STAS401 Tableau Project Report

**The Data Analysis and Visualizations of the Most Recent Updated Videos on Bilibili**

Boqing Zheng, Shuhe Wang, Xingyu Shen

**Introduction:**

The stream of Bilibili as a video platform has been brought to Chinese young people’s life for years and is increasingly flourishing. Bilibili is considered to be one of the most popular young-oriented video platforms among Chinese people (Song 2021). It contains diverse content encompassing different interests, from lifestyle, games, entertainment, anime, technology, and knowledge to many (“Corporate Information” n.d.). Many video beginners or young people decided to publish videos on Bilibili as a start to share their interests, daily life, and creative ideas. In this project, we scraped data of the most recent published videos from Bilibili and used data visualizations to analyze it. We are aiming to find the characteristics of the most recent videos and investigate the trend of interest among Chinese young people. The research questions are as follows: (1)What is the relationship between feedbacks of the audience and the number of views? (2) What are the common characteristics of the videos published recently? (3) What are the popular topics and keywords of the videos among young people?

**Methodology:**

1. **Data Collection**

We used a python request package to collect data from Bilibli. The data consists of three datasets: weekly popular (每周必看, 6559 videos in total), popular(综合热门, 500 videos in total), most\_recent(videos can be viewed before a certain date, 15000 videos in total). The columns we chose in our datasets are: tid(type id), tname(type name), title, desc(description), duration, stat(interactive information of the video), dynamic, rcmd\_reason, bvid(unique id of the video), owner, pub\_location. Since attributes of APIs for different Bilibili sections are slightly different, the columns of different datasets may also differ slightly.

1. **Data Cleaning and Processing:**

We wrote a python program to process the collected datasets.

We planned to make a word cloud to visualize the frequency of different words in columns of text. Since there is no space in a Chinese sentence, we used the jieba package to separate columns of text like title and description into word lists. In order to get rid of the influence of stopwords, we downloaded a stopwords list from github and made some adaptations to it. And finally, we processed all of these columns into two csv files. One is the original word list, and the other is the word list without stopwords.

We planned to create a recommendation rate column according to the recommendation rate, and we found a formula online to calculate it based on interactive information of the video (“大揭秘！B站‘热门’视频是怎么选出来的？” n.d.) (see Figure 1). To get the interactive information like coin, view number, reply, we applied eval function to each row to convert the string into a dict form, and then separate it into different columns. We also manually dropped a few rows with outlier data, whose data is relevant with that column.

In order to predict the trend. We include pubtime and ctime columns into our dataset. We called the to\_datetime function in pandas package to convert the timestamps into regular forms of time to make it easier to understand.

