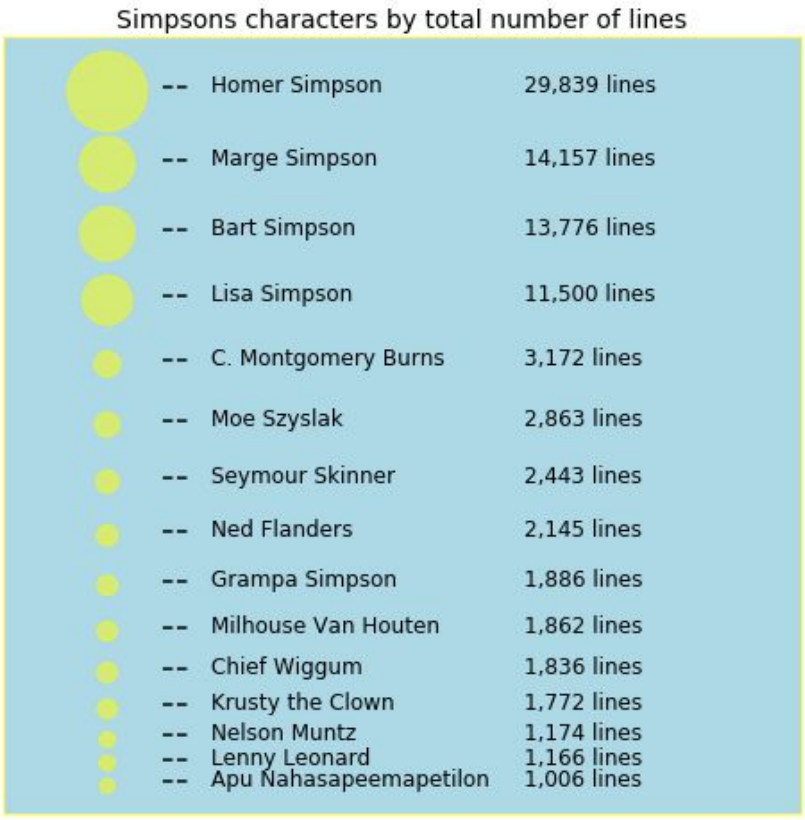
The Simpsons Meets Data Visualisation
Adam Reevesman
5/16/2019

Introduction

There are few things I love more than *The Simpsons*. With thirty seasons and over 600 episodes, the animated comedy show holds a special place in my heart. Naturally, when I discovered that I could download all of the episode scripts I could ever want (via Kaggle: https://www.kaggle.com/wcukierski/the-simpsons-by-the-data#simpsons_episodes.csv), I knew what I had to do. *The Simpsons* is one of those shows that I think about on a daily basis. Every so often, I find myself singing along when *Mr. Plow* or *Everybody Hates Ned Flanders* gets stuck in my head. Armed with access to just about any word that Homer has ever said, I couldn't resist putting on my data scientist hat in order tease out some insights and get another laugh out of one of the post prominent animated television shows of the last three decades.

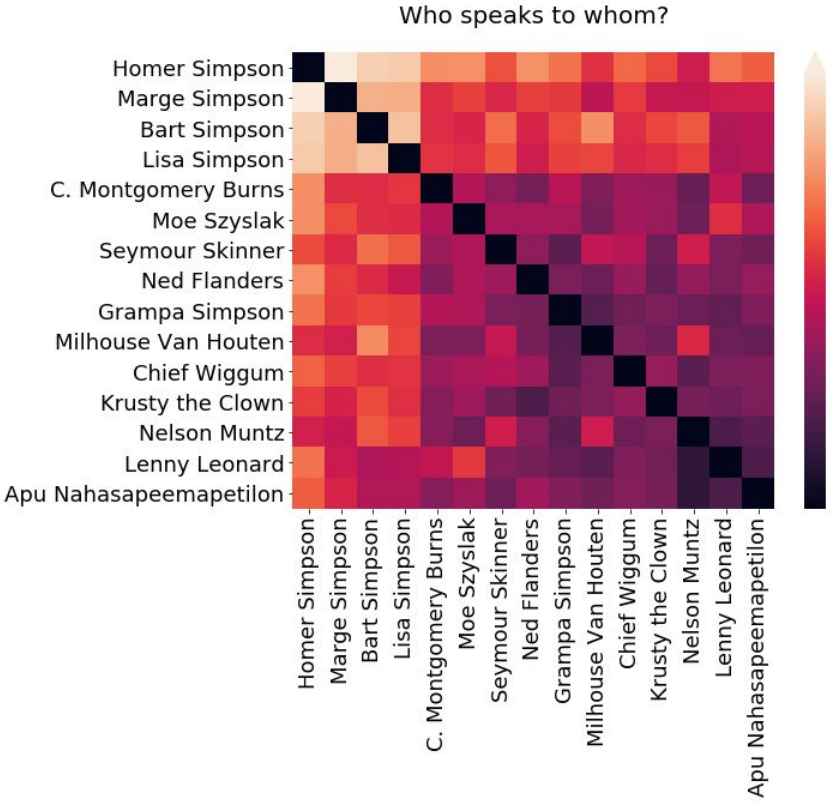
Summary and Storyline:

