Project 1 - Final Report

Eric Freitag, Miguel Roberts, Emily Robinson INFO3300

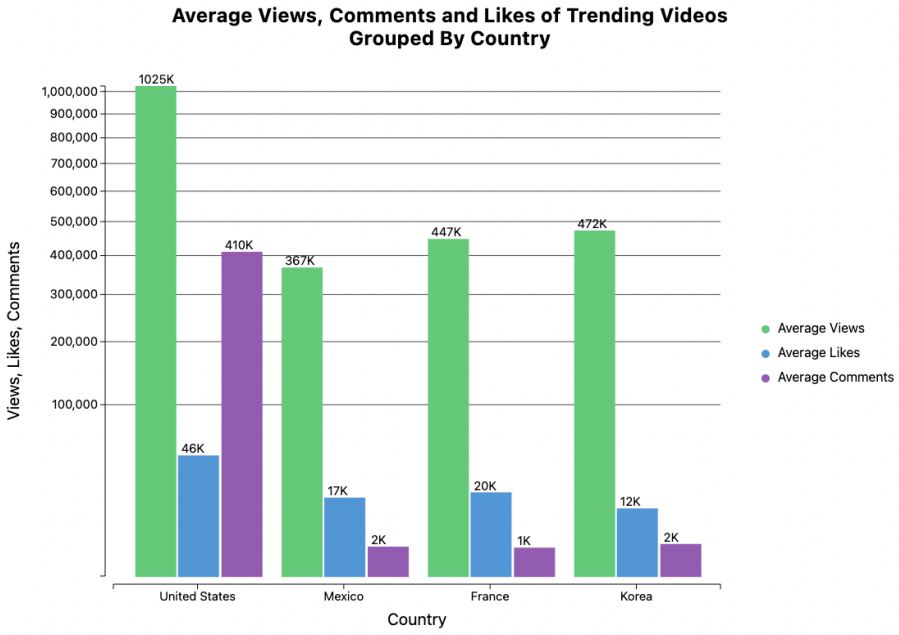
Description of data

We got our data from a page titled “Trending Youtube Video Statistics” on Kaggle.com. We downloaded data from this site in the form of CSVs. The link to this page is: https://www.kaggle.com/datasets/datasnaek/youtube-new/code?select=USvideos.csv. This site included information on trending youtube videos from 10 different countries, but we chose to work with data from just the United States, Korea, Mexico and France. We chose these four places because all of these locations had data from the same time range and we wanted to compare results from a diverse set of locations throughout the globe. After downloading these four datasets, we pre-processed the data within each dataset by trimming down the number of entries within each dataset from tens of thousands of data points to approximately 1,000. We chose to limit the number of data points in our datasets in order to improve the performance of our code and ensure that our live server could render our data visualizations quickly. To trim down our datasets, we analyzed trending youtube videos from just 5 separate days between November, 2017 and March, 2018, since each day contained approximately 200 lines of video data. We selected days that were a month apart in order to avoid situations in which videos were trending over the course of multiple days, causing a single video to be present in our dataset multiple times. The most important variables that we used from our data were the number of views, comments, and likes on each trending video as well as the publish date and trending date of each video. In order to process and compare dates, we made sure to parse the dates within our datasets using d3.timeParse according to the original format of the dates. We included a file

called types.txt in our code base to explain more about the names and types of all the relevant variables within our data.

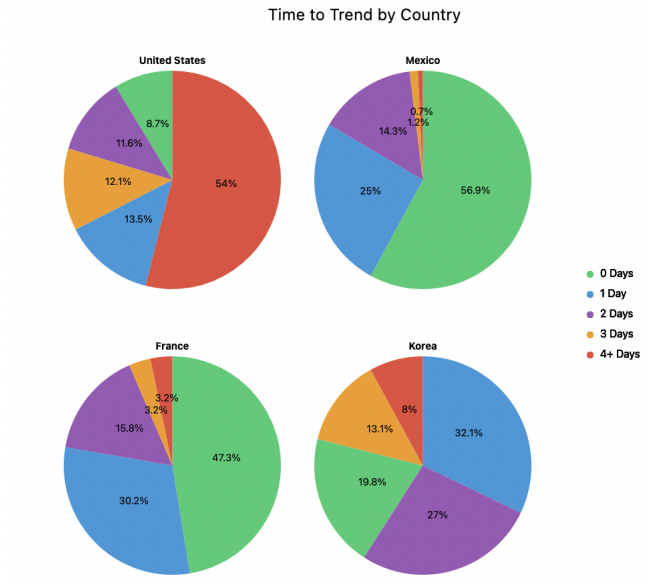
Design Rationale

In order to explore trends within our data, we created two data visualizations. Our first one, a multi-bar chart, focused on the average number of likes, comments and views that trending videos obtain across the four countries that we analyzed. Our second data visualization, a series of pie charts, focused on the amount of time that it takes a video to trend across the different countries. These visualizations used a wide variety of marks and channels to convey key differences in youtube trending data and as a result, we made a number of design decisions when creating each visualization.

