Project 3 Friends

Connor Hudziak

4/26/2022

Introduction

The friends data set is composed of information that is relevant to the hit sitcom tv series. It gives the information about each episode such as the date, writers, and even how the episode performed in ratings. For my analysis I wanted to focus closely on the ratings for each episode/season in two different categories. The two categories were an IMDB rating and the total United States views in millions. I was interested to see which season had the highest average in each category. I then wanted to find out if the total US views were related to the IMDB rating and how strong or weak their relationship is.

Data wrangling

In this section I wrangled the data to separate the two categories of IMDB rating and total US views (millions). I started with the IMDB ratings and grouped the data by seaosn and then selected the rows I wanted to look at which included the season, total views, and IMDB rating. I then took summarized teh average IMDB ratings for each episode in each season and produced the IMDB mean per season. I then completed these same steps for the total US views and found the average according to their season.

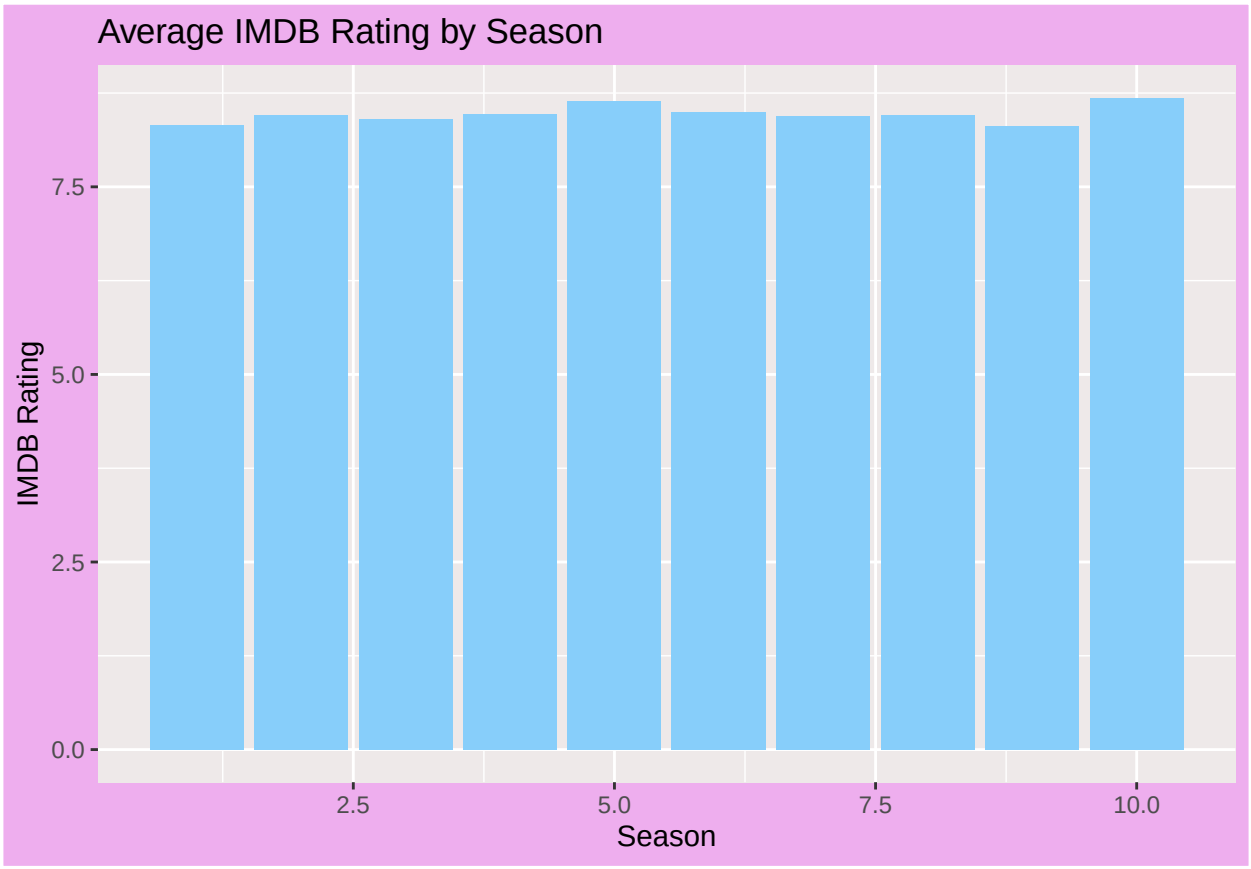
```
f_imdb <- friends %>%
  group_by(season) %>%
  select(1,7:8) %>%
  summarise(avg_imdb = mean(imdb_rating, na.rm = TRUE))

f_views <- friends %>%
  group_by(season) %>%
  select(1,7:8) %>%
  filter(season == 1 | 10) %>%
  summarise(avg_view_millions = mean(us_views_millions, na.rm = TRUE))
```