Let's look at the characters in the show. Below is a bubble chart that shows fifteen characters with the most lines and the total number of lines they have. Characters with larger bubbles speak more lines than characters with smaller bubbles.



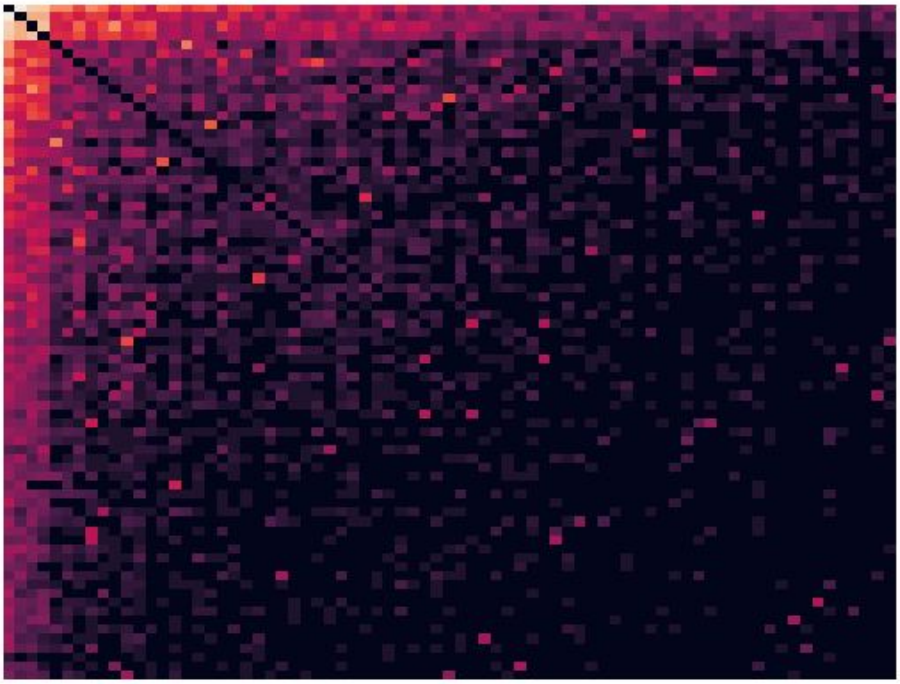
From this plot, it is clear that the main characters of the show are in fact Homer, Marge, Bart and Lisa Simpson. They have significantly more lines than the rest of the characters. It is interesting to note that even though Grampa is a member of the Simpson family and is often seen in the Simpson home, he does not speak as much as the rest of the family or even some other side characters.

Now that we've seen how much each of the characters speak in the show, let's go one step further by looking at how much the characters speak to each other. The following heat map will show this.

