Data Analysis in The Office

Dogan the Analyst

2025-01-10

```r
library(tidyverse)
```

Project Description

The purpose of this project is to use `ggplot` and its various geoms to answer questions about a data set by creating meaningful and aesthetically pleasing visualizations.

In particular, we will be analyzing data relating to the TV show *The Office*, which, as everyone knows, is the best show of all time. You can import the data set by executing all three commands in the following code chunk. (The data set is a compilation of the ones found here and here.)

```r
office_ratings <- readr::read_csv('https://raw.githubusercontent.com/jafox11/MS282/main/office_ratings.c
office_ratings$season <- as.character(office_ratings$season)
office_ratings$air_date <- as.Date(office_ratings$air_date, "%m/%d/%Y")
```

Data

Let's explore our data:

```r
glimpse(office_ratings)
```

```
## Rows: 186
## Columns: 7
## $ season      <chr> "1", "1", "1", "1", "1", "1", "2", "2", "2", "2", "2", "2"~
## $ episode     <dbl> 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1~
## $ title       <chr> "Pilot", "Diversity Day", "Health Care", "The Alliance", "~
## $ viewers     <dbl> 11.20, 6.00, 5.80, 5.40, 5.00, 4.80, 9.00, 7.13, 8.30, 7.6~
## $ imdb_rating <dbl> 7.6, 8.3, 7.9, 8.1, 8.4, 7.8, 8.7, 8.2, 8.4, 8.4, 8.2, 8.2~
## $ total_votes <dbl> 3706, 3566, 2983, 2886, 3179, 2852, 3213, 2736, 2742, 2713~
## $ air_date    <date> 2005-03-24, 2005-03-29, 2005-04-05, 2005-04-12, 2005-04-1~
```

The data set contains 186 observations and 7 variables.

