R Report

武岳 王嘉雪 谭致恒

January 2, 2024

Contents

1 Basic Information of Dataset

1.1 Introduction to Dataset

The dataset consists of observations from 8,360 Bilibili content creators encompassing 20 variables. These variables include the creator's ID (mid), number of followers (follower), gender (sex), number of videos published (video), number of featured works published (master), number of photo albums published (album), number of articles published (article), number of channels created (channel), average duration of recent videos (time ave20), average views of recent videos (play ave20), links to third-party platform information, personal tags, content categories (video tag combine), and the completion ratio of videos (video max ratio).

Further, the number of followers is the response variable under investigation. Gender and content category are categorical variables, while the number of channels created, average duration of recent videos, average views of recent videos, and the completion ratio of videos are continuous variables. The remaining variables are of count type. It is worth noting that the dataset does not contain any missing values.

Our goal is to delve into the intricate relationships among these variables to gain a comprehensive understanding of the dynamics within the dataset.

1.2 Correlation Heat Map

We generated a correlation heatmap based on the data, illustrating the relationships among various variables. In this graph [1], we consider correlation coefficients less than

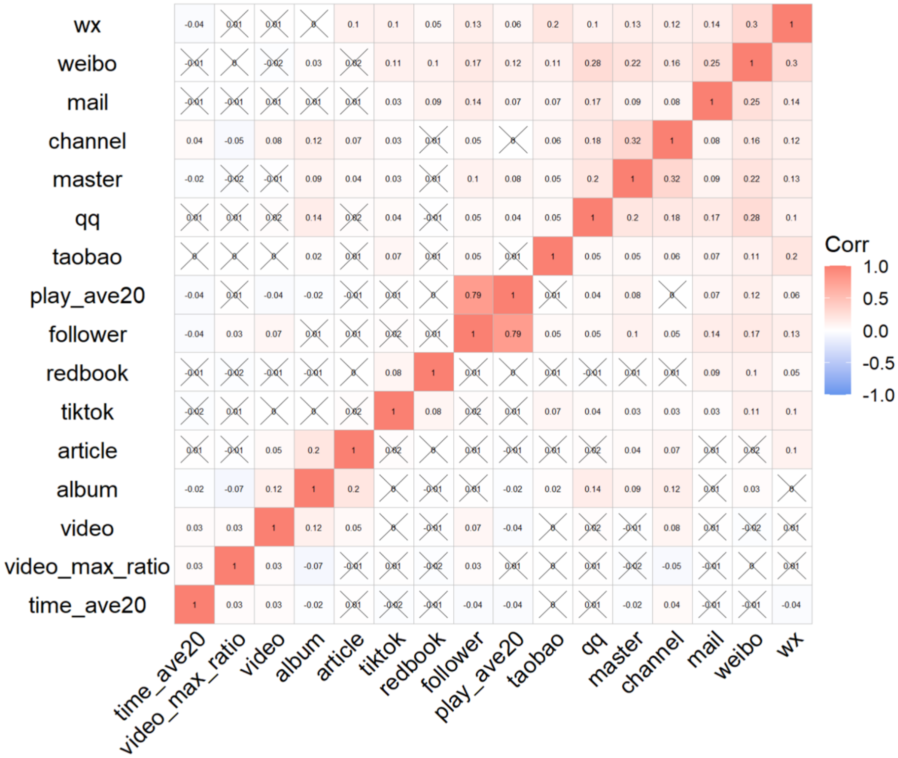


Figure 1: Correlation heat map

0.05 as statistically insignificant, denoting them with a '×' in the corresponding cells. Remarkably, we observed a strong linear correlation between the number of followers and the average views of recent videos, with a correlation coefficient of 0.79.

1.3 Log Transformation

Due to the subsequent analyses such as ANOVA and regression requiring the response variable to follow a normal distribution, we employed a log transformation to approximate its distribution to normality. As depicted in the figure [2], we used histogram smoothing and kernel smoothing methods to visualize the transformed histogram and density curve, comparing them with a normal distribution. Our observations reveal a close approximation of the log-transformed follower count to a normal distribution.

2 ANOVA

2.1 One-way ANOVA (on content category)

We initially conducted a one-way ANOVA on the categorical variable content category (video tag combine) and subsequently generated mean plot, violin plot, plot of differ-

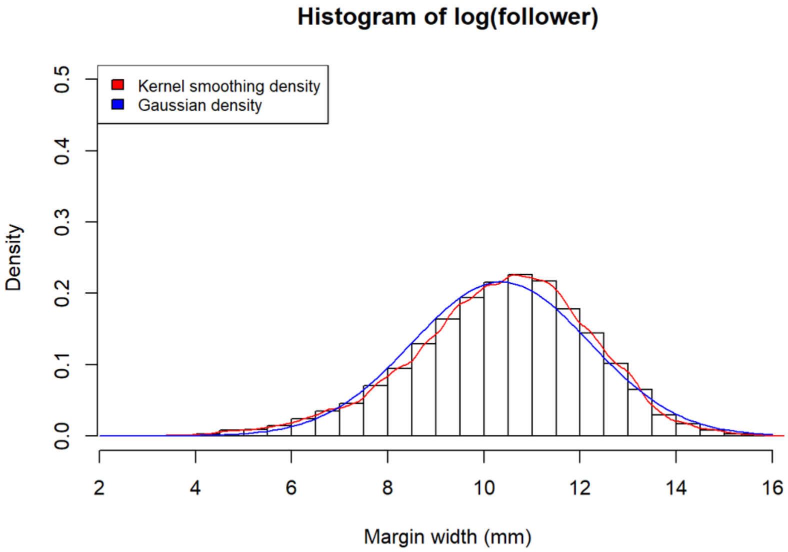


Figure 2: histogram smoothing

ences in mean levels with 95% confidence level and box plot (which are shown in figure [3]) to illustrate the mean differences caused by content category.

From the result plots [3], it is evident that the mean log(follower) of content creators vary across different content categories. Furthermore, based on the plot of differences in mean levels, we observe significant distinctions in log(follower) among content creators from different content categories at $\alpha = 0.05$ confidence level. Combining the analyses, we can conclude that the content category has a significant impact on log(follower). Additionally, the mean log(follower) of content creators in the Dance and Fashion zones are higher, while creators in the Music and Entertainment zones exhibit lower log(follower) counts.

2.2 One-way ANOVA (on sex)

Next, we conducted a one-way ANOVA on gender, employing a similar analytical approach. From the graphs [4], it is evident that all 95% confidence intervals does NOT encompass zero. This suggests that we can infer a significant difference in mean log(follower) counts among content creators of different genders; that is, gender has a notable effect on log(follower) counts. Furthermore, we discovered that female creators tend to have a larger number of followers.

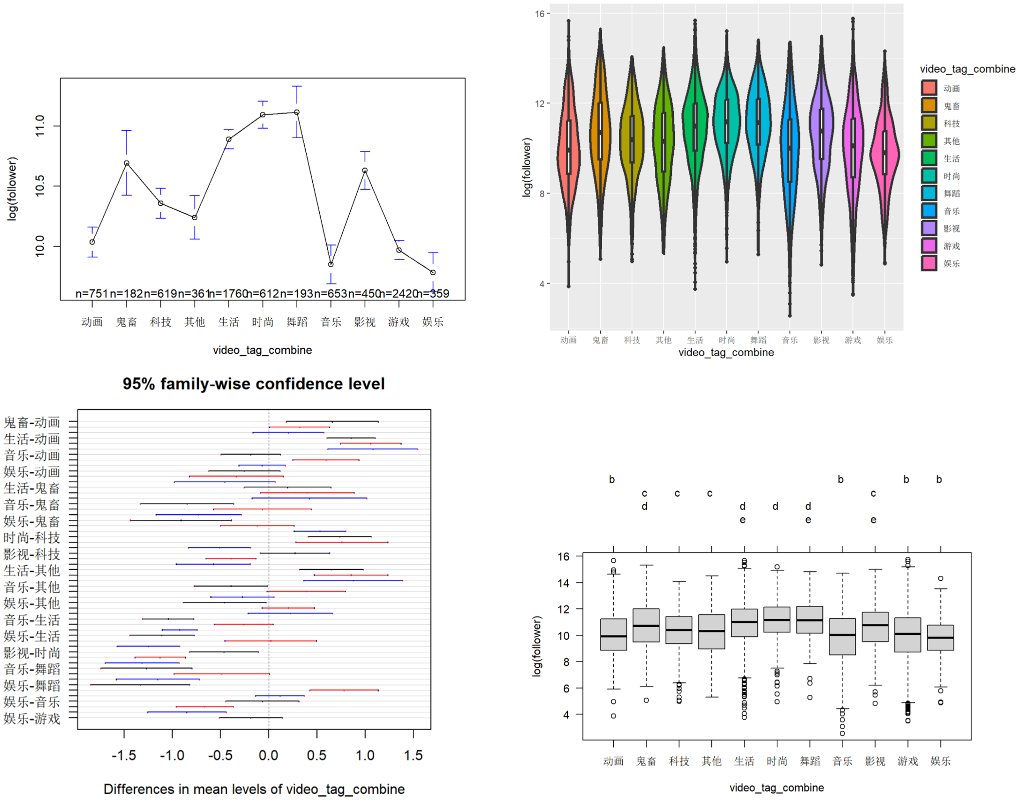


Figure 3: ANOVA (on content category) plot: mean plot (topleft), violin plot (topright), plot of differences in mean levels (bottomleft) and box plot (bottomright)

2.3 Two-way ANOVA

We have observed significant impacts of both gender and content category on the number of followers. Subsequently, we conducted a two-way ANOVA for gender and content category, producing graphical representations (top subfigures in [5]). From these graphs, it is apparent that the slopes and patterns of the curves differ among different genders, indicating an interaction between gender and content category. To further explore this interaction, we generated histograms (bottomleft subfigure in [5]) for each gender and content category, followed by radar charts (bottomright subfigure in [5]) that visualize the interactive effects based on the descending order of follower counts within each gender. The results, as illustrated in the figure, reveal distinctive patterns for different genders on the radar chart. Particularly noteworthy is the observation that content creators with a 'Confidential' gender label attract more followers with humorous and unconventional content, while male creators excel in producing cinematic content, and female creators garner more followers with dance-related videos.

