An Analysis of Episode Ratings of The Office (US) Throughout the 9 Seasons

The purpose of this analysis is to examine the viewing rates of different episodes of The Office (US). I will be analyzing this data to see if any other factors, such as the writers, season number, and other factors affect US viewing rates.

Webscraping:

The following data was retrieved from the Wikipedia page listing every episode of The Office (US) in a table for each season. For each episode, the following information was obtained: episode number overall, episode number in season, title, director, writer, release date, and US viewer counts (in the millions).

Dataframes:

The following code sets up the overall data frame for The Office episode data. There was a lot of clean up needed to format the data frame. I had to remove the webisodes and other unofficial episodes from the data frame. I also removed the "Prod. code" column because I did not think it added anything to the analysis.

There was also a slight issue with duplicate rows. Some episodes are split into two parts. Officially, the episode titles are "Title Pt.1" and "Title Pt.2" for example, and they are technically separate episodes. However, this website treated all two-part episodes as if they were one, which caused some misalignment with overall episode number and number of episode in season. Because of this, overall episode number sometimes skips over a few numbers. To solve this, I just removed any rows that didn't have a director or release date (which meant it was the second part of another episode). The viewer counts won't be affected by this because the website originally treated these two-part episodes as having one combined viewing count. This is just the reason why the actual total episode count does not align with the total number of episodes in this dataframe.

Other clean up I had to perform to format the data properly was to fix the formatting of the release date so that it was purely numerical. Also, during the webscraping, the US viewer count included a footnote number after each view count which I removed since it was not relevent to the data.

Out[3]:

	No_Overall	Season	No_Of_Season	Title	Director	
0	1	1	1	"Pilot"	Ken Kwapis	Gervais Merchar
1	2	1	2	"Diversity Day"	Ken Kwapis	E
2	3	1	3	"Health Care"	Ken Whittingham	Paul L
3	4	1	4	"The Alliance"	Bryan Gordon	Mich
4	5	1	5	"Basketball"	Greg Daniels	Gr
...	...	...	...	...	...	
207	195	9	19	"Stairmageddon"	Matt Sohn	Da
208	196	9	20	"Paper Airplane"	Jesse Peretz	Sulliva L
209	197	9	21	"Livin' the Dream"‡	Jeffrey Blitz	Niki
210	198	9	22	"A.A.R.M."‡	David Rogers	Brent
213	200	9	24	"Finale"*	Ken Kwapis	Gr