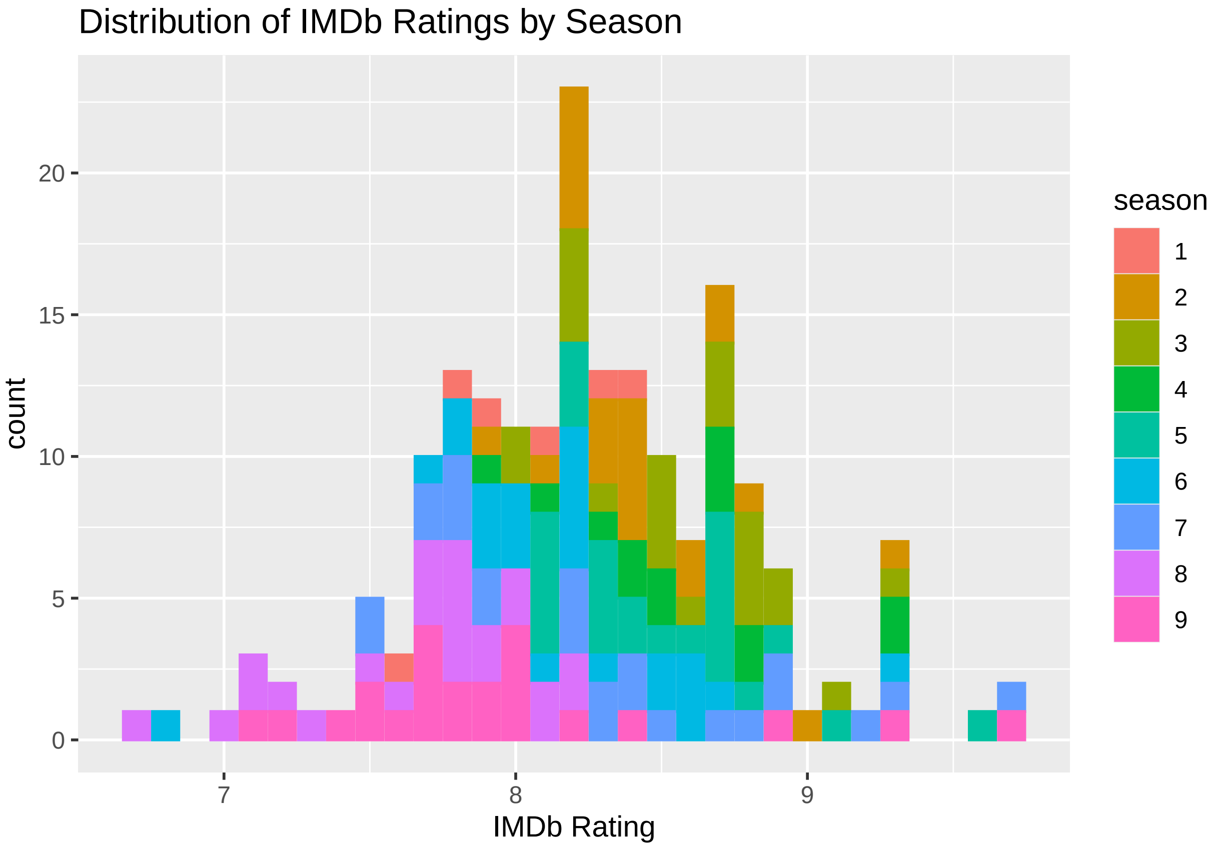
```r
tibble(office_ratings)
```

```
## # A tibble: 186 x 7
##    season episode title             viewers imdb_rating total_votes air_date  
##    <chr>    <dbl> <chr>               <dbl>       <dbl>       <dbl> <date>    
##  1 1            1 Pilot                11.2         7.6        3706 2005-03-24
##  2 1            2 Diversity Day         6            8.3        3566 2005-03-29
##  3 1            3 Health Care           5.8          7.9        2983 2005-04-05
##  4 1            4 The Alliance          5.4          8.1        2886 2005-04-12
##  5 1            5 Basketball            5            8.4        3179 2005-04-19
##  6 1            6 Hot Girl              4.8          7.8        2852 2005-04-26
##  7 2            1 The Dundies           9            8.7        3213 2005-09-20
##  8 2            2 Sexual Harassment     7.13         8.2        2736 2005-09-27
##  9 2            3 Office Olympics       8.3          8.4        2742 2005-10-04
## 10 2            4 The Fire              7.6          8.4        2713 2005-10-11
## # i 176 more rows
```

We have three continuous variables: `viewers`, `imdb_rating`, and `total_votes`. I will start data visualization with these.

