Module 3: Design Workflow

Casey Paul

Download this PDF under 4602 Module Assignment: <https://caseyjpaul.github.io/>

**Precondition**

1. One of my favorite and go-to shows when I’m bored is *The Office*. Something that I find interesting about this show is that it is so popular among teenagers and young adults today even though it first aired in 2005. The problem I want to solve is seeing if the show was as popular and great as people think it is nowadays.
2. The dataset I will be using to answer this problem is from [Kaggle](https://www.kaggle.com/nehaprabhavalkar/the-office-dataset). I chose to use a dataset on the show *The Office* because I have never analyzed data from a show before and I wanted to see what I could find out using this dataset. The dataset includes rows for every episode from each season, and is listed in order of release date. We can also see the rating, votes, viewership, and duration for each episode in separate columns. Bias this dataset might include is in the column viewership. This column contains the number of views the episode got when it was released, however, *The Office* was on Netflix for quite awhile, which will affect its number of viewers. Each episode's viewership is probably way higher than the dataset shows because of show reruns and streaming services. I am going to see what I can find out about *The Office* using the rating, viewership, and duration columns. Note that the ‘Viewership’ column is the number of viewers in the USA (in millions).

**Core**

1. Task 1: Does the number of views increase over the duration of the show being aired?

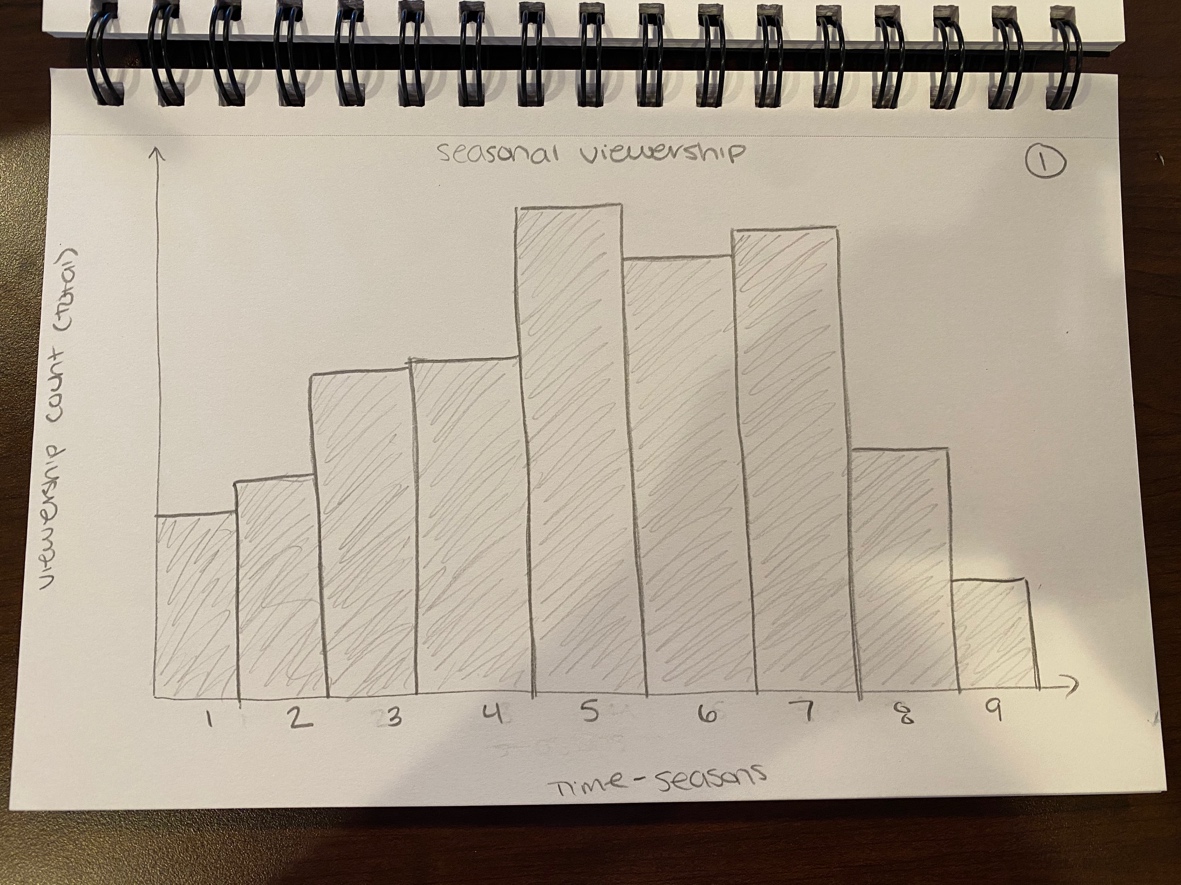
* Who: I will be executing the task.
* What: This task seeks to learn if the numbers of views per episode changed as the show progressed and aired more seasons over time.
* When: This task is performed after reading in the data and creating a dataframe with just the Date, Season, Episode Title, and Viewership columns. Then use Matplotlib to plot the data.
* Where: The task operates using the Date, Season, Episode Title, and Viewership data.
* Why: This task is pursued because it is important to know if your show is gaining or losing viewers overtime, which could ultimately affect how long the show airs.
* How: This task will be conducted using Python, Matplotlib, and possibly Altair in Jupyter Notebooks. I will look at the data and filter it to only include the data needed to complete this task.