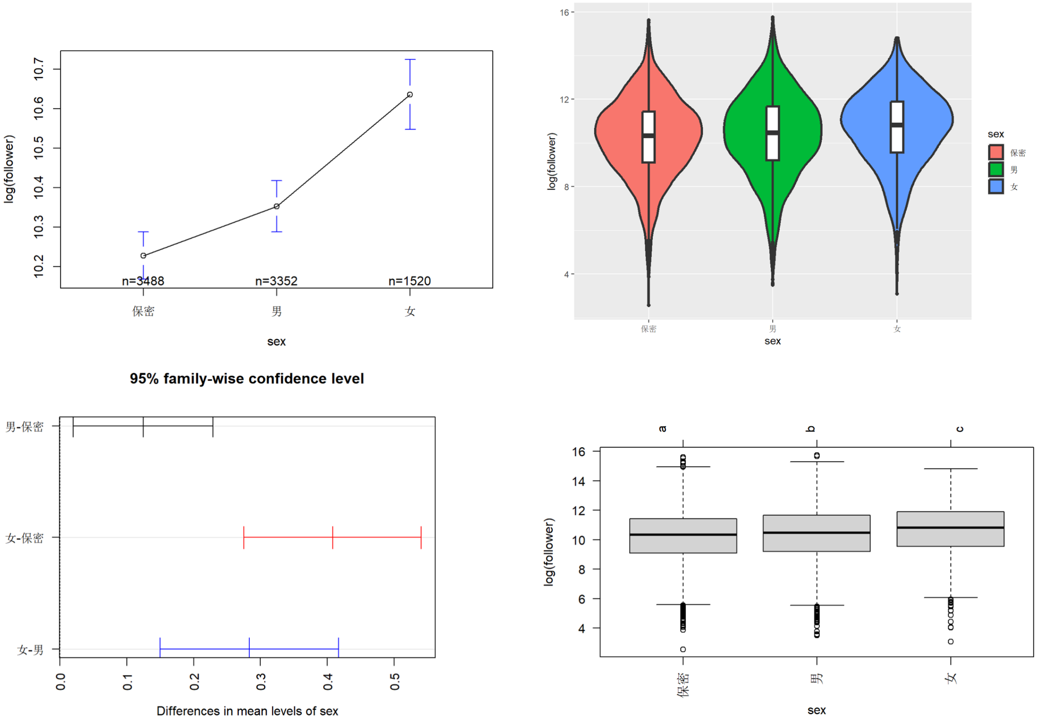


Figure 4: ANOVA (on sex) plot: mean plot (topleft), violin plot (topright), plot of differences in mean levels (bottomleft) and box plot (bottomright)

2.4 ANCOVA

We then examined the relationship between gender as a categorical variable and various continuous variables through analysis of covariance (ANCOVA), generating visual representations (shown in [6]) of how follower counts vary with these continuous variables across different genders. We observed that follower counts for different genders decrease with an increase in average duration; however, the extent of decrease varies with gender. Consequently, we infer an interaction between gender and average duration.

In the case of ANCOVA between gender and average view count, we found that follower counts increase with an increase in average view count across different genders, and the trend of increase is consistent. Therefore, we can conclude that there is no interaction effect between gender and average view count.

3 Variable Selection

In this section, we will employ three methods—Elastic Net, Random Forest, and DC SIS—to perform variable selection on the dataset. It is noteworthy that Elastic Net is particularly advantageous in selecting variables associated with linear relationships,

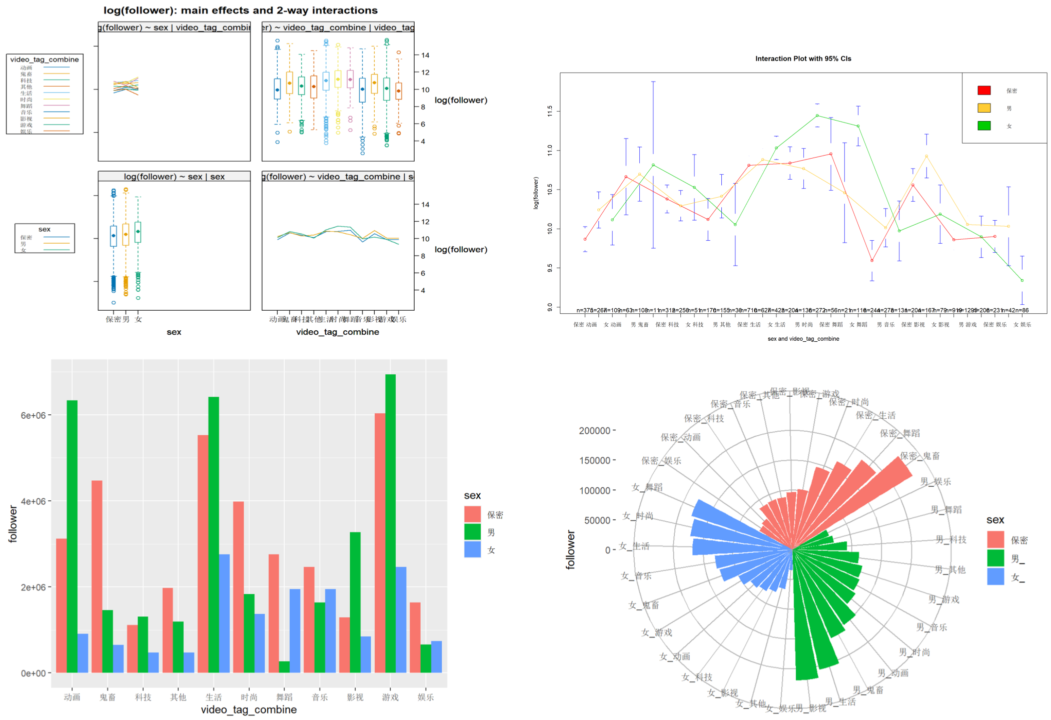


Figure 5: Two-way ANOVA plot: interaction plot (topleft), mean plot (topright), histogram (bottomleft) and radar charts (bottomright)

while Random Forest tends to favor non-linear relationships. On the other hand, DC SIS excels in handling both linear and non-linear relationships and is model-free. We will evaluate the performance of these three methods in conjunction with the actual characteristics of the data.

3.1 Elastic Net

To select variables, we configured the Elastic Net with an l_1 penalty of 0.8 and an l_2 penalty of 0.2, augmenting the l_1 penalty to facilitate variable selection. From the graphical representation in [7], it is evident that as the penalty coefficient increases, the coefficients of less significant variables gradually approach zero, eventually leading to the elimination of all variables. Therefore, by choosing an appropriate penalty coefficient, we can effectively eliminate less crucial variables while retaining important ones, achieving the objective of variable selection.

Based on the above method, the first seven variables selected through Elastic Net are weibo, master, wx, mail, qq, video max ratio and taobao.

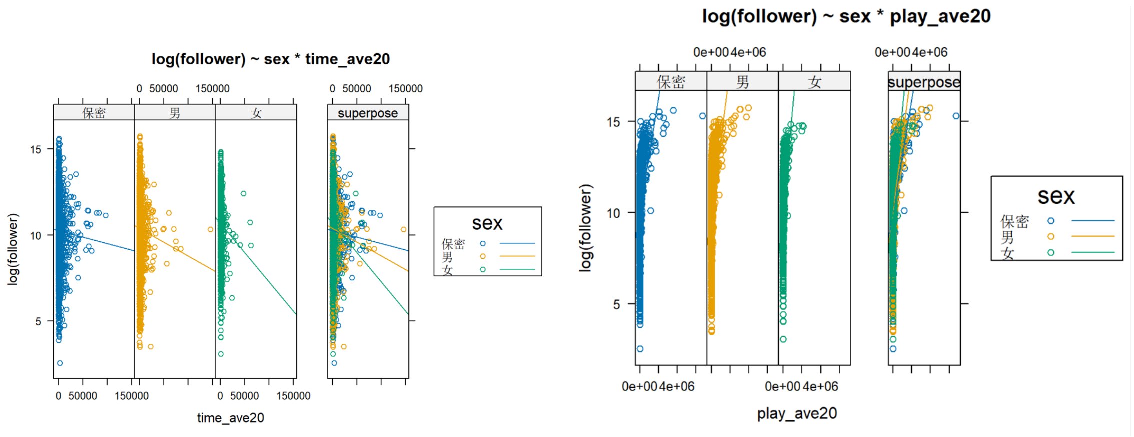


Figure 6: ANCOVA plot: ANCOVA on sex and average duration (left), ANCOVA on sex and average view count (right)

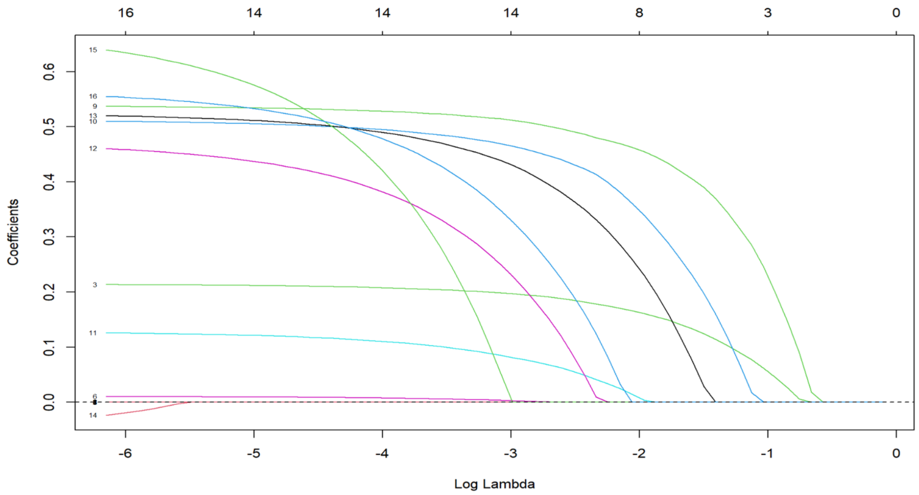


Figure 7: Variable selection by elastic net

3.2 Random Forest

After training a random forest, the evaluation of feature importance is based on measuring each feature's contribution to reducing overall impurity in the model's node splits. Specifically, features that are more frequently used for node splits and lead to a greater reduction in impurity are deemed more important. The importance of a feature is quantified by observing its contribution to the model's performance during the tree-building process. When selecting variables, we can rank features based on these importance values (as shown in [8]) and choose those with higher ranks as the most influential variables.