Distribution of IMDb Ratings by Season

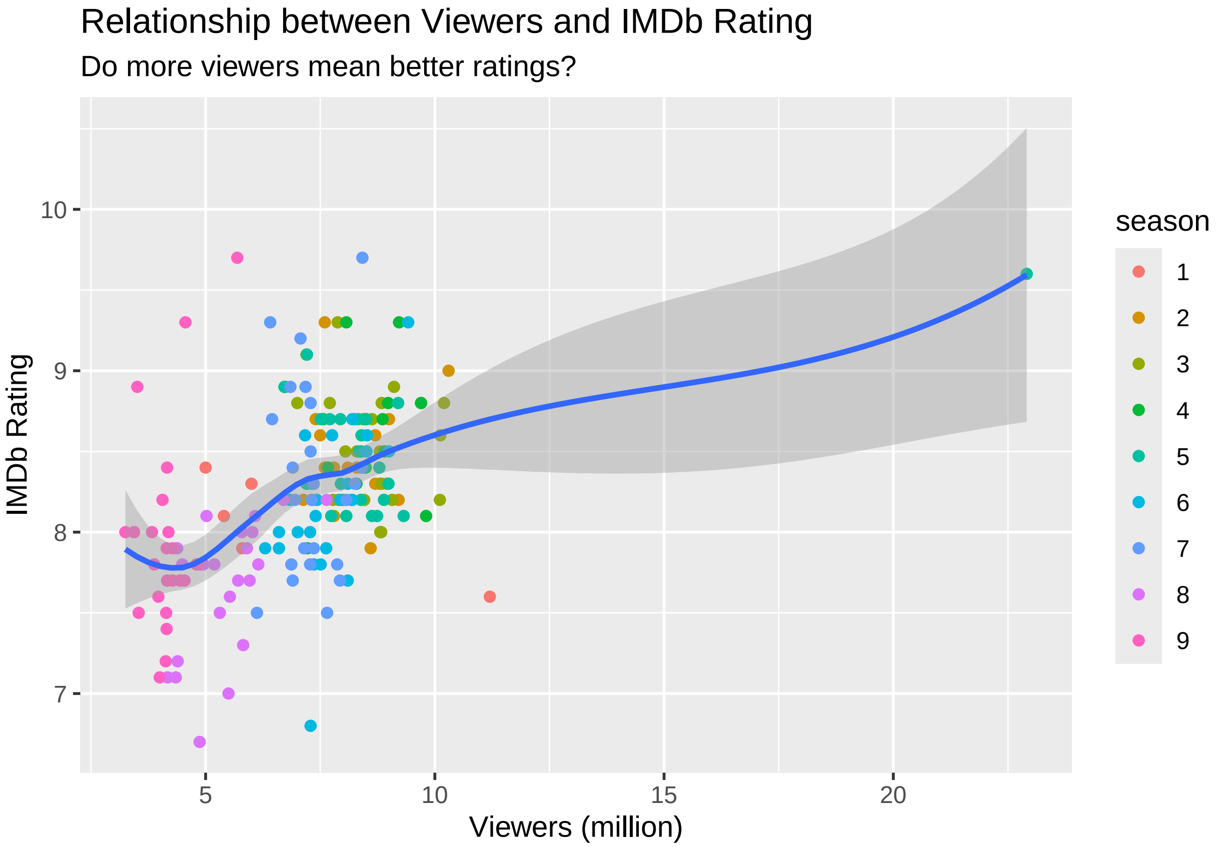
```
office_ratings %>%
  ggplot() +
  geom_bar(mapping = aes(x = imdb_rating, fill = season, color = season)) +
  labs(title = "Distribution of IMDb Ratings by Season",
       x = "IMDb Rating")
```



Based on the figure above, it can be said that Season 2 and Season 3 were the best seasons of the series in terms of the distribution of IMDb ratings. The eighth and ninth seasons are frankly not as good as the early seasons.

Relationship between Viewers and IMDb Rating

```
office_ratings %>%
  ggplot() +
  geom_point(mapping = aes(x = viewers, y = imdb_rating, color = season)) +
  geom_smooth(mapping = aes(x = viewers, y = imdb_rating)) +
  labs(title = "Relationship between Viewers and IMDb Rating",
       subtitle = "Do more viewers mean better ratings?",
       x = "Viewers (million)",
       y = "IMDb Rating")
```