186 rows × 8 columns

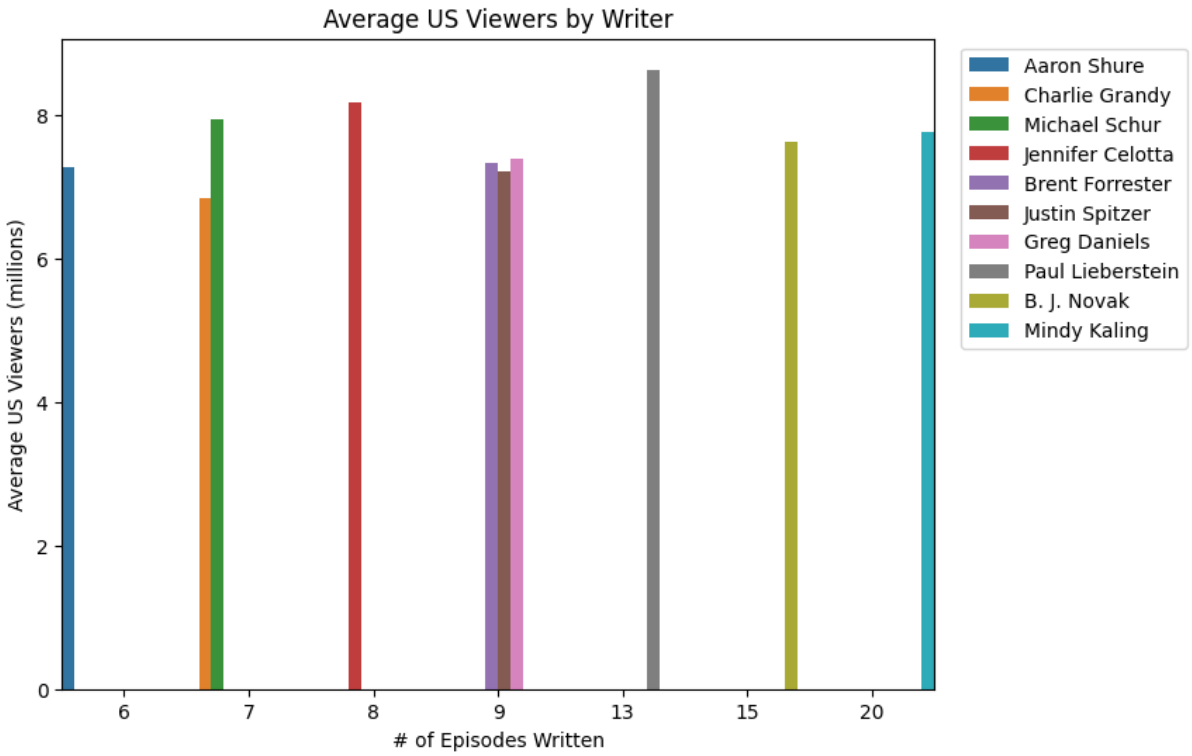
Average Viewers Per Writer

First, I want to analyze the average viewing rate of episodes according to the person who wrote the episode. I will group by writer and sort the data to show the top 10 writers who wrote the most episodes.

Out[4]:

	Writer	total_episodes	avg_viewers
36	Mindy Kaling	20	7.758500
4	B. J. Novak	15	7.610667
39	Paul Lieberstein	13	8.623077
5	Brent Forrester	9	7.326667
26	Justin Spitzer	9	7.214444
16	Greg Daniels	9	7.375556
20	Jennifer Celotta	8	8.168750
9	Charlie Grandy	7	6.841429
34	Michael Schur	7	7.940000
0	Aaron Shure	6	7.263333

Next, plot the data using a bar graph to show average viewers per writer. As can be seen in the following plot, the average US viewer counts do not drastically change according to which writer wrote the episode. While Mindy Kaling has written the most episodes, Paul Lieberstein has written episodes that had the highest average US viewer counts, 8.6 million viewers.