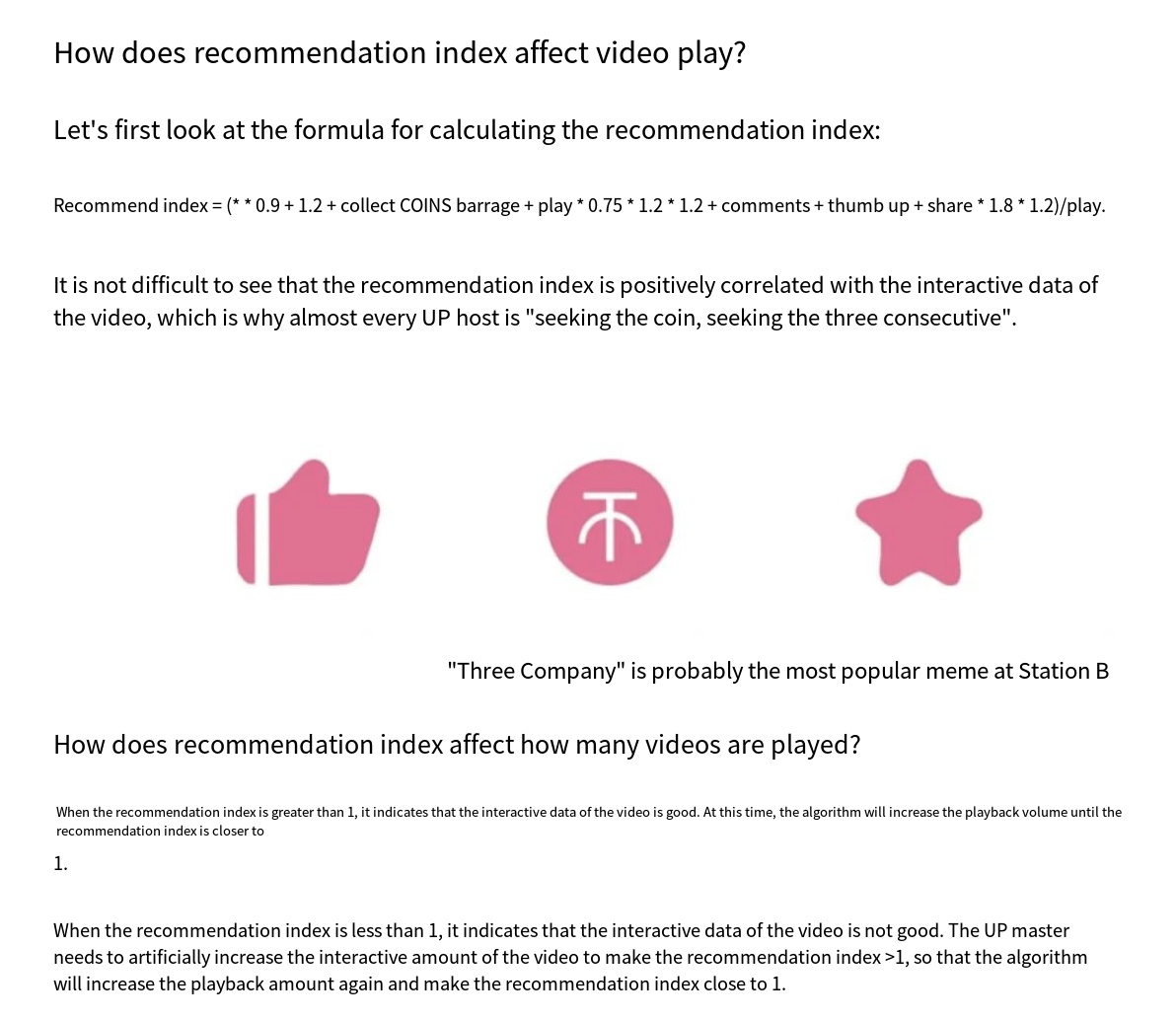


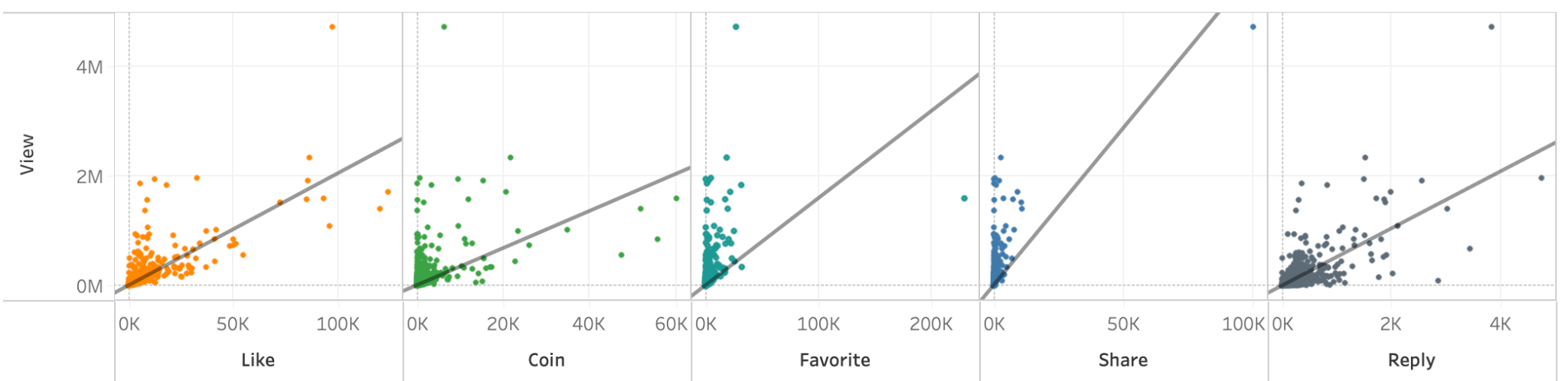
Figure 1. The formula of the calculation of recommendation rate and criteria

1. **Visualization**

We used Tableau as the visualization tool in this project. After importing the data source to Tableau, we made plots with different attributes, idioms, marks, and channels. For the results, we made six visualizations according to the research questions. The visualizations explored the relationships between different attributes, the characteristics of the videos according to the regions and time, and the frequently appeared elements.