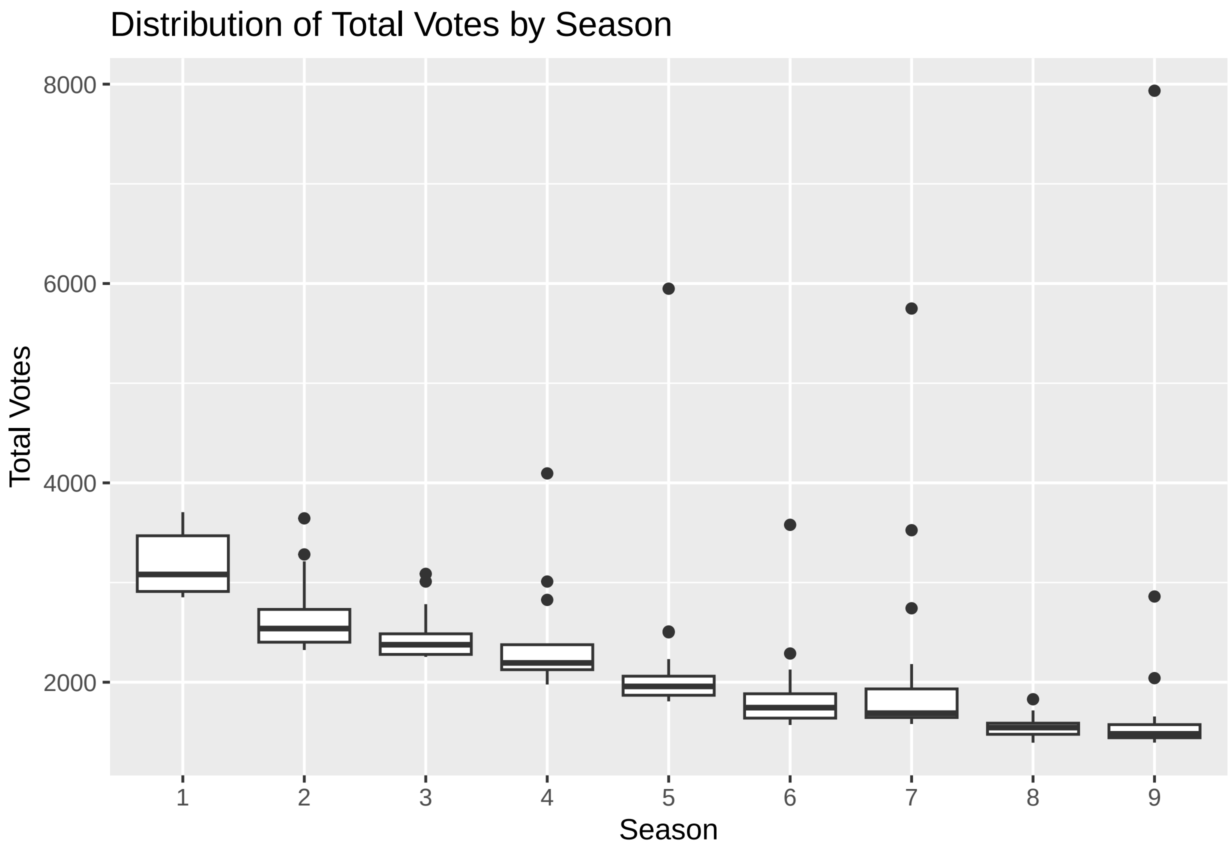


This figure shows a trend curve. Looking at it, we can assume that more viewers result in higher ratings. However, when examining the points on the plot, there are some points that align well with the IMDb ratings. The highest ratings are mostly between 5-10 million viewers.

Distribution of Total Votes by Season

On the other hand, the plot below also informs us that the popularity of the series is decreasing by season. If you take a closer look, you can see the outline, median, Q1 and Q3 values. For example, although first season's median is greater than fifth, seventh and ninth seasons, 8000 people voted for ninth season. But it doesn't mean that the ninth season was fans' favorite one.