From the graph above, there doesn't seem to be much of a difference or any

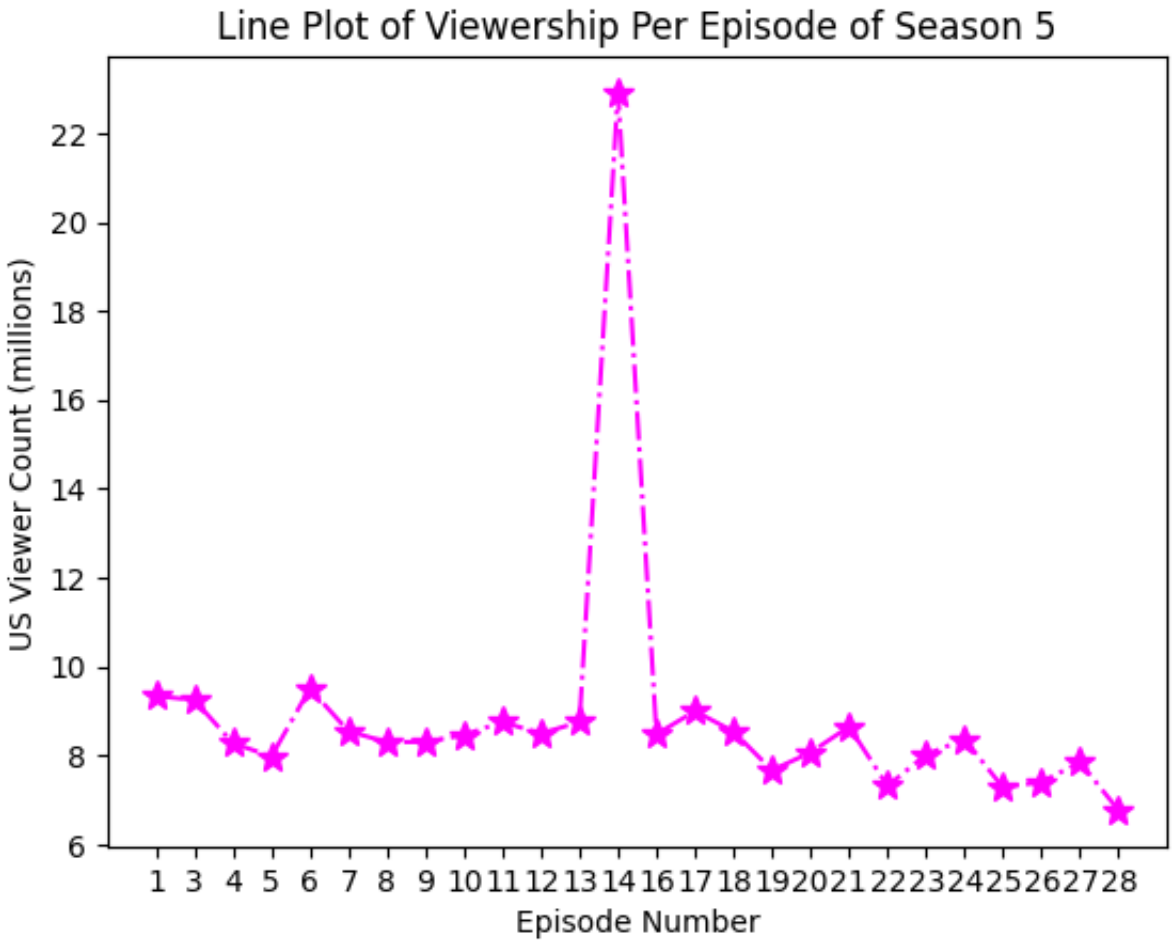
pattern in viewing rates and writers. The top 10 writers all have average viewer amounts around 7-8 million.

Viewership Throughout a Single Season

Next, I will analyze the season with the highest average viewers. First, I will find the season with the highest average viewers.

```
Season with highest average viewers:  5
```

Next, I will use a line graph to visualize how viewership fluctuates throught Season 5's run. I am interested in seeing if viewing rates are higher in the beginning and end of the season as opposed to the middle portion.



According to the line plot above, there is one very obvious outlier. Episode 14/15, "Stress Relief" had over 22 million views. This is likely the reason season 5 had the highest average viewer ratings. Other than this outlier, the distribution of viewing rates throughout the season doesn't seem to have any obvious patterns and the majority of episodes have a viewing rate of around 8-10 million.