By ranking the features based on importance values in random forest we can select the

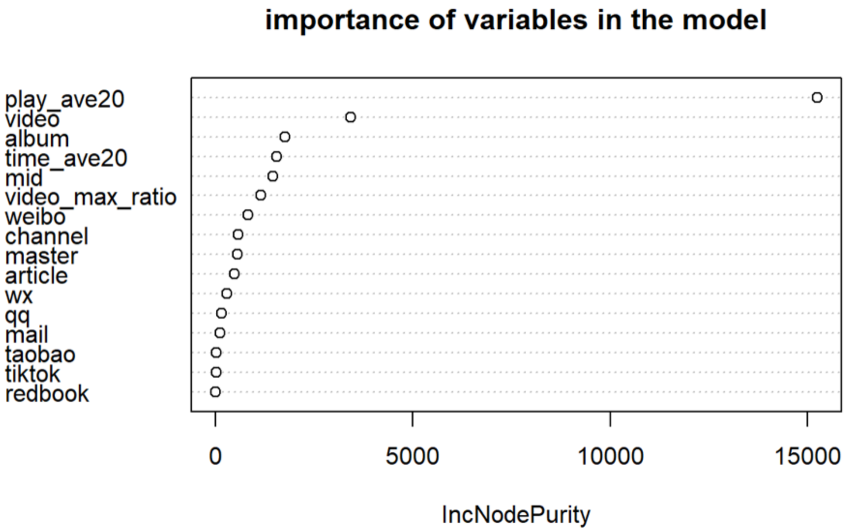


Figure 8: Variable selection by random forest

most several influential variables as follows: average view count, video, album, average duration, mid, video max ratio, weibo and channel.

3.3 DC-SIS

DC-SIS[1] first calculates the distance correlation between each feature and the response variable, which is log(follower) in this case. Subsequently, it then ranks the features based on the magnitude of their distance correlation and select the features with the highest distance correlation as the screening criteria. Traditional SIS[2] primarily relies on linear correlation for feature selection. In contrast, 'Sure Independence Screening via Distance Correlation' introduces distance correlation as a metric, enabling a better capture of nonlinear relationships among features. This makes it more effective in handling data with complex, **nonlinear** relationships. This approach enhances the ability to swiftly screen features in high-dimensional datasets, providing improved efficiency in subsequent modeling.

```
library(MFSIS)
DCSIS_result <- DCSIS(as.matrix(data), log(follower), nsis = 6)
colnames(data[ ,DCSIS_result])
```

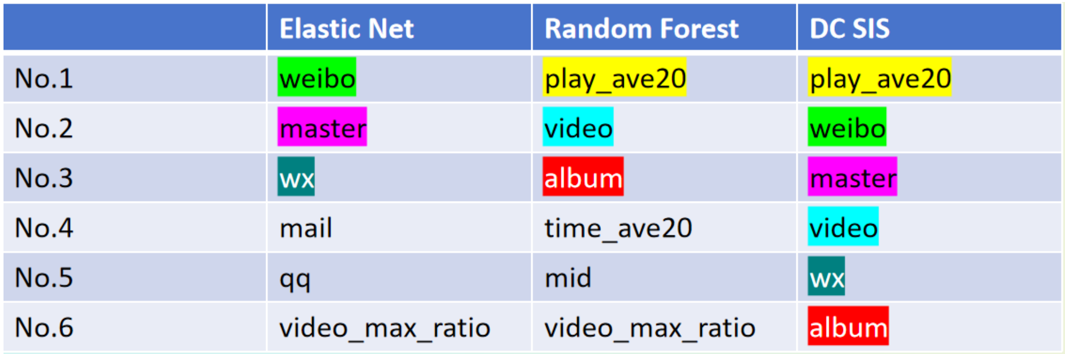
[1] DC-SIS is referenced in paper: Feature Screening via Distance Correlation Learning, which is an improvement based on the SIS method.

[2] SIS method was first published in Professor Fan's paper: Sure Independence Screening for Ultra-High Dimensional Feature Space.

By running the above code, we obtained the top six variables selected based on DC-SIS: average view count, weibo, master, video, wx and album.

3.4 Comparison of Different Methods

We list the top six important variables selected through different methods in this table. Subsequently, we will combine practical data analysis to assess the validity of these results and evaluate the performance of various methods.



	Elastic Net	Random Forest	DC SIS
No.1	weibo	play_ave20	play_ave20
No.2	master	video	weibo
No.3	wx	album	master
No.4	mail	time_ave20	video
No.5	qq	mid	wx
No.6	video_max_ratio	video_max_ratio	album

Figure 9: Table of results from different variable selection methods

It is noteworthy that Elastic Net excels in selecting variables with linear relationships, while Random Forest focuses on capturing nonlinear relationships. Here, we present an example to substantiate this observation. We conducted a NW regression on the response variable log(follower) with respect to log(average views) and employed cross-validation to determine the optimal bandwidth, resulting in the displayed graph [10]. We observed a linear relationship between the response variable and log(average views), indicating an exponential relationship with average views. This nonlinear relationship is effectively captured by Random Forest, as it selected the variable 'average views' with the highest importance ranking. In contrast, the Elastic Net method, which is advantageous for selecting linear relationship variables, did not identify this variable. Therefore, this confirms the conclusion that Random Forest and Elastic Net excel in capturing nonlinear and linear relationships, respectively.

Noticeably in table [9], among the top six variables selected by DCSIS, three originate from Elastic Net, and the remaining three from Random Forest. Moreover, their relative importance rankings remain consistent. This implies that DCSIS is indeed effective in capturing both linear and nonlinear relationships, aligning with the intended design of DCSIS.

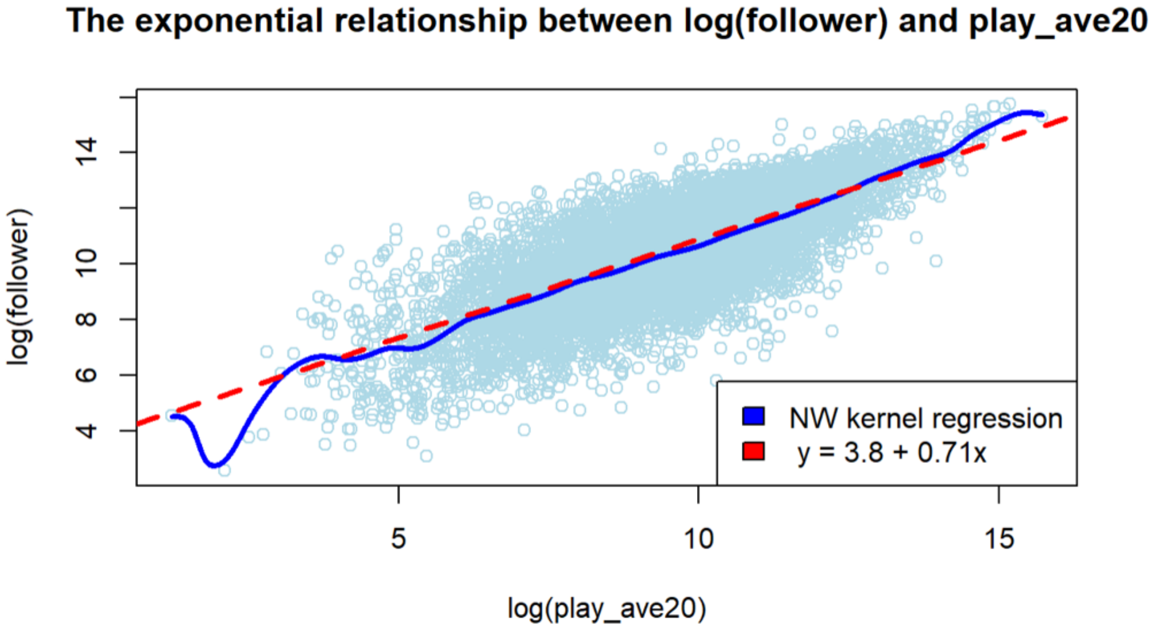


Figure 10: NW regression (Note: x-axis variable is log(play ave20) rather than play ave20)

4 Regression

We consider to model the follower using the other predictors. Notice that follower are all non-negative integers, it is count data with a thick-tailed distribution [11], and its right skewness is about 8.56. For count data, one way is to view it as a continuous variable, do log transformation, and then perform multiple linear regression, Or to fit a Poisson regression or negative binomial regression.

Data prepossessing:

- Given the variety of self-tags, for convenience, we only care whether this video has a tag or not, so we transformed the self-tags to a category variable;
- Set the category variables such as sex, video-tag-combine, self-tags as factors;
- Summarize the third party platform association information, simply superimpose the columns of Weibo, WeChat, QQ, etc., to obtain a new variable named num-platforms;

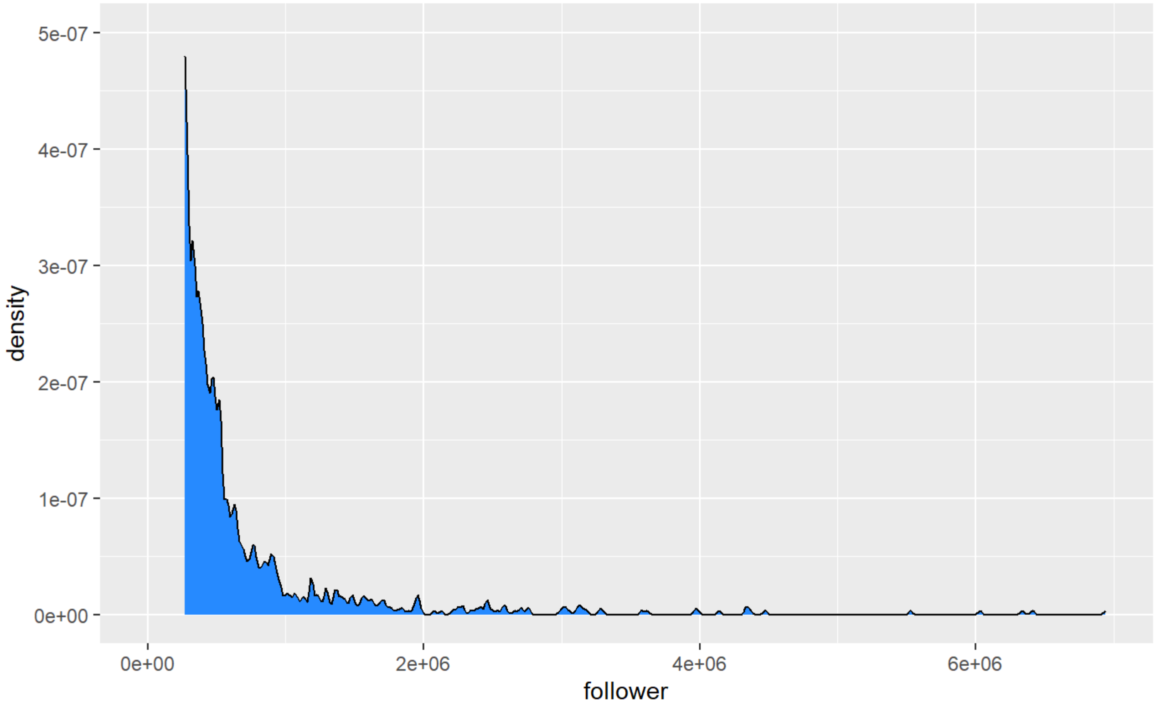


Figure 11: Thick-tailed distribution of follower

4.1 Multiple linear regression

Count outcome variables are sometimes log-transformed and analyzed using OLS regression. Many issues arise with this approach, including loss of data due to undefined values generated by taking the log of zero (which is undefined), as well as the lack of capacity to model the dispersion.

