Homework 3: ggplot2

This homework assignment is meant to get you comfortable with ggplot2. We will be using multiple data sets from the TidyTuesday website for convenience. Here is the order of data we will be using, along with the respective links:

1. Registered Nurses Data. Read in the data manually as shown on the Github page. Note that below is an example of the code you'll copy, but the link is too long to fit onto the page.

```
## this reads in the data
nurses <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

2. Netflix Titles. Read in the data manually as shown on the Github page.
3. TV's Golden Age. Read in the data manually as shown on the Github page.

Assignment

Registered Nurses Data

1. As a warm-up question, create the a lineplot of the average hourly wage over time using the Registered Nurses Data. Your graph should look like Figure 1.
2. For a slightly more complicated graph, let's see if we can make this more informative. While wages are increasing over time, it's important to know what is driving this (outliers?). For this question, you will need to create a boxplot graph over time, starting from the year 2000 onwards. As a hint, you may need to use the following function: `base::factor`. See Figure 2 for the final graph. Your job is to replicate this graph. You may also want to look into the `ggplot::theme()` function for tilting the axis text.
3. Since you will want to point out correlations in your presentation, Figure 3 will give you practice. Try to recreate this graph. Note that using `geom_smooth` will give you the line of fit. You can change the `method` argument in `geom_smooth` to be "lm" (linear model).

Netflix Data

1. For this question your goal will be to recreate Figure 4. This question should guide you through this process. First, using the `date_added` column, convert this to a date using the `lubridate` package. Next, since we want aggregations by year, use the `lubridate` package to get the year of each `date_added`. Once this is done, you should be able to `group_by` the `year` and `type` to get counts of TV shows and movies within each year. Now your job is to figure out how to indicate which year has the highest movie total and which year has the highest TV show total (you will want to do this programmatically). Once this is done, you should be able to get a close to the final figure. Note that I am using `geom_col` in addition to `geom_text` with a `facet_wrap`. If you need to get rid of a legend, you can look further into the `ggplot2::theme` function.

TV's Golden Age

1. *Challenge Question*: Making Figure 5 is challenging. Here is some code to get you "close" to the dataframe you need:

```
tv_ratings %>%
    group_by(title_id) %>%
    mutate(max_season = max(season_number), min_season = min(season_number)) %>%
    filter(season_number == max_season | season_number == min_season) %>%
    add_count() %>%
    filter(n > 1) %>%
    mutate(max_season_rating = ifelse(max_season == season_number,
        av_rating, NA), min_season_rating = ifelse(min_season ==
        season_number, av_rating, NA)) %>%
    tidyr::fill(max_season_rating, .direction = "up") %>%
    fill(min_season_rating) %>%
    distinct(title_id, title, max_season, max_season_rating,
        min_season_rating) %>%
    arrange(desc(max_season)) %>%
    head(20)
```

Go through each line of this code and write a down a sentence on what is happening. Why did I write the code like this?

2. To get the number of seasons for each of these titles "glued" together as they are in the graph, you will want to use the `glue::glue` function in conjunction with a `mutate`. Create a new column that can do this using these functions.
3. Create a new column that denotes if the minimum season is better or the last season is better.
4. Using `geom_segment` and `geom_point`, try to create this graph.
5. While this graph is visually appealing, it is actually not a good visualization. Take a second to think about how this visualization could be better. If you can make this visualization better, go for it!

