Uncovering statistical properties of YouTube videos

Son Le

1 Introduction

YouTube is the most popular video-sharing platform, hosting videos that can have up to billions of views. What makes YouTube interesting is that any video can "go viral" if it has the right recipes and its authors (called "YouTubers" or "content creators") either get lucky or know the inner workings of the platform. Usually, when a video gets popular in a short amount of time, it gets on YouTube's trending list and earns even more viewers. Although this trending list sounds like a goal for content creators, many of them, especially beginners, are unaware of how to get their videos on the trending list. In other words, it is unclear to them what makes a video popular and how different elements of a YouTube video interact with each other. With a goal to help such YouTubers, this project studies the relationships between different aspects of a YouTube video, such as its publish time, category (as indicated by the creators), its title and description, and its "statistics" (views, likes, dislikes, comment count, and how long the video was trending - creators cannot control these measurements).

The dataset in this project is collected by J [1] on Kaggle. The dataset contains daily statistics on videos featured in YouTube's trending list for several countries/regions, as the trending list is different for each region, over a time period from 2017 to summer 2018. See here for more information on the dataset. This project uses a modified version of the original dataset, and only considers the region of Great Britain and videos published in the first six months of 2018.

Since I am interested in the relationships between the characteristics of a video, my research questions are:

- Can variances in a video's statistics (views, likes, dislikes, number of comments, and how many days the video was trending) be explained in few dimensions (as there are many variables just to describe user interaction)?
- How do aspects of a video relate to each other, especially between a video category and its view count?

The rest of this report is organized as follows. Section 2 gives a description of the variables, some summary statistics, and some visualizations. In section 3, more visualizations are given, and some analysis on bivariate dependencies are conducted. Section 4 gives the descriptions, implementations, and results of PCA and MCA. Section 5 provides a summary of the main findings, and finally section 6 and gives critical evaluations on the whole analysis.

2 Univariate analysis

The original dataset contains several records for one video over the days it stays on the trending list. In the modified version, each video only has one record; for each video, the time series over the different days the video is trending is summarized as follows: the interaction statistics are the numbers seen on the video's last trending day, and the trending day timestamps are summarized as the number of "trending days" for the video. In addition, each video's publish timestamps is separated into its month, day, and hour components (all videos analyzed are published in 2018). A day is described as its day of week (e.g., Monday), and whether it is a weekend or a weekday.

Since analyzing raw text data is difficult, an external algorithm [2] was used to compute sentiment scores for the titles and descriptions. These scores range from -1 to 1; the higher a piece of text's sentiment score, the more positive the text. Based on thresholds for sentiment scores in the same paper, a text can be classified into positive, neutral, or negative classes. Although the (in)accuracy of the algorithm might invalidate the analysis, I checked its results and most of the time they agree with my opinions. Furthermore, since many pieces of text receive a sentiment score of 0 (or very close to 0), the analysis might conclude that there is little association between, e.g., title sentiments and views.

The variables considered in this project are: `cat`, the category of a video (e.g., `Comedy`); `tags`, the number of tags of a video (creators can assign tags to their videos in hope of reaching out to more viewers); `pub_dow`, `pub_hr`, the publish

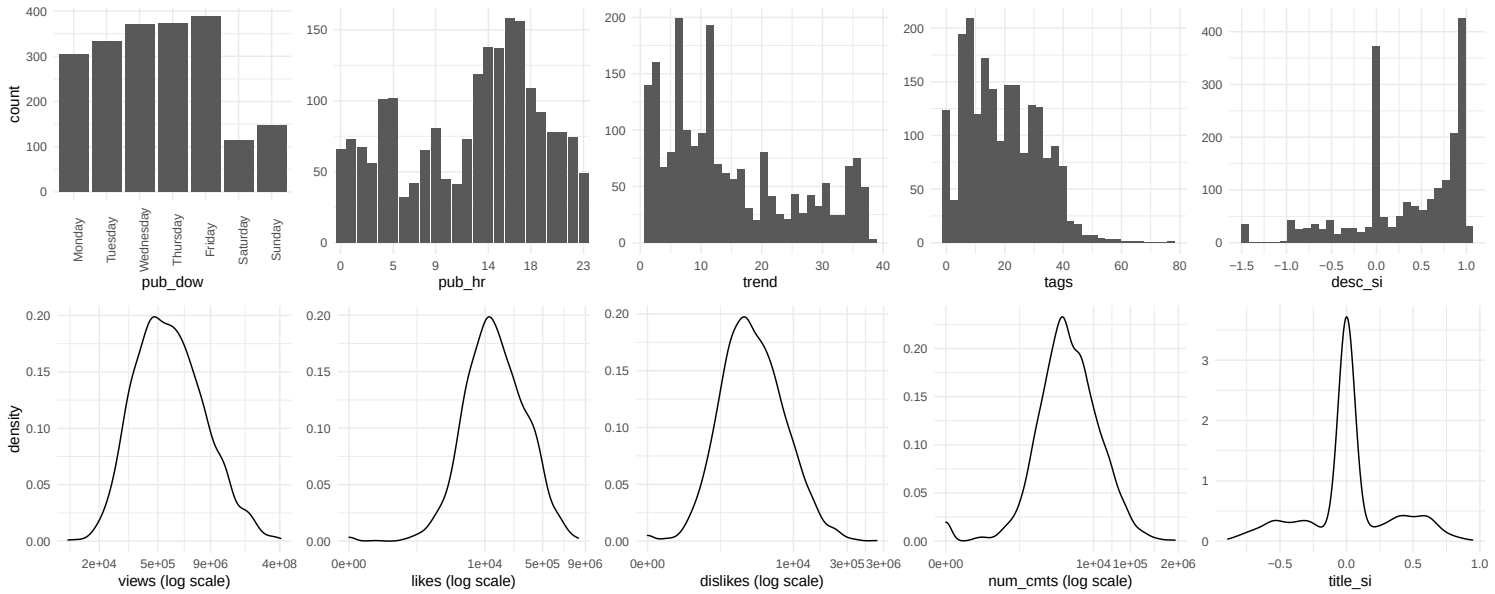

Figure 1: Univariate plots.