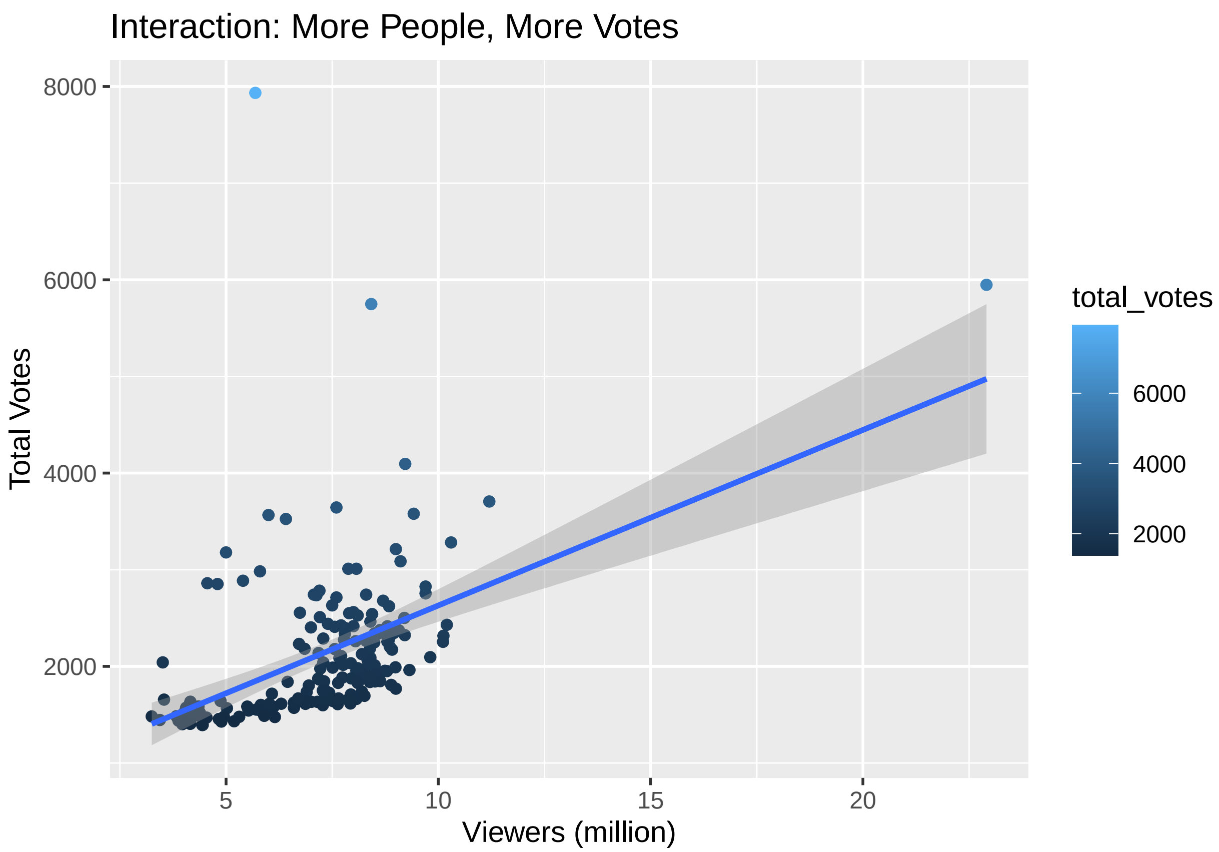
```
office_ratings %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = season, y = total_votes)) +
  labs(title = "Distribution of Total Votes by Season",
       x = "Season",
       y = "Total Votes")
```



Interaction: More People, More Votes

Let's learn "The more people watch an episode, the more people leave an IMDb rating?":

```
office_ratings %>%
  ggplot() +
  geom_point(mapping = aes(x = viewers, y = total_votes, color = total_votes)) +
  geom_smooth(mapping = aes(x = viewers, y = total_votes), method = "lm") +
  labs(title = "Interaction: More People, More Votes",
       x = "Viewers (million)",
       y = "Total Votes")
```

