Final Project Report

Members: 窦宇佳 11911038 郭城 12112504 李瑞旻 12110448

Introduction and Preparation

We chose bilibili_data.cvs as our datasets, which is the data of publishers (i.e. ups) of the video website bilibili. After data cleaning, we process the cvs file into a dataset contains 20 columns of information of more than 8000 ups who publish more than 50 videos**.**

Persona of Ups

In this part, we focused on the personal information and homepage information of the ups as a whole, and analyze the distribution characteristics of variables.

- **Sex**

According to the column ['sex '], we know that the gender of ups is divided into three categories: male, female and unknown. By drawing a pie chart **(Figure 1)**, we found the number of "unknown" people is the largest, and the male female ratio is about 2.21. Therefore, it can be basically judged that the ups are mostly male.

By grouping the data according to different video tags, it can be seen that the gender distribution of ups in each video partition is obviously different **(Table 1)**. The male to female ratio in *remix-themed*(鬼畜区) and *life*, *game* and *technology* area are more than 5, and the three zones are dominated by male ups. The ratio in *fashion*, *entertainment* and *dance* area is 0.5 or less, which are dominated by female ups.

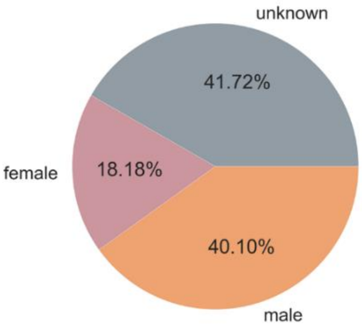
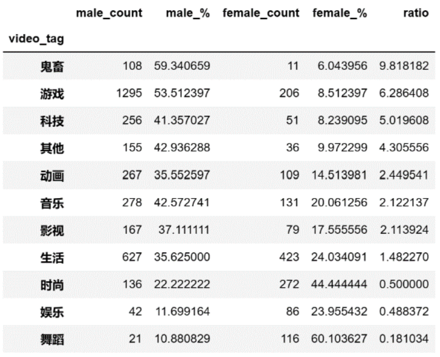


Figure 1

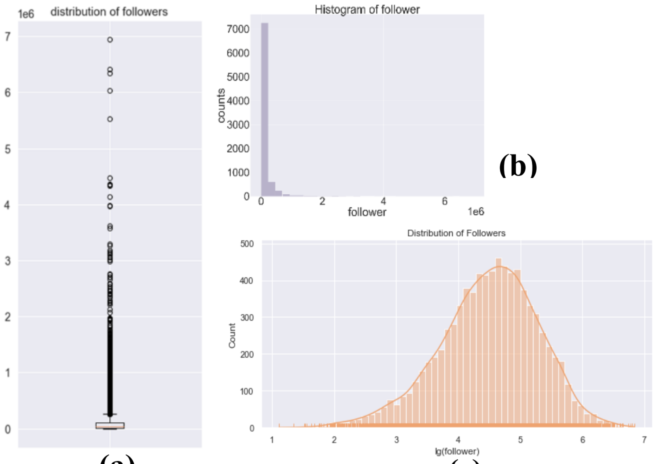


video_tag	male_count	male_%	female_count	female_%	ratio
鬼畜	108	59.340659	11	6.043956	9.818182
游戏	1295	53.512397	206	8.512397	6.286408
科技	256	41.357027	51	8.239095	5.019608
其他	155	42.936288	36	9.972299	4.305556
动画	267	35.552597	109	14.513981	2.449541
音乐	278	42.572741	131	20.061256	2.122137
影视	167	37.111111	79	17.555556	2.113924
生活	627	35.625000	423	24.034091	1.482270
时尚	136	22.222222	272	44.444444	0.500000
娱乐	42	11.699164	86	23.955432	0.488372
舞蹈	21	10.880829	116	60.103627	0.181034

Table 1

- **Followers**

According to the box graph **(Figure 2a)** and histogram **(Figure 2b)** of the number of followers, we can find that the number of followers is under heavy-tailed distribution. The distribution characteristics of followers can be seen more intuitively after log transformation **(Figure 2c)**. Most of the ups has only few followers, with 75.28% of ups has followers above 10000, 27.17% of ups has followers above 100000, and only 1.81% of ups hasfollowers above 1000000.



(a) **(b)** **(c)**

Figure 2

Similarly, the distribution of followers in each partition is studied **(Figure 3)**, and the distribution type of followers in each partition is not very different

