A. A description of the data. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. Editing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection. (10 pts)

Being such frequent users of the World Wide Web, we were curious to find data that might unpack more about its structure. Particularly, we wanted to see what the web looks like around the world. We found a Kaggle dataset, *Popular Websites Across the Globe* (https://www.kaggle.com/bpali26/popular-websites-across-the-globe), that gave us a lot of information about the most popular websites viewed in each country, the number of visitors to each of these sites, and also the hosting location of each site. There was also data about the trustworthiness and child-friendliness of each site, which was of less interest to us.

Additionally, we used map data (world-50m.json) from class notes, as well as country capital, longitude, and latitude data from https://gist.github.com/alexwebgr/10249781. This was used in our map visualization to plot the points (arbitrarily) at the capitals of each country.

We realized that we were really curious about the hosting location of websites; which countries hosted popular sites? Across the world, are the sites we see the sites that are popular around the world? Or do other countries use other websites?

We wanted to represent host countries of popular sites and the countries that use these sites. With so many countries, it would be very difficult to create a non-interactive visualization of the data from every country. Many countries also hosted very few or no popular websites. So, we brought our focus towards two countries that would be interesting to compare: the United States and Russia. Both are among the top hosts of popular websites. Also, these countries have very different rules about internet censorship, with the United States being relatively lenient and the Russian Federation having much stricter censorship rules.

The variables we chose to represent these ideas are as follows:

- Host Country
- Viewing Country: Countries whose top viewed websites are hosted by U.S./Russia
- Websites: The top 5 most visited websites from each country (U.S./Russia)

- Average Daily Visitors: The number of people on average visiting each website each day

Because the data focused on the viewing countries instead of the hosting countries, we had to reformat the data.We deleted all irrelevant columns from the large dataset to speed up processing; leaving columns Website, Avg_Daily_Visitors, Location, Country. First, we used regular expressions to remove spaces in the numbers, which appeared between every three digits. We then copied these columns so we could process the data for each of the two visualizations. For our map visualization, we considered columns Avg_Daily_Visitors, Location, Country. We wrote and ran a Python program to identify countries that host popular websites (countries listed in the Location column) by picking top 50 countries with the most Avg_Daily_Visitors of all sites hosted by that country. We then grouped by the host countries (Location) and Country, and summed up Avg_Daily_Visitors for each. This gave us the number of visitors from each country to host sites. We also considered our originally selected four columns and considered each unique website listed; for these we summed the Avg_Daily_Visitors grouped by Location; United States and Russian Federation. These were used to give context to the data; we display these numbers so the viewer gets a better picture of the comparison between the United States and Russia. In terms of additional data, in order to represent location in the map visualization, we combined the main dataset described above word-topo.json (https://gist.github.com/alexwebgr/10249781), which describes the capital coordinates (longitude and latitude) of all countries in the world. The Python script also included the coordinates of country capitals into the cleaned data and associated them with the names of hosting countries and other site visiting countries. Lastly, we used word-50m.json (found from lecture notes) as the shape file for visualizing a world map.

For our second visualization, we considered columns Website, Avg_Daily_Visitors, Location. We selected only rows where the host country (Location) was “United States” or “Russian Federation”. Then, we grouped by and output rows with unique Website values. Manually, we wrote the names of the websites, as the original data was presented at links to the actual data. We then selected the top five most visited websites from each country.

We selected the data as described because it best represented what we wanted to know. There were also many columns with similar meanings, but the values for these columns were confusing. For instance, there were columns measuring percent visitors to websites, with negative values. It was important to us that we understood the meaning of the data we chose, and so we focused on the columns whose entries made more numerical sense. We also were uninterested in “trustworthiness” and “child-friendliness” of websites.

These metrics did not answer questions we had, and we were unsure about how the websites were categorized.

B. A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts)

One of the scales we used was position. We represented host and viewing websites as dots on the map, positioned arbitrarily at their capitals. The United States and Russian Federation are highlighted by larger circles, so they stand out visually. They are also identified by coloring for more visual impact. The geographic mapping helps give context to the data, as the reader should be familiar with the world map and understand approximately where each country it. Lines go from each viewing country to the host country, again colored to correspond to the host country from which it originates. The opacity of the lines is set with a log scale transformation; this is because of the extreme difference between the numbers of viewers. This shading is very subtle. We also use geography to show other host countries; countries that host popular websites are denoted by small dots placed at the capital.

Text is also used in the first visual to give a context about the total number of viewers along all lines and the number of lines, which would be difficult or impossible for the reader to discern with lines alone.

We also used rings to represent the number of visitors to the major websites hosted by the U.S. and Russian Federation. There, the thickness of the rings corresponded to the number of users; this is visually easier to compare than an area might be. Here a linear scale is used to illustrate the number of users for each. Each ring is set to a shade of the same color, so that the rings are distinct. This color is not significant otherwise.

C. The story. What does your visualization tell us? What was surprising about it? (5 pts)

Most straightforwardly, our visualization shows which countries use websites hosted by the United States and by the Russian Federation. It shows how many people visit the websites hosted by these countries. It also shows other countries that host popular websites. In the second figure, our visualization shows the top five most visited websites hosted by the U.S. and Russian Federation. It shows relatively how popular each of these websites is to the others.

Our visualization also shows how dominant the United States is in terms of the popularity of the websites it hosts. This is evident by the number of countries that use its sites and the number of people who use these sites (almost 100 times more!). With websites like Google, Youtube, and Wikipedia being hosted in the United States, this is not entirely surprising. However, looking more closely at the data, we see some things that are more interesting.

For instance, note that many of the European countries are also major web-hosts, as illustrated by the bright points. Asia is also host to many websites. Africa and South America, on the other hand, appear to host many fewer popular websites. We also see that the latter continents' populations do not visit websites hosted by Russia. Indeed, Russia's websites are almost entirely visited by European and Asian countries, with few exceptions. This indicates that there are some websites that are only visited by select countries. Given that the websites we use most are Google, Facebook, etc. this feels surprising; in America, our websites are more universal in that very few countries do *not* visit websites hosted here.

However, we see that the country that hosts a website is not always the country that visits this website. This is by our second visualization, showing the most popular websites from the United States and Russia. Note that from the United States, the fifth-most-popular website is QQ. With even a small amount of experience using the internet, most Americans would recognize that QQ is not a popular website here, as many would not even know what QQ is. In fact, QQ is an instant messaging website popular in China. So, we realize that a host country of a website does not have to be the location where that website is most popular.