THE UNIVERSITY OF SYDNEY

Assignment Coversheet – GROUP ASSIGNMENT

Please fill in your details below. Use one form for each group assignment.

Personal Details of Students

Group Name/Number					
Family Name	Given Name (s)	Student Number (SID)	Unikey	Contribution + percentage	Signature
Chunliang	Pan	470160239	cpan7779	Initial report, presentation, Introduction of final report-100%	Chunliang Pan
Mingjie	Shi	470193259	mshi3248	Initial report, presentation, conclusion and summary of final report-100%	Mingjie Shi
Xiaochuan	Xia	490353617	xxia2851	Initial report, presentation, Task B of final report-100%	Xiaochuan Xia
Wenjing	Deng	470349281	wden4101	Initial report, presentation, Evaluation of final report-100%	Wenjing Deng
Yue	Yang	490284768	yyan8094	Initial report, presentation, Task C of final report-100%	Yue Yang
Zhecheng	Zhong	490319299	zzho7727	Initial report, presentation, Task A of final report-100%	Zhecheng Zhong
Zekun	Tao	490573260	ztao9690	Initial report, presentation,Task C of final report-100%	Zekun Tao

Assignment Details:

Assignment Title	Visualization of Trending YouTube Video				
Assignment number	**Assignment2**				
Unit of Study Tutor					
Group or Tutorial ID	**Group 15**				
Due Date	**07/11/2019**	Submission Date	**07/11/2019**	Word Count	**4732**

Declaration:

1. I understand that all forms of plagiarism and unauthorised collusion are regarded as academic dishonesty by the university, resulting in penalties including failure of the unit of study and possible disciplinary action.
2. I have completed the **Academic Honesty Education Module** on Canvas.
3. I understand that failure to comply with the Academic Dishonesty and Plagiarism in Coursework Policy can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the ***University of Sydney By-Law 1999*** (as amended).
4. This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.
5. The assessment has not been submitted previously for assessment in this or any other unit, or another institution.
6. I acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment may:
   a. Reproduce this assignment and provide a copy to another member of the school; and/or
   b. Use similarity detection software (which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking).
7. I have retained a duplicate copy of the assignment.

Please type in your group number here to acknowledge this declaration:	Group 15

COMP 5048 Assignment 2 - final report

Date: 06/11/2019

Group 15

Visualization of Trending YouTube Video

Team Members:

Chunliang Pan – 470160239

Mingjie Shi – 470193259

Xiaochuan Xia – 490353617

Wenjing Deng – 470349281

Yue Yang – 490284768

Zhecheng Zhong – 490319299

Zekun Tao– 490573260

Content

1.Introduction

It helps humans perceive more subtle and sensitive information, and through the visualization of the data, this information will be used to reveal concealed secrets to guide real practice (Steed et al., 2014).

With the continuous advancement of Internet infrastructure and information technology, the network video portal, such as YouTube, Netflix, has become one of the main sources of people acquiring dynamic information. In other words, if the analyst can mine, analyse and visualize the dataset from the Petabyte (PB) level of video interaction to a specific client, it will undoubtedly support the business decision-making. In the article, with the assistant of programming languages and analytics tools, the visual analysis could process the selected dataset in a variety of ways. The tasks and solutions will be demonstrated in the detailed examples below to improve the efficiency of visual analysis.

1.1 Dataset and Tasks

1.1.1 Dataset

Based on the pre-defined data sets from the assignment requirement, the analysis team selected 'Trending YouTube Video Statistics' as the analysis object. The data set records the top trending of YouTube videos in 10 different geographic locations. To retrieve the categories for a specific video, one JSON file is attached a single dataset. One such file is included for each of the five regions in the dataset. The following table shows the metadata dimensions of the selected dataset.

Dimensions	Description
video_id	Unique id for each video
trending_date	Date that the video starts to be popular
title	Title of each video
channel_title	Channel name of each video
category_id	Category id of each category
publish_time	Publish time of each video
tags	Tags of video
views	Views number of the video
likes	Likes number of the video
dislikes	Dislikes number of the video
comment_count	Comments number of the video
thumbnail_link	Thumbnail link of the video
comments_disabled	Comment state of the video
ratings_disabled	Rating state of the video
video_error_or_removed	Video state of the video
description	Description of the video

Table 1-1 Summary of factors in the YouTube Trending data set

Due to problems with language and information invalidity, the analysis team found that not all data dimensions in the metadata can be used. Therefore, after cleaning the dataset, the analysis team simply sorts out the data categories of the metadata and obtains the following table.