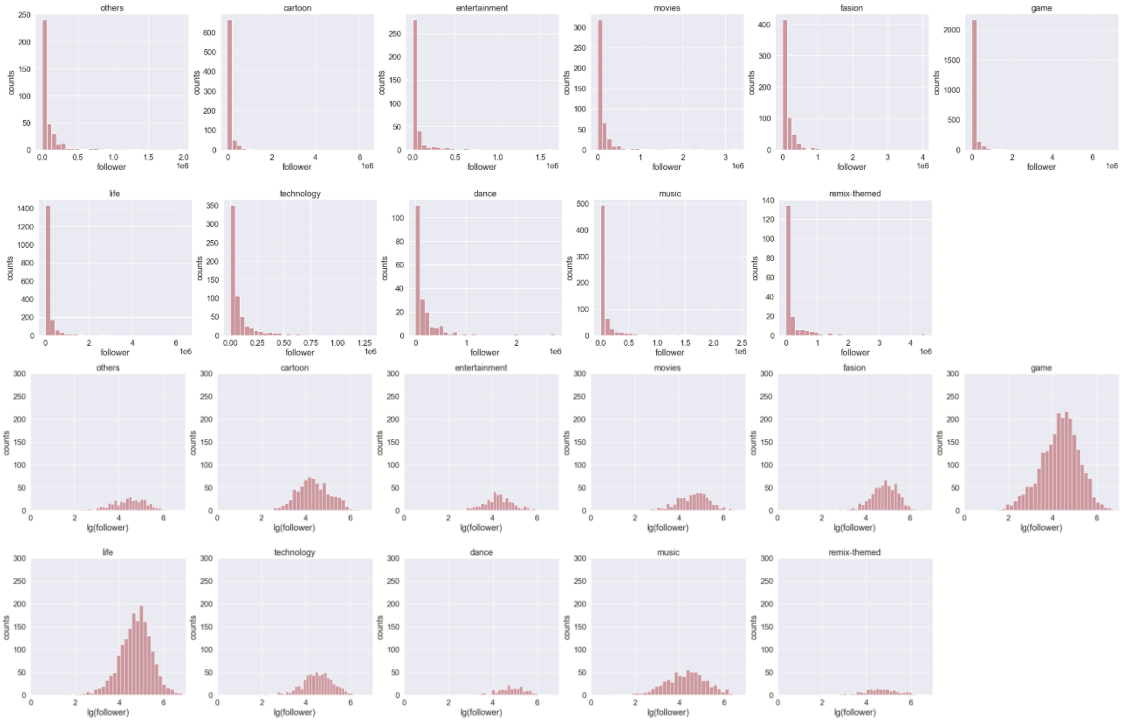
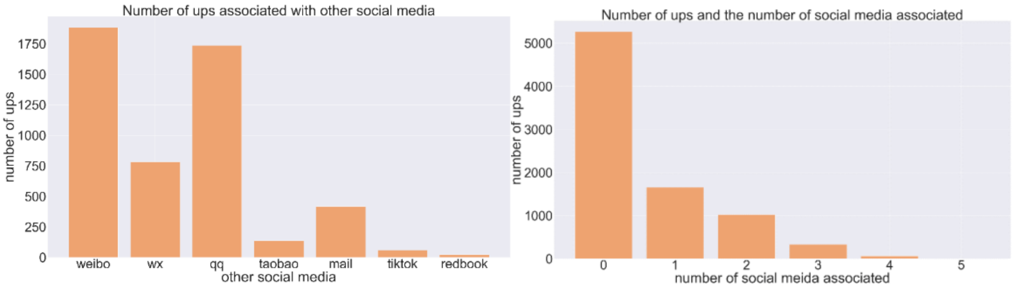


Figure 3

- **Other Media Associations**

We have made both horizontal and vertical analysis of other media associations of up. From the aspect of different media, and the number of people associated with a single media does not exceed 23%. Among all ups, the number of ups associated with *Weibo* and *QQ* is the largest **(Figure 4a)**. From the aspect of ups, 63% of ups are not associated with other social media, only 7 ups associated with 5 social media **(Figure 4b)**. In general, the overall relevance of ups to other social media is not high.



(a) **(b)**

Figure 4

- **Representative works and Self Tag**

Ups can set representative works and customized personal tags on their homepage. Among all ups, 80.42% have representative works. Among the ups that have set representative works, 81.8% of them have set three representative works, and only 7 ups has 4 representative works set **(Figure 5)**. This is a very interesting phenomenon, because ups can set up at most four representative works on the bilibili website, but most people set up exactly three. As for the personal tags, 32.22% of all ups have set them. 29.71% of all the ups set representative works and personal labels at the same time, which means that most ups who set personal tags have also set representative works, and the setting of personal tags can better reflect ups' creation of home pages and profiles.

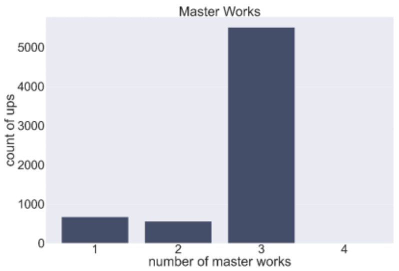


Figure 5

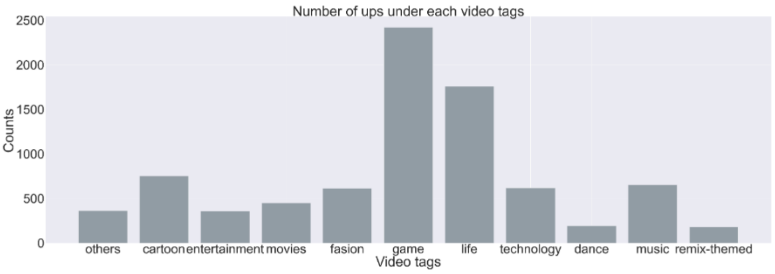


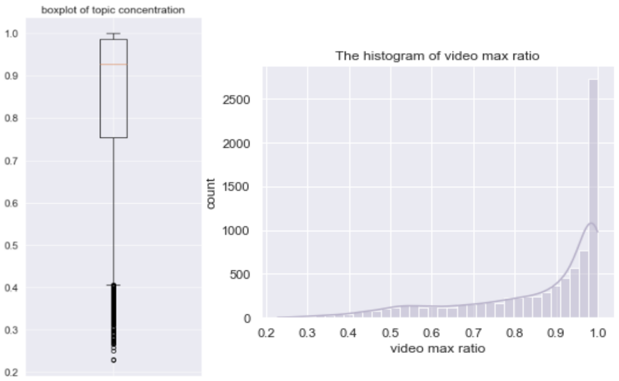
Figure 6

- **Video Tags (Partition)**

According to the column ['video_tag_combine '], we know that the main gender of up is divided into 11 categories. Among them, the *game* area and *life* area have the most ups. The *remix-themed, fashion* and *technology* are also relatively popular partitions, with more than 500 up players. **(Figure 6)**

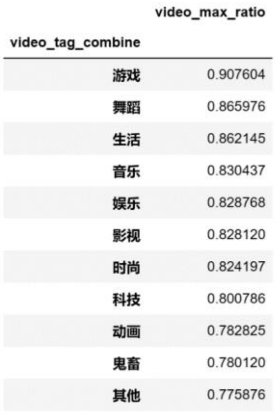
- **Video Theme Concentration**

The variable 'video_max_ratio' means the proportion of videos which belong to the main area of the up in all the videos of the up, we can also call it video theme concentration. In general, the video theme concentration of the ups' videos is high, and 55.96% of the main up videos are more than 0.9, only 5.67% below 0.5 **(Figure 7)**. This variable is slightly different in each partition, but the video theme concentrations of each partition are high **(Table 2)**.