Table 1: Summary statistics (there are 2032 videos in the dataset).

	tags	title_si	desc_si	pub_hr	views	likes	dislikes	num_cmts	trend
Min.	1	-0.89	-2.00	0	3.6e+03	0	0	0	1
1st Qu.	9	0.00	0.00	7	2.5e+05	5e+03	1.9e+02	6.4e+02	6
Median	18	0.00	0.49	14	8.6e+05	1.8e+04	6.7e+02	1.9e+03	11
Mean	19	0.02	0.34	12	6.3e+06	1.2e+05	6.9e+03	1.2e+04	15
3rd Qu.	29	0.00	0.89	17	3.4e+06	7.9e+04	2.9e+03	6.9e+03	22
Max.	78	0.95	1.00	23	4.2e+08	5.6e+06	1.9e+06	1.6e+06	38

day of week and hour of the video, respectively; `is_weekend`, an indicator whether a video is published on a weekend or not; `title_si`, `title`, the sentiment score and sentiment class of a video's title, respectively; `desc_si`, `desc`, the sentiment score and class of a video's description, respectively; `views`, `likes`, `dislikes`, `num_cmts`, how many views, likes, dislikes, and comments a video has, respectively; and `trend`, the number of days the video was trending.

Figure 1 and Table 1 summarize variables mentioned above. We see that few videos were published during the weekend. There were more videos uploaded in the afternoon than in other times in a day. Many videos trended for about 15 days or less. Most creators used 30 or fewer tags for their videos. Many descriptions have sentiment scores of zero (`desc_si`) or very positive scores. This pattern does not apply to `title_si` as many titles have sentiment scores around zero. Note that there are 36 videos without descriptions, which appear as -1.5 in the histogram. This value is only for visualization purposes, and the class of videos without descriptions would be "unknown".

In the bottom row of Figure 1, the first four plots show that those four variables seem to follow a log-normal distribution. However, this is not the case; for the logarithm of each of these variables, I conducted the Shapiro-Wilk normality test, and the results were that the null hypothesis of normality is rejected.

Figure 2 illustrates the proportions of the modalities for each categorical variable. We observe that more than half of the videos belong to either the Entertainment or the Music categories. The "Other" category consists of categories to which very few videos belong (under 30 for each of these categories). More than half of videos have positive descriptions whereas about this proportion of videos have neutral titles.

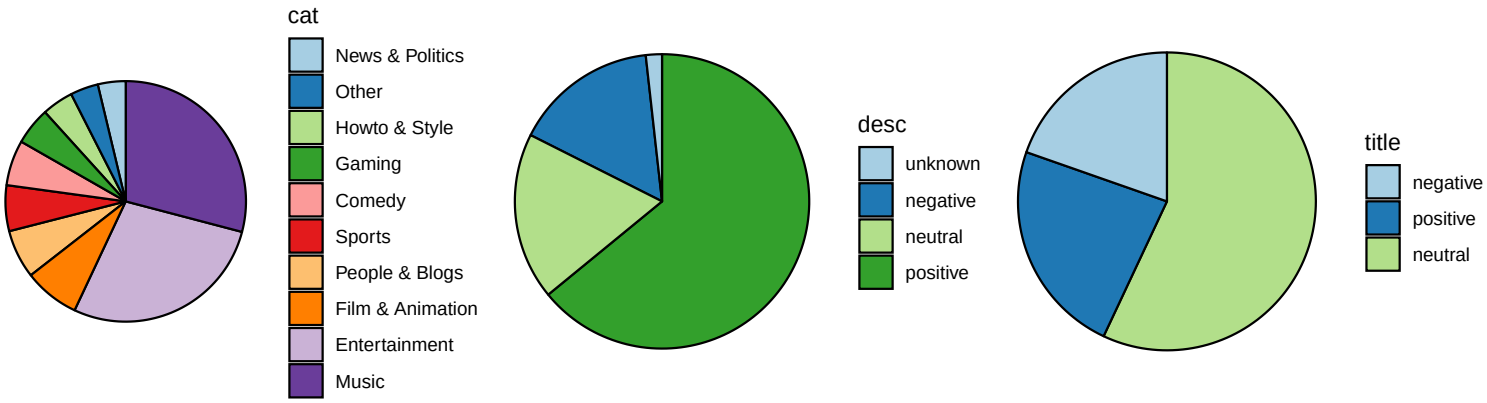

Figure 2: Distribution of some categorical variables.

3 Bivariate analysis

In this section, we discuss bivariate relationships. Figure 3 depicts some plots on relationships between two categorical variables, between a categorical variable and a numerical variable, and between two numerical variables. In the top row, the first plot shows that visually categories have similar distributions in publish hours. One category that is quite different from the others is `Comedy`: many videos of this category were uploaded earlier in a day. The `Howto & Style` category has the opposite pattern. The second plot shows that no matter the category, more than half of the videos have neutral-sounding titles. Whereas `Howto & Style` have significantly more positive-titled videos than ones with negative titles, other categories have similar proportions of videos with positive or negative titles. As for the box plots, it is evident that `Music` videos have both the largest median number of trending days and view counts. On the other hand, the "outliers" shown in these plots might be simply because there are fewer videos belonging to those categories in the dataset. That said, `Howto & Style` has the most outliers with regard to the number of trending days.