Dimensions	Description
video_id	Unique id for each video
trending_date	Date that the video starts to be popular
title	Title of each video
channel_title	Channel name of each video
category_id	Category id of each category
publish_time	Publish time of each video
tags	Tags of video
views	Views number of the video
likes	Likes number of the video
dislikes	Dislikes number of the video
comment_count	Comments number of the video

Table 1-2 selected dimensions of a cleaned dataset

1.1.2 Tasks

Based on the initial report, we have merged some tasks together and improved those tasks into a more commercial orientation, which could support decision-making in current businesses directly. In this section, the analysis team focuses on three hypothesises, which are respectively 'What is the market size?', 'How is the flavour of the audience?', 'When is the best time to upload a video', to extend three relevant tasks.

Task	Description
Task A	Identify the most influential category of YouTube videos around the world
Task B	Identify which is the most popular channel based on different geographic locations
Task C	Identify the best video launching time and evaluate the relationship between category and how long it takes for the related video to become prevalent

Table 1-3 Task list

1.2 Aims and Contribution

According to the three hypothesises and related tasks that are listed forward by the analysis team, the business investment-oriented data visualization analysis is obvious. We aim to use visual analysis to find out the categories of video programs, the market where the data sets are located, the tastes of users, the release time and the popularity of time to guide the commercial investment of venture capital in various video channels. At the same time, we also aim to provide decision-making support for non-commercial video bloggers to help them understand when the best releasing time is, which could possibly associate them make better productions.

Through the visual analysis of the team, combined with further natural language processing, or modelling of big data. We could build an analysis tool in the future to automate the media companies to achieve efficient and accurate data analysis. Furthermore, these analyses would mean a lot to the development of the market for improving the developing speed and help the companies to find more opportunities in the fierce competition.

2.Design

2.1 Analysis

Task A
This task is to analyse the most influential category of YouTube videos in different area.

First, the task will focus on market size of YouTube worldwide, it is used to decide which region has more potential viewers for the trending videos. To achieve the estimation of market size in different region, the number of views of all videos in one region are added together. The result is the total number of views and their corresponding country, it is regarded as potential market sizes for each region worldwide.

Second, the task will go deep to each region to identify their favourite video categories. It will determine the top 3 influential categories of YouTube videos, which is calculated by the number of views for each category in one nationality. The influence of category equals total count of views, then the numbers will be sorted in descending order. Only the first three categories are retained in order a legible visual analysis.

Finally, the task will combine multiple dimensions to analysis the relationship between numbers of views and likes for individual category, it determines the repeat view rate of video category. The YouTube videos can be viewed many times, but a video only can be liked one time for each personal user. The ratio of views to likes to indicate that whether the high number of views because of multiple replays by individual users. It will as a support analysis for above two tasks to reduce the estimation error of repeat views. The computing results equal ratio the number of views to likes, and the higher ratio means more replays.

These three methods could be combined to find out the most influential category of YouTube videos in different area and could be used to support each other to further evident the conclusion.

Task B
To find out the most popular channels and the relationship it has with categories, we analysed the data in some representative area to find the relationship between the two factors. YouTube categorized all video into 31 categories. The list of US trending shows that the most popular video category is Entertainment and the second is music. However, the most popular channel or the channel has most videos on the list is ESPN which all its videos are sports videos. By doing the same analysis to all other region it is clear to find all the top channels are making the same category videos. Many channels have videos on different country's list. We also want to know how different areas like the channel by looking at the videos at their lists.

Task C
When we discuss the value behind a video, the amount of views is the most intuitive and important evaluation criterion. A recent study suggests that the best time to post on Instagram is between 9am - 11am EST (Loren, 2019). If bloggers upload their social media at that reference time, they are more likely to get more views.

Views of a video is the currency in the social media market. It is vital for both the advertisement investors and youtubers to look at the number of views to determine their financial decision. While investors are putting more money in these industries, youtubers stand tremendous chances to enlarge their profits by improving the views on their video. In this section, based on the history data we have, we are trying to find some invisible patterns that, giving the same video, whether there is a certain amount of uploading time (a day in a year or an hour in a day), that if the video is launched at that time of period, it can obtain more views. Considering the feature of the data set, we separate this task into two sub-topics.

The first topic is to find the general pattern on the trending duration between different categories and in different countries.

To solve this, we need to calculate the duration (Incubation Period) of different kinds of videos. Therefore, it is necessary to determine the category ID, the publish time, and trending date that also contained the information of ending time among the dataset. Then, we use Incubation Period as a standard to count the time required for each kind of video to become prevalent, get the number of times that different Incubation Period values appear under a certain video category, and calculate their average, median, mode and standard deviation. It should be emphasized that the videos which take a long time to become prevalent (Incubation Period is more than 10 days) should be ignored when we calculate the duration, and these data should be analysed separately.

The second topic is to find the relatively best time for uploading a video for a specific category.