(a) **(b)**

Figure 7



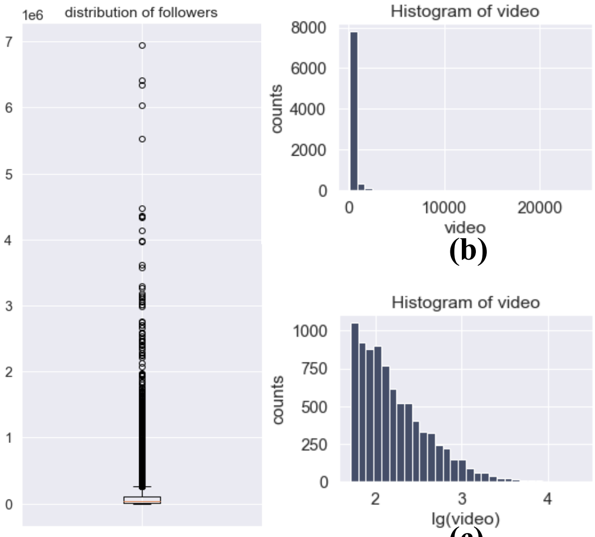
video_tag_combine	video_max_ratio
游戏	0.907604
舞蹈	0.865976
生活	0.862145
音乐	0.830437
娱乐	0.828768
影视	0.828120
时尚	0.824197
科技	0.800786
动画	0.782825
鬼畜	0.780120
其他	0.775876

Table 2

- **Number of Videos**

According to the box graph **(Figure 8a)** and histogram **(Figure 8b)** of the number of videos, we can know the distribution of videos. The distribution characteristics of videos can be seen more intuitively after log transformation **(Figure 8c)**. Most of the ups have a small number of videos, with 62.61% of ups has videos more than 100, 5.07% of ups has videos more than above 1000, and only 9 ups has videos more than 10000.

Similarly, the distribution of each partition is studied, and there is little difference in the video quantity distribution types of each partition **(Figure 9)**.



(a) **(b)** **(c)**

Figure 8

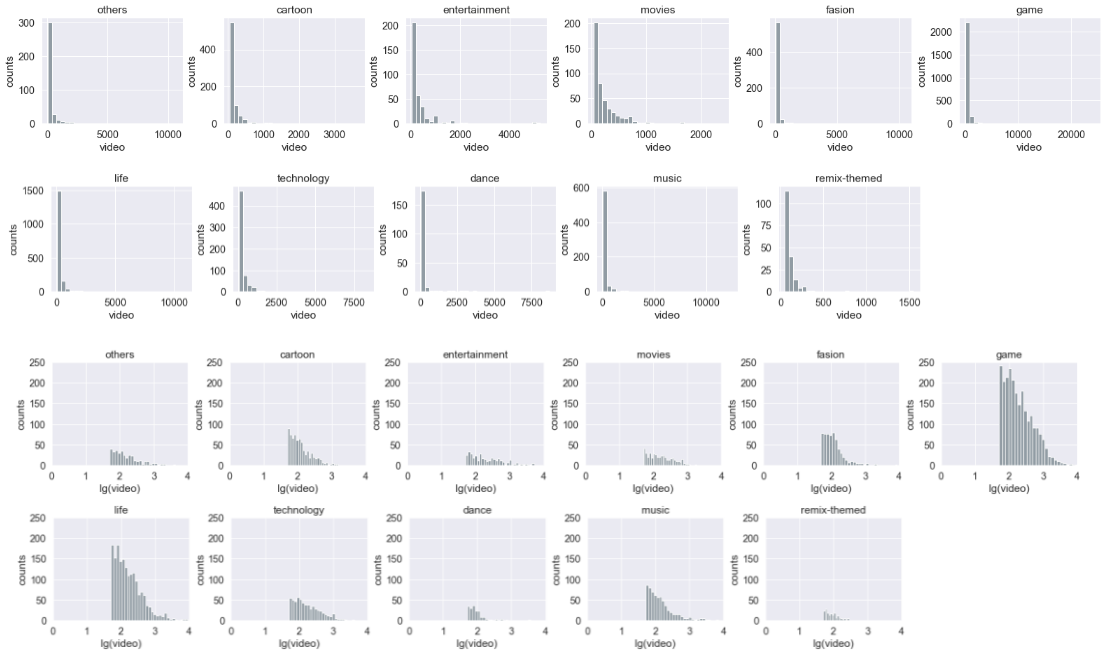


Figure 9

- **Number of Albums and Articles**

The number of albums **(Figure 10a)** and articles **(Figure 10b)** also show a significant heavy-tailed distribution.

- **Recent Video Information**

The distribution of average length and average playback of recent videos and the distribution after log transformation are as shown in **Figure 11** and **Figure 12**.