Average IMDB Rating by Season

In this column chart,I have depicted the average IMDB ratings according to their season. I was initially surprised by the amount of large IMDB results that consistently appeared on the visualization. All of the shows seasons were over a 7.5 average rating. Referencing back to a previous project of mine that included IMDB ratings of Scooby Doo episodes showed a bit more drastic differences between different seasons and/or networks. I think that this adds to the significance of the show Friends and how successful and popular it really was. Looking at the graph it does appear that season 9 had the lowest average and season 10 had the highest which could be from it being the finale of the show in the last season. I will explain further in the next graph on more specific and obvious treads by season when looking at the total viewer count.

```
ggplot(f_imdb, aes(x = season,
                   y = avg_imdb)) +
  geom_col(fill = "lightskyblue") +
  labs(title = "Average IMDB Rating by Season",
       x = "Season",
       y = "IMDB Rating") +
       theme(plot.background = element_rect(fill = "plum2"),
       panel.background = element_rect(fill = "snow2"))
```

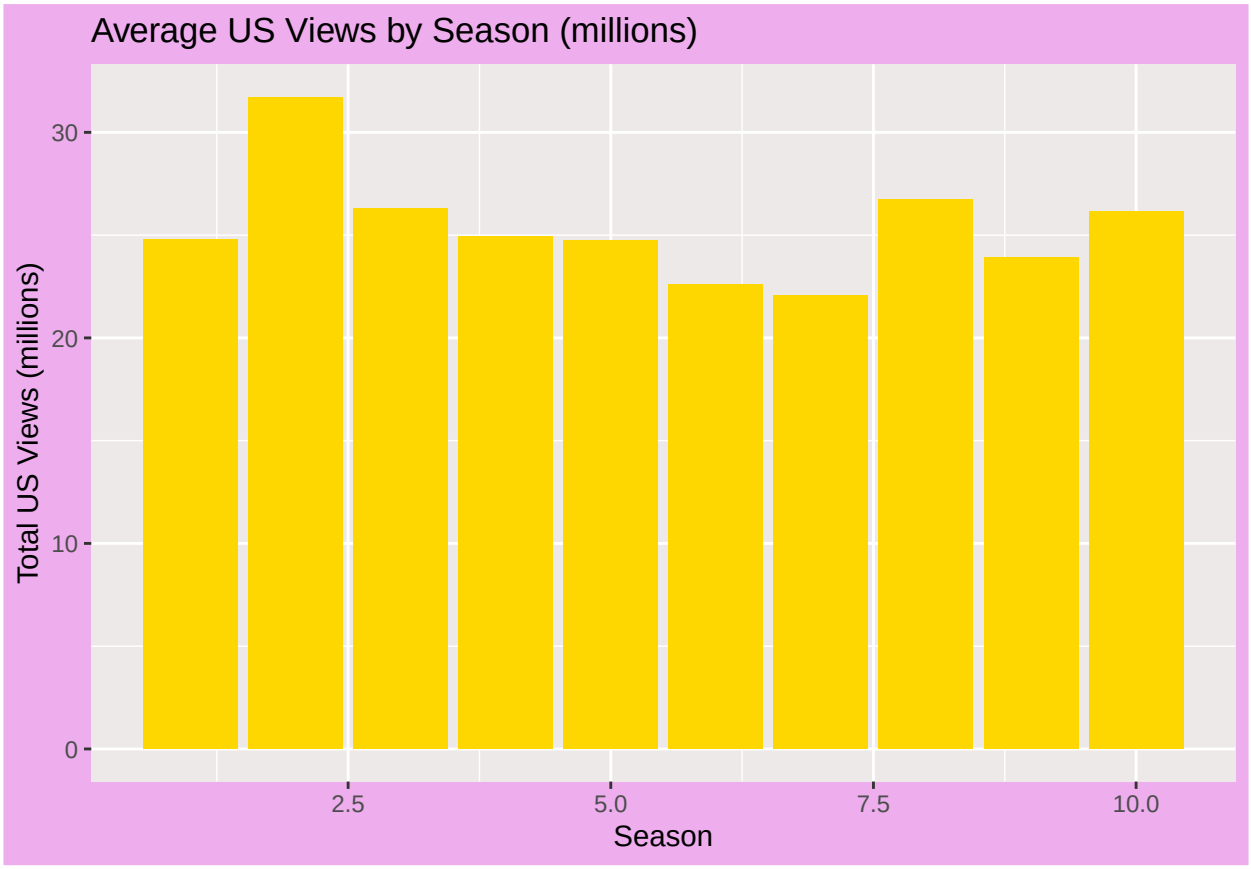


Average US Views by Season (millions)

In this graph it shows the average total US viewer count in the millions. This is a bit more accurate of determining which seasons appeared to be favorites and which ones were least watched. I noticed a large dip in viewership during season 7 and decided to further my research as to why this would've occurred. I concluded that during this season the overall direction of the show was altered in te way that two characters had started becoming closer in a relationship and may have turned some fans off. On the other hand, I have read an article saying that season 7 was a favorite by many because it concluded a wedding season finale. Altogether I believe that the filler episodes in the middle of season 7 were not as lively as the rest of the show. This resulted in lower viewership and even the last episode of the series could not make up for the lost views.

```
ggplot(f_views, aes(x = season,
                    y = avg_view_millions)) +
  geom_col(fill = "gold") +
```

```
  labs(title = "Average US Views by Season (millions)",
       x = "Season",
       y = "Total US Views (millions)") +
        theme(plot.background = element_rect(fill = "plum2"),
        panel.background = element_rect(fill = "snow2"))
```