Log transformation:

Follower, video, album, article, time-ave20 and play-ave20 both have a right-biased distribution, so we consider to take log to remove the skewness. However, there exists a lot of zeros in article, from the result of variable selection, article is not an important variable, so we did not take log transformation to this variable. Even so, our observed values have decreased from the original 8360 to the current 7094.

Scatter plot:

Based on the transformed data, we plotted scatter plots[12] for these continuous predictors against follower and fitted a regression line to figure out the linear relationship between the continuous predictors and follower. Combined with the previous variable selection results obtained by DC-SIS, there exists a clear relationship between playave20, num-platforms, master, video, album and follower. In particular, the plot shows there exists a negative linear relationship between the time-ave20 and follower.

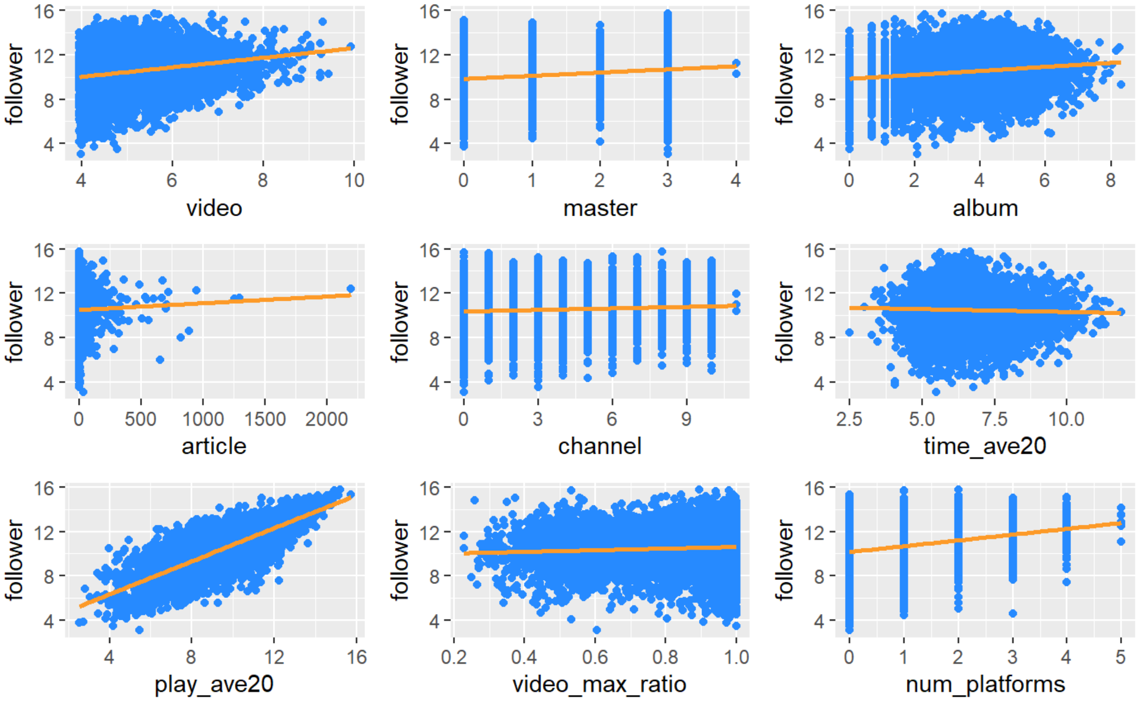


Figure 12: Scatter plot

Fit model(OLS):

We conducted initial fitting for the above five variables, play-ave20, num-platforms, master, video and album, along with a categorical variable video-tag-combine. All variables' p-value are statistically significant and the model's R-square reached 0.7279. AIC value is 19219. The p-value of F statistic is very significant, this proves the utility of our model.

Step BIC:

Then we perform the step BIC, add two more variables, time-ave20 and video-max-radio, these two variables also appeared in the previous variable selection results. Now, use these eight variables to fit a new linear regression model. The model's R-square has improved a little bit, approximate 0.7353, AIC value is 19027.

An unusual thing is that the coefficient of time-ave20 is estimated to be positive, but the correlation between the time-ave20 and follower is negative. So we guess there may exist col-linearity issue. So we computed the vif for each variable, it figures out that all variables' vif are roughly one, far smaller than ten. The col-linearity may not occur.

Validate the assumptions

- Normality

  We plot the Q-Q plot[13] and found it performed well, so the normality is satisfied.

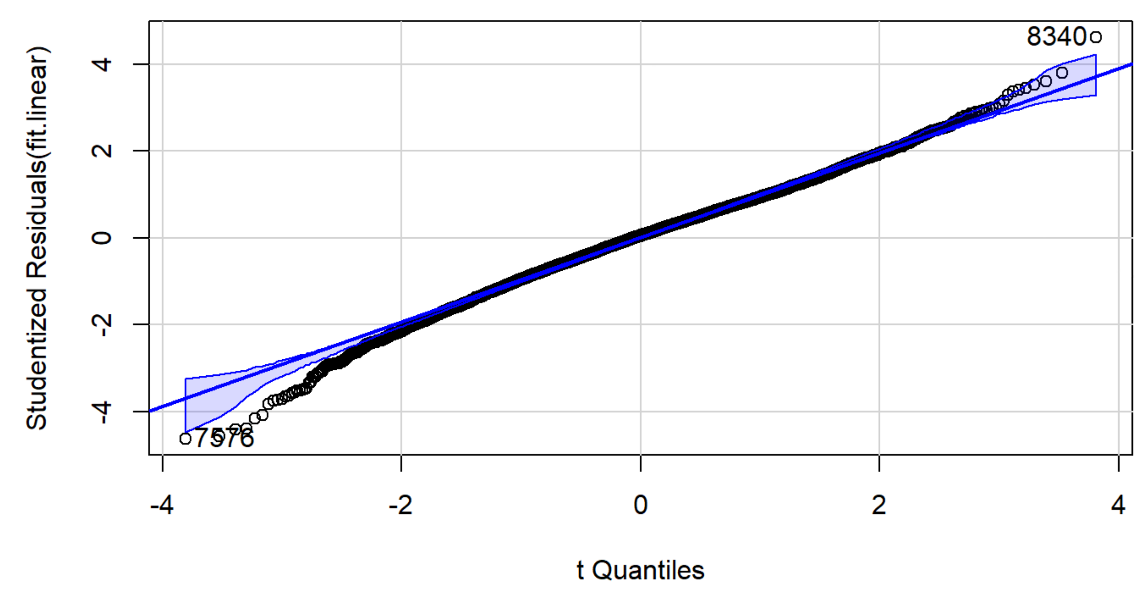


Figure 13: Q-Q plot

- Independence

  We did the durbinWastonTest to check the independence between the observations. The p-value is 0.054, not very significant, means a lack of auto-correlation between the observations, so the independence also satisfied.

- Linearity

  To check the linearity between the predictors and the follower, we drew crplots[14] . The blue line and the red line are almost overlap, so the linearity also satisfied.

- Homoscedasticity

  However, a constant variance suggests that the points in the Scale-Location graph[15] should be a random band around a horizontal line. So, the variance of the residuals are not a constant. What's more, we did the ncvTest, the p-value is very small, means the heteroscedasticity occur.

Fix heteroscedasticity

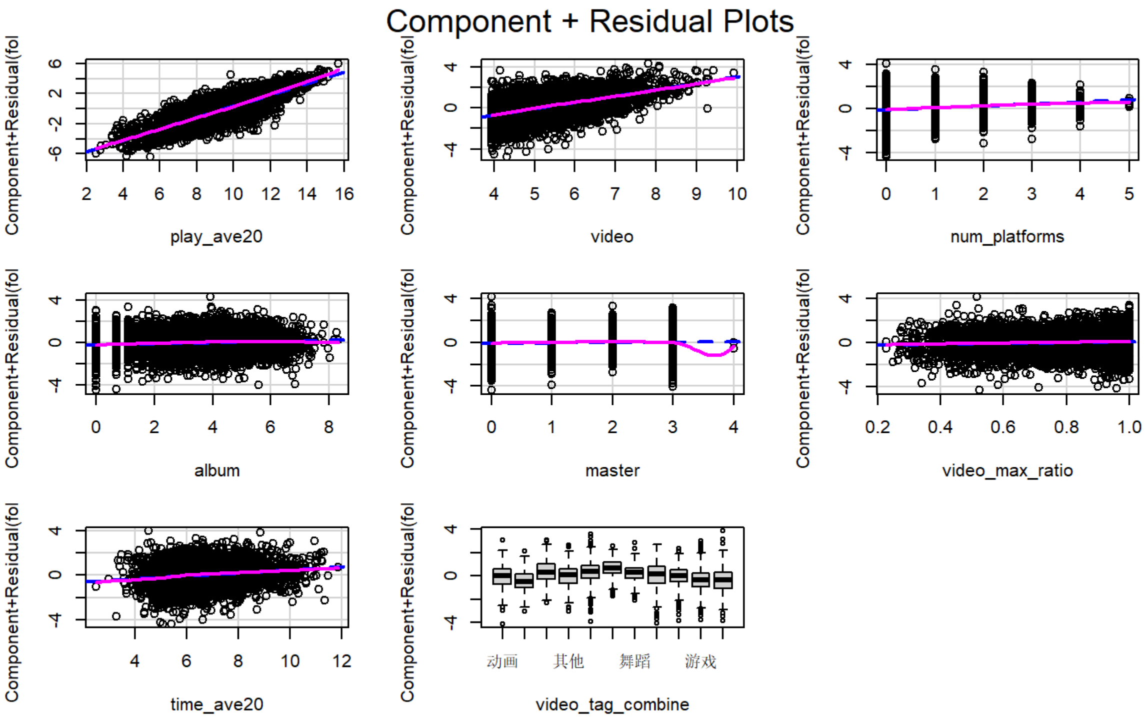


Figure 14: Linearity check

- One approach is to use the generalized least square(GLS), in this model, we assume the variance of ε is $\sigma^2 V$ rather than $\sigma^2 I$, to obtain the estimators by minimize $(Y - X\beta)' V^{-1} (Y - X\beta)$. The AIC of the GLS model is 18847, decreased a little bit than previous OLS model.

