Morgan Keeton

STA 402

22 November 2021

An Exploratory Analysis of Global YouTube Data

An Introduction –

Many countries have various forms of video streaming sites and pages, but one could argue that YouTube is perhaps the most widely used video-streaming service on a global scale. We have seen YouTube videos used for educational purposes in large amounts of our classes, as well as for our own personal entertainment and education. YouTube is easily accessible in the United States especially, and used largely for a content-creator platform. It has been a growing industry over the last 10 years with access to music videos, vlogs, various types of tutorials (makeup, hair, sewing, how to change a tire, etc.). Many YouTubers actually create content full-time and are able to make substantial amounts of money off of the content that they create.

Because of this, researchers amongst the various fields in humanity (think: fields like sociology, psychology, linguistics) might wonder what types of videos people across the globe, in each respective country, or in each respective continent are watching. Due to this curiosity, data such as the one used in this exploratory analysis is widely available and widely used to draw conclusions on the types of videos that people across the globe navigate towards or stray away. Additionally, viral videos (ones with large amounts of views) paint a history of the global and respective community culture as time goes on. For example, one of the most watched videos (as you will see in the later sections) is a video that Logan Paul uploaded (and later took down) that created world-wide uproar for its insensitivity. It would be easy to argue that given our age and generation, most of us have heard of this video and are unsurprised that researchers might want to pinpoint these types of statistics.

A Description of the Data –

This data was taken from ten different countries across the world, which included Great Britain, Canada, the United States, Mexico, Korea, Germany, Japan, Russia, India, and France. Each respective dataset includes information about YouTube videos from each respective country. This information was categorized into different variables, including a thumbnail link to the video, the publishing time, a description of the video, the number of likes, views, dislikes and comments, and a set of tags that correspond with each video. These tags are what are used to control the auto-play on YouTube. So when you are watching a video, it has certain tags that cue a video similar to play after it. The likes, dislikes, views, and comment count are all variables that are accessible on a specific video. The data is categorized by these tags, and that is how we organize the data for our entire analysis.

Since each dataset is from a different country, some of the datasets were written in different languages. SAS was able to handle these alphabetic differences through the Unicode version, which accepts input from other languages than English. Additionally, each one of these respective datasets was rather large with around 4,000 data entries. Some of the variables, including tags, thumbnail link, and title had data entries that were larger than any datasets we had handled in class thus far. This created a problem with reading in the dataset without formatting any of the variables.

Initially, it was difficult to realize how each dataset was named. While the dataset “USvideos.csv” was obviously a dataset on the United States, it was not as easy to grasp that “DEvideos.csv” was a dataset from Germany. In doing research, it was discovered that each dataset was categorized by their respective ISO country codes. For example, DE is the ISO country code for Germany, as CA is to Canada, and as MX is to Mexico. After learning this fact, understanding the data and its respective countries was easier.

Aside from the various datasets provided, there were an additional ten JSON sets provided that were used to categorize each respective video. The JSON’s were used in a format which assigned the various variables to each video. For example, a few of the variables for each respective dataset are in a binary format (namely, true or false). Specifically, these are the variables that correspond to having the comments disables, ratings disabled, or having any type of video error / removal. These categories are imperative for drawing further conclusions and also looking at the skewness of the data. For example, if we are looking into the relationship between number of likes and number of comments, we need to consider those videos that have the comments disabled, as this would skew the data. Having zero comments does not necessarily mean that no one commented under the video, but could mean that the producer disabled this function. These same parameters coincide with any of these respective variables, and the JSON documents for each country shed light to these ideas.

The Strategy Employed –

One of the biggest problems with these respective datasets was dealing with the proper data steps to read in the data. By using PROC IMPORT first, this difficulty was easily managed. This beginning step allows for the user to see in the log output how the dataset should be formatted in an in- file statement. By doing this for each country, the in-file and formatting was able to be modified so that it could handle each one of these ten datasets in their true format. Because these datasets have over 4,000 observations / entries, it was easier in the first exploratory coding to only look at a small chunk of the data. By looking at only the first 10-20 data entries, it was less time-consuming to see how to tweak the code in order to properly accomplish the task at hand. The log provided helpful suggestions as far as output goes as well. Since some of these numbers of views, comments, dislikes, and likes are quite large, changing the graphics side of the ODS format was necessary in order to paint a full picture. If you look at the beginning of the exploratory analysis part of the code, you will see a statement that specifies the graphics size. This was imperative for the PROC SERIES graphs ran to look at the relationship between views and each respective integer variable as a whole.

