Module 3: Design Workflow

**Precondition**

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.

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**

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.

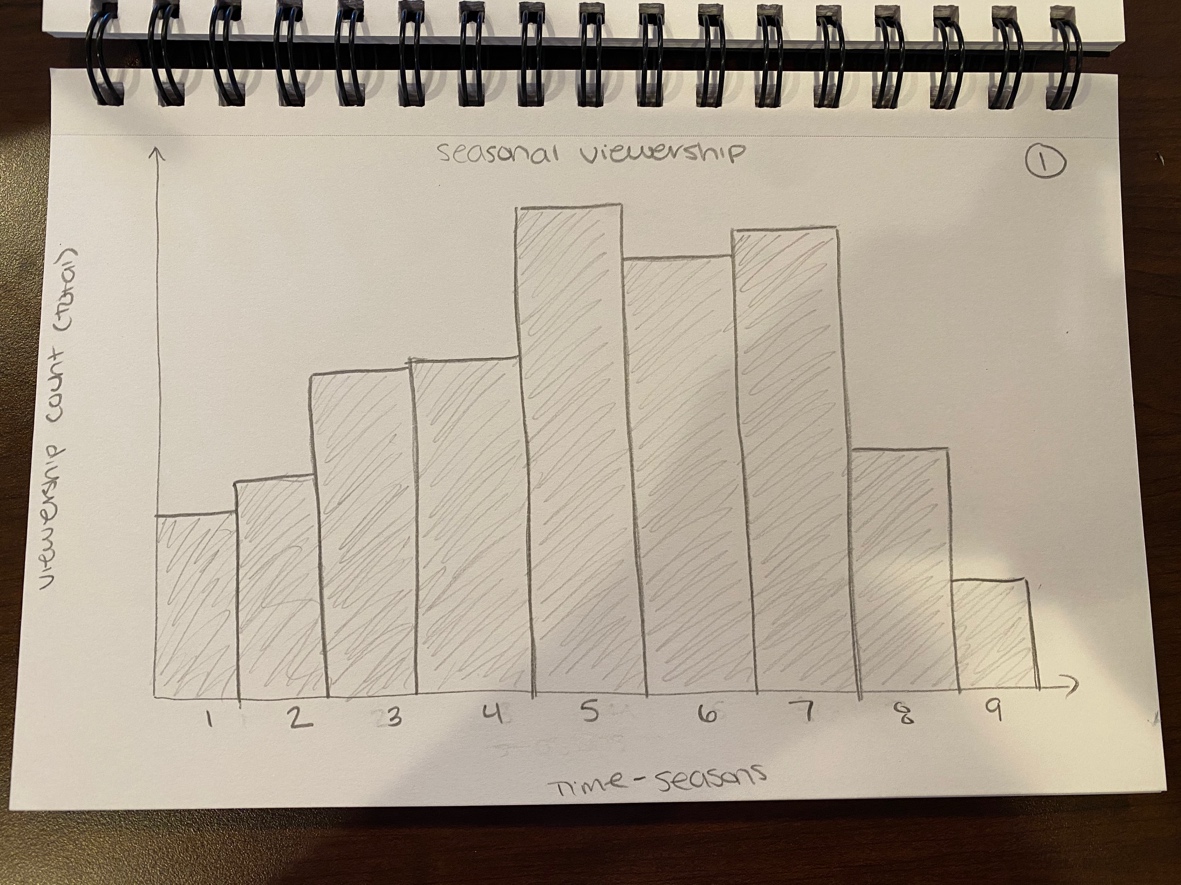
Task 2: Does the number of views correlate with the number of rating votes an episode will get?