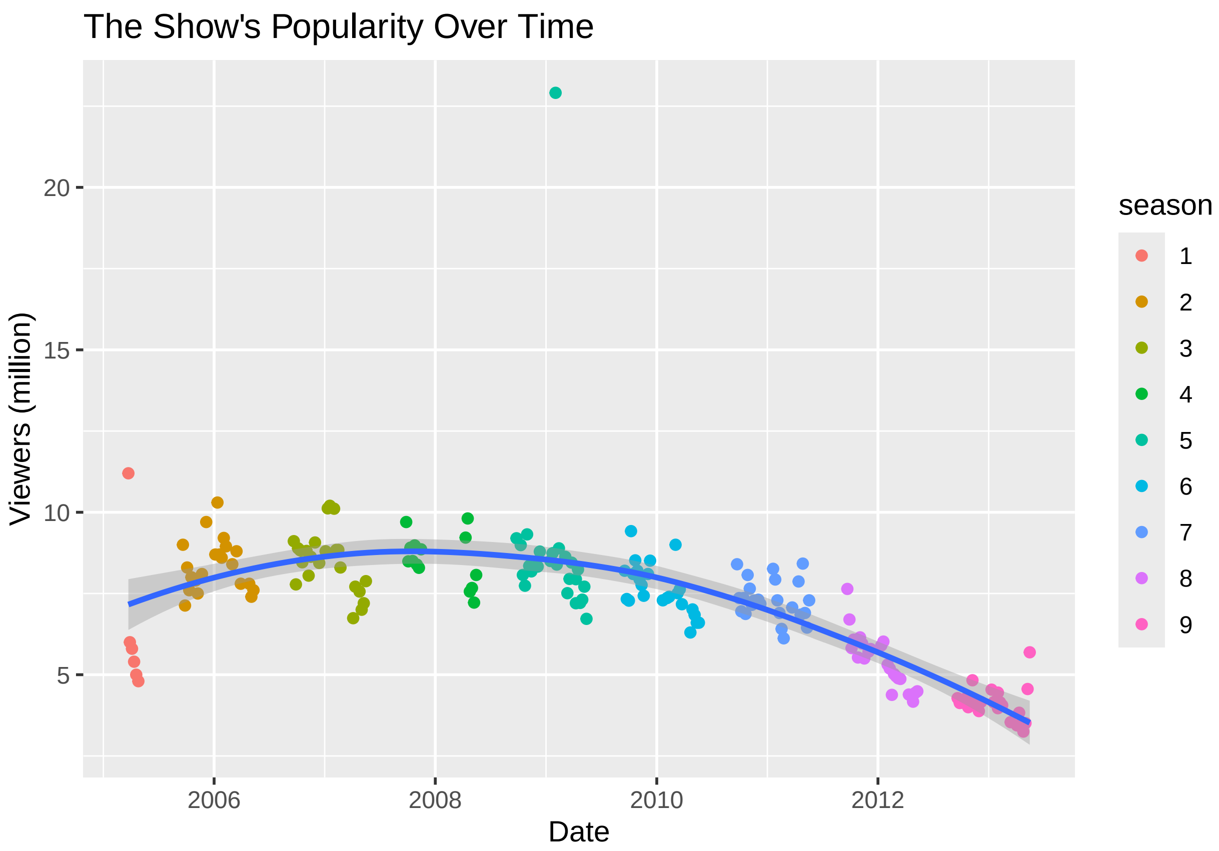


Looking at the line, one might say 'Yes,' but there are some exceptions. For example, 6 million viewers result in 8000 votes, which is higher than the number of votes from episodes with over 20 million viewers.

The Show's Popularity Over Time

The next question is how did change *The Office*'s popularity over time. To answer it, we can create another plot by using `air_date`, and `viewers`.

```
office_ratings %>%
  ggplot() +
  geom_point(mapping = aes(x = air_date, y = viewers, color = season)) +
  geom_smooth(mapping = aes(x = air_date, y = viewers)) +
  labs(title = "The Show's Popularity Over Time",
       x = "Date",
       y = "Viewers (million)")
```