For the assigned part of the analysis, using PROC IMPORT was crucial in creating code that was able to use macro variables to switch between each respective country. Once I was able to use the import statement on each dataset, the in-file statement was able to be modified to fit each country as a whole. Macro variables were then implemented so that the user could decide first, the country of choice; second, the variable of interest; and third, the number of observations wanted or needed. From these macros, we were able to cut the total amount of code in half. This made the code not only easier to read, but also have a faster run-time. Each macro was then called for each respective dataset to achieve the results below. The user is able to specify the country that is of interest along with variables and the number of observations. With the number of observations, this displays the respective number of tags that the user will see in the following output. This correlates to the number of tags that was specified with respect to the sorted dataset by the variable(s) that the user specifies. Whichever variable the user specifies first is the variable that SAS will automatically sort by. This same macro is able to handle more than the 10 data sets that were provided, provided that the format of the name of the file is of the same ideal as the ones provided. Additionally, the user is able to set more than one variable of interest to compare, say, likes and dislikes based upon the tags for each respective country. This way, the code is able to be used in a greater format than the exploratory data analysis that is touched upon below.

A Look into the Results –

From looking at Figure One, we can see that the views variable has the most interaction from all of the variables used, with a maximum value of 424.5 million views, compared to the comment count variable which has a maximum value of 1.6 million comments. This seems intuitive, as it is often that we watch a video without doing anything further. From this same figure, we are also able to see that there was not a single video out of these ten datasets that had less than 233 views. The average number of views globally for a YouTube video from one of these ten datasets is around 2.7 million views. This seems exponentially higher than one might imagine for a larger population (given the number of videos released daily), but one would imagine that viral videos have an effect on the distribution of the data.

In looking into each respective country’s output data, I chose to focus in on the United States, Mexico, and Russia. Beginning with the United States (Figure Two), we can see that videos in the United States get an average of 681,816 views per each video, with a range of 549 to 225.2 million views. As mentioned in the introduction, the video with the most dislikes from the United States is the Logan Paul video (Figure Five, tags include Logan Paul, apology, suicide forest) that most of us who were in the United States at the time remember going viral. The idea that this is the video with the top number of dislikes (1,674,420 dislikes) is not surprising, considering the aforementioned insensitivity. This is followed by a YouTube rewind 2017 video with 1643059 dislikes.

In looking at Russia’s data (Figure Three), we can see that videos in Russia got an average of 240,715.15 views per video, which about a third of the amount of views videos in the United States received. The comment count (or number of comments) variable was of particular interest here, with an average of 2,039.66 comments per video, with a range of 0 to 905,925 comments. The video with the top number of comments has tags that include BIGHIT. However, the third video with the top number of comments is actually the second most disliked video in the United States: the YouTube Rewind 2017 video.

In Mexico, we can see that videos get an average of 342,381.97 views, which seems to split the average number of views for Russia and the United States. We were particularly interested into looking at the tags that correspond to the videos with the largest number of views in Mexico. The same YouTube Rewind video that appeared in both Russia and United States' top dislikes and comment count videos appears here with the top number of views of 100.9 million views. Interestingly enough, the Logan Paul apology video does not appear in the top number of views for the Mexico data as it did for likes and comments for the United States and Russia.

Overall we can see that there is some sort of relationship between the respective countries, as several of the tags reappear in each. This correlates to what was aforementioned in the introduction (the idea of viral videos). Because some tags reappear with number of comments, views, likes and dislikes, we can explore that relationship. That is discussed in the conclusions below.

Further Conclusions –

Further conclusions could be drawn by running various types of statistical tests, such as t-tests or any number of nonparametric statistics tests based on the overall skewedness of the data that we have been able to get a glimpse at with the QQ plot and Histogram (Figures Eight and Nine) that were mentioned in our previous analysis. In Figures Eight and Nine, we explored the overall distribution of the data with respect to likes and dislikes from all ten countries datasets. By looking at the Normal QQ plot (Figure Eight) for the likes variable, which corresponds to the number of likes, we can see that the data does not follow a diagonal straight line whatsoever. This indicates that to some degree, we can assume that the likes variable is not normally distributed. Further analysis into boxplots or histograms could show the degree to which the likes variable violates any normality assumption, but nonparametric tests could still be used for further analysis here. Using Figure One to further expand on this idea, we see that the average number of likes for a video from one of these ten respective countries is 78709.79 likes, while the median number of likes is 1433.50 likes. The average is over 50 times that of the median, indicated positively skewed data.