- Another way is to fit the weighted least squares(WLS), in this model, we try to minimize $\sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$, where the w_i represents the weights, here we use one over the square of residuals that produced by the previous linear model. The R-square of the WLS model is 0.8224, AIC value is 12449, having an significant improvement.

4.2 Generalized linear model

For count data, we usually do Poisson regression or negative binomial regression. We first compute the mean and variance of the follower, its mean is 141311.6 far less than its variance 123439106868, but Poisson regression assumes the mean of the outcome variable should equal to its variance. We did the qcc.overdispersion.test to check the existence of over-dispersion. The p-value is zero, strongly suggesting the presence of over-dispersion. So we choose to fit a negative binomial regression.

NB regression

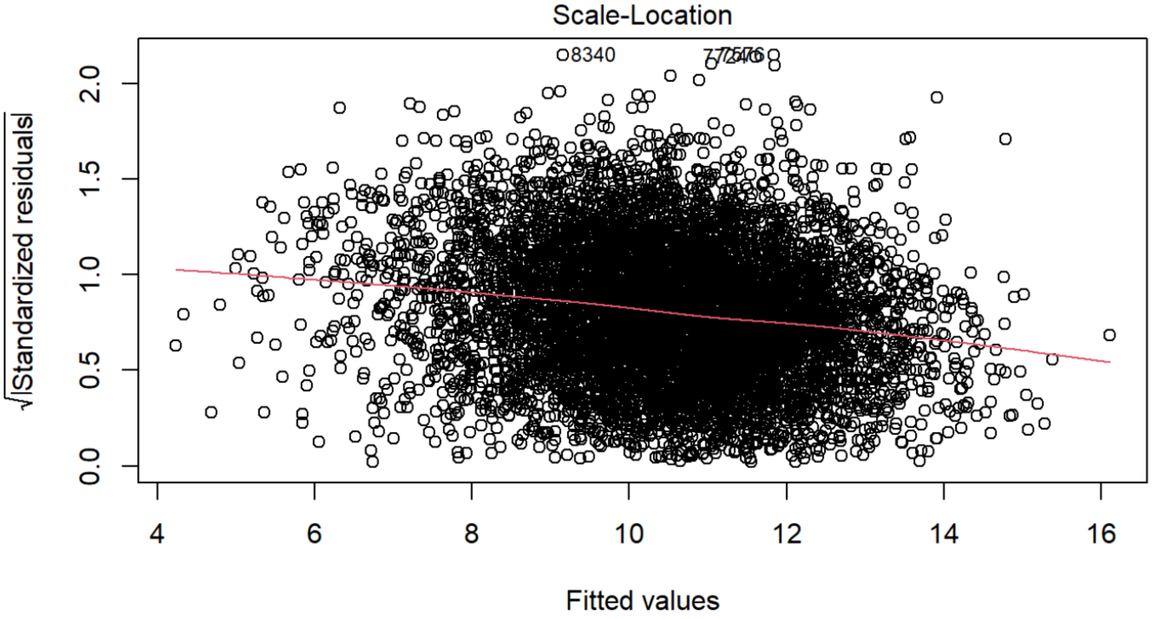


Figure 15: Scale-Location graph

The negative binomial(NB) distribution arise from a Poisson-Gamma mixture model. Suppose U is a gamma-distributed random variable with $E(U) = u$ and $Var(U) = \alpha u^2$, and that $Y|U$ Poisson(U), then the marginal distribution of Y is negative binomial with mean u and variance $u + \alpha u^2$, $\theta = \frac{1}{\alpha}$. When $\alpha = 0$, it is Poisson model.

The NB regression can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion. The mean structure is $log(u) = X\beta$, that is $log(u)$ is a linear combination of predictors. The dispersion parameter does not effect the expected counts, but it does effect the estimated variance of the expected counts.

After fitted, the estimated θ value is 1.387, and so $\alpha = \frac{1}{theta} \neq 0$, it is suitable for us to choose the negative binomial model rather than the Poisson model. However, the AIC of the NB model is much higher than the previous model, suggesting a poor fitting performance. Possibly because the time of each event varies and we did not add the time line in this NB model.

4.3 Model interpretation

- GLS

$$\begin{aligned} log(follower) = intercept + 0.77 * log(playave20) + 0.6 * log(video) + \\ 0.17 * numplatforms + 0.05 * log(album) + 0.05 * master + \\ 0.43 * videomaxratio + 0.14 * log(timeave20) \end{aligned}$$

- WLS

$$\begin{aligned} log(follower) = intercept + 0.65 * log(playave20) + 0.39 * log(video) + \\ 0.48 * numplatforms + 0.27 * log(album) + 0.05 * master - \\ 0.09 * videomaxratio - 0.04 * log(timeave20) \end{aligned}$$

- NB

$$\begin{aligned} log(follower) = intercept + 0.72 * log(playave20) + 0.6 * log(video) + \\ 0.14 * numplatforms + 0.03 * log(album) + 0.02 * master + \\ 0.08 * videomaxratio + 0.1 * log(timeave20) \end{aligned}$$

Consider the above three models, the GLS and WLS models treat the follower as a continuous variable, involving a log transformation to remove skewness. On the other hand, the NB regression assumes that log(u) is a linear combination of predictor variables.

Due to the highest correlation between follower and play-ave20, based on the fitted regression model, we quantitatively analyze the impact of changes in play-ave20 on follower. When play-ave20 increase by 100, there is an approximate increase of 25 fans; when play-zve20 increase by 1000, there is an approximate increase of 126 fans. And from the coefficients of the video-tag-combine, we found that parody videos(鬼畜视频) relatively do not attract followers, while fashion videos relatively attract followers.

5 Cluster analysis

5.1 Data preparation

Our data is about the master of Bilibili up, which is rich in multi-dimensional information. In order to better understand the data, group analysis of the classes after unsupervised classification is a good choice.We filtered the data according to the previous work, and selected the most important 7 variables(”play-ave20”,”video”,”time-ave20”,”album”,”video-max-ratio”,”channel”,”follower”) according to their importance, normalized them, and then performed unsupervised classification.Considering that the data will have isolated points and noise, and the size of the data is not too large, we used the data of K-medoids algorithm, which is more robust than K-Means, for classification. Before classification, we calculated the size of K in a variety of ways. The flow chart of our work is as follows.

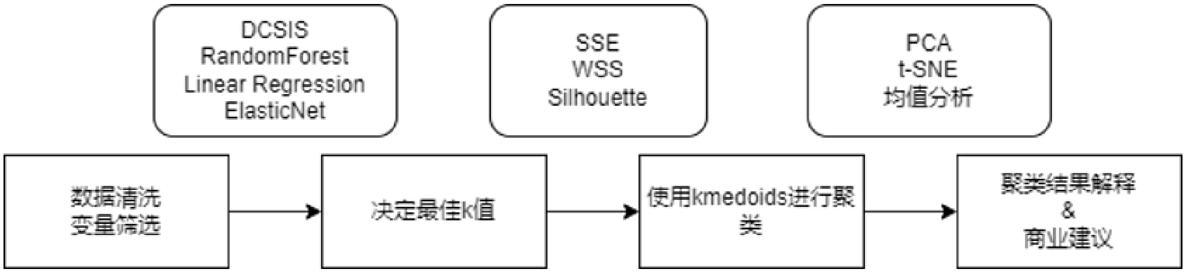
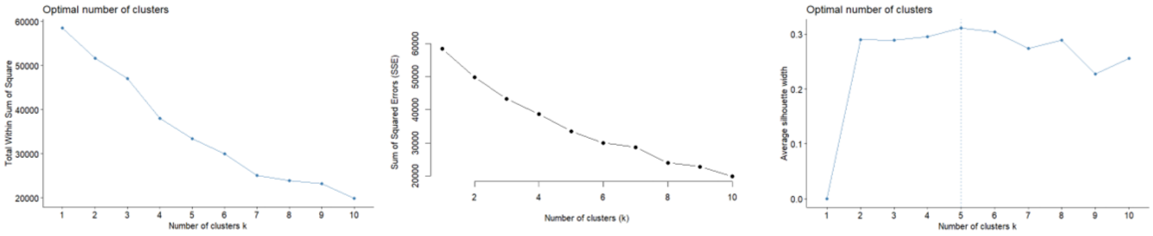


Figure 16: Roadap

5.2 Model Training and Visualization

We made total within sum of square plots, sum of squared errors plots, and silhouette score plots to help us decide the best K. It is found that K=5 has the highest silhouette score, and it also has more elbow features, so K=5 is chosen



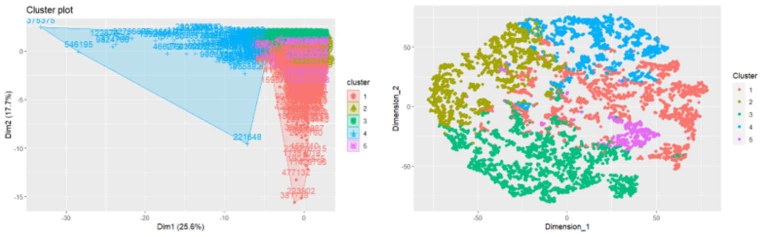
WSS SSE silhouette score

Figure 17: Decision of the optimal k

After classification, we used PCA and t-SNE methods to reduce the dimensionality of the data to two dimensions for visualization, and found that our grouping had a good effect, and the five groups were clearly separated.

5.3 Analysis of results

Next, we plot the five groups of views and followers and the within-group means of each feature for each group.