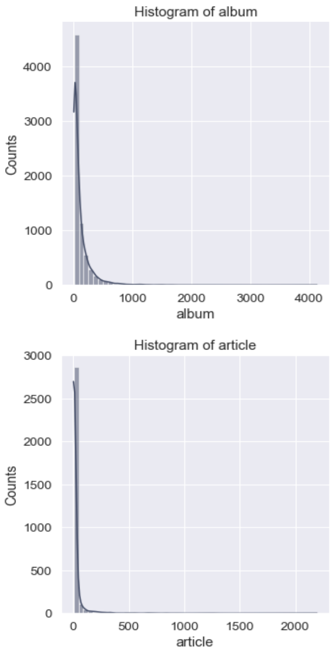


Figure 10

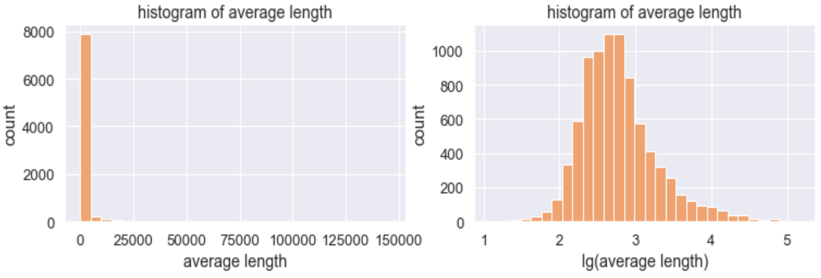


Figure 11

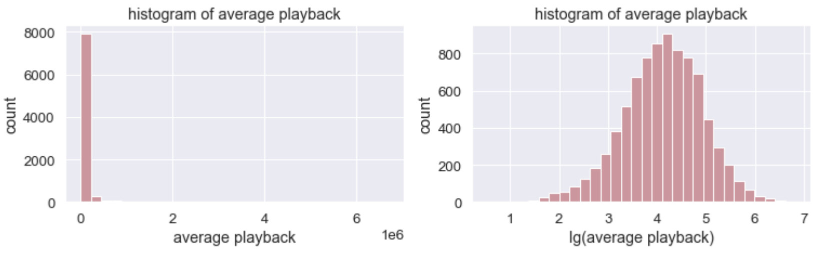


Figure 12

Data Relationship

- **Data cleaning**

Converting data of "sex" and "self_tags" into **0** or **1**. For "sex", male into **1**, female into **0**, and unknown into **0.5** (for further progress). For "self_tags", has-tags into **1** and no-tags into **0**. Correspondingly into column “sex01” and “self_tags01” to help finding digital relationships.

- **Correlation (coefficient)**

After data cleaning, the very first relationship between data is their correlation coefficient. Coefficient table is plotted below. (**Table 3**)

Through observation, most of which are “Dark”, which means low correlation except for **follower** between **play_ave20**.

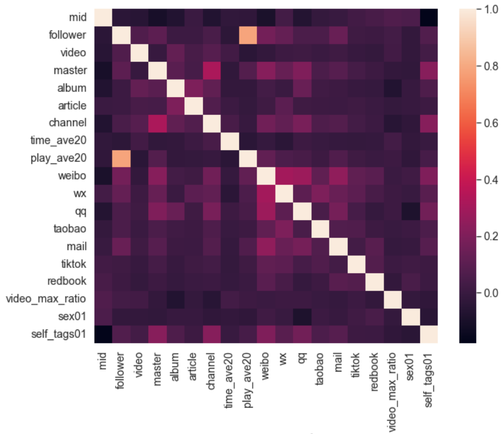


Table 3

It is observed that the correlation coefficient between the **follower** and **play_ave20** is extremely large, and the number of **channels** is highly correlated with the number **master** works, while the correlation between **Weibo, WeChat** and **QQ** is relatively large.

Relatively high relation with the **follower** (the absolute value of the correlation coefficient which greater than 0.1, weak correlation): **master, play_ave20, Weibo, WeChat** and **mail**.

After observing the coefficient, there may hidden some unobservable relationship under it. It is necessary to make boxplot and scatter plot to find further relationship.