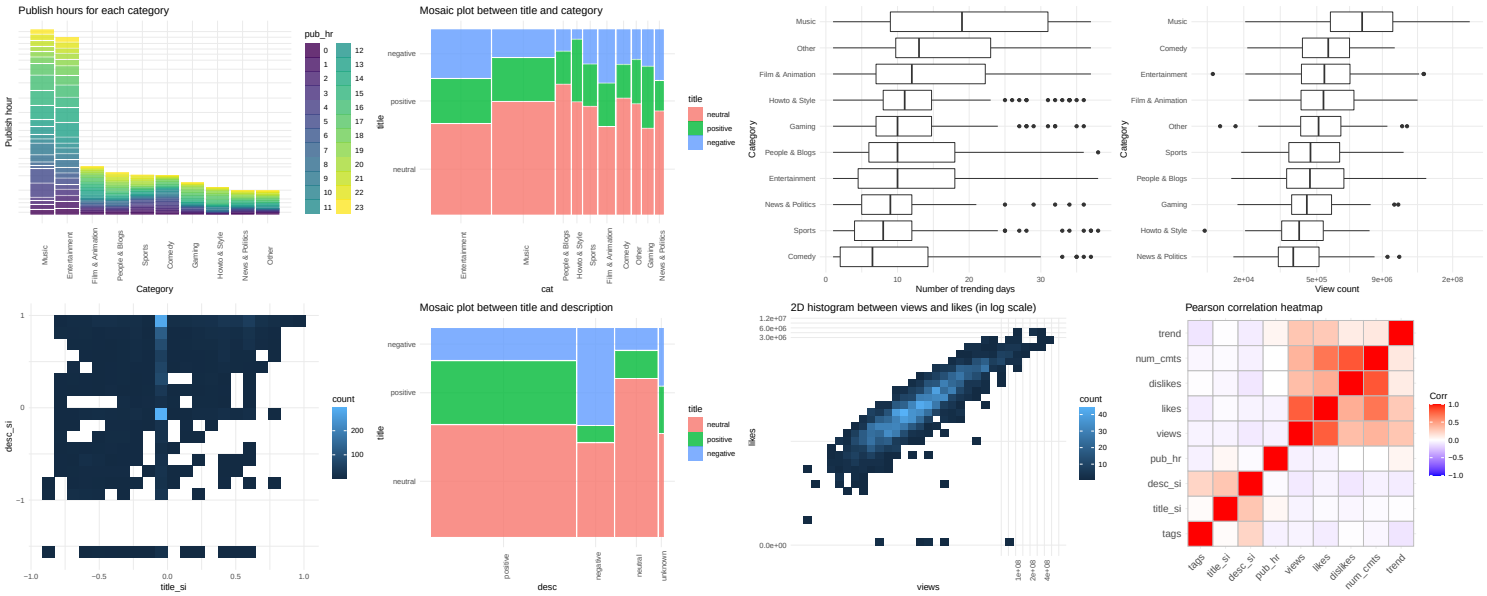

Figure 3: Some bivariate relationships.

In the bottom row of Figure 3, the first plot does not show any clear pattern between sentiment scores of titles and descriptions. However, in the second plot, videos with negative-sounding descriptions rarely have positive-sounding titles. Similarly, positive descriptions go with positive titles more often than negative titles. The third plot shows that the logarithm of `views` and `likes` seem to have a linear relationship. According to the final plot, `title_si` and `desc_si` have a weak positive correlation whereas the five "statistics" variables in the upper right corner have mild to strong positive correlations.

Figure 4 reveals more about the linear relationships between those five variables. First, we notice that the graphical plots are in normal scales, showing potential outliers. The true shape of the distribution of these variables are depicted in the diagonal kernel density estimate plots (compare these with the KDE plots in Figure 1). Next, the highest correlation occur between `num_cmts` and `dislikes`. This is no surprise as videos with high dislike counts tend to be controversial, thus generating much discussion in the comment section. On the other hand, more views often means more likes.