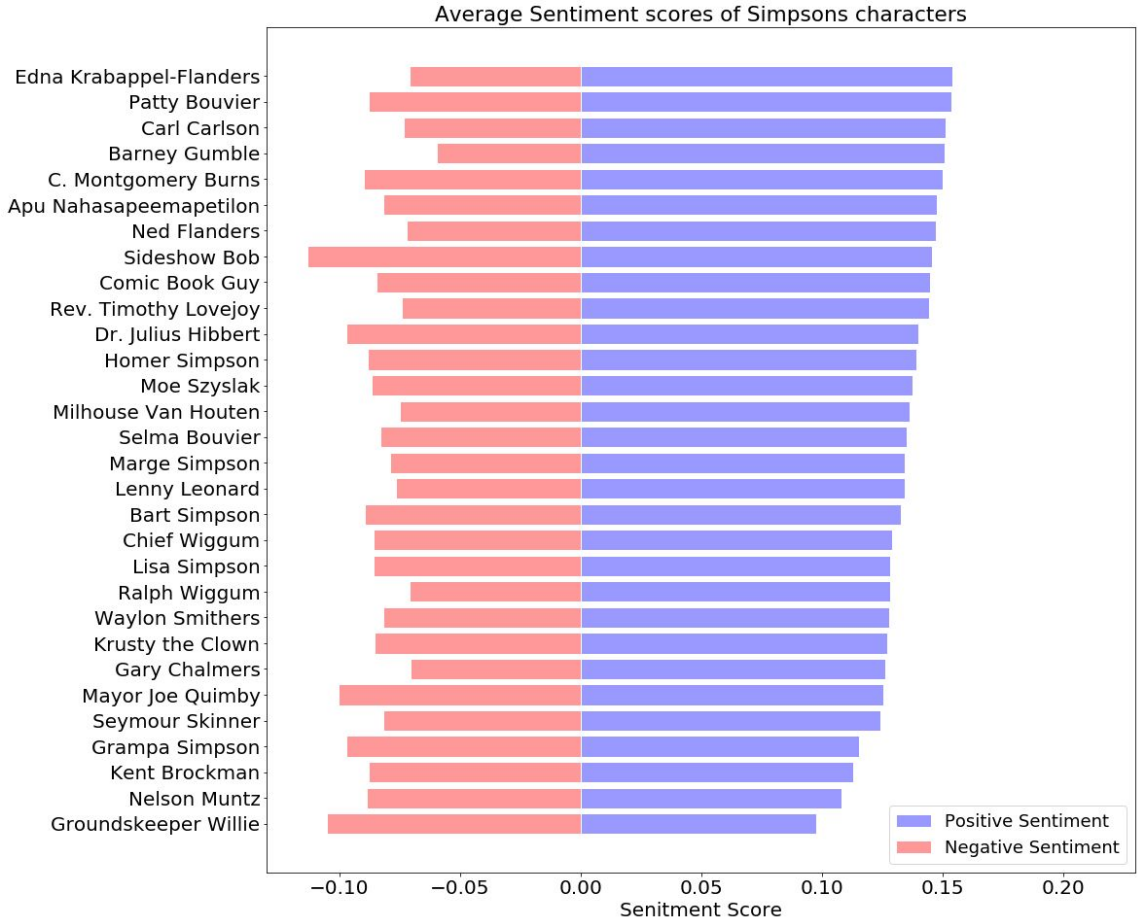


The color of each square represents the number of lines that the character in each row said to the character in each column. Lighter squares mean that more lines were spoken. The plot shows us three main kinds of interactions between the characters. First, there are many conversations internally between the Simpson family. Next, there are a medium amount of conversations between Simpson family members and the side characters. Finally, there are relatively few conversations that do not involve the Simpson family. The neat thing about this plot is that it highlights some interesting character dynamics. For instance, Krusty the Clown's run ins with the law are reflected by the large amount of conversations with Police Chief Wiggum.

I made another heat map where I included 75 characters. It makes a pretty picture, but it also shows that most conversations happen between just a few characters (i.e. the data has a “long tail”).