In looking at the dislikes variable, which corresponds to the number of dislikes, we see that the data is extremely skewed. The histogram (Figure Nine) does not appear to be normally distributed at all. Using the same idea that we explored with the likes video, in figure one we can see that the average number of dislikes for a video is 4359.62 dislikes, while the median is 486 dislikes. The median is over 8 times that of the mean, which indicates negatively skewed data. One would imagine that this is ideal for content-creators, as it seems that people are disliking videos less than they are liking them.

Furthermore, it would be interesting to run various forms of relationship testing and confidence intervals to see and explore the relationship between each respective country. For example, if we compared the top ten most viewed videos and their tags for each respective country, we would have a larger grasp on what it means for a video to go viral. Considering that these datasets are from around the world, we could draw conclusions with respect to sociology on the matter in which people watch videos, interact with videos, and run some sort of experiment given these parameters.

Numerical Analysis of Country Video Data (Figure One)
The MEANS Procedure

Variable	Label	Minimum	Median	Mean	Maximum
likes	Number of Likes	0	14333.50	78708.79	5613827.00
dislikes	Number of Dislikes	0	486.0000000	4359.62	1944971.00
views	Number of Views	223.0000000	438165.00	2739027.18	424538912
comment_count	Number of Comments	0	1366.00	8353.86	1626501.00

5 Number Summary of US Data (Figure Two)
The MEANS Procedure

Variable	Label	Minimum	Mean	Median	Maximum
likes	Number of Likes	0	74266.70	18091.00	5613827.00
dislikes	Number of Dislikes	0	3711.40	631.0000000	1674420.00
views	Number of Views	549.0000000	2360784.64	681861.00	225211923
comment_count	Number of Comments	0	8446.80	1856.00	1361580.00

5 Number Summary of RU Data (Figure Three)
The MEANS Procedure

Variable	Label	Minimum	Mean	Median	Maximum
likes	Number of Likes	0	12435.22	1880.00	4470923.00
dislikes	Number of Dislikes	0	1475.20	128.0000000	884967.00
views	Number of Views	117.0000000	240715.15	66316.00	62796390.00
comment_count	Number of Comments	0	1775.23	309.0000000	905925.00

5 Number Summary of MX Data (Figure Four)
The MEANS Procedure

Variable	Label	Minimum	Mean	Median	Maximum
likes	Number of Likes	0	15861.84	1246.00	4470923.00
dislikes	Number of Dislikes	0	747.1603916	63.0000000	1353667.00
views	Number of Views	157.0000000	342381.97	56973.00	100912384
comment_count	Number of Comments	0	2039.66	196.0000000	905925.00

Top 10 Dislikes from US Video Data (Figure Five)

Obs	Tags	Number of Dislikes
1	"logan paul vlog"|"logan paul"|"logan"|"paul"|"olympics"|"logan paul youtube"|"vlog"|"daily"|"comedy"|"hollywood"|"parrot"|"maverick"|"bird"|"maverick clothes"|"logan paul apology"|"suicide forest"|"japanese suicide forest"|"suicide"|"logan paul suicide"|"suicide apology"	1674420

Obs	Tags	Number of Dislikes
2	"Rewind"|"Rewind 2017"|"youtube rewind 2017"|"#YouTubeRewind"|"Rewind 2016"|"Dan and Phil"|"Grace Helbig"|"HolaSoyGerman"|"Lilly Singh"|"Markiplier"|"Swoozie"|"Rhett Link"|"Liza Koshy"|"Dolan Twins"|"Lele Pons"|"Jake Paul"|"Logan Paul"|"KSI"|"Joey Graceffa"|"Casey Neistat"|"Poppy"|"Niana Guerrero"|"Daddy Yankee"|"Luis Fonsi"|"Ed Sheeran"|"Kendrick Lamar"|"Stephen Colbert"|"Fidget Spinners"|"Slime"|"Backpack Kid"|"April the Giraffe"|"#Rewind"|"Despacito"|"Shape of you"|"YouTubeRewind"|"I'm the One"|"Humble"	1643059
3	"logan paul vlog"|"logan paul"|"logan"|"paul"|"olympics"|"logan paul youtube"|"vlog"|"daily"|"comedy"|"hollywood"|"parrot"|"maverick"|"bird"|"maverick clothes"|"logan paul apology"|"suicide forest"|"japanese suicide forest"|"suicide"|"logan paul suicide"|"suicide apology"	1611043