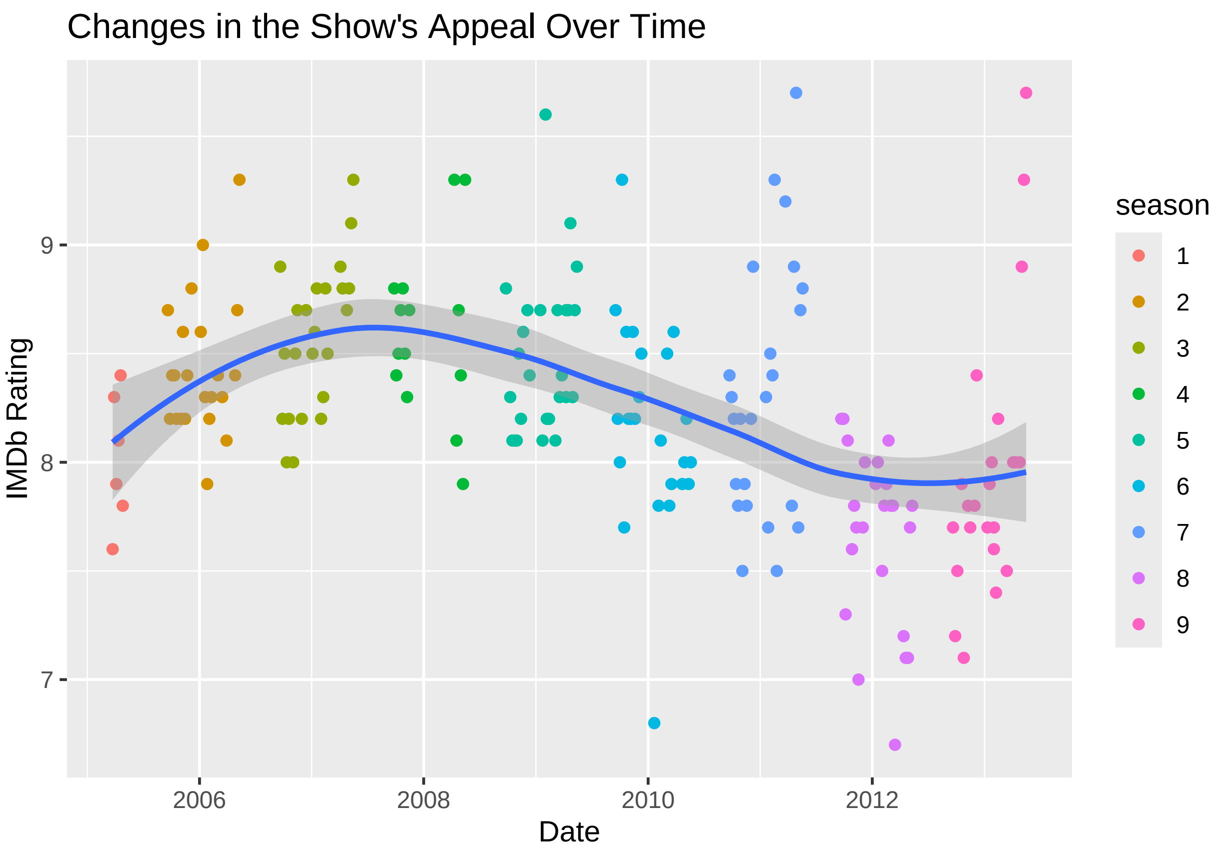


It seems the show's popularity has decreased since its peak in early of both 2007 and 2008. During that time period, *The Office* averaged 8-9 million viewers, not below 6 million. Again, while outliers exist, we are speaking in general terms.

Changes in the Show's Appeal Over Time

The previous plot focused on popularity, but this one will center on appeal. Therefore, `imdb_rating` and `air_date` will be used.

```
office_ratings %>%
  ggplot() +
  geom_point(mapping = aes(x = air_date, y = imdb_rating, color = season)) +
  geom_smooth(mapping = aes(x = air_date, y = imdb_rating)) +
  labs(title = "Changes in the Show's Appeal Over Time",
       x = "Date",
       y = "IMDb Rating")
```



The ratings in the first years were above 7.5, but over time, they dropped below 7. Of course, we can still observe some outliers in the figure.

Conclusion: Popularity vs. Appeal

The difference between popularity and appeal can depend on several factors, such as episodes with special guests, filler episodes, episodes where beloved characters (like Michael Scott) said goodbye to the show, finales, etc. As a result, both the number of viewers and the ratings increased in the early years (2005-2008) of the show and it can be considered its golden years. However, in the following years, the show lost its audience.

Now let's take a look at the appeal plot. We can see the increase in IMDb ratings during the same period. However, if you look at the ratings, you can see that the show still maintained ratings of 8+ or higher throughout the later years. While the popularity of the show declined over the years, it seems that the appeal of the show was not directly affected by this decline.

After all, some episodes managed to get high ratings with fewer viewers, showing that not every popular episode will have a high IMDb score, and not every highly rated episode will be very popular.