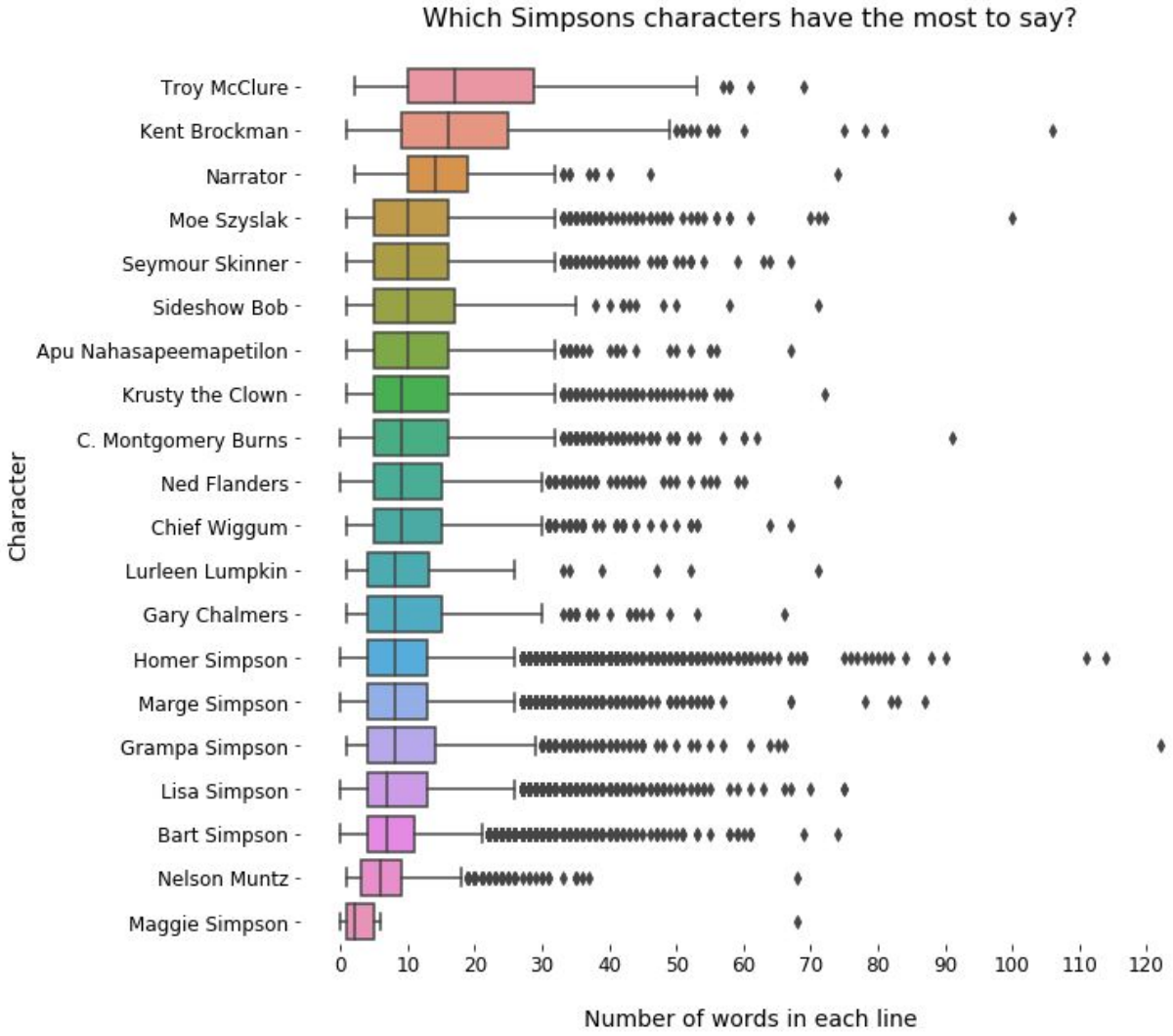


Can we get get a better idea of what these characters are saying? One way to do that is to look at the sentiment of the words they speak. We can quantify the emotion of each line using natural language processing technique, VADER Sentiment Analysis (http://datameetsmedia.com/vader-sentiment-analysis-explained/). One caveat to note is that this approach is optimized for social media text, so we should take the results with a grain of salt.



Now we can see the thirty characters with the most lines ordered by an estimation of how positive they are. It seems that Edna Krabappel (Bart's teacher) and Patty Bouvier (Marge's sister) are the most positive characters. However, I never found them to be particularly pleasant. On the other hand, it is reassuring to see to Ned Flanders near the top of the chart and Grampa, Nelson and Groundskeeper Willie near the bottom. Additionally, it is interesting that Sideshow Bob is both very positive and very negative. He is Krusty the Clown's sidekick, so he should be a positive guy, but when we remember all of his murderous plots, it is not surprising to see such negativity.