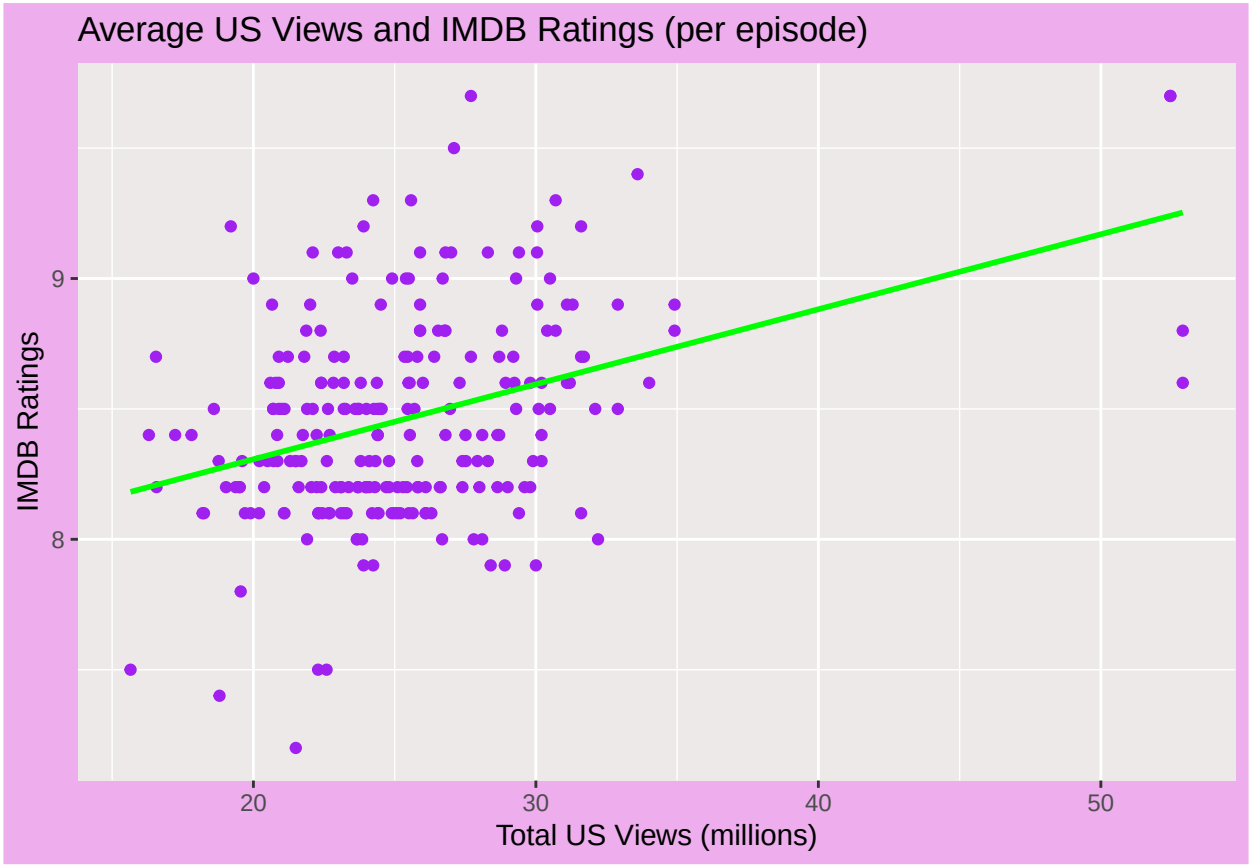


Statistical inference

In this last visualization I wanted to address my main question of whether the IMDB ratings and US viewer count were related to each other. To do this I first wanted to make a point plot to depict the relationship visually. I also used the "lm" method to fit the model and prepare to find the linear regression of the data. Already looking at the plots of the graph I can tell that there is not a strong relationship visually which changes my initial hypothesis of that they would be similar. Down below the graph I wil explan my results from the correlation test.

```
ggplot(friends, aes(x = us_views_millions,
                    y = imdb_rating)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  labs(title = "Average US Views and IMDB Ratings (per episode)",
       x = "Total US Views (millions)",
       y = "IMDB Ratings") +
        theme(plot.background = element_rect(fill = "plum2"),
        panel.background = element_rect(fill = "snow2"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Correlation Test

I had prepared the visual graph above for a regression analysis and will use a correlation test to get the cor value of the relationship between IMDB and views. This test will give me a value between -1 and 1 and the closer to those values the stronger the relationship will be. The cor value was .3774125. This means that the correlation between the IMDB and views was not very strong and would not be considered significant to relate both of these categories to each other.

```
cor.test(friends$us_views_millions,friends$imdb_rating)
```

```
##
##  Pearson's product-moment correlation
##
## data:  friends$us_views_millions and friends$imdb_rating
## t = 6.2344, df = 234, p-value = 2.097e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2623565 0.4818880
## sample estimates:
##       cor
## 0.3774125
```

Conclusion

In all, I was surprised to see the difference in averages of IMDB to be so closely related as well as seeing that IMDB and total views were not related as they were two categories that measure a similar aspect of the show. To further my interest in this data set I might look more closely to the writers of each episode and see if that impacted the results in any way.

Sources: 1. https://github.com/rfordatascience/tidytuesday/issues/254 2. https://friends.fandom.com/wiki/Season_7 3. https://collider.com/friends-seasons-ranked-from-worst-to-best/