**Results:**

1. The dot plots of relationship between the audience feedback and number of views:



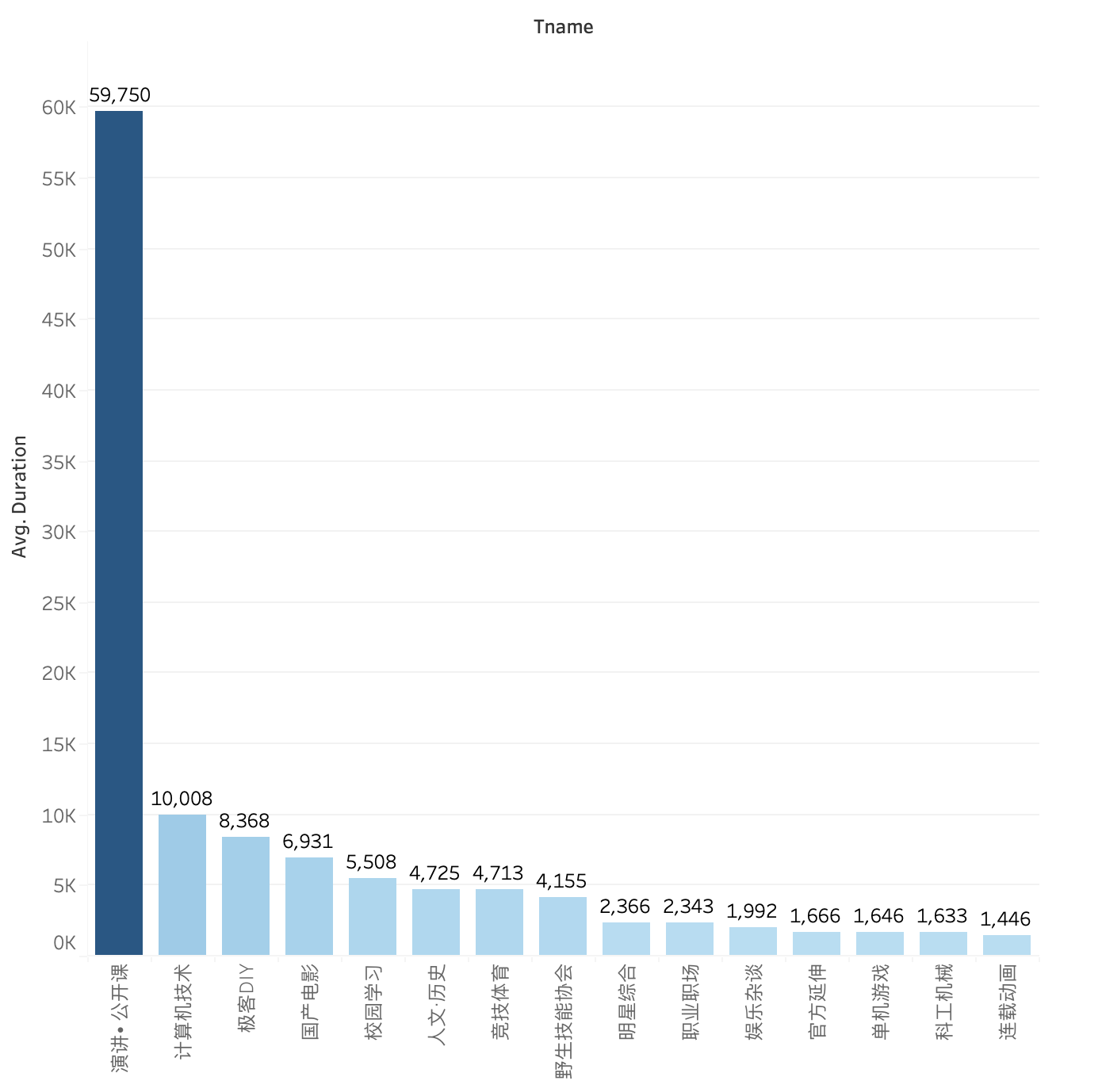
**User task:** Find the relationship between elements of the audience feedback and the number of views.