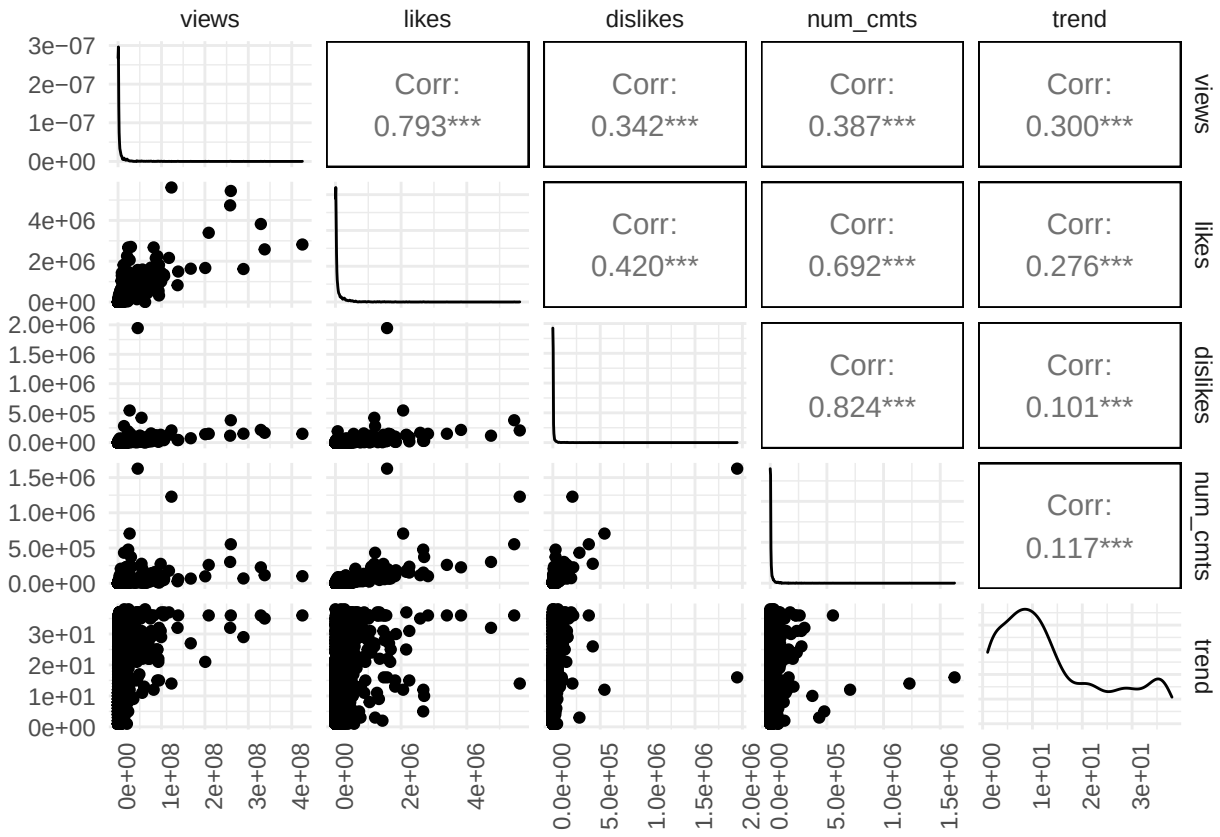

Figure 4: Pairs plot between five statistics variables.

4 Multivariate analysis

4.1 Principal component analysis

We first answer the first research question with principal component analysis (PCA). In short, PCA seeks a projection of the data cloud onto an orthogonal basis, in which the first dimension explains the greatest variance in the data; the second dimension contains the second largest variance, and so on. These dimensions are called principal components (PCs), each of which is uncorrelated with all other PCs. As the data is projected onto this new basis, we could perform "dimension reduction" by only keeping the first PCs containing the desired amount of variance and discard the rest.

Since we are trying to summarize the five "statistics" variables (for want of a better term) in fewer dimensions, PCA would be a suitable method for this task. In addition, we have noticed mild to high correlations between these variables, further motivating the use of PCA as it is an algorithm that retains as much information as possible in as few dimensions as possible. The method was implemented using the `princomp` function in the `R` programming language, with `cor = TRUE`, the option to perform PCA based on the correlation matrix of the data. The theoretical derivations of the technical implementation can be found in Ilmonen [3] and Ilmonen [4].

The top row of Figure 5 shows a summary of the PCA. The red line in the first plot corresponds to a component with eigenvalue 1, or the average percentage of explained variance (20% in this case since there are five variables). About 90% of the variance in the data is explained by the first two PCs, the first of which explains over 50% of the variance. On the other hand, according to the (simple) Kaiser criterion, we should keep all PCs with eigenvalues above 1, i.e., higher than the red line in the first plot, i.e., `PC1` and `PC2`. And we shall do so for visualization purposes.

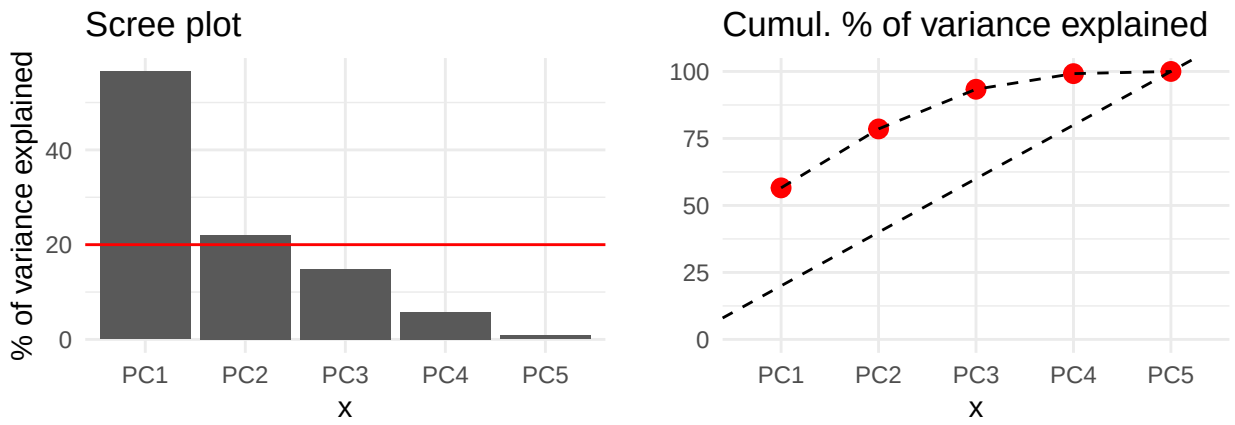

Figure 5: PCA percentage of explained variance (individual PCs and cumulative).

Table 2: Loadings for each variable on the principal components

	views	likes	dislikes	num_cmts	trend
Comp.1	0.46	0.53	0.45	0.51	0.21
Comp.2	0.37	0.18	-0.47	-0.39	0.68
Comp.3	0.47	0.35	-0.36	-0.18	-0.70

Table 2 contains the loadings of each variable on the first 3 PCs. Together with Figure 6, we can interpret the first two components as follows. All variables except for `trend` have relatively large positive association with the first PC, so this component likely measures the volume of interaction on each video. On the other hand, the `trend` variable has the largest coefficient on PC2, to which only `dislikes` and `num_cmts` is negatively associated. Thus, PC2 probably tries to separate trendy videos with high views and likes but with lower dislikes and number of comments (relative to the view and like counts), or vice versa. Although PC3 looks similar to PC2, `trend` is highly negatively associated with the former component. As such, PC3 potentially distinguishes between trendy videos with high dislikes and comment counts (controversial videos) but lower views and likes (relative to the dislike and comment counts), or vice versa.