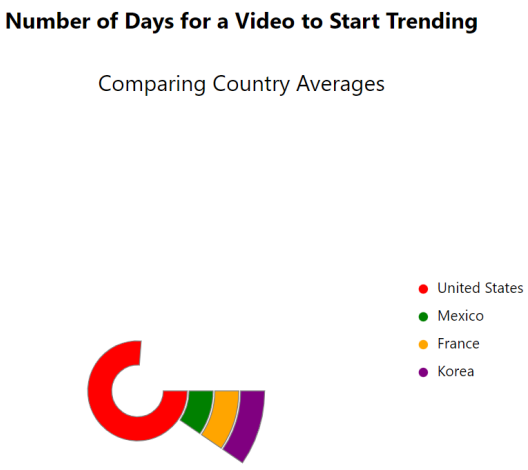


Our first visualization aimed to convey insights regarding the levels of interaction necessary for a YouTube video to be considered ‘trending’ in a given country. The main ways to interact with a video are through views, likes, and comments, so we decided to graph how these

interactions varied by country. In deciding on a graph, we made 2 different visualizations, but agreed upon the second one–a multi-bar chart to represent the average number of views, likes, and comments a trending video receives in different countries. The first graph we created was a stacked bar chart to try and capture 4 different values: 1) Average views 2) Average likes 3) Average comments 4) Average total interactions. While this chart was visually appealing, there was one major flaw, which clearly became evident: in each country, except the United States, the number of likes and comments were so small relative to the y-scale and the views, that they were practically unobservable. The likes and comments made up such a small percentage of the total interactions that the two bars stacked on the views bar for each country were practically invisible. This also caused the labels for the likes and views to stack on top of each other, further reducing visual clarity. After ideating through the second visualization, the multi-bar chart, we quickly realized this had much more visual clarity, because we could use an exponential scale (with a fractional exponent) to make the height from 0 to 100,000 much larger than the heights for the following values which were not as significant. In doing so, we achieved a final bar chart that had each of the interactions (views, likes, comments) distinguishable and comparable.



Our second visualization aimed to compare how long it takes a video to ‘trend’ in different countries. In order to illustrate this comparison, we created a pie chart for each country that we were analyzing. We chose to use pie charts to display this data because pie charts allowed us to clearly illustrate the percentage of videos that began to trend within certain time ranges and by making a separate pie chart for each country, we were able to clearly illustrate differences in trending time across countries. The main design decisions that we made for this visualization are in our selection of time ranges and chart colors. We made sure to choose the same colors across each pie chart and choose colors that were very distant from each other in order to provide a clear contrast between the different time ranges in which a video could began trending. In addition, we chose to split each of our pie charts into 5 categories because we didn’t want to clutter our pie charts with too many categories and we noticed that the vast majority of videos began to trend within the first 4 days of the video being published. Finally, we added text labels to our pie charts in order to convey the percentage of videos that began to trend in certain time ranges because this was the main focus of these visualizations. In order to avoid labels overlapping one another when slices of the pie chart got especially thin, we transformed the vertical position of some of our labels to prevent these overlaps.



We were going to add an additional chart that was a spiral visualization of the average number of days for a video to trend across each country. However, we decided to not include it since, we believed that without the numerical labels, it would be unclear.

The Story

Our visualizations seek to answer the question of what criteria determine whether or not a video will trend on YouTube across four countries: The United States, France, Mexico, and Korea. This topic fascinates us because we're currently living in a world where people invest a lot of time into both media creation and consumption. One of the greatest sources of such media is YouTube. So, by analyzing and graphing patterns in 'trending' YouTube videos, which are generally some of the most influential videos on the platform, we can discover useful insights into the prevalence of YouTube within each country and the formula that YouTube uses to select trending videos.

Our first visualization, the multi-bar chart, synthesizes information regarding the average number of views, likes, and comments that trending videos receive in different countries. One of the most notable results that we can see from this visualization is that trending videos in the U.S. receive the greatest number of comments, views, and likes for Youtube videos, averaging 1025k views, as opposed to the other three countries that were between 367k - 472k. This may suggest that people in the U.S. spend more time on YouTube and rely more heavily on the platform to gain information or for entertainment. Furthermore, the number of comments on videos in the United States are drastically higher–over 200x more than Mexico and Korea, and 400x that of France–and the likes are nearly twice that of videos in all other countries. This suggests that people in the United States are more likely to engage in discussion and interact with content where as videos in other countries are more passively watched. Therefore, a trending video in the

United States needs to be more than just watched by a lot of people, but rather also needs to incite a certain degree of interaction by its viewers ,which is not necessarily the case for videos in other countries.

The following graphs looked at the time it takes for a video to become trending after initially published, by which we sought to see if there were any patterns regarding when a video was most likely to be trending and whether or not there was a ‘deadline’ for videos to receive interaction. In creating the pie charts that measured average time to trend for videos in each country, we found that in the United States, videos were most likely to trend after 4 days, where as in other countries, videos were more likely to trend before the 4 day mark. This lets us know that if a video is going to be trending in other countries, it will most likely do so before the 4-day mark, after which it is much less likely. This was rather surprising because in the United States, it is almost expected that a video will need some time to reach a large audience, however, in France, Korea, and Mexico, at the very least, videos seemingly take no time, most of which trend on the day they were published or within the following day or two.

Team Contributions

We worked highly collaboratively throughout the course of this entire project. As a result, we met up both in person and over zoom to select a topic for our project, chose relevant datasets, plan the designs for our data visualizations, and complete the first two milestones. Once we finalized the datasets that we were using and completely planned out our designs together, we split up the process of creating our data visualizations. Since we were planning to make 3 separate visualizations, we each took on the task of creating a single visualization, with Miguel creating the multi-bar chart, Eric creating the pie charts, and Emily creating the spiral chart. After creating all of our visualizations, we worked together to debug any problems that we

encountered in our code, edit our visualizations based on the feedback we received in the project showcase, and write the final project report as a group. During this step of debugging and review, we came to the agreement that the spiral chart, while unique in how it attempted to convey data, did not add much more to the understanding of the viewer that the pie charts did not already add. As a result, we just kept the first two and included what we learned from the process in this document. The initial planning and data selection phase of our project took approximately 5 hours working together, the creation of our data visualizations took approximately 7 hours each, and the process of editing our visualizations and creating our report took approximately 3 hours working together. In total, about half of our time was spent planning and writing the final report and the other half of our time we spent writing the code for the visualizations themselves.