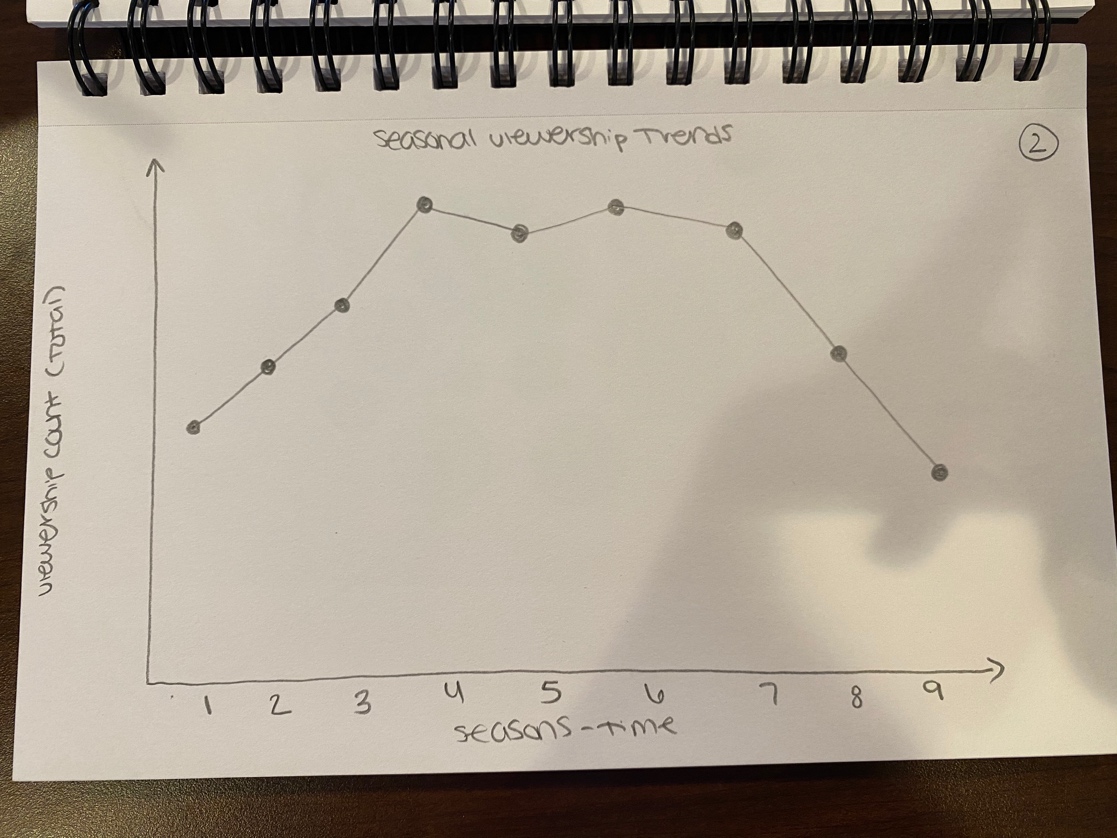
* Who: I will be executing this task.
* What: his task seeks to learn if the numbers of views per episode affects the number of rating votes it receives.
* When: This task can be performed two ways. One way is to read in the data and create a dataframe with just the Date, Episode Title, Votes, and Viewership 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.

4. *Design:* Create a sketch of your initial prototype. Note that this sketch can be drawn from your work either in Assignment 2 or in the source article, but should include some design iteration to support your target tasks. Add brief design justifications and discussions of the trade-offs of key design choices in the sketch, being sure the discussion is closely tied to your target tasks. Make sure to include a copy of the sketch in your write-up.