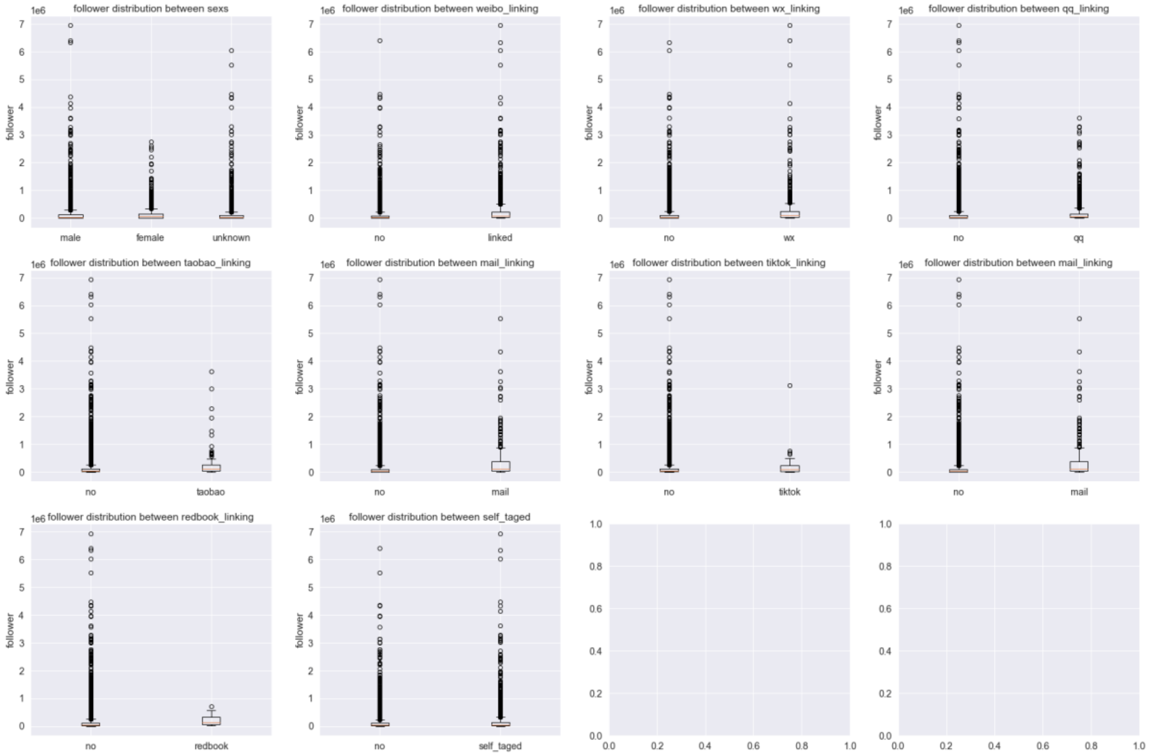


Figure 14

We have known that **follower** is heavy-tailed distribution and there are 11 discrete distribution. It is better to make boxplot for discrete ones. (**Figure 14**) The other is log-transform and horizontal dispersion scatter plot. (**Figure 15**)

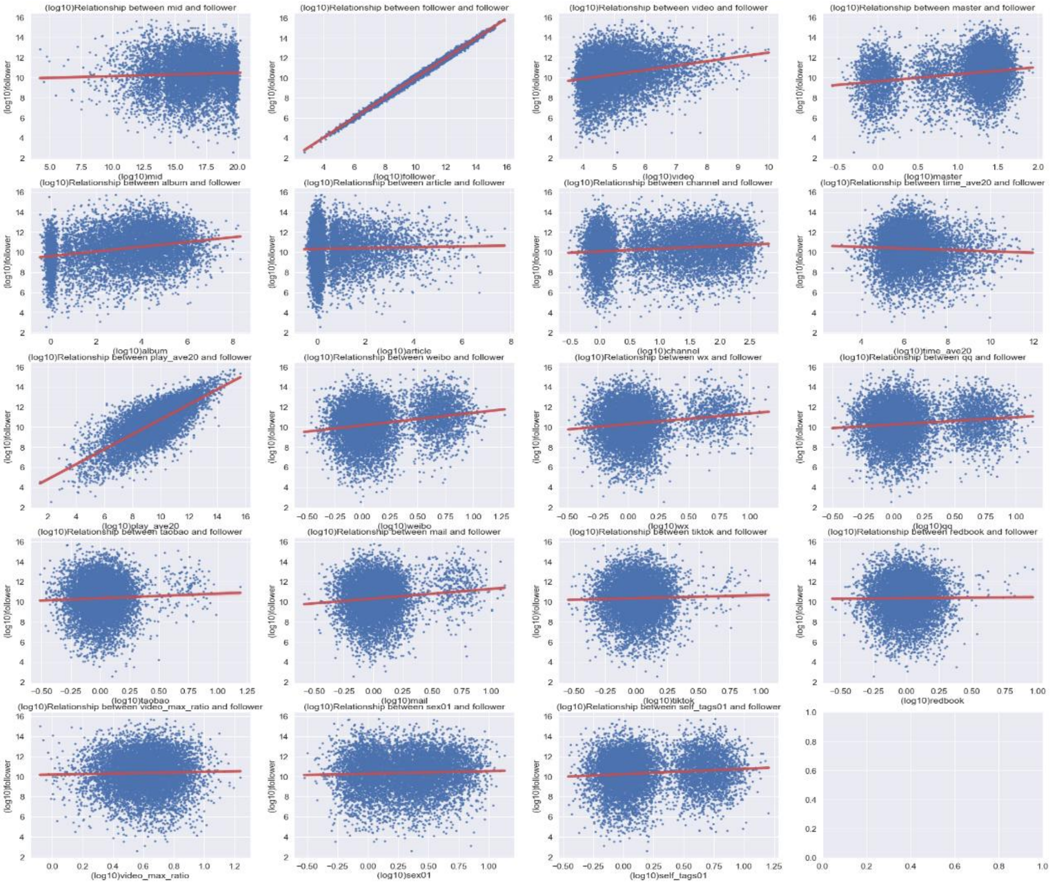


Figure 15

Still, there is no more significant relationship performed. Most fix lines are nearly horizontal, which slope is less than 0.1. Also, the distribution of points is not typical too. Detailed transformation is needed to get possible relationships.

- **Correlation (one to one)**

① **play_ave20** and **follower** (normal and log): (**Figure 16**)

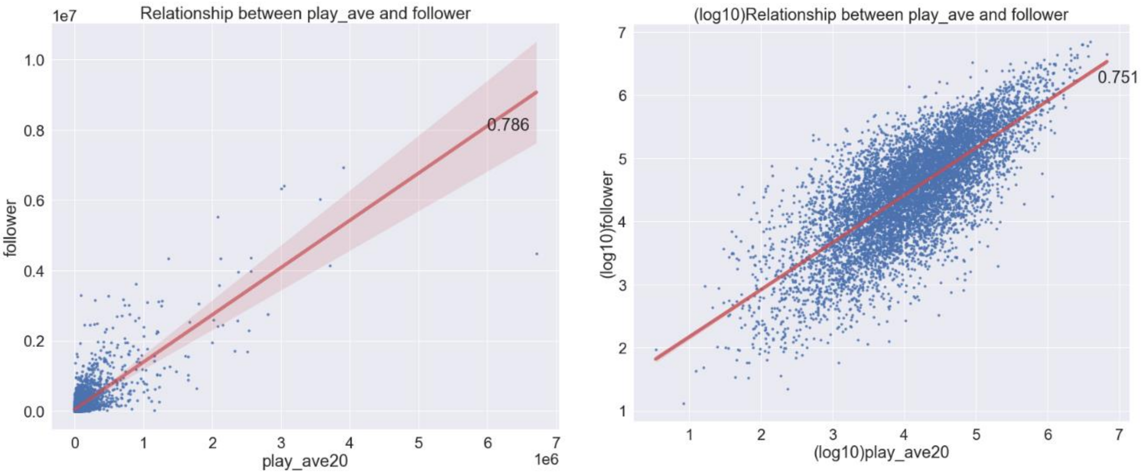


Figure 16

Figure above shows strong correlation of more than 0.7 coefficient. It conforms to a general cognitive rules that more plays brings more fans.