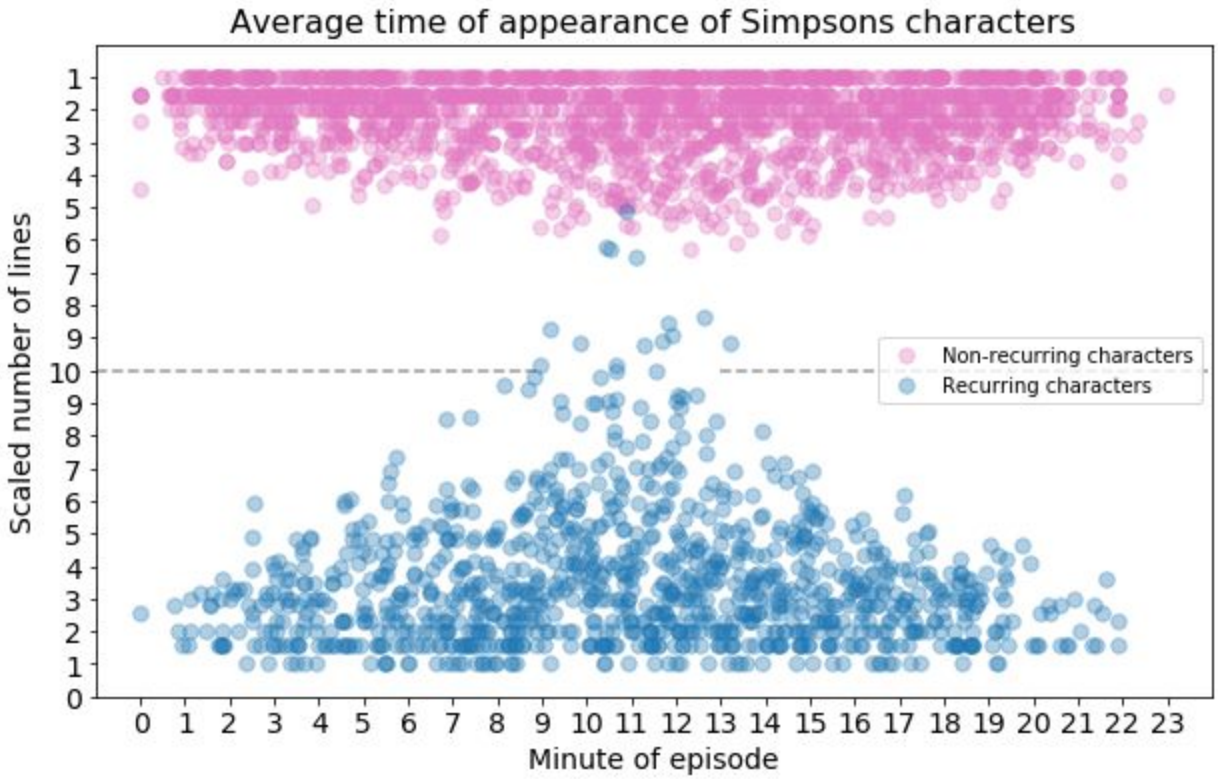
It would also be helpful to see if some characters have longer lines than others. It may be that Homer has the most lines because he is the main character, but maybe he only says a few words at a time like “Mmm donuts” or “Why you little!” The boxplots below will help us understand which characters speak the most on a per line basis.



It turns out that only a few characters have a median number of words per line that is more than ten, but many have at least a few lines where they talk for a long time. Maggie Simpson rarely speaks at all (she is in fact a baby and she is not a *Family Guy* character), but she does have one long rant. Troy McClure and Kent Brockman have the largest median number of words per line, which makes sense when we remember that they are both television stars (Kent Brockman delivers the news while Troy McClure is a washed up actor from the 70s). Grampa Simpson has the longest line of the show. I will leave it here for your enjoyment.

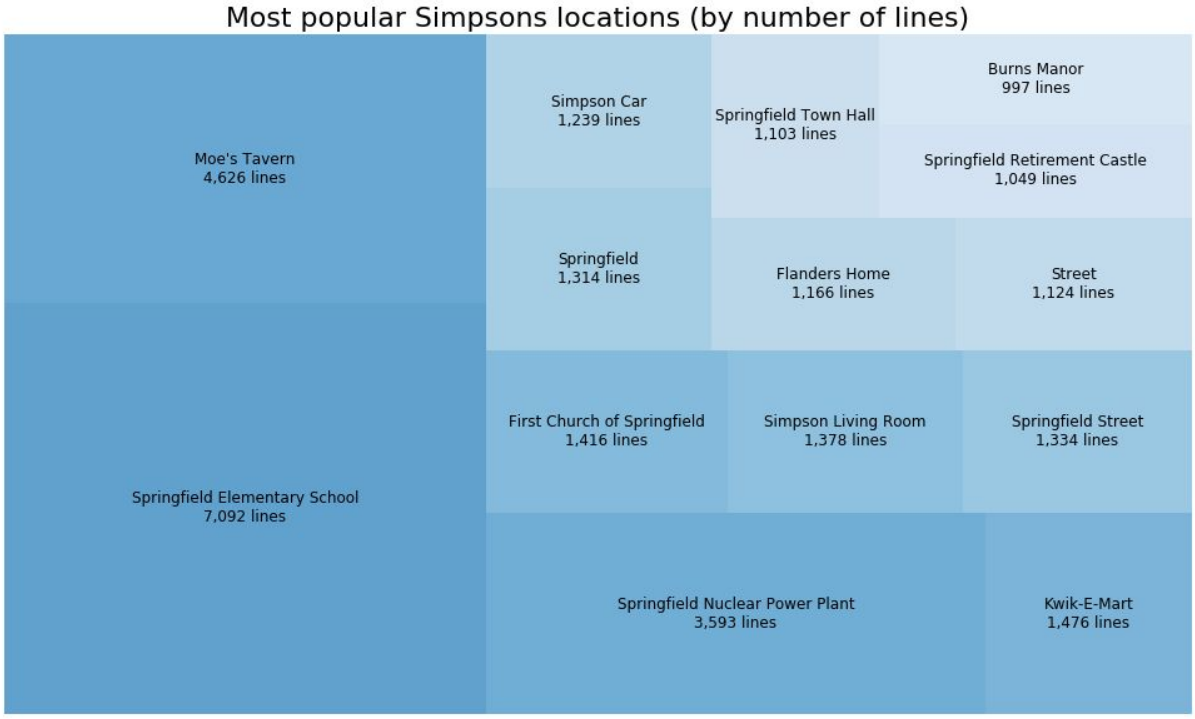
One trick is to tell them stories that don't go anywhere... Like the time I caught the ferry over to Shelbyville. I needed a new heel for my shoe, so I decided to go to Morganville, which is what they called Shelbyville in those days. So I tied an onion to my belt, which was the style at the time... now to take the ferry cost a nickel, and in those days nickels had pictures of bumblebees on them. "Give me five bees for a quarter" you'd say. Now, where were we? Oh yes, the important thing was that I had an onion on my belt, which was the style at the time. They didn't have white onions, because of the war…