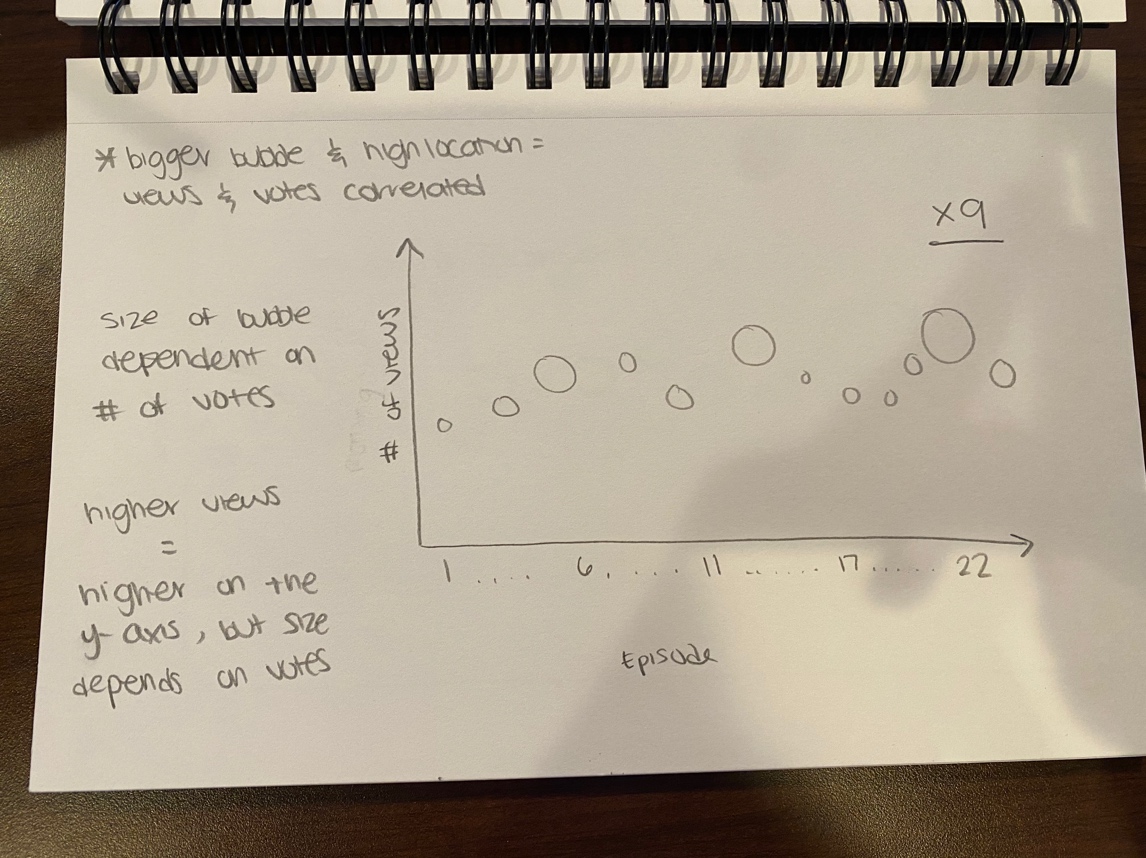
**Design**

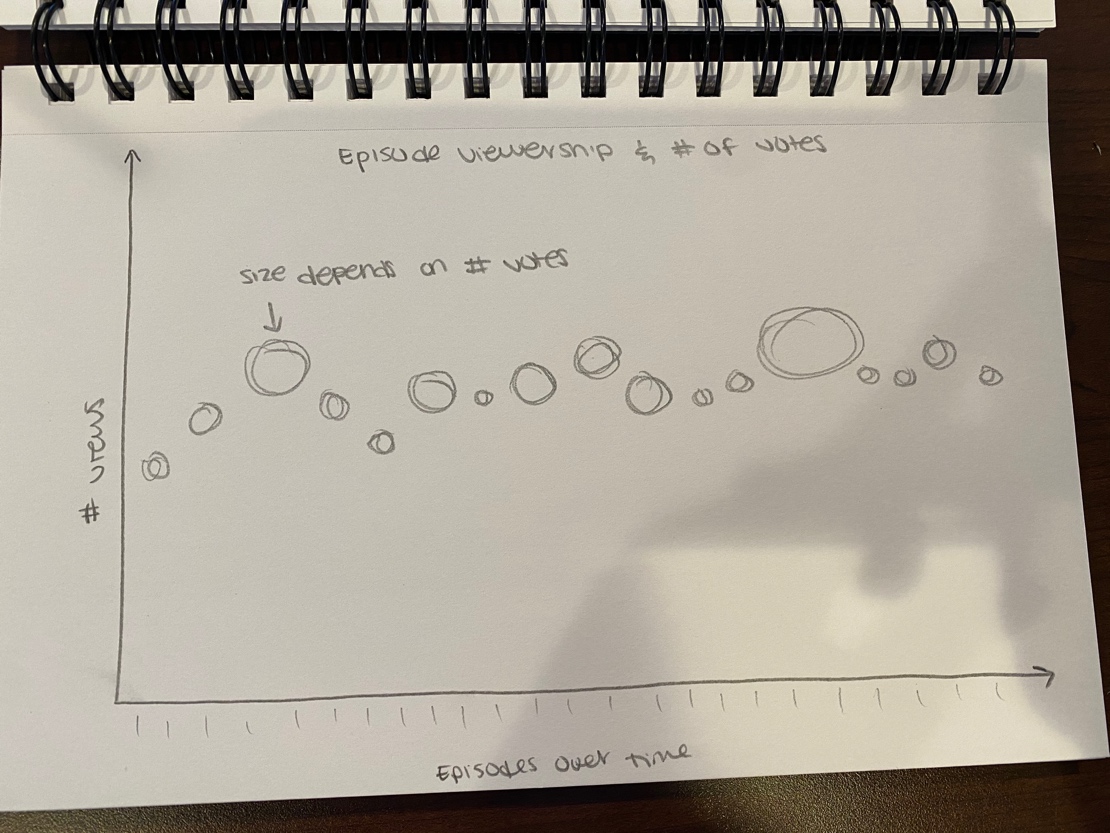