② **album** and **article** (normal and log): (**Figure 17**)

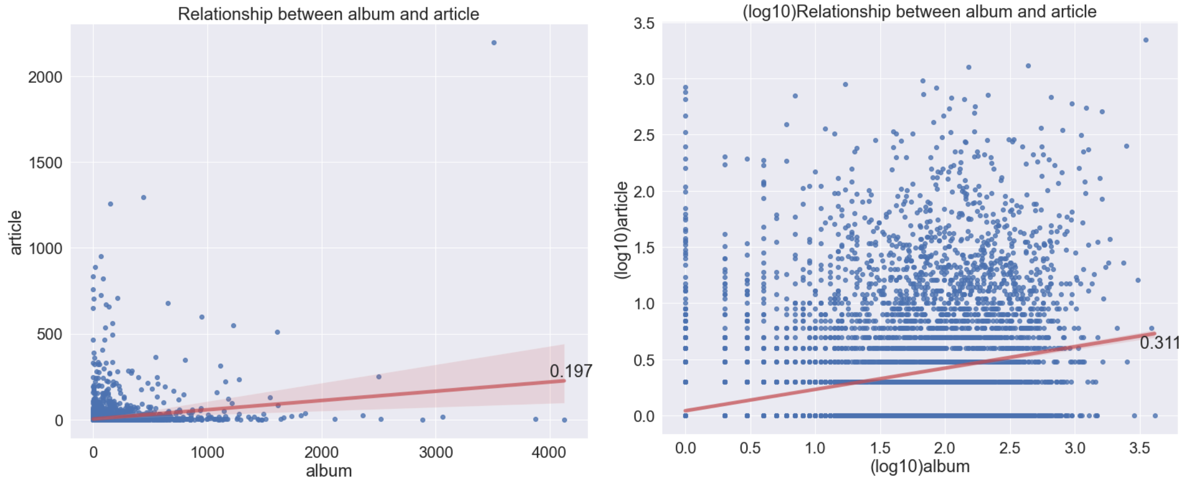


Figure 17

After log transformation, it shows high correlation. (More than 0.3) It indicates the social contact elements in bilibili has median positive correlation.

③ **master** and **follower** (normal, horizontal dispersed and log): (**Figure 18**)

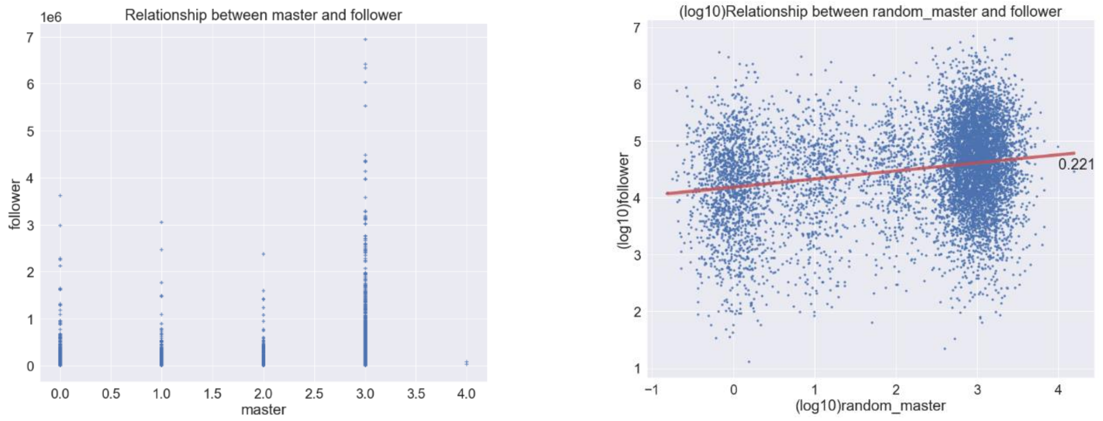


Figure 18

For discrete case, dispersing makes it more visible. From **Figure 18**, it shows that more **master** work might bring more **follower**, but it is a weak positive correlation. From 0 to 3, every discrete value has full range of distribution of follower. It is possible to get more follower in every master group.

④ **time_ave20** and **follower** (normal and log): (**Figure 19**)