The points in both plots in Figure 6 are the projected data points onto the first two PCs. Both of these plots show one clear outlier (at least according to the five variables used in the current analysis), namely the point at the bottom right corner. This point corresponds to a video regarded as highly controversial at the time. This video has a like count lower than both comment and dislike (highest in the entire dataset!) counts, which is a rare phenomenon in YouTube context. The second plot best describes this point. The other outlying points mainly correspond to trendy music videos with very high like-to-dislike ratio and high views. The first plot best describes these points.

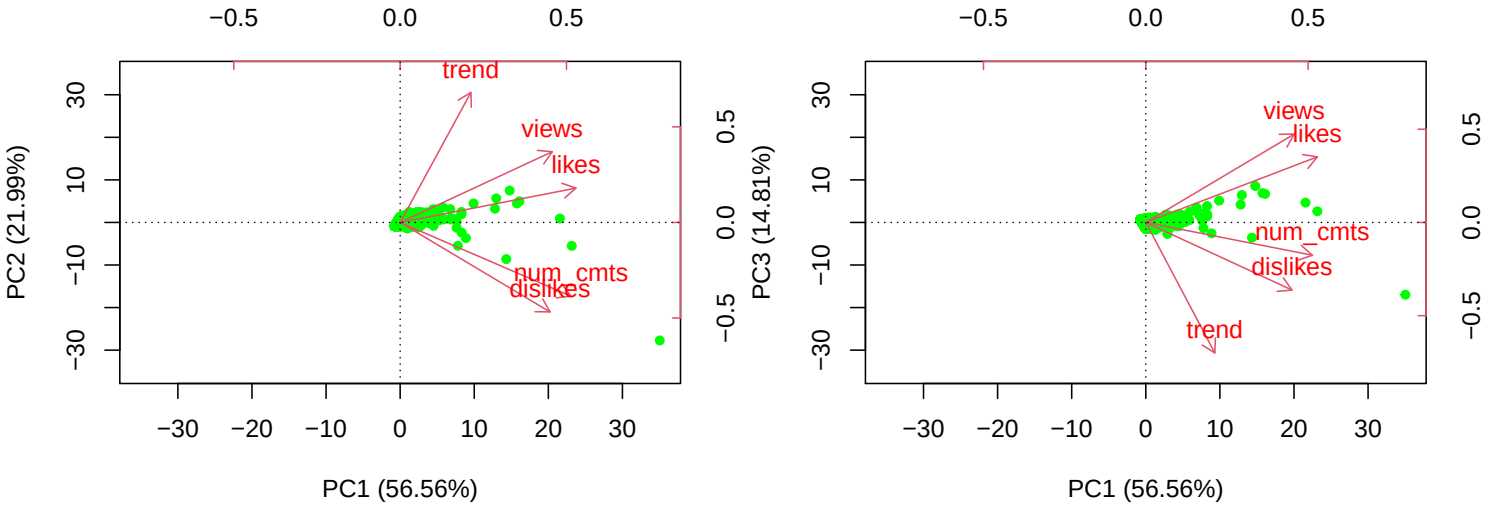

Figure 6: PCA scores and loadings (red axes depict coordinate scales for loading vectors).

4.2 Multiple correspondence analysis

Next, we answer the second research question with multiple correspondence analysis (MCA). Correspondence analysis (CA) is similar to PCA but it analyzes categorical data. In short, CA tries to summarize categorical data in as few dimensions (called principal axes) as possible, each axis being orthogonal with all other axes, and describes possible dependencies between categorical variables, resulting in a two-dimensional graphical display which allows for convenient interpretation of the data. MCA is an extension of bivariate CA to the case of more than two categorical variables. Whereas bivariate CA analyzes the contingency (frequency) table of two categorical variables, the version of MCA used in this project applies CA on the complete disjunctive table of multiple categorical variables. This table can be thought of as containing all one-hot label encodings for the categories.

As the second research question pertains to categorical variables, particularly `cat` the category of YouTube videos, MCA is naturally a suitable method. The method was implemented by using the `R` programming language's `ca` package containing the function `mjca` with option `lambda = 'indicator'` for performing MCA based on the complete disjunctive table (indicator matrix). The theoretical derivations for this particular MCA method could be found in Ilmonen [5]. For the analysis, we use the following eight variables, the first four of which are `cat` (10 video categories shown in the first pie chart in Figure 2), `title`, `desc` (the classes shown in the last two pie charts in Figure 2), and `is_weekend` (is the publish day of the video a weekend?). The last four variables are `tags`, `pub_hr`, `views`, and `trend`. Each of these numerical variables has been binned into three categories with comparable counts to avoid a modality being "rare". All in all, there are 31 modalities among all eight variables. In addition, I only use `views` as a representative of the four user interaction variables (the other three being `likes`, `dislikes`, `num_cmts`) since we discovered in section 4.1 that they can be explained pretty well with the first principal component.

First, we look at Figure 7. Total inertia is the χ^2-statistic divided by the total sample size. Similar to how in PCA each PC explains some of the variance, each principal axis in CA (and MCA) explains some of the total inertia (called principal inertia). To be more specific, the sum of the eigenvalues corresponding to each principal axis is the total inertia. One may look at Figure 7 and think that the first axes contain very low inertia; however, the advantage of MCA is visualizing categorical data in two dimensions. The red line in the first plot, again, correspond to an axis with the average inertia. We have 23 axes, where 23 is the total number of modalities minus the number of variables.