Task 1





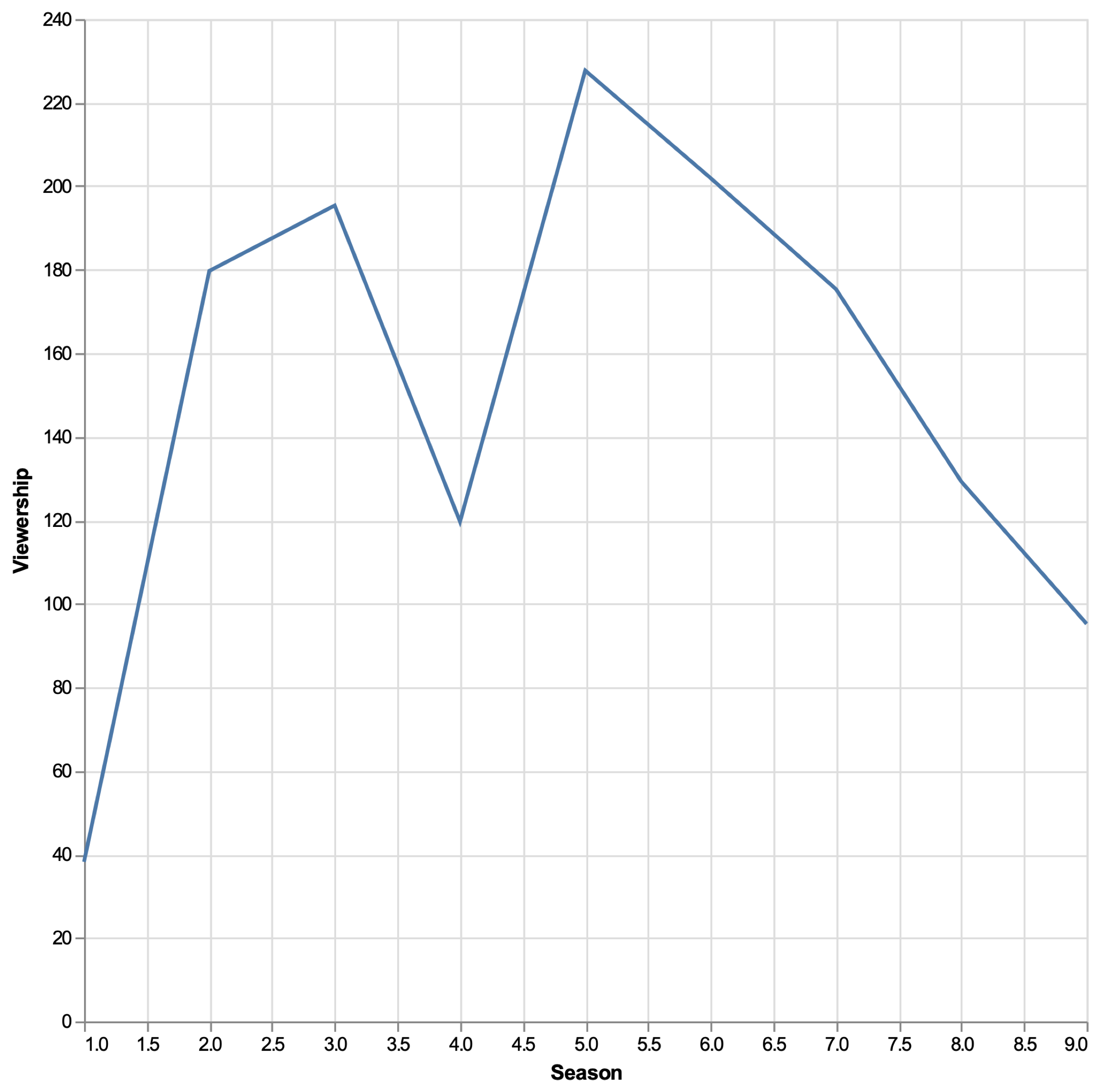
Task 2





**Implementation**

Task One

****