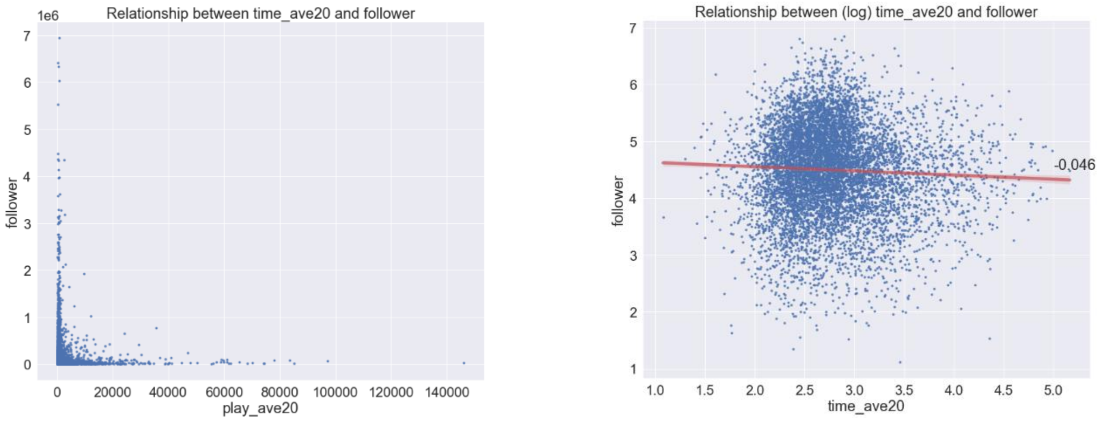


Figure 19

It didn’t show much correlation. But we know that the length of a video is divided into short-medium video (0-10min) and long video (>10min). So it might makes some difference if we break the time slot into two part, under 10min and more than 10min. (**Figure 20**)

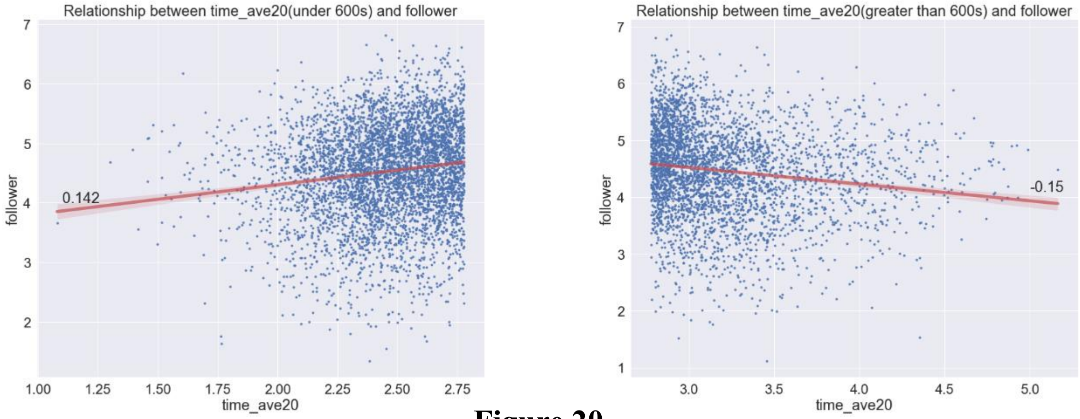


Figure 20

This time, it shows weak positive correlation when video time is under 10min and weak negative correlation above 10min. But, it is not a necessary trend according to the large degree of dispersion in every time slot. For detailed analysis is left in modeling part.

⑤ **play_ave20** and **time_ave20** (same as ④) (**Figure 21**)

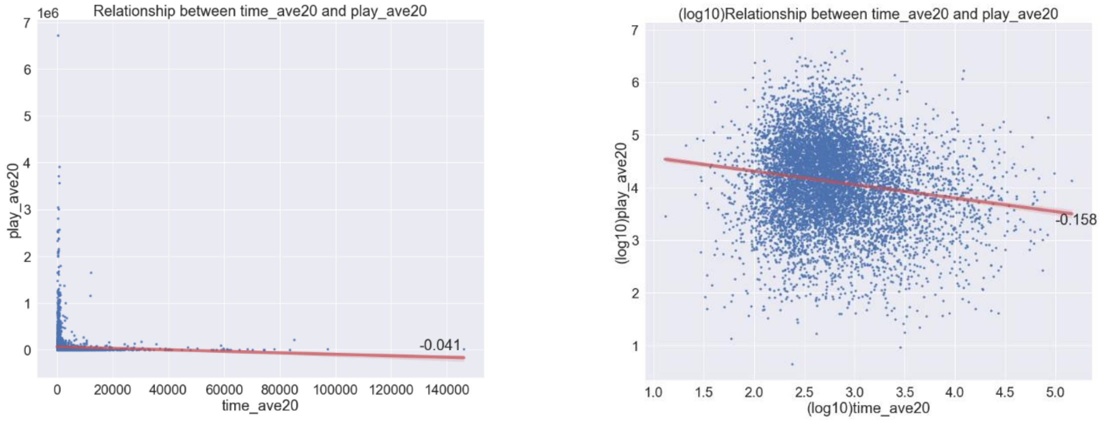


Figure 21

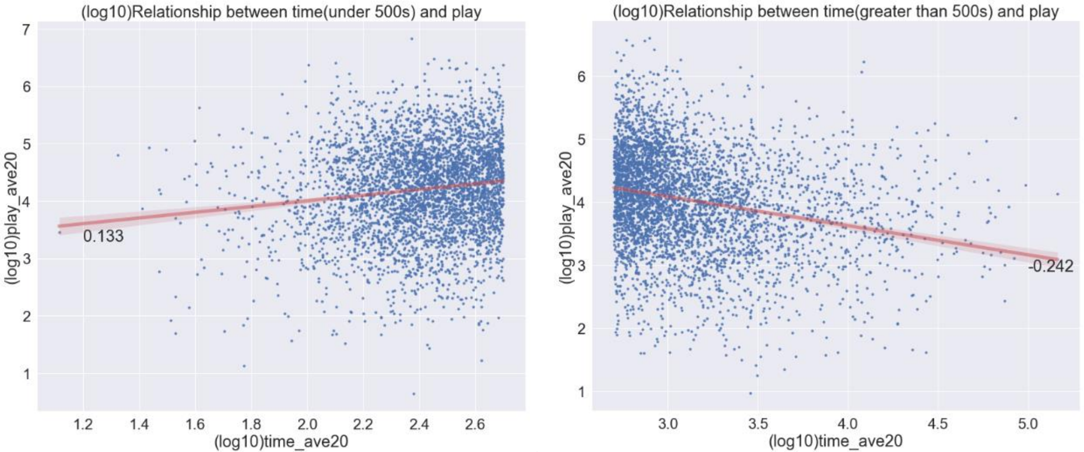


Figure 22

For play volume and video length, we can do the same as ④. It also shows familiar relationship between 500s. It is possible to find the best range of video length to gather more follower and play volume. According to Figure 19, 20, and 21, the distribution of follower and play volume shows large scale distribution in video length. The majority of time, follower and play distributed in 100s to 3000s, 1000 to 1000000 and 1000 to 300000. (**Figure 22**)