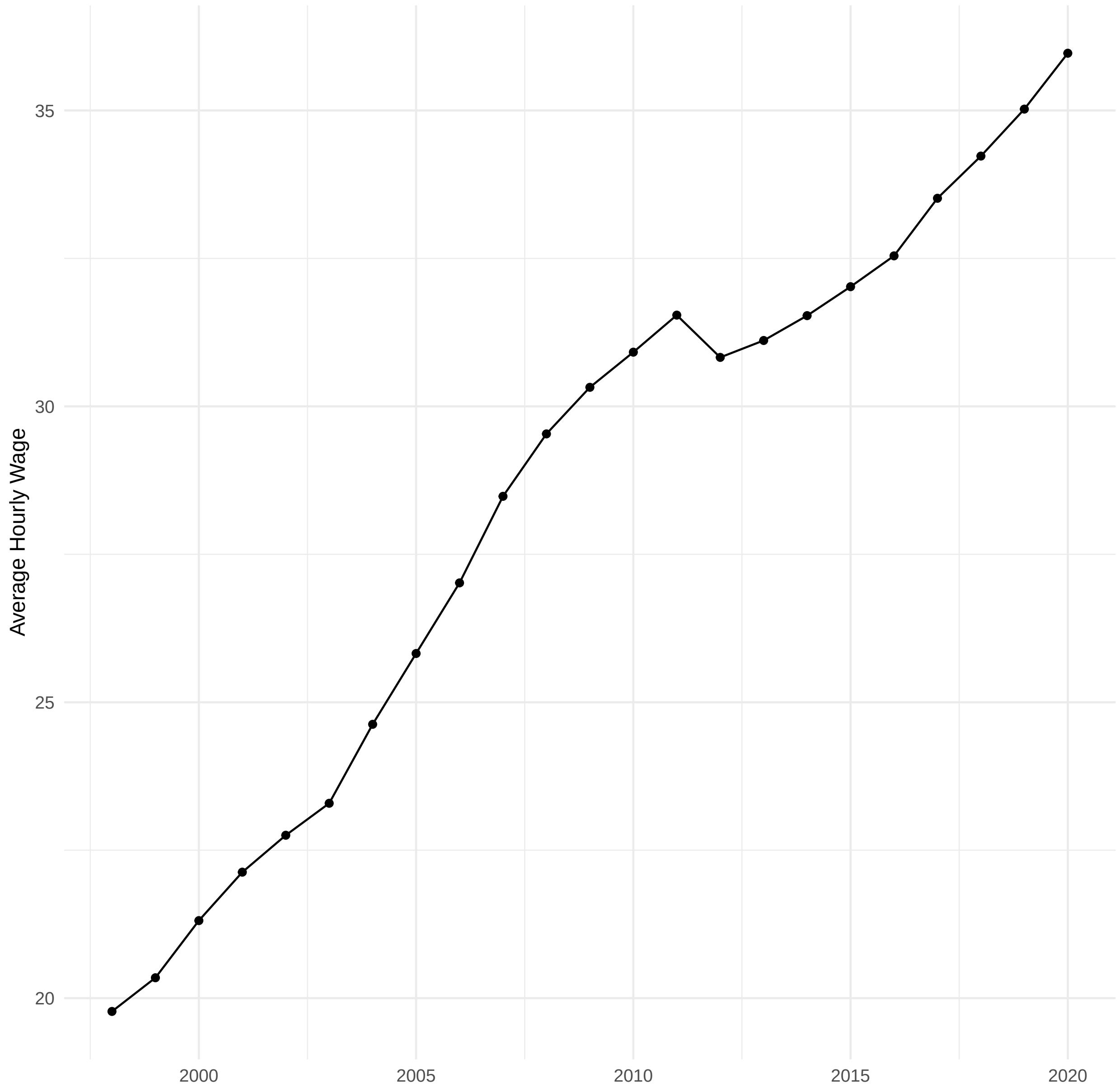

Figure 1: Average Hourly Wage of Nurses Over Time in the US