PCA t-SNE

Figure 18: Visualization

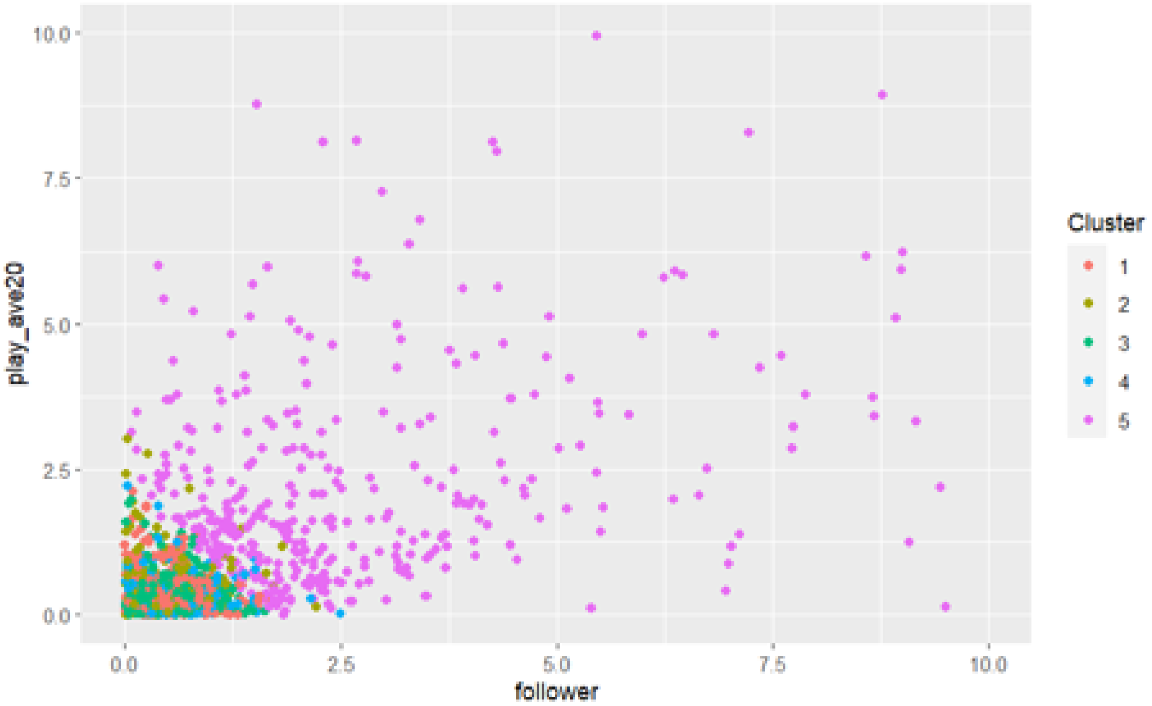
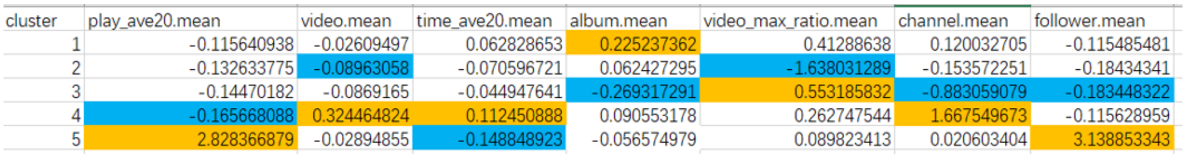


Figure 19: Amount of play - Follower

We can see the following conclusions:

- The up in cluster5 either has a lot of followers or a high play, so we classify them as big up masters, who are the best candidates when placing ads.
- There are a lot of videos in cluster4, and the average duration is relatively high, and there are a lot of channels. We believe that they are senior up masters and have their own loyal fan groups.
- Although cluster3 has a small number of fans, its completion rate and play volume are relatively high. We believe that they are a potential group with the potential to explode in popularity. When advertising, if the budget is not high, they will



cluster	play_ave20.mean	video.mean	time_ave20.mean	album.mean	video_max_ratio.mean	channel.mean	follower.mean
1	-0.115640938	-0.02609497	0.062828653	0.225237362	0.41288638	0.120032705	-0.115485481
2	-0.132633775	-0.08963058	-0.070596721	0.062427295	-1.638031289	-0.153572251	-0.18434341
3	-0.14470182	-0.0869165	-0.044947641	-0.269317291	0.553185832	-0.883059079	-0.183448322
4	-0.165668088	0.324464824	0.112450888	0.090553178	0.262747544	1.667549673	-0.115628959
5	2.828366879	-0.02894855	-0.148848923	-0.056574979	0.089823413	0.020603404	3.138853343

Figure 20: Clusters' feature

be a good choice.

- cluster2 has a relatively poor number of up plays, a relatively small number of videos, and a relatively low completion rate. We think that they may be new up masters, and we do not recommend you to choose them when placing advertisements.

6 Tag analysis

When we are engaged in the up main industry, the determination of the video partition and the length of the video is very critical. It also contains a lot of information, such as what partition is prone to pink? What partition has the highest maximum number of fans? What partition is suitable for men and women? What lengths of videos are more popular? In the following, we analyze the gender as well as the video partition and video length.

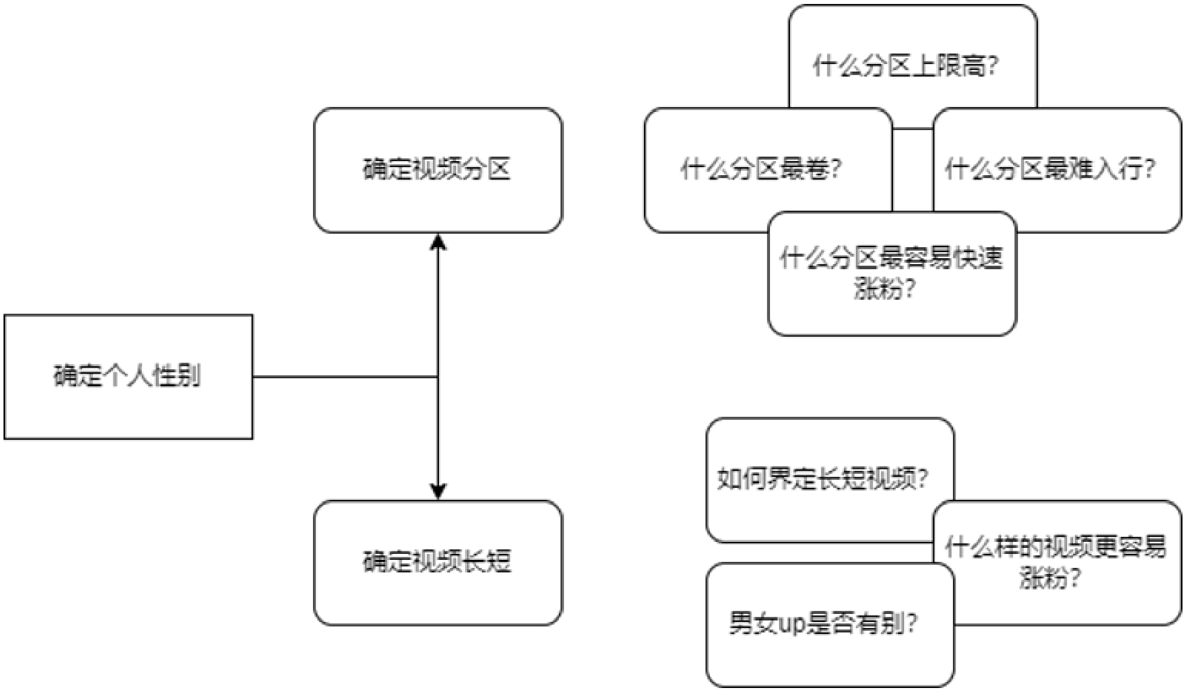


Figure 21: Questions

6.1 ANOVA

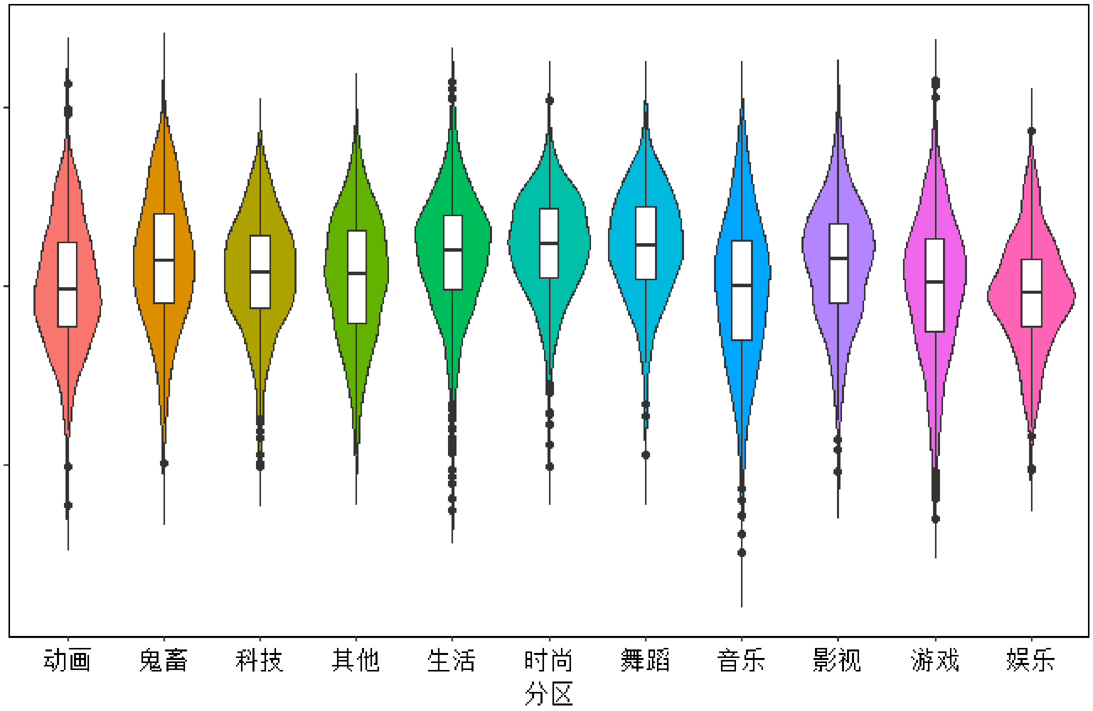
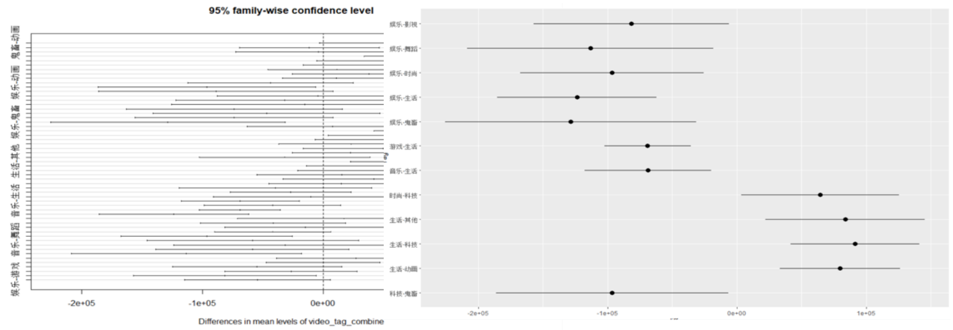