In addition to learning how much each of the characters speak, I was also curious about when they spoke. Each line in the data included a within episode timestamp, which allowed my to answer this question. I considered the 3,000 characters with the most lines.



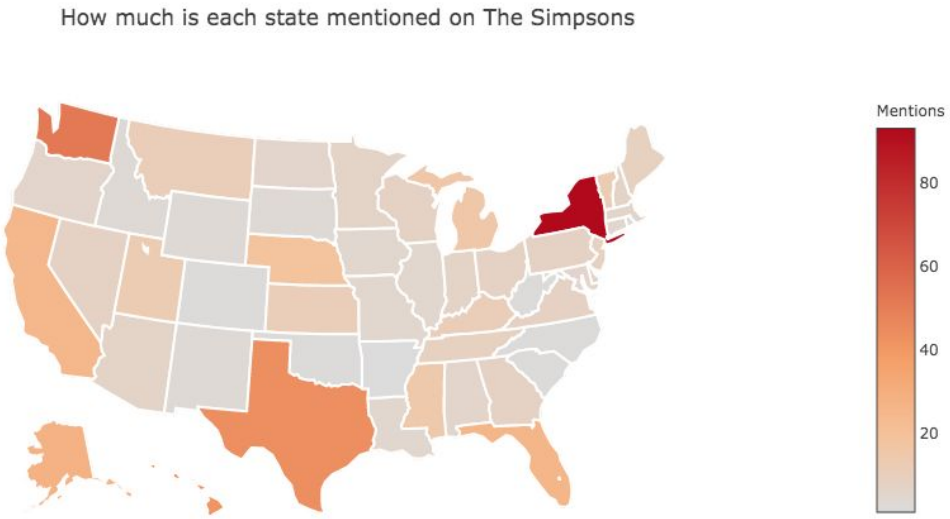
I separated the characters into two categories, those that appeared in only one episode and those who appeared in more than one. Plotting the (scaled) number of lines for each character on the y axis, it became clear that characters with more lines tend to appear in the middle of the episode, on average. This makes sense because they speak so much all the time that their average time will be near the middle of an episode. It is interesting to note however, that in the upper-right corner, it looks as though characters that only appear in one Simpsons episode often speak about two thirds of the way into the episode. This seems reasonable, since we probably would expect to meet new characters somewhere in the middle of the episode rather than the very beginning and they probably speak more lines as we get to know them.

So far, we have only looked into the characters and their lines. We have not yet looked into the locations of the Simpsons episodes or their relationships. Consider the plot below.



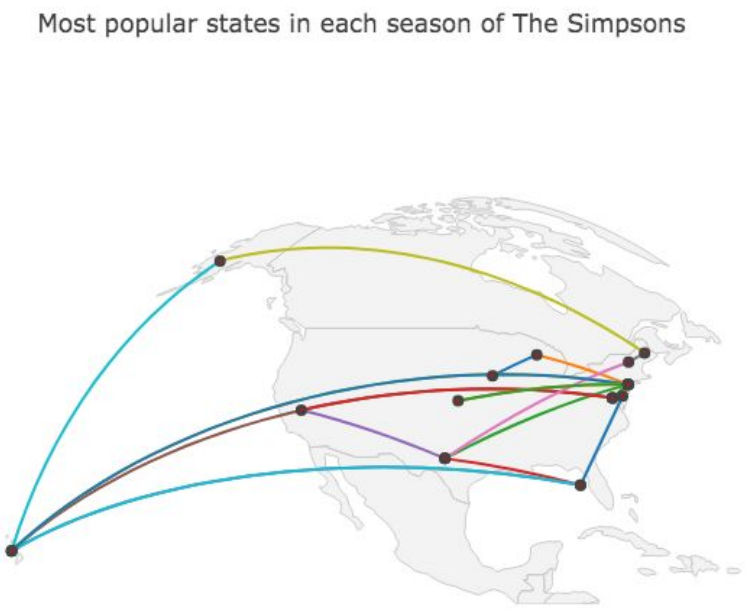
We see that Springfield Elementary School, Moe's Tavern, and the Springfield Nuclear Power plant (where Homer works) are the most common location in the show. I will point out that I removed the Simpson home from this plot because otherwise it would have dominated the space. Other popular locations are church, the living room, the Kwik-E-Mart and even Burns Manor. The color and size of each of the boxes corresponds to number of lines that were spoken at each place.