Next, Tables 3, 4, and 5 displays various statistics on the column profiles. The `mass` of a column profile is the marginal relative frequency of that profile ($f_{.pl}$ in Ilmonen [5]). The `qlt` of a column profile is the quality of display of the profile in the two-dimensional map (in the tables, the first dimension is the first principal axis; the second dimension is the second axis). To be more specific, each `qlt` equals the sum of squared cosines [6] of the angle between a profile and each axis in the subspace (in the tables: `cor + cor.1 = qlt`). The inertia `inr` of each column is the inertia it contributes to the total inertia. `k=1` and `k=2` are the scores of the profiles in the subspace spanned by the first and second principal axes. The contribution `ctr` of a modality is the amount of inertia it contributes to the construction of the corresponding axis. These terminologies are further explained in Greenacre [7].

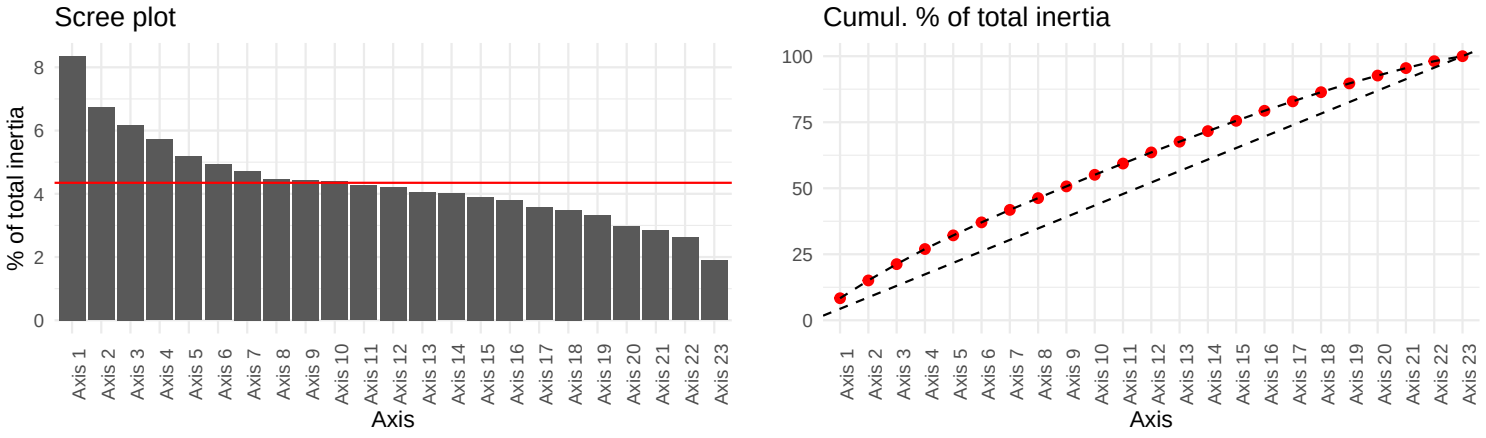

Figure 7: Percentage of inertia contained in each principal axis.

The statistics of interest in Tables 3, 4, and 5 are the column masses `mass`, qualities of display `qlt`, `cor`, and `cor.1` as these numbers directly relate to the plots in Figure 8. `desc:positive` has the highest mass, and `cat: Music` has the

Table 3: Column profiles summary 1 (‰).

	cat:Entertainment	cat:Music	cat:People & Blogs	cat:Howto & Style	cat:Sports	cat:Film & Animation	cat:Comedy	cat:Other	cat:Gaming	cat:News & Politics
mass	35	36	8	5	8	9	8	5	6	5
qlt	174	503	110	118	73	1	97	24	52	78
inr	31	37	38	41	39	37	39	39	39	40
k=1	-324	1104	-285	-1067	-927	-109	-483	-182	-813	-660
cor	41	500	6	50	56	1	15	1	35	17
ctr	15	185	3	25	28	0	7	1	17	8
k=2	-588	78	1223	1238	-503	52	-1120	760	573	1259
cor.1	134	3	105	68	17	0	82	22	17	62
ctr.1	62	1	63	42	10	0	49	14	11	38

Table 4: Column profiles summary 2 (‰).

	title:neutral	title:positive	title:negative	desc:positive	desc:negative	desc:neutral	desc:unknown	tags:[1,11]	tags:(11,25]	tags:(25,78]
mass	71	29	25	80	20	23	2	42	43	39
qlt	78	8	105	86	26	160	44	264	18	415
inr	18	32	36	16	37	36	40	30	27	31
k=1	102	-160	-104	-192	37	640	-40	396	5	-429
cor	14	8	3	65	0	92	0	80	0	85
ctr	3	3	1	12	0	39	0	28	0	30
k=2	220	8	-648	-109	-372	549	1554	601	185	-845
cor.1	64	0	103	21	26	68	44	184	18	330
ctr.1	18	0	53	5	14	36	28	79	8	146

highest quality of representation in this subspace (spanned by axes 1 and 2). Meanwhile, this subspace fails to display `cat:Film & Animation` as the `qlt` of this modality is only 1‰, meaning that most of the inertia of this modality lies in some other subspace. We also notice low `qlt`s of several other variables. To simplify this report, I only consider principal axes 1 and 2 although the next few axes explain a similar amount of inertia as axis 2.

Perhaps the most interesting part of MCA are the graphical displays. The first plot in Figure 8 displays the scores of the column profiles in the two-dimensional subspace spanned by principal axes 1 and 2. The size of the dots is based on the mass of the corresponding column profiles. The transparency of the lines and the dots is based on the quality of display of the corresponding column profile in the subspace in the plot. The first axis is most noticeable for its clear separation of Music from all other video categories as well as the separation of the highest `trend` and `views` bins from the lower bins. These modalities themselves are attracted to each other (small angles between the lines), indicating that videos with the Music category tend to have very high view counts and stay on the trending list for many days. Music videos also tend to have neutral titles low tag count, and neutral titles, which is not true because more than half of Music videos have positive descriptions. The general pattern in YouTube confirms these interpretations except for the last one. Note that `title:neutral` and `title:positive` is close to the origin, indicating that these modalities are very similar to the average modality, or that they are poorly represented in the subspace. Furthermore, since there are so few videos without descriptions in the dataset, the modality `desc:unknown` lies the furthest from the origin.