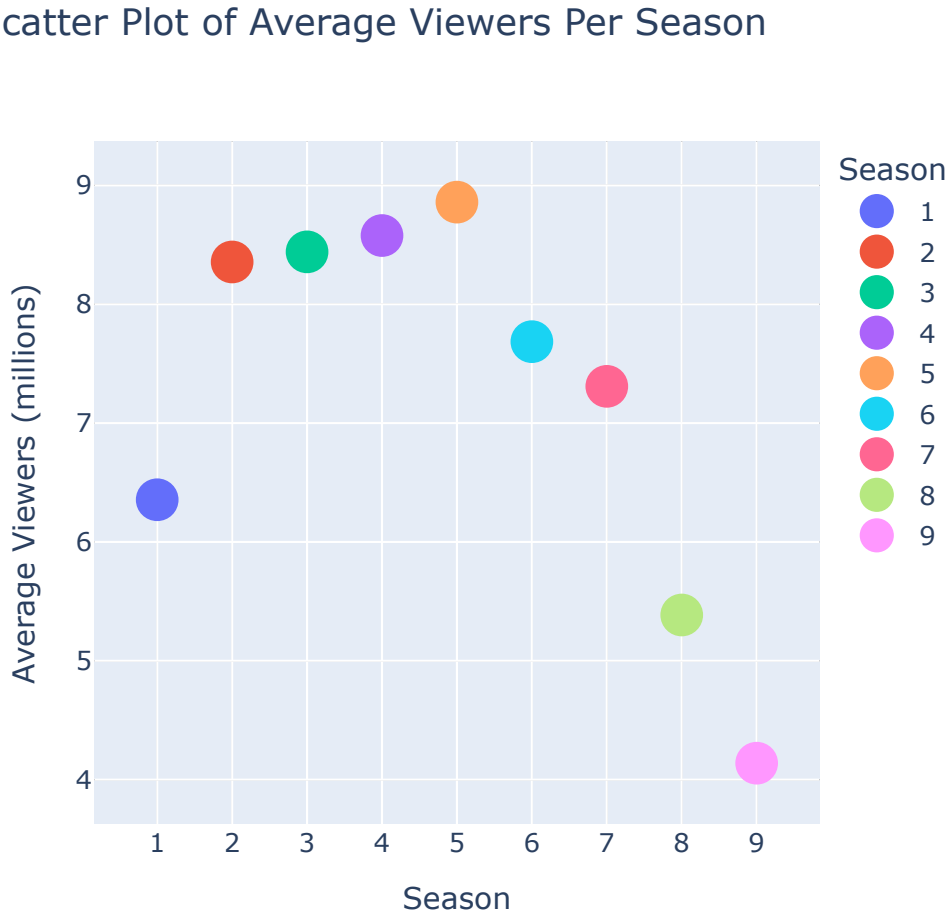
Average Viewers Per Season

From the analysis above, I have already created a dataframe that groups according to season and finds the average amount of viewers per season. We will use this same dataframe for this analysis.

Out[8]:

	Season	avg_viewers
0	1	6.355000
1	2	8.355455
2	3	8.440435
3	4	8.577857
4	5	8.858846
5	6	7.685000
6	7	7.309167
7	8	5.385417
8	9	4.138696

Next, plot the average viewers per season using a scatter plot. The reason why I decided to plot this data comparison was to see if there was a significant difference between the earlier seasons and the later seasons. Specifically, I wanted to see how much the viewing rates dropped after Steve Carell left the show in season 7.