Top 10 Videos with Highest Comment Count from RU Video Data (Figure Six)

Obs	Tags	Number of Comments
1	BIGHIT|"빅히트"|"방탄소년단"|"BTS"|"BANGTAN"|"방탄"|"FAKE LOVE"|"FAKE_LOVE"	905925
2	BIGHIT|"빅히트"|"방탄소년단"|"BTS"|"BANGTAN"|"방탄"|"FAKE LOVE"|"FAKE_LOVE"	905925
3	Rewind|"Rewind 2017"|"youtube rewind 2017"|"#YouTubeRewind"|"Rewind 2016"|"Dan and Phil"|"Grace Helbig"|"HolaSoyGerman"|"Lilly Singh"|"Markiplier"|"Swoozie"|"Rhett Link"|"Liza Koshy"|"Dolan Twins"|"Lele Pons"|"Jake Paul"|"Logan Paul"|"KSI"|"Joey Graceffa"|"Casey Neistat"|"Poppy"|"Niana Guerrero"|"Daddy Yankee"|"Luis Fonsi"|"Ed Sheeran"|"Kendrick Lamar"|"Stephen Colbert"|"Fidget Spinners"|"Slime"|"Backpack Kid"|"April the Giraffe"|"#Rewind"|"Despacito"|"Shape of you"|"YouTubeRewind"|"I'm the One"|"Humble"	702790

Top 10 Videos with most views from MX DATA (Figure Seven)

Obs	Tags	Number of Views
1	Rewind|"Rewind 2017"|"youtube rewind 2017"|"#YouTubeRewind"|"Rewind 2016"|"Dan and Phil"|"Grace Helbig"|"HolaSoyGerman"|"Lilly Singh"|"Markiplier"|"Swoozie"|"Rhett Link"|"Liza Koshy"|"Dolan Twins"|"Lele Pons"|"Jake Paul"|"Logan Paul"|"KSI"|"Joey Graceffa"|"Casey Neistat"|"Poppy"|"Niana Guerrero"|"Daddy Yankee"|"Luis Fonsi"|"Ed Sheeran"|"Kendrick Lamar"|"Stephen Colbert"|"Fidget Spinners"|"Slime"|"Backpack Kid"|"April the Giraffe"|"#Rewind"|"Despacito"|"Shape of you"|"YouTubeRewind"|"I'm the One"|"Humble"	100912384
2	Rewind|"Rewind 2017"|"youtube rewind 2017"|"#YouTubeRewind"|"Rewind 2016"|"Dan and Phil"|"Grace Helbig"|"HolaSoyGerman"|"Lilly Singh"|"Markiplier"|"Swoozie"|"Rhett Link"|"Liza Koshy"|"Dolan Twins"|"Lele Pons"|"Jake Paul"|"Logan Paul"|"KSI"|"Joey Graceffa"|"Casey Neistat"|"Poppy"|"Niana Guerrero"|"Daddy Yankee"|"Luis Fonsi"|"Ed Sheeran"|"Kendrick Lamar"|"Stephen Colbert"|"Fidget Spinners"|"Slime"|"Backpack Kid"|"April the Giraffe"|"#Rewind"|"Despacito"|"Shape of you"|"YouTubeRewind"|"I'm the One"|"Humble"	75969469
3	marvel|"comics"|"comic books"|"nerdy"|"geeky"|"super hero"|"superhero"|"avengers: infinity war"|"avengers"|"infinity war"|"marvel studios"	74789251