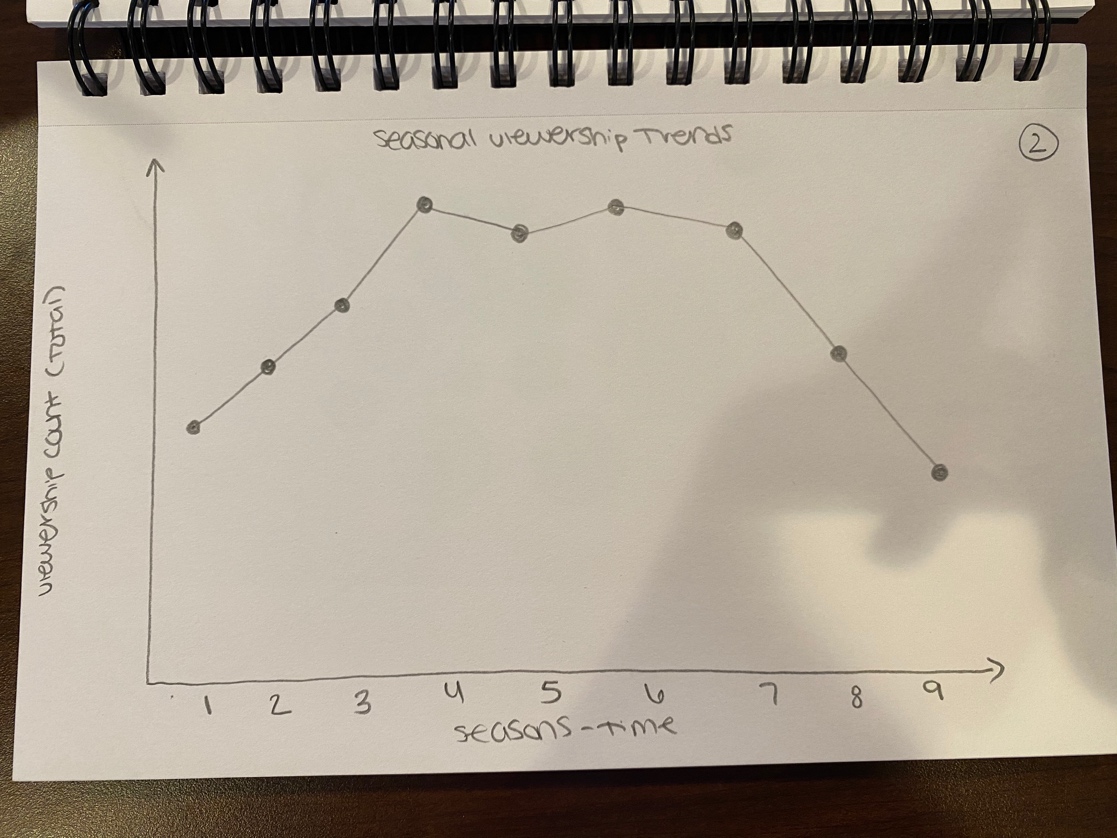
Task 2: Does the number of views correlate with the number of rating votes an episode will get?