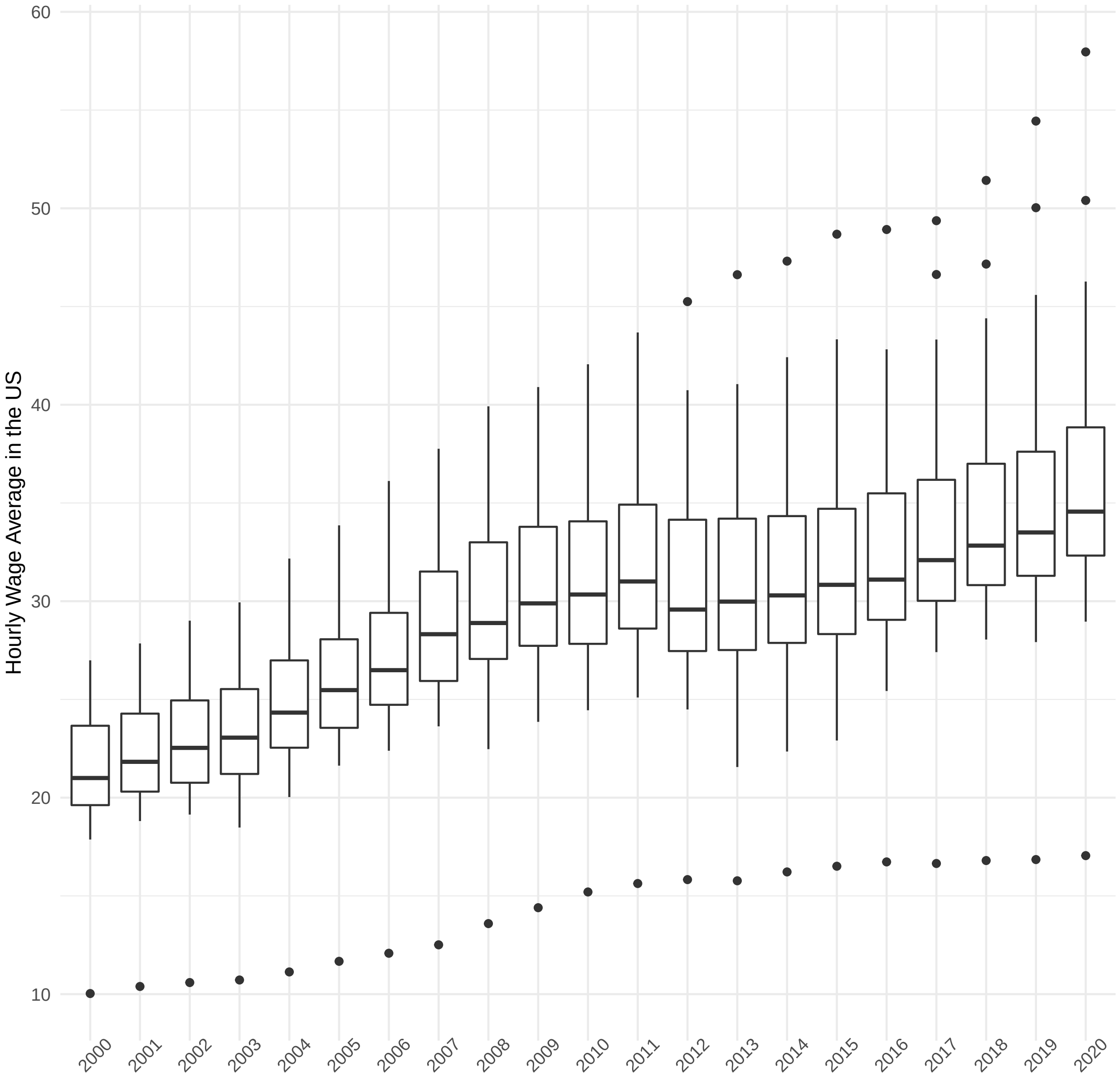

Figure 2: Hourly Wage of Nurses Over Time in the US (2000-2020)