Figure Eight

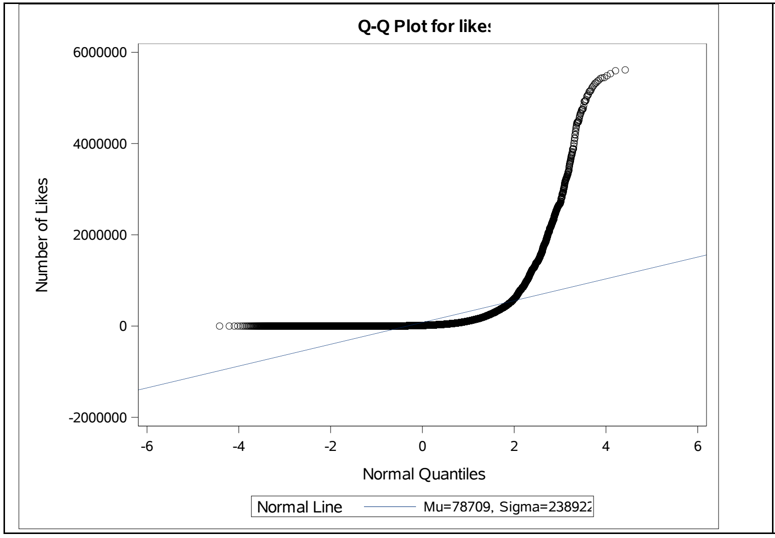
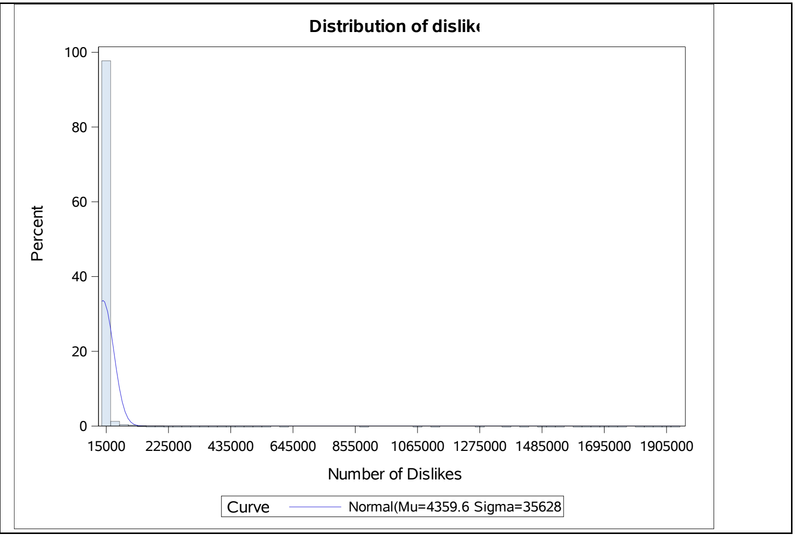