Modeling and Interpretation

Video view count is considered as an essential evaluation measurement of the popularity and influence of bilibili video creators and is directly related to their income. Therefore, with an aim to give bilibili video creators some instructional suggestions on increasing video view counts, we are highly interested in how other factors affect video view count (i.e., *play_ave20* variable in the dataset). In order to verify the conclusions in previous analyses and have a deeper insight into the contribution of certain variables to *play_ave20*, we further apply ridge regression modeling on log transformation of *play_ave20* as predicted variable.

Before modeling, we conducted data cleaning and no missing value is found. Explanatory variables are selected and transformed. For numeric variables, we take log-transformation and standardization on variables *follower*, *video*, *album*, *article* and *time_ave20* due to their skewness, and standardization on variable *master*. For categorical variables, we select *weibo*, *wx*, *qq*, *taobao*, *mail*, *tiktok*, *redbook* and *sex* as explanatory variables. Dummy variables *male* and *female* are created based on sex with 'unknown sex' taken as baseline value. Furthermore, according to previous analyses, the patterns how *time_ave20* affect *play_ave20* of short videos (*time_ave20*<500s) and long videos (*time_ave20*>=500s) are different. Thus, we divide the data into short video group and long video group according to *time_ave20*, and conduct modeling separately on the two groups. The penalty coefficients alpha is selected as 100 (**Figure 23**). Regression coefficient estimates are derived as **Table 4**.

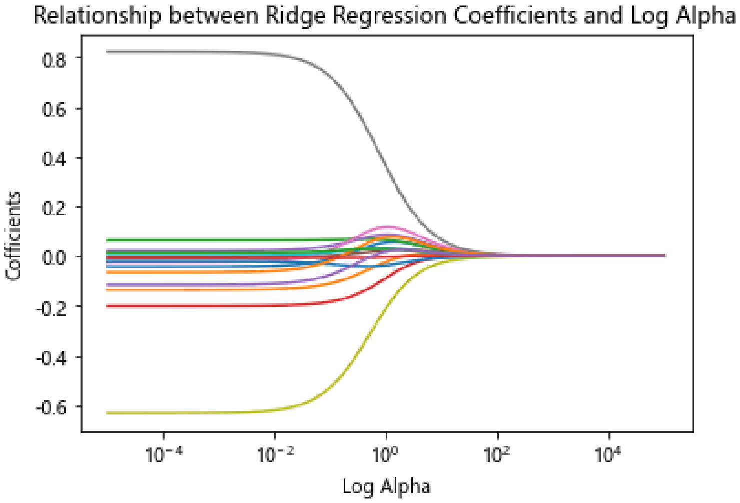


Figure 23

We have some intuitional findings as regard to the coefficient estimates. Concerning sex, female bilibili ups have slightly lower video view counts compared to males. Concerning number of followers, *log_follower* has greatest influence on *log_play_ave_20*, and they are positively related. Concerning number of videos, *log_video* has negative influence on *log_play_ave_20*, indicating that too many videos would distract fans and thus decrease view count. Concerning average time length of videos, the estimated coefficients of log_time_ave_20 for short video group and long video group suggest that videos of moderate time length are preferred by viewers. There is also finding that we think difficult to explain. Among all the linked sns, linking to wechat somehow distinctly decreases the view count of short videos.

Variable	**Short Video (N=4104)**	**Long Video (N=4256)**
Intercept	4.1817	4.0927
weibo	-0.0442	0.0143
wx	-0.1319	-0.0296
qq	0.0479	-0.0330
taobao	-0.0869	-0.0358
mail	0.0176	0.0202
tiktok	0.0040	-0.0195
redbook	-0.0021	-0.0140
male	0.0387	0.0566
female	-0.0462	-0.0437
log_follower	0.6360	0.6492
log_video	-0.2601	-0.2172
log_album	0.0058	-0.0228
log_article	-0.0234	-0.0220
log_time_ave20	0.0035	-0.0772
master	0.0107	0.0067

Table 4 Coefficient Estimates of Ridge Regression on Log Average View Count

Conclusion and Limitations

By descriptive and exploratory analysis of *bilibili.csv* dataset, we draw conclusions regarding the content ecology of bilibili. Overall, bilibili displays a role of a user-generated-content video platform loved by the young. The contents are diverse in types, and incline to the interests of teenagers and young adults. Game, Life, Anime, Technology and Fashion are most popular types of contents in bilibili. Moreover, most of the video creators are concentrated on their main types of contents. On top of that, compared to the short-video platforms that are now gaining popularity, bilibili insist on videos of longer time length.

We also explore the relationship between various variables using data visualization and modeling methods. Based on the findings, we suggest that the video creators should, on one hand, enhance number and viscosity of followers with high-quality video contents; on the other hand, control the time length of videos, preferably 5-15 minutes.

It should be noted that our analysis is time-limited because the data is not collected up-to-date. Furthermore, we do not include text variables and interaction terms in modeling. More study is needed to identify other potential explanatory variables and enhance the effectiveness of the model.