**What:** This dot plot shows the relationship between the response of the audience towards the videos and the number of views.

**Why:** Seeing from the plot, users can figure out that the number of views is positively correlated to the number of the elements included in the response of the audience. Although there are some expectations and extreme values, the trend is quite clear.

**How:** The audience can grasp the information by observing the distribution of the dots, finding the relationship between its abscissa and y-coordinate, and according to the trend line.

1. The bar plot of average duration of the videos from different categories:



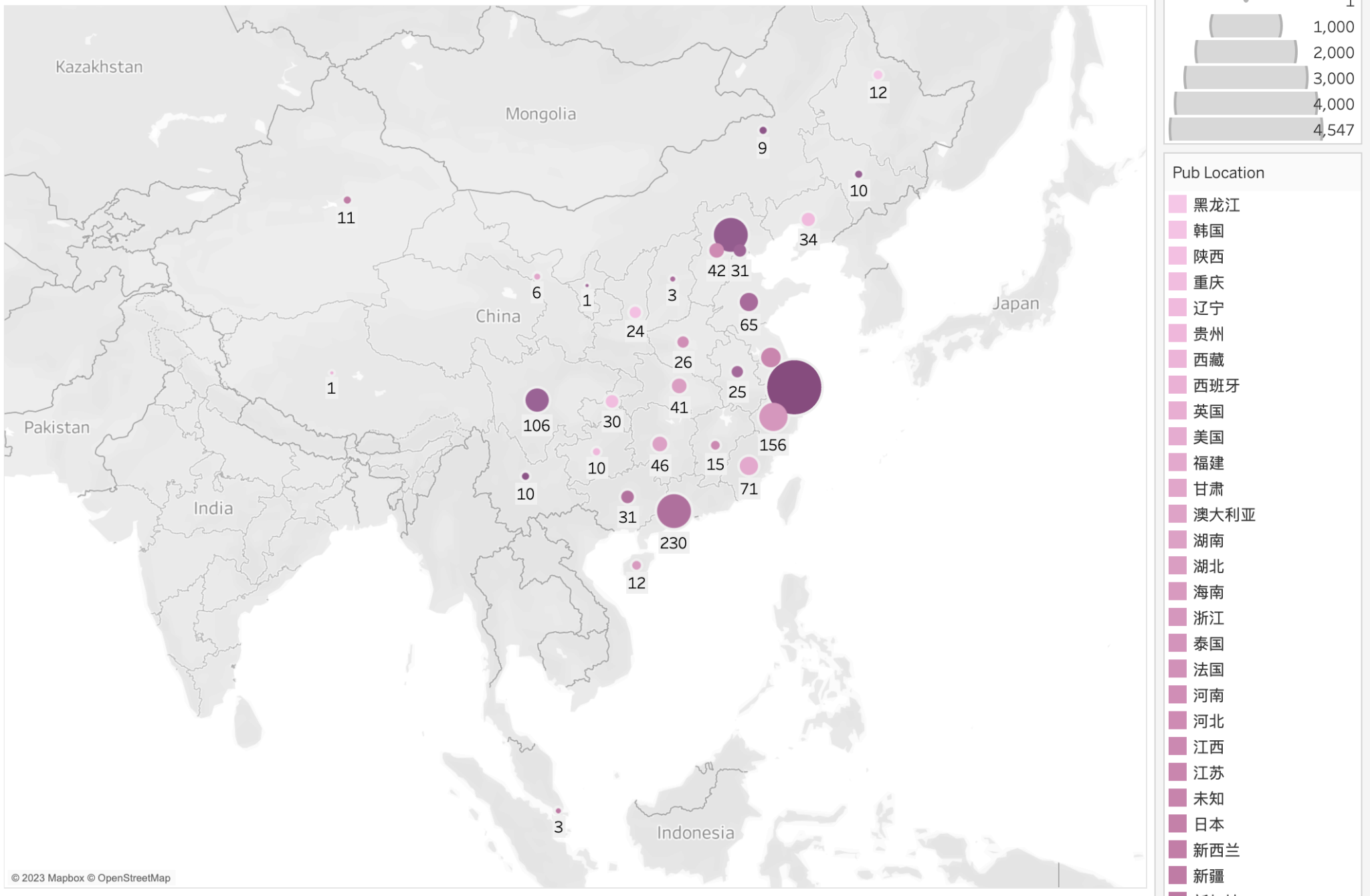
**User task:** Retrieving value and characterizing the distribution of different categories sorted with top ten average video duration.