* Who: I will be executing this task.
* What: this task seeks to learn if the numbers of views per episode affects the number of rating votes it receives.
* When: This task can be performed two ways. One way is to read in the data and create a dataframe with just the Date, Episode Title, Votes, and Viewership columns. Then use Matplotlib to plot the data. Another way to do this is to read in the data and use Altair to specify which columns to plot.
* Where: The task operates using the Date, Episode Title, Votes, and Viewership data.
* Why: This task is pursued because TV shows want as many viewers as they can get, and they also want the highest rating they can get. The more viewers a show has the higher chance for lower ratings, which could bring down a shows overall rating.
* How: This task will be conducted using Python, Matplotlib, and possibly Altair in Jupyter Notebooks. I will look at the data and filter it to only include the data needed to complete this task.

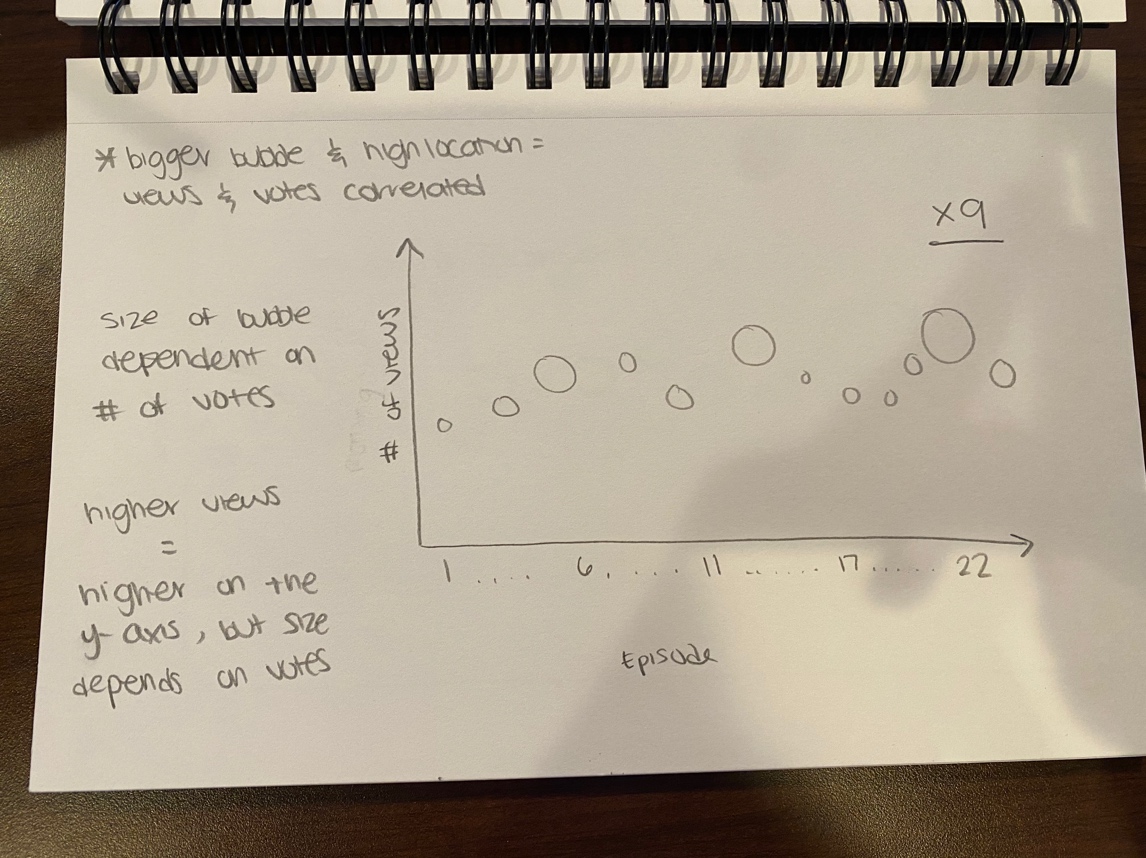
1. Design

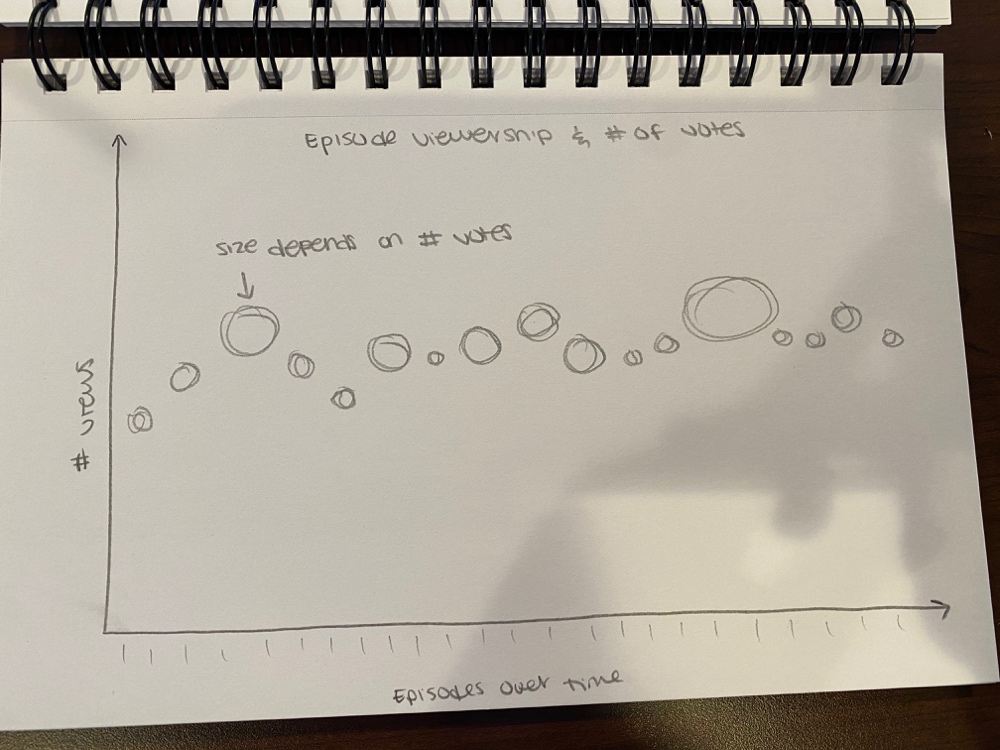
Task 1





Task 2



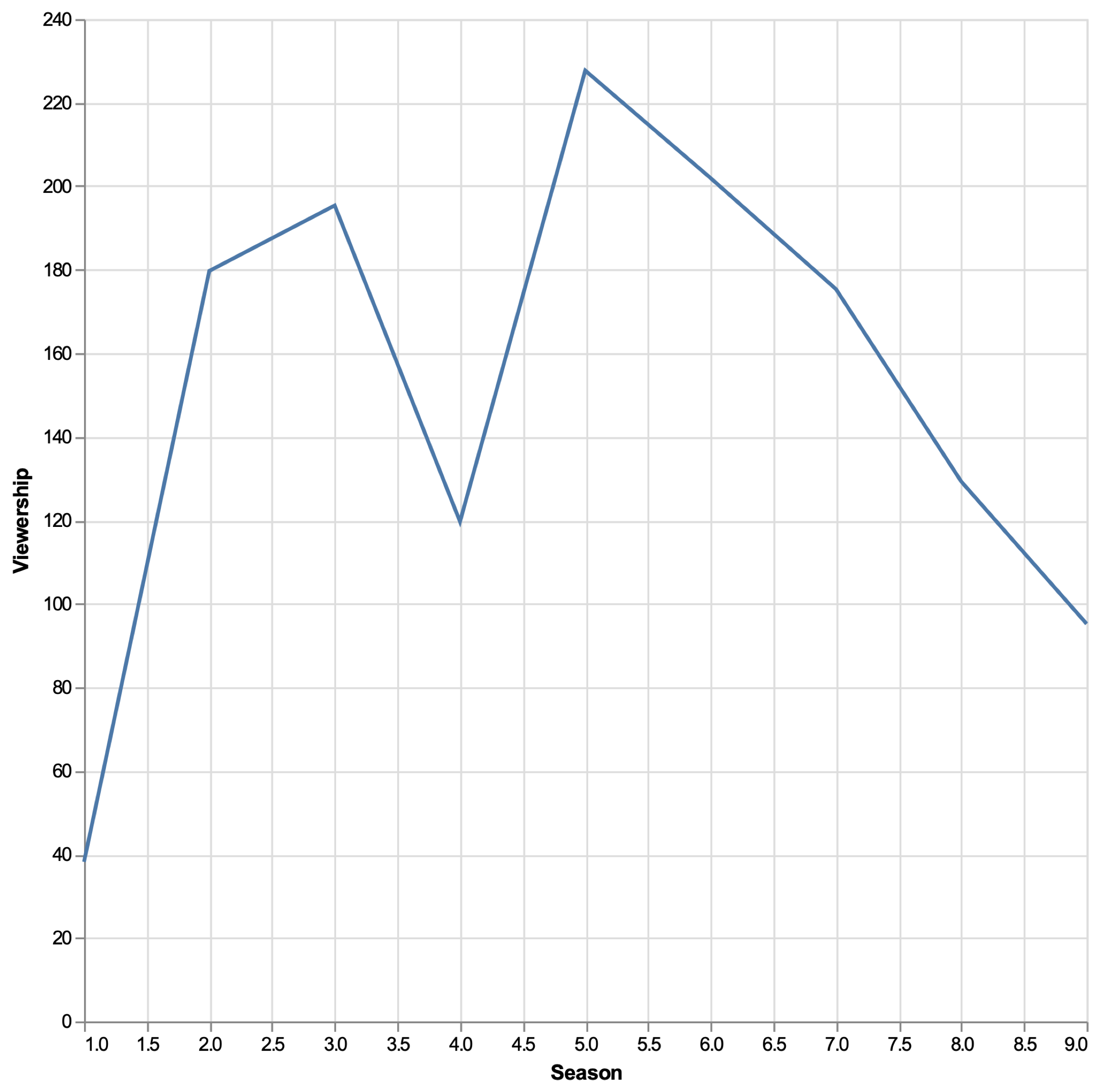


My first iteration of task one was a bar chart, but it was only until after I finished that I realized I wanted to look at the viewership trend, and trends are commonly represented as line graphs. I then sketched out what I would want each seasons total viewership to look like as a line graph.

For task two, I made two charts of a similar style and tried to make both to see which I liked better. I was unable to figure out how to make the second one, so the first iteration is what stuck. I felt like these two visuals represented my target tasks very well.