Meanwhile, the four video categories People & Blogs, News & Politics, Howto & Style, Gaming, and Other usually have low views, average trending day count, and published late in the day. These observations are reasonable, except

Table 5: Column profiles summary 3 (‰). Low, medium, and high views correspond to the ranges [3.57e+03,3.65e+05], (3.65e+05,2.07e+06], and (2.07e+06,4.25e+08] respectively.

	pub_hr:[0,9]	pub_hr:(9,16]	pub_hr:(16,23]	is_weekend:FALSE	is_weekend:TRUE	views:low	views:medium	views:high	trend:[1,8]	trend:(8,17]	trend:(17,38]
mass	42	44	39	109	16	42	42	42	46	39	40
qlt	106	23	226	115	115	470	65	531	278	108	480
inr	28	27	29	5	36	34	28	36	30	29	36
k=1	90	130	-243	58	-391	-844	-173	1018	-646	-276	997
cor	4	9	27	23	23	357	15	518	242	34	475
ctr	1	3	10	2	10	124	5	180	80	12	167
k=2	-448	-160	661	-117	791	477	-316	-161	-250	405	-104
cor.1	102	14	199	93	93	114	50	13	36	74	5
ctr.1	44	6	88	8	52	49	21	6	15	33	2

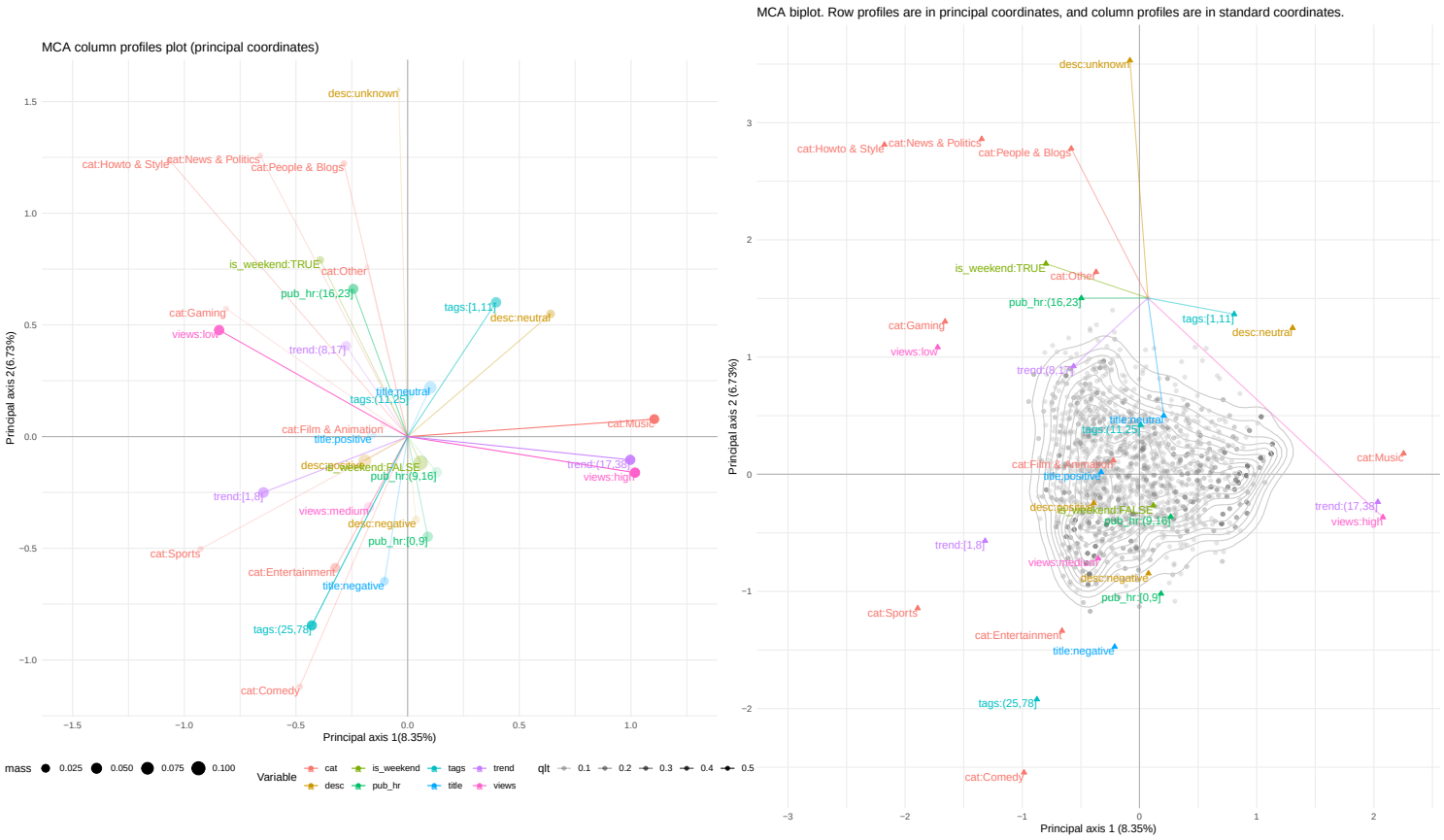

Figure 8: Column profiles plot and the MCA biplot.

that I previously thought Gaming videos should have `medium` view counts (refer to the caption of Table 5 for the cutoffs). After examining the data, indeed more than half of Gaming videos have low view counts. One conflicting observation is that `cat:News & Politics` is attracted strongly to `trend:(8,17]` because among all videos in this category, videos which trended for 8 days or less account for the largest proportion. This conflict is simply because `trend:(8,17]` is not that well represented in the subspace. One more observation is that videos published on the weekend tend to have low views, and this is confirmed by the data. This might be because British people have things to do other than browse YouTube on the weekend.