Figure 22: Violin plot

First, we want to investigate whether different partitions have a significant impact on the number of fans. We ran an ANOVA analysis on the partitions to see if the mean number of followers was significantly different across the partitions. But in practice, you'll find that both the normality test and the uniform variance test fail. Therefore, we chose permutation ANOVA for further analysis, and obtained the following results:



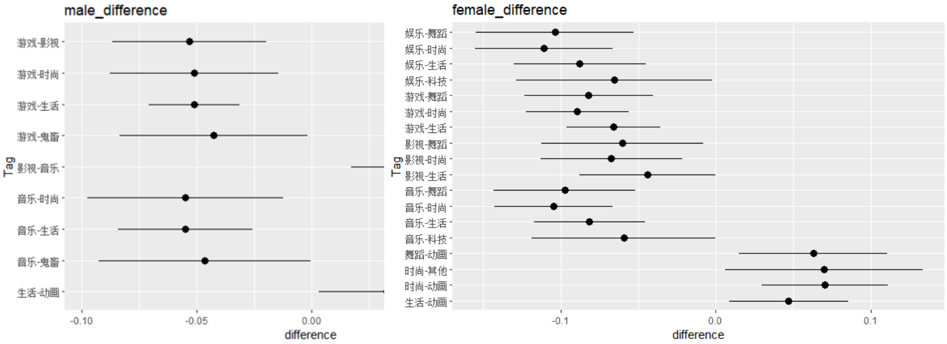
ANOVA

Significant plot

Figure 23: Permutation ANOVA for all data

We split the data into male and female and do the same for both.

The following conclusions were found: For men up:



Male Female

Figure 24: ANOVA for male and female

- For men up:
    - Not recommended game area, music area;
    - Suggest living area, fashion area
- For female up
    - Not recommended: game area, entertainment area, film and television area, music area
    - Suggestions: Living area, fashion area, dance area

6.2 Research on tag and different amount of fans

We also divided up into five main categories according to the number of fans, namely 1w-, 1W-10W, 10w-50w, 50w-100w and 100w+. After that, according to the proportion of different sections of up in different categories, we made the following heat map:

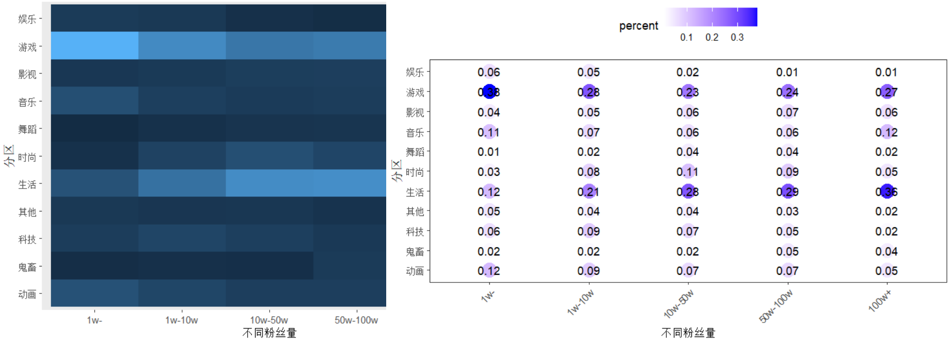


Figure 25: Heatmap for different amount fans

Each column represents the proportion of different partitions in this order of magnitude of fans. Here is our analysis of the results and suggestions for partitioning:

- It is found that the proportion of game area decreases with the increase of the number of fans, and the proportion is the largest in the low fan area. It can be seen that the game area is relatively closed and competitive.
- The proportion of living area increases with the growth of fans, and the proportion of high fans area is the largest. It can be seen that the upper limit of the living area is relatively high, and it is easier to obtain more fans.
- The proportion of autotune remix area and entertainment area has been relatively small. The autotune remix area is especially small in the low fan area, and the autotune remix area is difficult to get started. The proportion of entertainment area is especially small in the high fan area, which shows that the upper limit is relatively low. Neither of these areas is highly recommended.
- The proportion of dance area and fashion area increased first and then decreased. It can be seen that these areas are easier to obtain more fans and become the middle up, but the upper limit is not so high.

6.3 Underestimation and overestimation

Through the above analysis, it can be found that the number of plays and the number of fans are very relevant. After we fit a linear model to it, we can classify the data into underestimation and overestimation. Those ups that are underestimated (that is, the actual number of followers is greater than the estimated number) have something to learn from, and those that are overestimated (that is, the actual number of followers is less than the estimated number) have something to avoid. We made bar charts of the proportion of different partitions of up in underestimate, overestimate, and overall, as shown below:

It can be found that some partitions seem to be more likely to be underestimated, such as life and fashion, and some partitions are more likely to be underestimated, such as games and animations, which is also in line with our cognition. However, we want to quantify the degree of underestimation, so we design an underestimation parameter as follows:

$$\frac{\alpha_{under} - \alpha_{all}}{\alpha_{all}}$$

This reflects the degree of improvement in the undervaluation of the partition compared

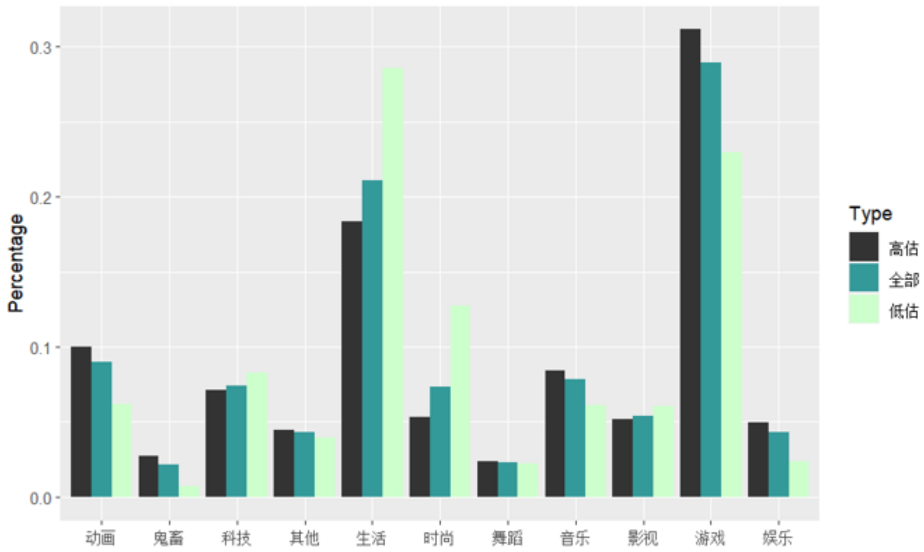
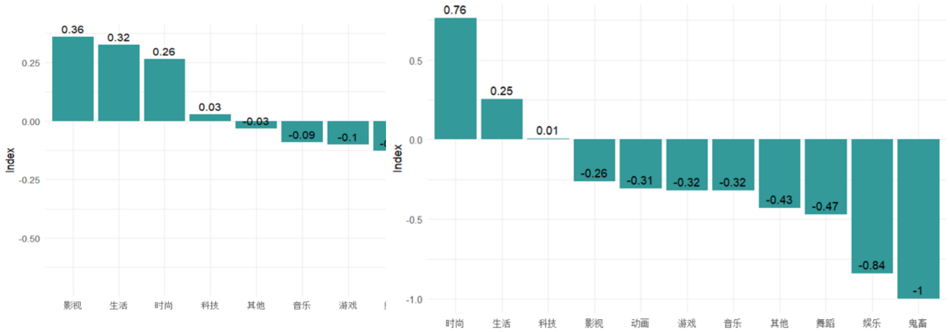


Figure 26: Bar chart

to the overall proportion. We plotted the descending exponential ranking of all up's and the descending exponential ranking of men's and women's up's as follows:



Male

Female

Figure 27: Index of male and female

The following conclusions are obtained:

- Generally speaking, it is more recommended to fashion, life, film and television area
- Not recommended for autotune remix , entertainment, animation areas
- For men, suggest film and television, life, fashion; Do not recommend autotune remix , animation, entertainment, dance.

- For women, fashion, Life science and technology are recommended, autotune remix , entertainment, dance, music are not recommended

7 Time analysis

When we become up, the choice of video length is very important. Naturally, we will divide videos into three categories: long, medium and short, but how should they be distinguished? Is three classes enough to characterize the distribution of videos of different lengths? What method should we use to distinguish? In the following, we analyze these issues.

7.1 About video length division

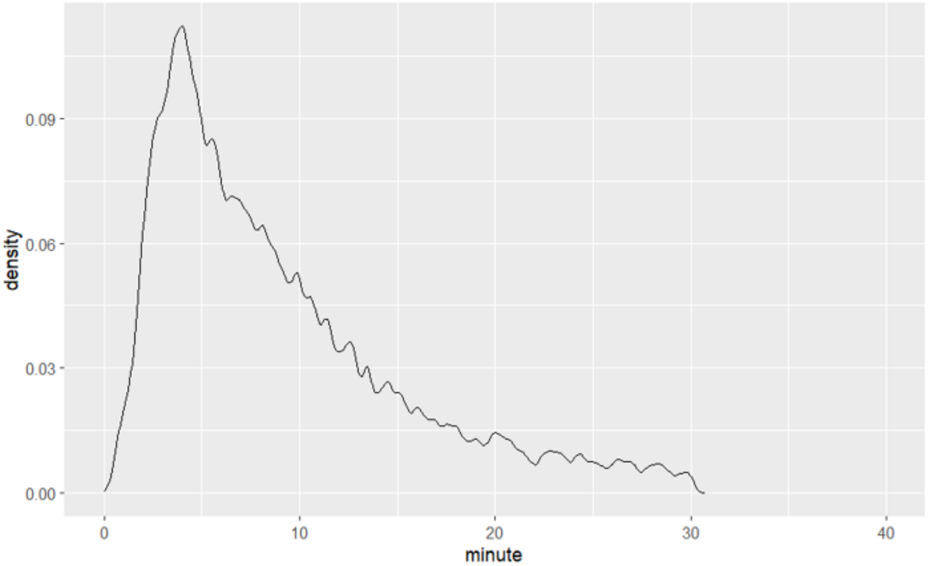


Figure 28: Nonparametric regression density

For the sake of interpretability and intuition (peak followed by a gentle decline and a thick tail), we believe that the current density distribution is the result of the addition of multiple Gaussian distributions, and consider using the GMM method to classify it.