**What:** This bar plot shows the top 10 average video duration and the name of corresponding video categories.

**Why:** The audience can discover the distribution and rank of the average video duration of different categories and their specific average duration values. We found that educational videos tend to be longer, while entertainment videos tend to be shorter.

**How:** The audience can look at the X-axis to determine categories, and the Y-axis to determine the average video duration in that category. Among them, the height and color intensity of the bar express the average length of video time for its category. At the same time, the categories are sorted in order from longest to shortest.

1. The symbol map of the number of videos published in different regions:



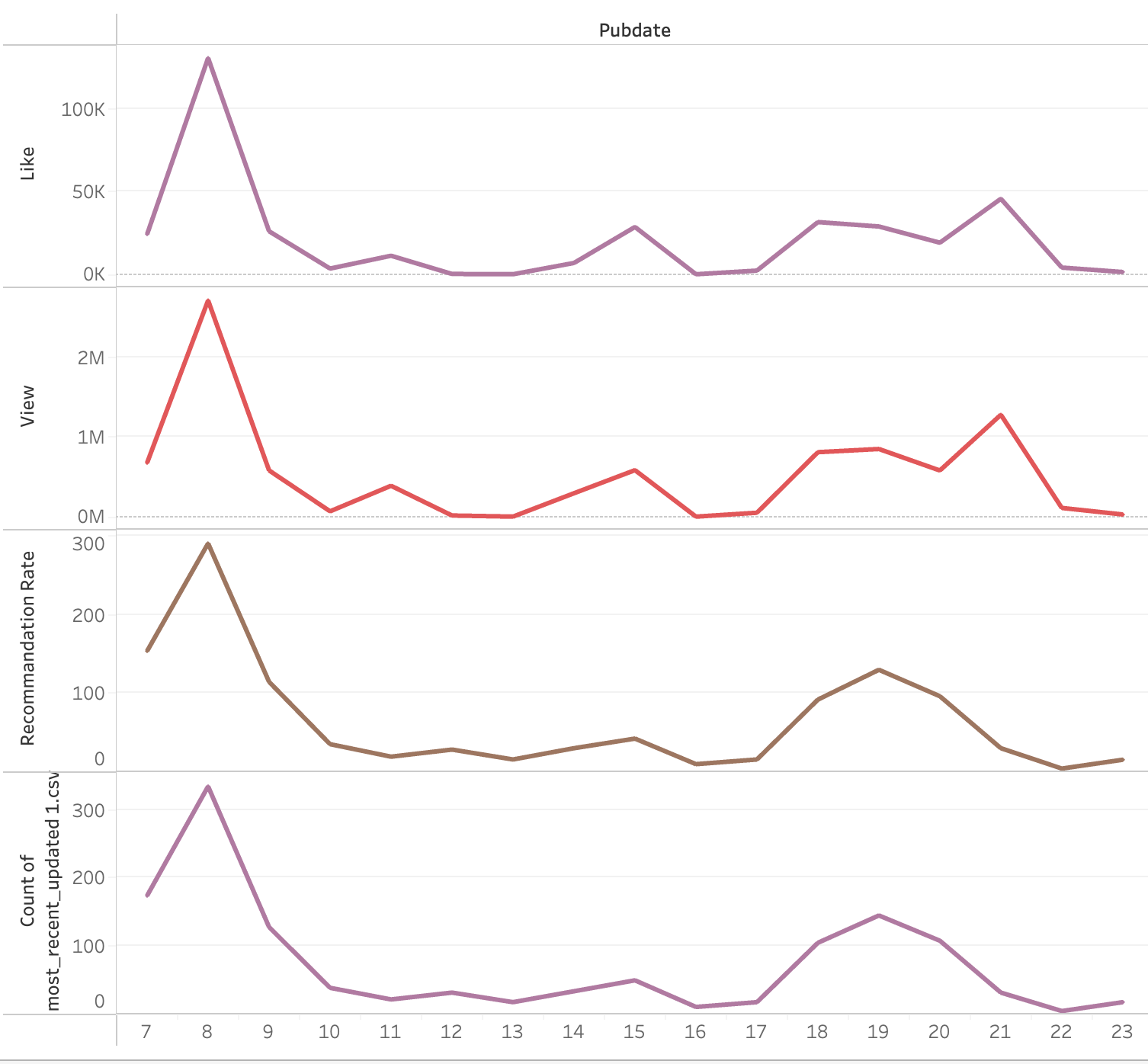
**User task:** Observe the amount of videos published on Bilibili in different regions.