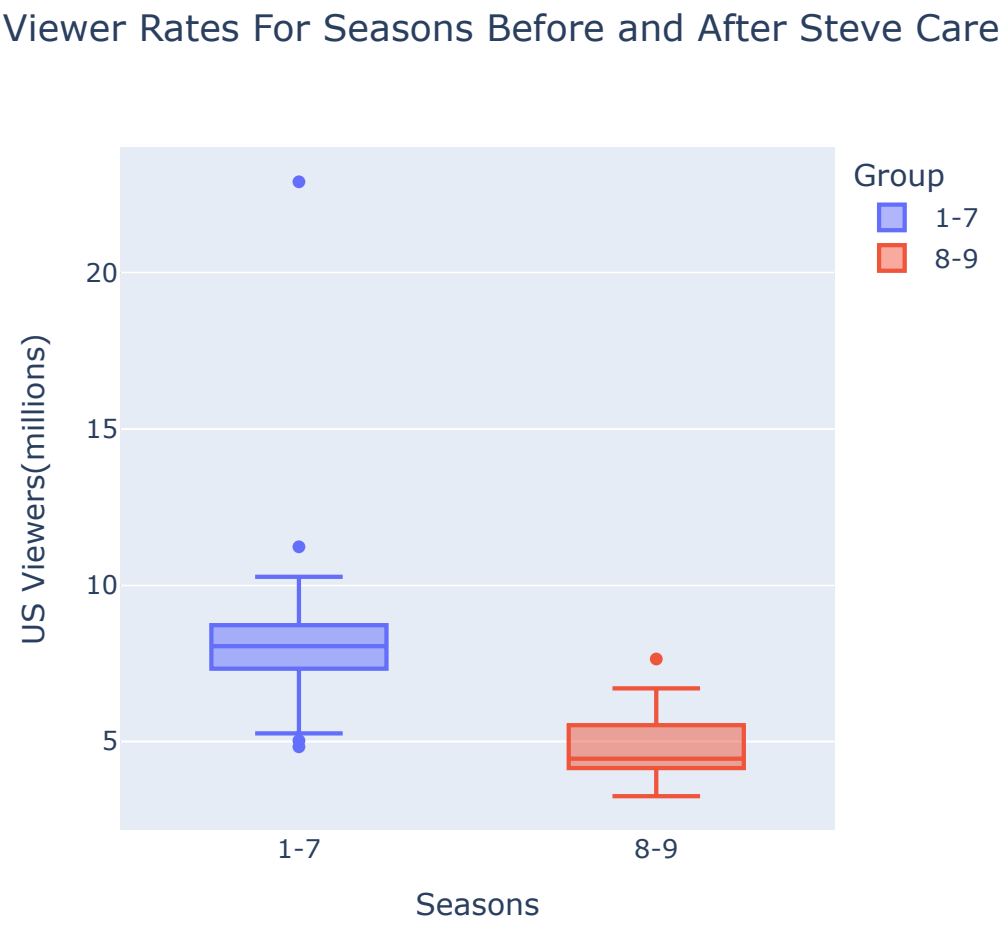
Clearly, the scatter plot shows that there was a signicant decrease in viewers once the main character, Michael Scott, had left the show permanently in season 8. I will perform a statistical analysis test next on this difference to see if there is any statistical evidence to back up what we can see visually in the plot.

Is there evidence that the amount of viewers in seasons 1-7 of The Office (US) differ from the amount of viewers in seasons 8-9?

First, I will combine seasons 1-7 and combine seasons 8-9.

```
/tmp/ipykernel_1433455/3833611096.py:3: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_1433455/3833611096.py:4: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

Now, I will plot box plots for each grouping of seasons to visualize the data distribution before performing an analysis.



From the box plots above, it does visually appear that there were more viewers for seasons 1-7 than there were for seasons 8-9. The mean, standard deviation, and sizes of each dataframe is shown below. These will be used in the following t-Test to be more sure about the difference in viewer counts.

```
The following shows the mean, standard deviation, and counts respective
ly for each group of seasons:
```

Out[12]: `{'Seasons 1-7': [8.1, 1.63, 139], 'Seasons 8-9': [4.78, 0.94, 47]}`

t-Test to Analyze Whether Steve Carell's Absence from The Office caused a Decrease in US Viewership

Population 1 will be Seasons 1-7 (seasons with Steve Carell) and population 2 will be Seasons 8-9 (seasons without Steve Carell).

First, find the test statistic:

Out[13]: 17.143158132871296

Find degrees of freedom and p-value:

```
Degrees of freedom:  139.28649698605258
P-value:  0.0
```

The p-value is extremely small, 0.0, meaning there is evidence that there was a significant decrease in viewership in the seasons after Steve Carell left the show. This conclusion backs up my original observation that the viewing rates decreased drastically in the later two seasons, likely due to this absence.

Follow Up

Follow up analysis ideas: If I had access to more data involving the actual ratings of each episode rather than just the viewing counts, I would want to do more analysis on the relationship between viewership and ratings.

I also think it would be interesting to have data regarding the demographics of the viewers and analyze if a certain age group or a certain location is more likely to watch the show.

Another analysis that would be intersting to see is comparing The Office (US) with the original The Office (UK) and see the differences in viewing rates.