1. **Implementation**

Task One

****

The solution to task one is shown above. I implemented this using Altair in Jupyter Notebooks. My first iteration of this was a bar chart for each singular episode. I wanted to do each season instead so I found the sum of views for each season and plotted that as a line graph. Line graphs usually indicate a trend over time, and since seasons air over time, I figured this was a good way to show viewership over time.

My solution to task two was a bubble chart showing the number of views for each episode where the bubble size depended on the number of votes. If the bubble was higher on the chart that meant it had more views, and if the bubble way big it meant it had more votes, signifying some sort of correlation between the two.

These address my core tasks by answering my task questions, which would ultimately let a TV show producer (or whoever is in charge of how long shows air) know how well the show is doing.

1. **Deploy**

I have deployed these visuals to my GitHub page and can be found at this link:

* <https://caseyjpaul.github.io/4602%20Mod%203%20Visuals.html>

1. **Iterate**
   * One new task I would conduct now that I have had a chance to investigate my dataset would be to see if the duration of each episode changed over time (seasons) and if this effected viewership and ratings. In order to accommodate this task I would plot the views for each individual season on a separate visualization so each episode could be looked at more granularly. In Altair I could do something like make a visual for each season and display them next to each other, or I could layer them on top of one another.
   * I would also like to compare this to a dataset from Netflix or something that has the viewership of the show now. That would ultimately answer my target problem, however, the show is streamed so frequently and on so many different platforms it is hard to accurately get that data.

**Analysis**

1. My solution tells me a lot about my target problem. My target problem was to see if the show The Office was as popular and great as people think it is nowadays. Based on the visualizations, The Office did in fact get millions of views and was very popular when it first aired. Like most shows, it took a bit to get to its highest viewership number and then slowly decreased from there. One thing to note about the amount of votes the show got compared to its viewers was that there was a huge difference and did not often correlate with one another. This could be because if people did not feel strongly about a specific episode they may not have given the show a rating vote.
2. My solution tells me there are many ways to go about designing visualizations for this problem but it may be hard given that not all of the data is out there. I think visuals are a great way to solve my target problem because one could analyze numbers all day, however, looking at a visualization answers the problem at a glance and gives the audience an answer in a matter of seconds.