**What:** This is a map with symbols of the number of videos published in different regions.

**Why:** Users will see that Beijing, Shanghai, and Guangdong are the three regions that have the most publications. It shows a trend that the regions with high developing speed and advanced technology may have more publications on Bilibili.

**How:** Users can see the amount of videos published in each region by the area and color saturation of each dot.

1. The line chart of the characteristics of the videos published in different hours:



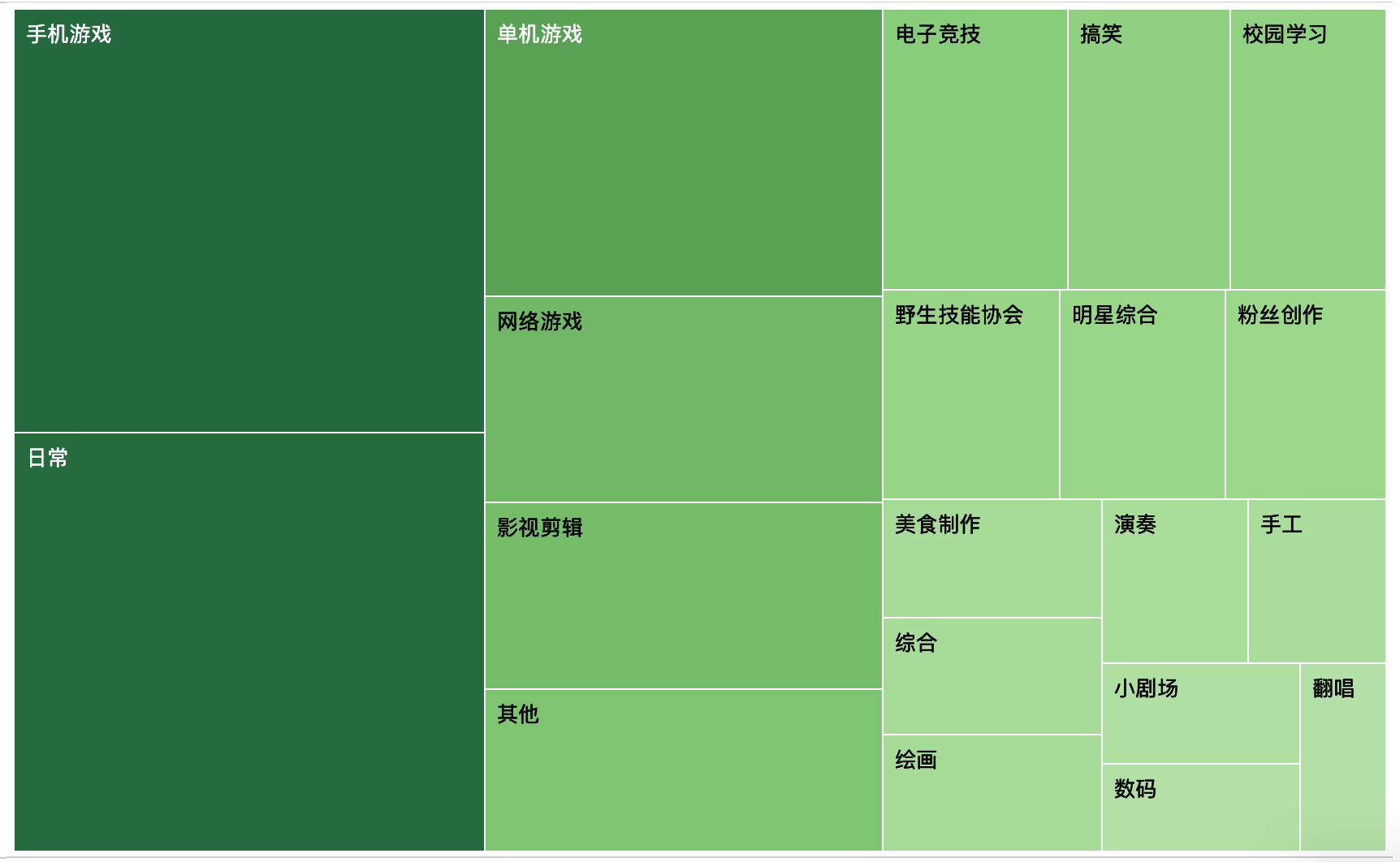
**User task:** Find the characteristics of the videos published in different hours on a specific day.