In the bottom left corner of the first plot in Figure 8, videos with categories Sports, Entertainment, and Comedy tend to have medium view counts, trended for only about a week or less, very high tag counts, and positive-sounding descriptions. Although `title:negative` looks most attracted to `cat:Comedy`, according to the data the Entertainment category proportionally has the most videos with negative titles. Again, this conflict is because `cat:Comedy` is not as well represented as `cat:Entertainment`. In addition, since `cat:Sports` and `views:medium` are also not well represented, the former should be more attracted to `views:low`, not `views:medium`. Since `desc:negative` is not well represented, it is unclear which categories tend to have negative-sounding titles. The data says Entertainment, but the second largest group of videos belong to this category.

The second plot in Figure 8 overlays the scores of row profiles (the videos) onto the subspace spanned by the first and second principal axes. Note that the column profiles are now in "standard coordinates" (standardized scores to have mean 0 and variance 1) and row profiles are in principal coordinates [see 8]. This means that both row and column profiles are now in the same coordinate system. In this plot, each video point is the average of the coordinates of the modalities that the video takes [9]. An example is the point connected by the colored lines. The density contours are there because a lot of points share the same coordinates (due to the nature of the indicator matrix). There are no clear groups of videos in the plot, but the darker the point is the more videos lie there. Some of the darkest points are close to the rightmost modalities, i.e., trendy and highly popular Music videos. Many video scores do not differ much from the "average" video. We cannot infer much from this biplot because of the low percentages of principal inertia percentages (8.35% for axis 1 and 6.73% for axis 2).

5 Conclusion

This report has presented univariate, bivariate, and multivariate statistical analysis on a dataset about YouTube videos. The most interesting finding of the univariate analysis was that more than half of the videos have positive descriptions or neutral titles. The main finding of the bivariate analysis was that the four user interaction variables `views`, `likes`, `dislikes`, and `num_dislikes` have mild to strong positive linear correlations with each other. The principal component analysis (PCA) helped us answer the first research question by summarizing the four interaction variables in the first principal component and the `trend` variable in the second and third components. PCA also discovered how `trend` interacted with the other four variables and additionally revealed some outliers. Because these variables are correlated, PCA was able to explain most of the variation with just two or three dimensions. Meanwhile, multiple correspondence analysis (MCA) partly helped us answer the second research question by analyzing the relationships between categorical variables and also between categorical variables and numerical variables by dividing the numerical variables into bins. Although MCA was not too successful because of the low percentages of inertia explained by the principal axes, we still discovered general patterns that hold true after cross-checking with the data, such as the pattern of Music videos having high view counts. The interpretations of the plots in Figure 8 were intuitive and fell in line with my knowledge about YouTube.

6 Critical evaluation

The first and the most significant source of bias is my domain knowledge of YouTube. Although I have watched videos on the platform for a long time, I myself am not a creator, making my understanding of YouTube restricted to the perspective of a viewer. In this project, I sometimes use this knowledge to interpret results that seem counterintuitive at first. Someone without any knowledge of YouTube might come to slightly different conclusions when looking at the results of the analyses (e.g., the plots).

The second limitation of this project is about the PCA. Because of the outliers, it was difficult to analyze the remaining projections of the datapoints. Thus, I did not mention anything about possible patterns or clustering of datapoints based on the PCA. The PCA could be improved by leaving out the most extreme outliers and performing robust PCA.

The third limitation is about the MCA. The numerical variables were categorized into very few bins (only 3 bins for each of them), hindering possible findings on the values in the interquartile range as well as the more extreme values. However, I did this in order to keep Figure 8 the report manageable. In addition, the low principal inertias and quality of display values could make the findings questionable, thus emphasizing the importance of cross-checking with the data. There are other MCA techniques that can result in higher principal inertias (e.g., MCA based on the Burt matrix), and I leave the implementation of these methods to future work.

The final source of bias is the modified dataset used in this project. First is how I analyzed textual video titles and descriptions by using sentiment scores and classes as I chose to believe in Hutto and Gilbert [2]. There might be better ways of working with text data in tandem with categorical and numerical data. In addition, I disregarded the time series structure, which might contain interesting patterns on how videos accumulate, e.g., views and likes, over time while they were trending. Furthermore, since I only used videos published during the first six months of 2018 by Great Britain-based creators, the results might not be applicable to other regions and time periods.

References

[1] M. J, *Trending YouTube Video Statistics*, https://www.kaggle.com/datasnaek/youtube-new, 2017.

[2] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014.

[3] P. Ilmonen, *Lecture 2: Principal Component Analysis*, https://mycourses.aalto.fi/pluginfile.php/1400532/mod_resource/content/1/2021Mult2.pdf, 2021.

[4] ——, *Lecture 3: Principal Component Analysis - part II*, https://mycourses.aalto.fi/pluginfile.php/1400533/mod_resource/content/2/2021Mult3.pdf, 2021.

[5] ——, *Lecture 7: Multiple Correspondence Analysis*, https://mycourses.aalto.fi/pluginfile.php/1400545/mod_resource/content/2/2021Mult7.pdf, 2021.

[6] ——, *Lecture 6: Bivariate Correspondence Analysis - part II*, https://mycourses.aalto.fi/pluginfile.php/1400544/mod_resource/content/2/2021Mult6.pdf, 2021.
[7] M. Greenacre, *Correspondence analysis in practice*. CRC press, 2017.
[8] ——, in *Correspondence analysis in practice*. CRC press, 2017, ch. 9.
[9] M. J. Greenacre, in *Biplots in practice*. Fundacion BBVA, 2010, ch. 10.