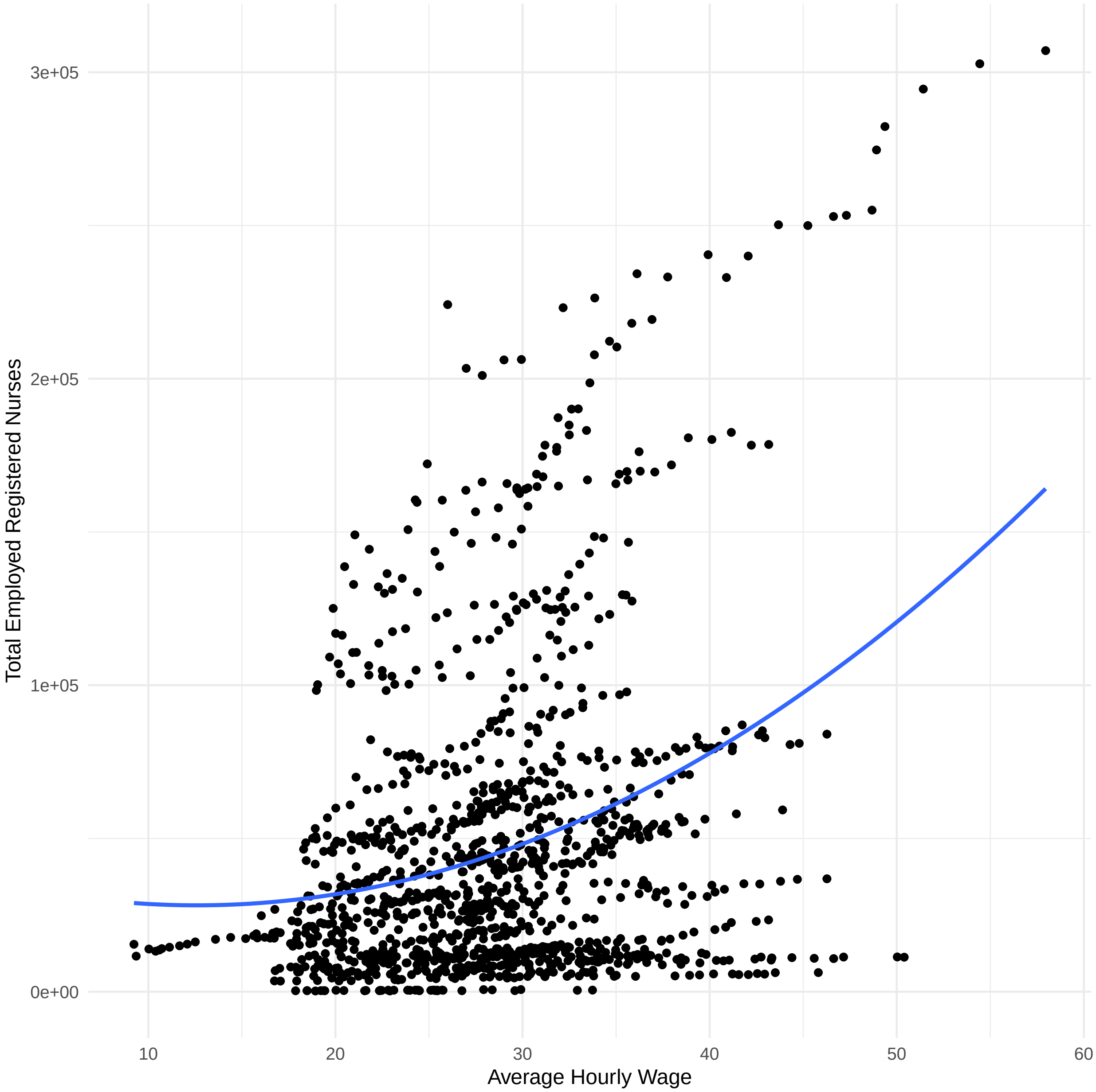

Figure 3: Correlation between average hourly wage and total registered nurses