**What:** This line chart shows the distribution of likes, views, recommendation rate, and number of publications along all the hours on a specific and typical day.

**Why:** From the plot, users can get the information that most people would like to publish new videos on Bilibili during morning or evening.

**How:** Users can get the information by observing the local maximum and minimum of the line, seeing the trend of the line, and focusing on the corresponding x-axis as the timeline.

1. The treemap of the top 20 popular video categories:



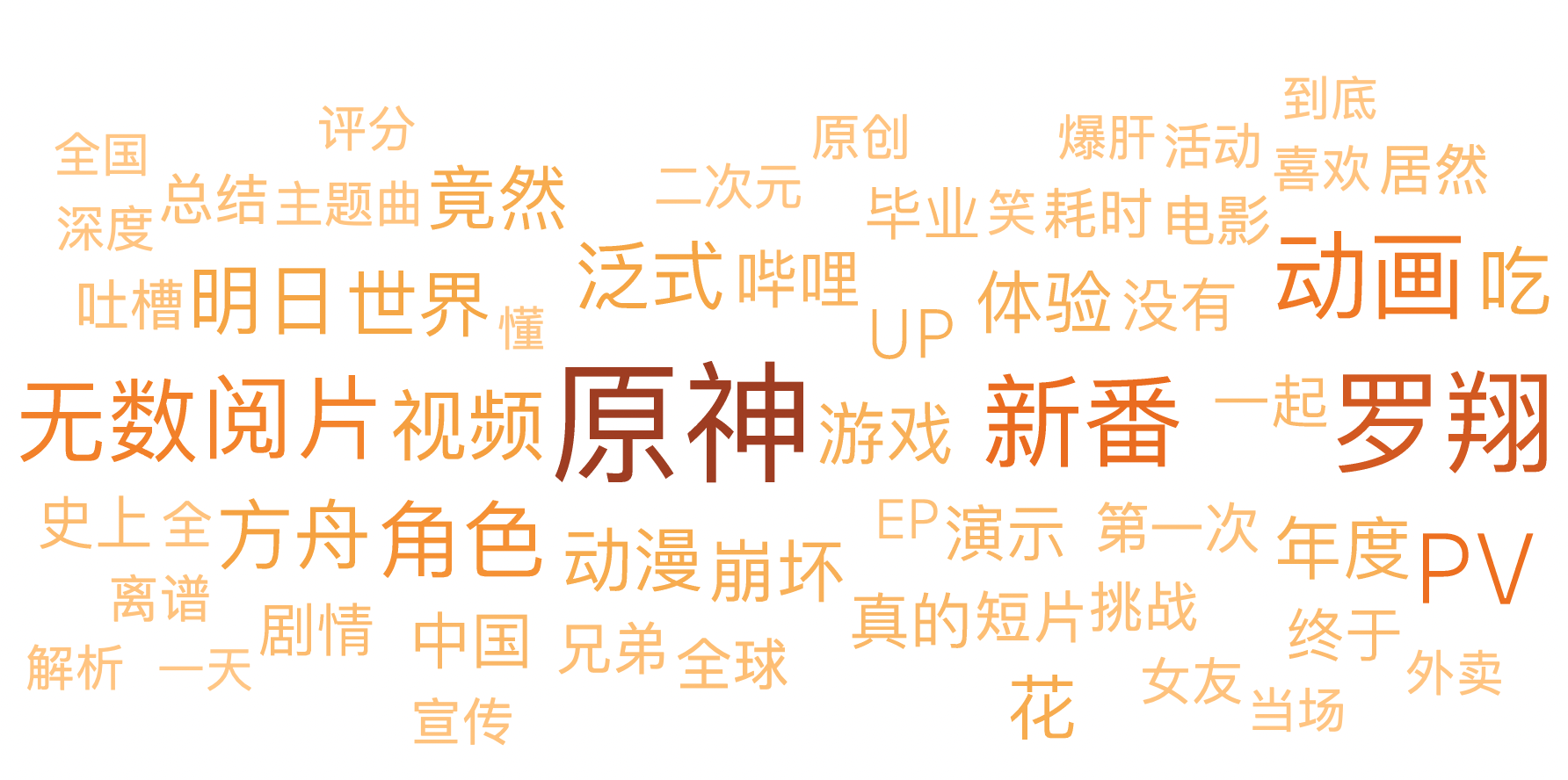
**User task**: The graph shows the total number of videos posted in different video categories and is ranked.