First, we called the mclust package to see the optimal number of Gaussians and found that the optimal number was 4, but we chose k=3 because of interpretability and because 3 and 4 are not very different.

We can then classify them using GMM and obtain the following result:

The original distribution is well characterized. We choose the intersection point of

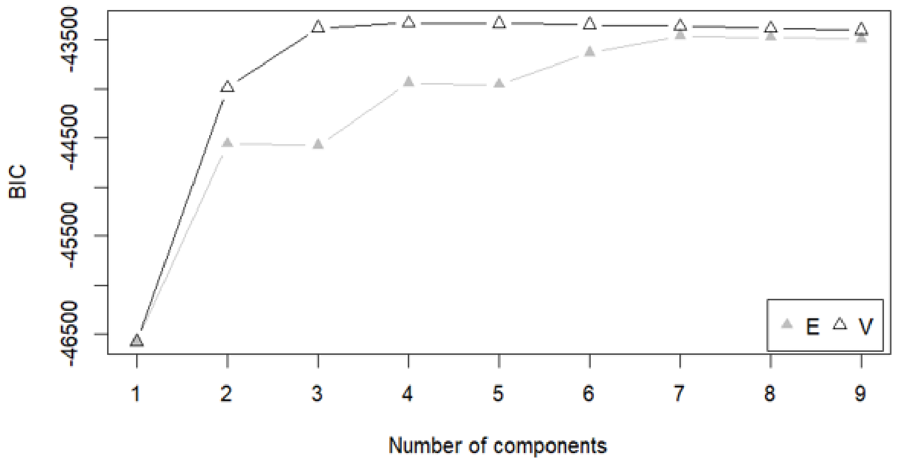


Figure 29: K-decision

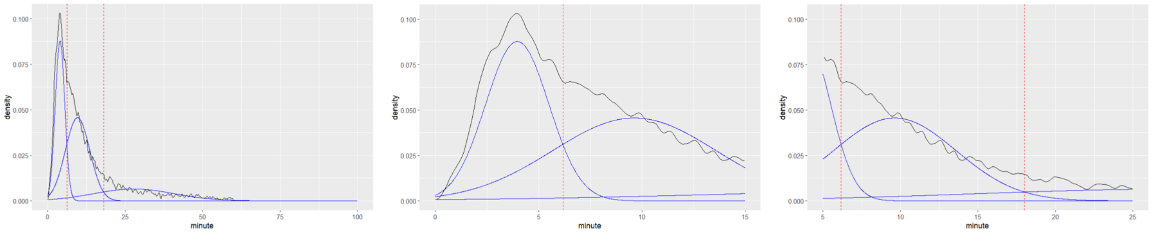
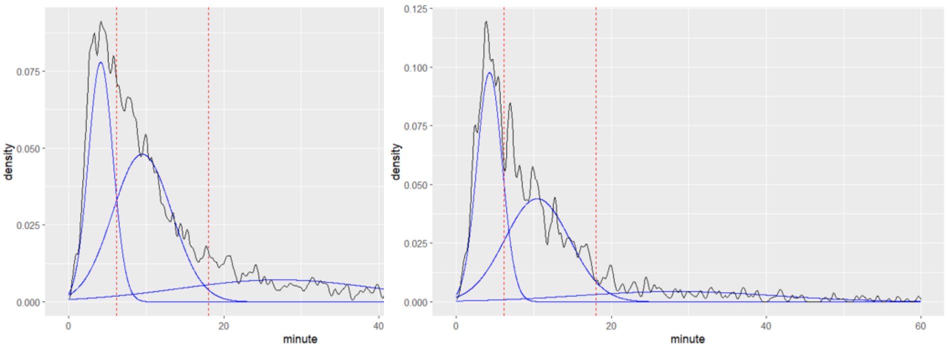


Figure 30: GMM result

three Gaussian distributions as our division and find that they are 6.17min and 18min respectively, that is, the short video is 0-6.17min, the medium video is 6.17-18min, and the long video is 18+min, which is also consistent with our cognition. At the same time, it is also found that the mean value of short video is 4mean, medium video is 9.6min, and long video is 27.9min, which can help us to better determine the length of the video.

We also plot the density distribution of different video lengths for male and female up as follows:

It can be seen that the boundaries of women are generally backward, so we have reason to believe that women can appropriately extend the length of their videos.



Male Female

Figure 31: Density of male and female

7.2 Research on video length and number of followers

Next we wanted to see, do different video lengths lead to different numbers of followers? We first made a violin plot for visualization to help us understand intuitively

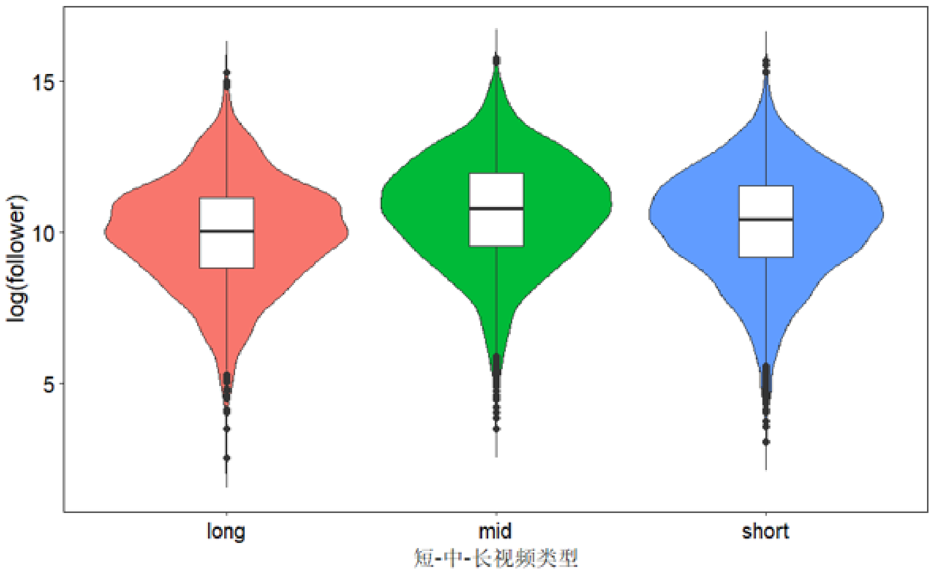


Figure 32: Violin plot

It seems that medium videos have the highest number of followers, followed by short videos, and then long videos. Using permutation ANOVA for quantitative analysis, we get the following results:

It can be seen that we recommend giving priority to short and medium videos and less to long videos.

We also make the following partitions for different video durations:

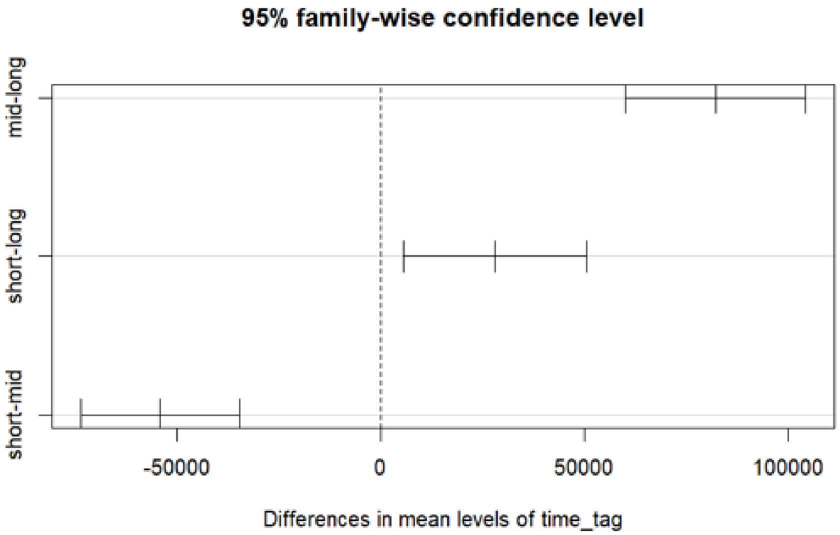
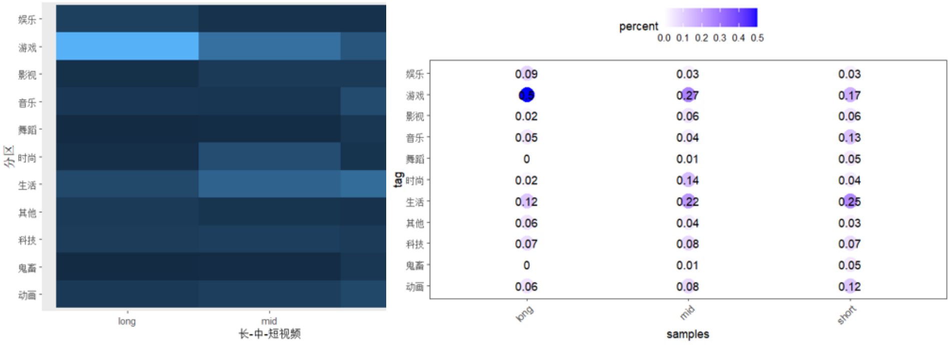


Figure 33: ANOVA result



Male Female

Figure 34: Heatmap of time

It can be seen that the game area has the largest proportion of long videos, while the dance area has a large number of short videos, which is reasonable, because the game area is usually recorded and broadcast, and the strategy is generally relatively long, and there are few dances longer than 18 minutes in the dance area. This is all very much in line with reality.

8 Conclusion

Finally, based on the above analysis, we give relevant suggestions and analysis for engaging in the up industry as follows:

- Competition is fierce in the game area, there are many bottom up, please choose carefully.

- The autotune remix area is difficult to get started, and the overall easy to overestimate; The entertainment area has a low ceiling and is easily overestimated, so be careful.

- The dance area and fashion area are easy to be underestimated, which can quickly acquire a large number of fans and become the middle up, but the upper limit will not be very high.

- The best suggestion is the living area, the upper limit is high, the entry threshold is low and easy to be underestimated, the fan increase speed is also relatively fast, is a good choice.

- For men, the film and television area is relatively a good choice, and women have more advantages in science and technology

- In terms of video length, it is recommended that you send short and medium videos (less than 18min), and it is not recommended to send long videos

Finally, data may be cold, but I hope that through our analysis, everyone can create successful and engaging videos. I also wish all the students who aspire to become popular (referred to as 'up') can enter the top 100 rankings soon