What about real places? How much do the characters in the show know about our world? One thing we can do is look at how much they mention each US state.



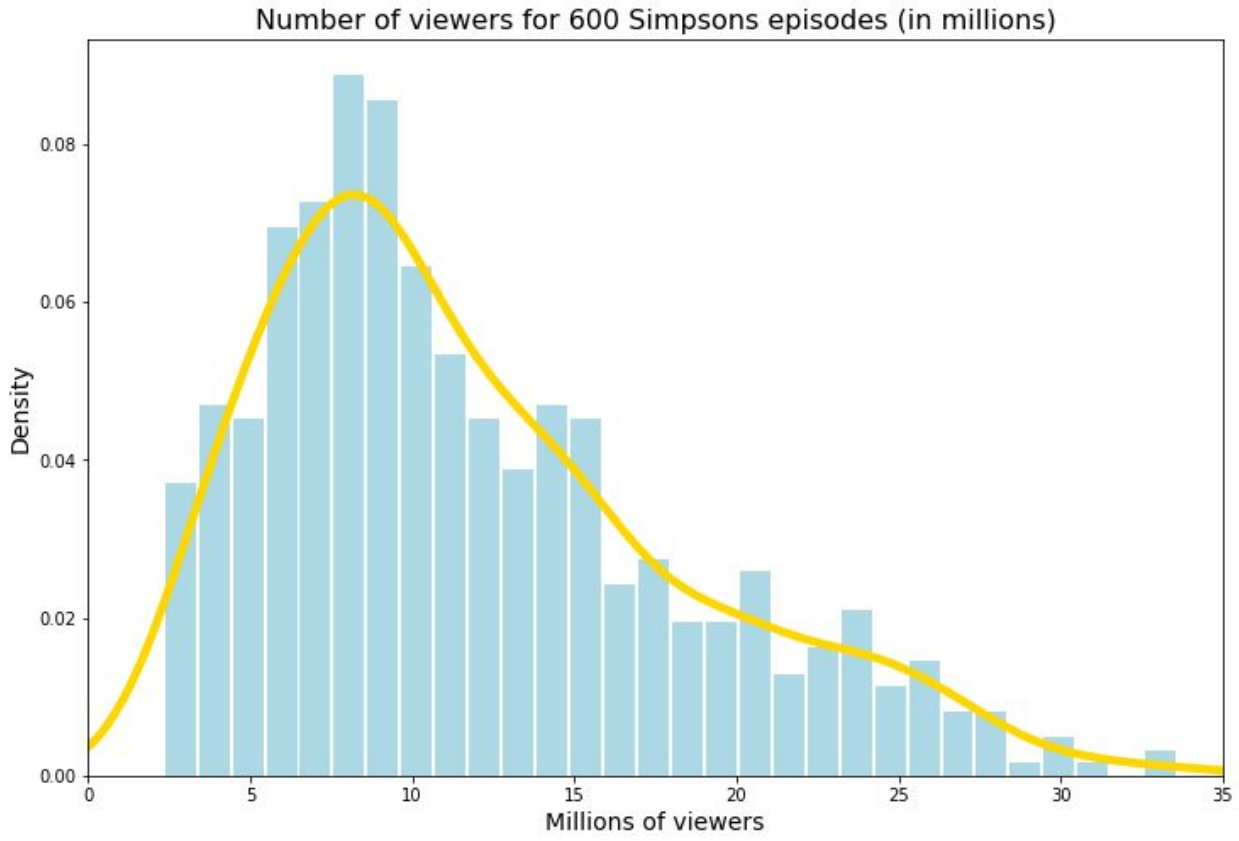
We can see that they often talk about New York, Texas, California, Alaska, Florida, Washington and even Nebraska. I should point out that we should look at the data for Washington more closely. I filtered out lines that contained “Washington D.C.” or “George Washington”, but there may be other instances where they weren’t necessarily talking about the state. This is nice because it is not subject to to typical problem with heat maps of the United States. These plots often turn into population density plots because the thing we want to plot is usually related to the population. However, since we are looking at a metric that is not really related to the population, we should not expect (and did not have) this problem.

