scaled the same. The previous plot was an area chart combined in a stacked nature with percentage labels. These new graphs are line graphs show less of a change than previously indicated.

General Instructions

(i) Create a single pdf document, using R Markdown, knitr, or Sweave. When you take this course at the 6000–level, you have to use LaTeX in combination with knitr or Sweave. You only have to submit this one pdf document to Canvas.

(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. Your answer to question (i) should start on page 2!

(iv) Show your R code and resulting graph(s) [if any] for each question part!

(v) Before you submit your homework, check that you follow all recommendations from the tidyverse style guide (see `https://style.tidyverse.org/`). The easiest way to obtain properly formatted R code is via the *styler* R package (see `https://styler.r-lib.org/`). Moreover, make sure that your R code is consistent, i.e., that you use the same type of comments and the same type of quotes throughout your entire homework.

(vi) 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). Make sure to properly format code from such external sources via *styler*.

(vii) If you have used ChatGPT or any other AI tool, you have to acknowledge its use. Be specific for which question/question parts it has been used — and for which purpose, e.g., to write certain code for you or just fix the grammar in your text.

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

(ix) 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!

(x) Submit your single pdf file via Canvas by the submission deadline. Recall the **No Excuse Needed Late Homework Submission Policy** from the syllabus. Make use of your three tokens wisely.