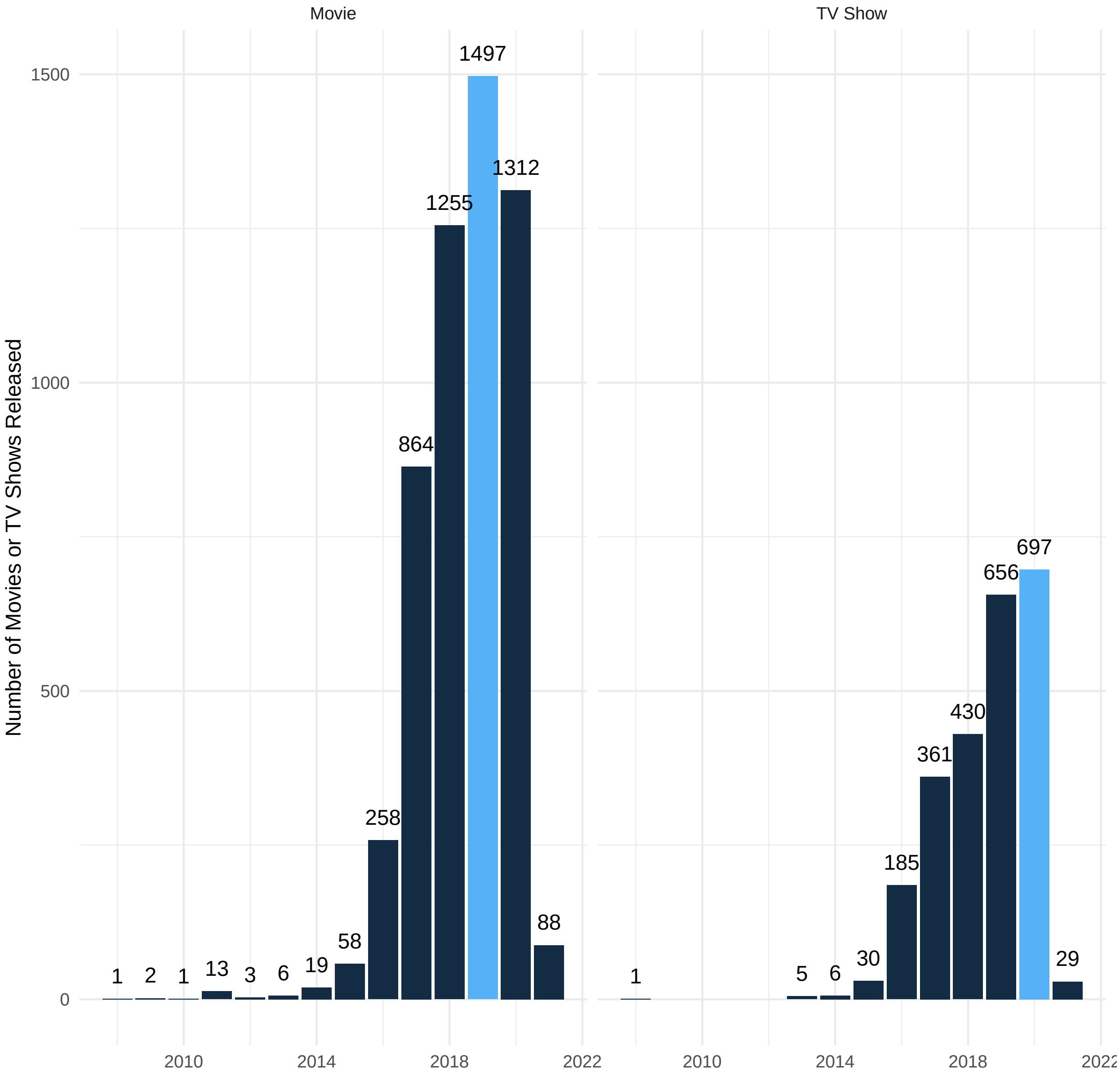

Figure 4: Number of Movie and TV Shows released on Netflix over time.