Now let’s look on a per season basis. Let’s find the most popular state in each season and send a connection from the most popular state in each season to the most popular state in the next season. Doing that results in the following map.



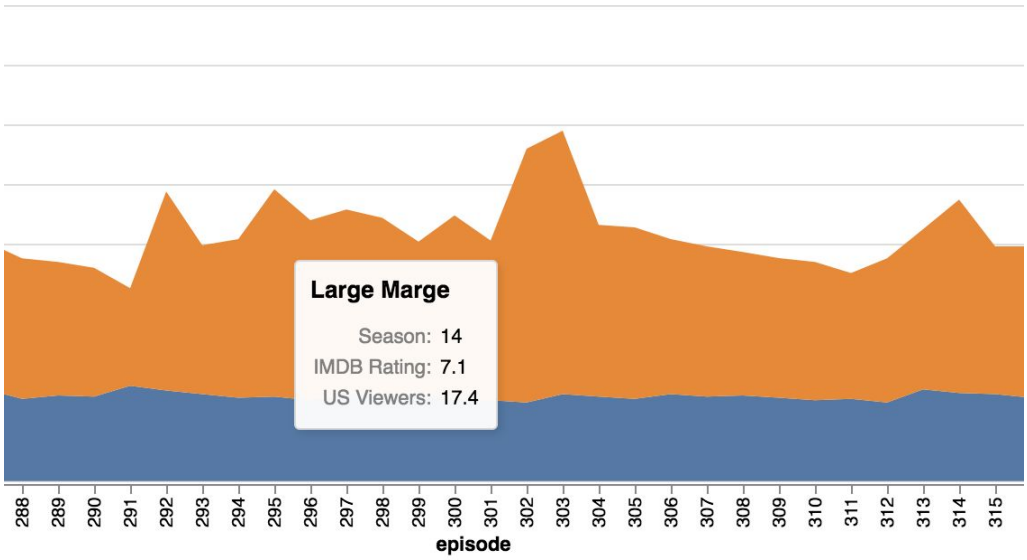
The most concentrated area for popular states was New York and the New England states. One explanation for this is just the fact that they are smaller so there is more room to put points next to each other. At the same time, most of the states are still on the Eastern half of the US. It is interesting that the only states on the western side are popular travel destinations.

The script lines have allowed us to better understand the dynamics of the people and places *The Simpsons.* Now let's change gears and examine the show's audience. The following shows the distribution of the number of viewers in the United States across all episodes.



Most episodes are watched by eight or nine million viewers, but some have been seen thirty million or more! It also much more common for an episode to have less than fifteen million viewers than it is for them to have more.

This does not account for time, however, even though we would expect that this would be an important factor. As my next plot will show, the popularity of *The Simpsons* has changed over the course of thirty years.



This visualization is interactive and you can get the full version online here: https://github.com/areevesman/the_simpsons/blob/master/plots/stream.html.
It is an html file that you can download and open in your browser. If that doesn't work then you can re-run this notebook: https://github.com/areevesman/the_simpsons/blob/master/notebooks/viewers-histogram_and_stream_plot.ipynb.
The plot shows the IMDB rating (blue) and number of US views (orange) for every episode. In addition, it displays the name of the episode and the season when the mouse hovers over. Scrolling horizontally shows the full timeline of all of the episodes.

It is clear that in the beginning, *The Simpsons* had lots of viewers. Their popularity lasted in fact until the end of eight seasons. During seasons nine, ten, and eleven, there were very few viewers. There was a significant boost once season twelve began, and It just so happens that "Worst Episode Ever" had one of the largest audiences that season. Since then, *The Simpsons* has experienced a steady decline. Are less people watching? Maybe. However, less people are watching television networks in general. The ratings are pretty close to where they have always been, so I am going to say that *The Simpsons* are alive and well.

Conclusion

In this report, we have seen a lot about *The Simpsons* that even dedicated viewers like myself may not have thought about before. We have learned who talks the most, who talks to whom, and what these characters are saying. We have also learned when and where they speak in addition to what places they speak about. Finally, we examined their place in popular culture and how they have survived for so long. If you have made it to the end, then you should be in one of two categories. Either you never want to hear about *The Simpsons* again, or you'll crack open a cold Duff beer go start an episode right now!

Appendix

All code is available on Github here: https://github.com/areevesman/the_simpsons.