Figure Nine



```
/*Produces tables with the top number of specified values

(chosen by the user) for the country (user specified. Then does
an exploratory analysis of the video data.*/
ods html style=htmlbluecml;
/*user specifies country from the list of 10 countries
user specified variables from the list of variables
user then specifies the number of desired observations*/
%macro videos(country = , var = , num = );
data &country;
      %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
     infile "M:\STA402\Project\Data\&country.videos.csv"
            delimiter = ','
            MISSOVER
            DSD
            lrecl=32767
            firstobs=2 ;

      *used PROC IMPORT to get the format;
      format video_id $11. trending_date $8. title $88.
            channel_title $23. category_id best12.
            publish_time B8601DZ35. tags $538.
            views best12. likes best12. dislikes best12.
            comment_count best12. thumbnail_link $46.
            comments_disabled $5. ratings_disabled $5.
            video_error_or_removed $5. description $2259.;

      input video_id $ trending_date $ title $ channel_title $
                  category_id publish_time tags $ views likes
                  dislikes comment_count thumbnail_link $
                  comments_disabled $ ratings_disabled $
                  video_error_or_removed $ description $;
       if _ERROR_ then call symputx('_EFIERR_',1);
run;

data work.&country;
      set &country;
      keep tags views likes dislikes comment_count;
run;

ods rtf bodytitle file = "M:\STA402\Project\&country..rtf";

*prints figures 2 through 4 for respective countries;
title "5 Number Summary of &country Data";
proc means data = work.&country min mean median max;
      var likes dislikes views comment_count;
      label likes = "Number of Likes"
                  dislikes = "Number of Dislikes"
                  views = "Number of Views"
                  comment_count = "Number of Comments";
run;

proc sort data = &country out = sorted;
      by descending &var;
run;

*only keeps 10 observations;
data sorted2;
      set sorted (obs = &num);
      keep tags &var;
      label tags = "Tags";
      label  &var = "Number of &var";
run;

*prints the data;
*prints figures 5 through 7 for respective countries;
title "Top 10 Video Tags with highesy &var";
title2 "from &country Video Data";
proc print data = sorted2 label;
run;

ods rtf close;
%mend videos;
*call macro below;
/*%videos(country = US, var = dislikes , num = 3)
%videos(country = RU, var = comment_count, num = 3)
%videos(country = MX, var = views, num = 3)

*/

/*This begins our overall exploratory analysis*/

*the following macro reads in the data for each country;
*it then sorts the data by likes;
%macro exploratory(data = ); *data = user specified country;
data &data;
      %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
     infile "M:\STA402\Project\Data\&data.videos.csv"
            delimiter = ','
            MISSOVER
            DSD
            lrecl=32767
            firstobs=2 ;
      *used PROC IMPORT to get the format;
      format video_id $11. trending_date $8. title $88.
                  channel_title $23. category_id best12.
                  publish_time B8601DZ35. tags $538.
                  views best12. likes best12. dislikes best12.
                  comment_count best12. thumbnail_link $46.
                  comments_disabled $5. ratings_disabled $5.
                  video_error_or_removed $5. description $2259.;

      input video_id $ trending_date $ title $ channel_title $
                        category_id publish_time tags $ views likes
                        dislikes comment_count thumbnail_link $
                        comments_disabled $ ratings_disabled $
                        video_error_or_removed $ description $;
       if _ERROR_ then call symputx('_EFIERR_',1);
run;
%mend exploratory;
*calls each country;
%exploratory(data = US)
%exploratory(data = RU)
%exploratory(data = CA)
%exploratory(data = KR)
%exploratory(data = JP)
%exploratory(data = DE)
%exploratory(data = GB)
%exploratory(data = MX)
%exploratory(data = IN)
%exploratory(data = FR)

%macro sortem(country2 = );
proc sort data = &country2 out = &country2_sorted;
      by likes;
run;
%mend sortem;
%sortem(country2 = US)
%sortem(country2 = RU)
%sortem(country2 = CA)
%sortem(country2 = KR)
%sortem(country2 = JP)
%sortem(country2 = DE)
%sortem(country2 = GB)
%sortem(country2 = MX)
%sortem(country2 = IN)
%sortem(country2 = FR)

data all;
      merge  US_sorted RU_sorted MX_sorted KR_sorted
                  JP_sorted IN_sorted GB_sorted FR_sorted
                  DE_sorted CA_sorted;
      by likes;
      keep tags views likes comment_count dislikes;
run;
ods rtf bodytitle file = "M:\STA402\Project\graphs2.rtf";
title "Relationship of Views and Likes";
title2 "from all video country data";

proc sgplot data = all;
```

```
      series x = views y = likes / markers;
      label views = "Number of Views"
                  likes = "Number of Likes";
run;
title "Relationship of Views and Comment Count";
title2 "from all video country data";
proc sgplot data = all;
      series x = views y = comment_count / markers;
      label views = "Number of Views"
                  comment_count = "Number of Comments";
run;
title "Relationship of Views and Dislikes";
title2 "From all video country data";
proc sgplot data = all;
      series x = views y = dislikes / markers;
      label views = "Number of Views"
                  comment_count = "Number of Comments";
run;
ods rtf close;

ods rtf bodytitle file = "M:\STA402\Project\numbersanduniv.rtf";
*let's look at a number summary;
title "Numerical Analysis of Country Video Data";
proc means data = all min median mean max;
      vars likes dislikes views comment_count;
      label likes = "Number of Likes"
                  dislikes = "Number of Dislikes"
                  views = "Number of Views"
                  comment_count = "Number of Comments";
run;
*let's explore the distribution of our data;
title "Univariate Exploration of Country Data";
*Figure Eight and Nine;
proc univariate data = all;
      var likes dislikes views comment_count;
      histogram likes / normal;
      qqplot likes / normal(mu=est sigma=est);
      histogram views / normal;
      qqplot views / normal(mu = est sigma = est);
      histogram dislikes / normal;
      qqplot dislikes / normal (mu = est sigma = est);
      histogram comment_count / normal;
      qqplot comment_count / normal (mu = est sigma = est);
      label likes = "Number of Likes"
                  views = "Number of Views"
                  dislikes = "Number of Dislikes"
                  comment_count = "Number of Comments";
run;
ods rtf close;
```