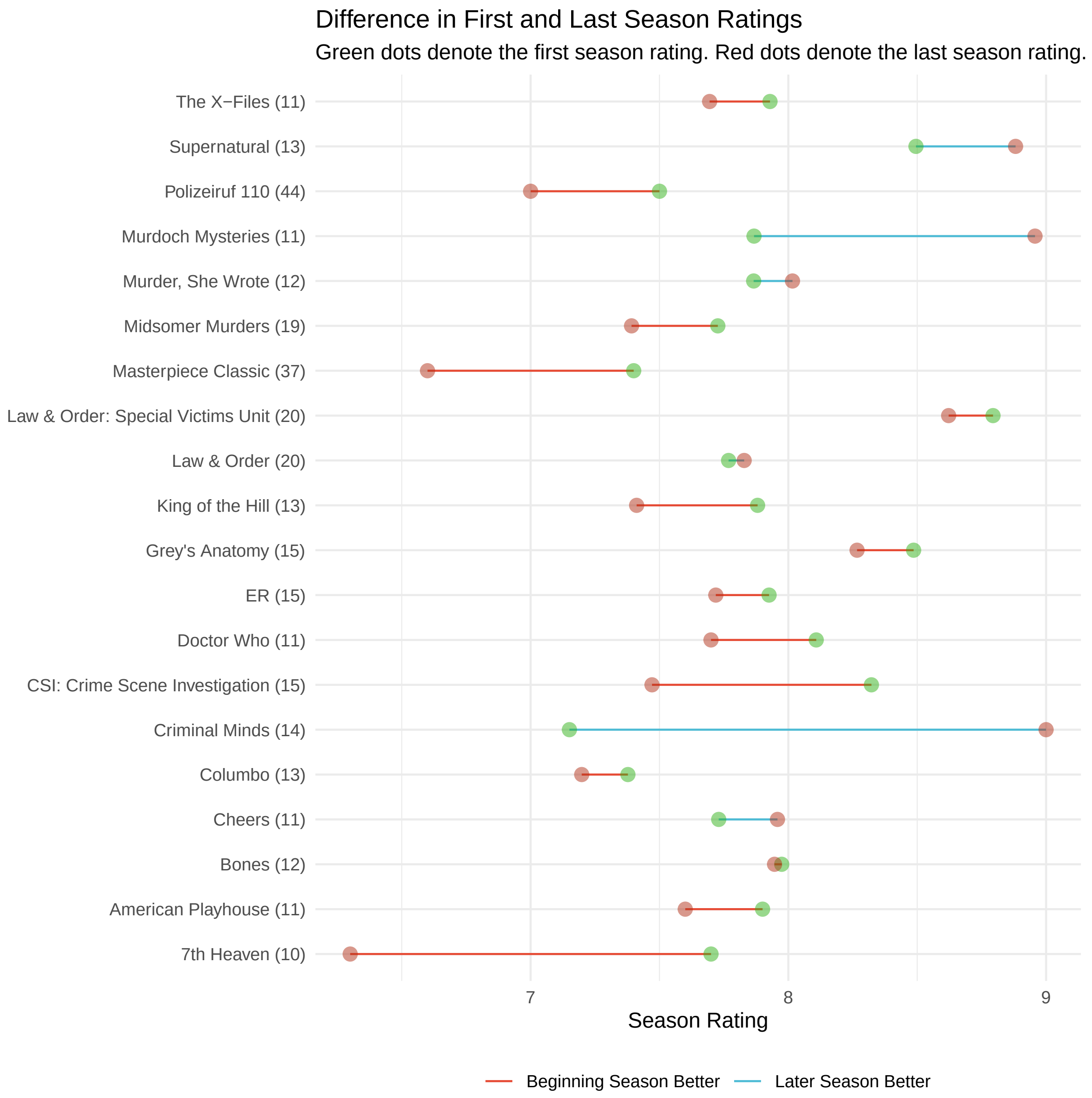

Figure 5: Difference in first and last season ratings.