**What**: This graph contains categorical data for different categories and quantitative data for the number of videos posted in a certain category. It shows the total number of videos posted on each category.

**Why**: Viewers can find rankings of the number of titles published in different categories to gauge their popularity. We found that game videos and daily record videos are more published.

**How**: On each block, text was written to indicate what category it is, and the larger and darker the color, the more videos were posted in that category.

1. The word cloud of the most frequently used keywords in the publisher’s posts:



**User task:** In all categories, sort and find the most popular keywords in publishers’ posts.

**What:** The word cloud shows the keywords and the number of times in the copywriting when a publisher posts a video.

**Why:** Audiences can study publishers' preferences by looking at word cloud to see which keywords appear most often. We found that video creators are very fond of content related to “动画/新番” (animation), “原神” (a mobile game) and “罗翔” (a famous law professor), and because word cloud has a large number of exaggerated adverbs or adjectives, publishers like to use exaggerated words such as “无数” (countless) to promote.

**How:** The text displays the words that appear in the copy. The darker the color, the larger the font, the more times the word appears.

**Discussion:**

Our visualizations perfectly show the data we collected and answer our research questions. However, when we were analyzing the publication time in the visualization (4), we found that most of the 15000 videos were published on the same day. Based on annual new submissions of videos on Bilibili, users publish around 31,729 videos every day, which is much larger than the number of videos in our datasets. So, the data we collected may not precisely show the actual trend, especially when we want to make predictions. Also, the jieba package has limitations on separating some of the commonly used words by bilibili users. Thus, for future experiments, researchers can collect more videos and try to make the videos evenly distributed on different days. And they can also use a better python package for separating chinese words.

**Conclusion:**

This project provides effective ways to scrape and analyze data from the videos published recently on Bilibili. We dealt with a large amount of data and analyzed it from different aspects using Tableau as the visualization tool. From the results, we successfully got the information about different characteristics of the recent videos and summarized the trend of the popular topics. Afterwards, we will improve the way of scraping data to get more complete data though the timeline. Overall, this study provides valuable ways and possibilities to catch the trend among young people in China by analyzing the most recent videos published on Bilibili.

**Appendix:**

* + - 1. **Work Distribution:**

Xingyu Shen: data collection, data cleaning, data processing, collaborated on report and slides

Shuhe Wang: data visualization, collaborated on report and slides

Boqing Zheng: data visualization, collaborated on report and slides

* + - 1. **Code and dataset:**

https://github.com/shenxingy/stats401\_TableauGroupProject

* + - 1. **Bibliography:**

“Corporate Information.” n.d. Accessed February 5, 2023. https://ir.bilibili.com/en/corporate-information/.

Song, Shilin. n.d. “What Is Bilibili? A Look into One of China’s Largest Online Video Platforms - WalktheChat.” Accessed February 5, 2023. https://walkthechat.com/what-is-bilibili-a-look-into-one-of-chinas-largest-online-video-platforms/.

“大揭秘！B站‘热门’视频是怎么选出来的？.” n.d. 知乎专栏. Accessed February 5, 2023. https://zhuanlan.zhihu.com/p/490919116.