In order to achieve the two topics illustrated above, we define some analysis rules here. First of all, views are in proportion to the trending duration. The data set records daily trending YouTube videos. If a video maintains more time on the data list, it has more trending duration and it is more likely to gain more views. Here, we define:

Incubation Period (IP) = FT - PD (1)
Trend Duration (TD) = LT - FT (2)

Where incubation periods standing for the time period between the first trend date (FT) and the publish date (PD). And trend duration is the time difference between the latest trend date (LT) and the first trend date (FT).

In general, if a video is ready to launch, we want to minimize the period between the day when the video become a trend and the day when the video is launched, which is the incubation period. Also, we want to maximize the trending duration to have a sizable profit. Here, we generate the two variables by combining them into one:

DI-value = Trend Duration - Incubation Period

However, the trend duration and incubation period defined in formula (1) and (2) are lack of comparison standard. In the real case study, we minus each by the mean of them. By doing so, the DI-value is more comparable and more objective.

DI-value = (TD -TD mean) - (IP - IP mean) (3)

$$\text{TD - TD mean} = \frac{x \cdot f(x)}{\sum_{i=1}^{\infty} f(x_i)} - x \qquad (4)$$

2.2 Visualization

Task A
Figure 2-1(see next page) demonstrates the market size of YouTube trending videos worldwide. It displays the geographical location of individual region in a map, and the darker colour means more numbers of views, its range is from 5.3 billion to 2 trillion. Japan is the country with the smallest amount of play and Great Britain has the largest number of video views. According to the data analysis, a unexcepted result is that the view number of Great Britain is more than twice as much as the second USA's view number. Thus, it also means Great Britain has the largest market size of YouTube trending videos in the ten countries.

Then we focus on the 'what are the top 3 favourite video categories in different regions' (Figure 2-2). It is a side-by-side bar chart,

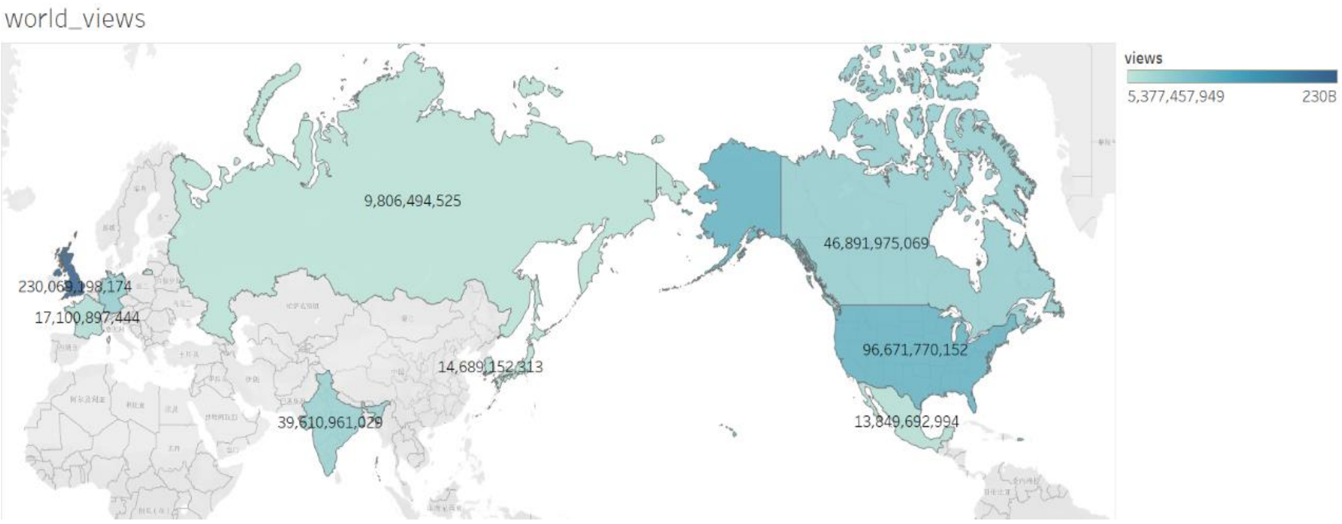


Figure 2-1 YouTube Trending videos worldwide

demonstrate the sum of views for each video category and broken down by country. The category is filtered on number of views and only keep the top 3 categories, the colours explain different video category, are divided into blue (Film & Animation), orange (Music) and green (Entertainment). As the result of visual analysis, the top 3 video categories for all countries are Film & Animation, Music and Entertainment, and Music and Entertainment videos are stable in the top 2 position. Another analysis is attracted by the view number of Music videos in Great Britain. The value is several times than other countries' value. A result can be concluded, trending music videos may have numerous potentials users in Great Britain.

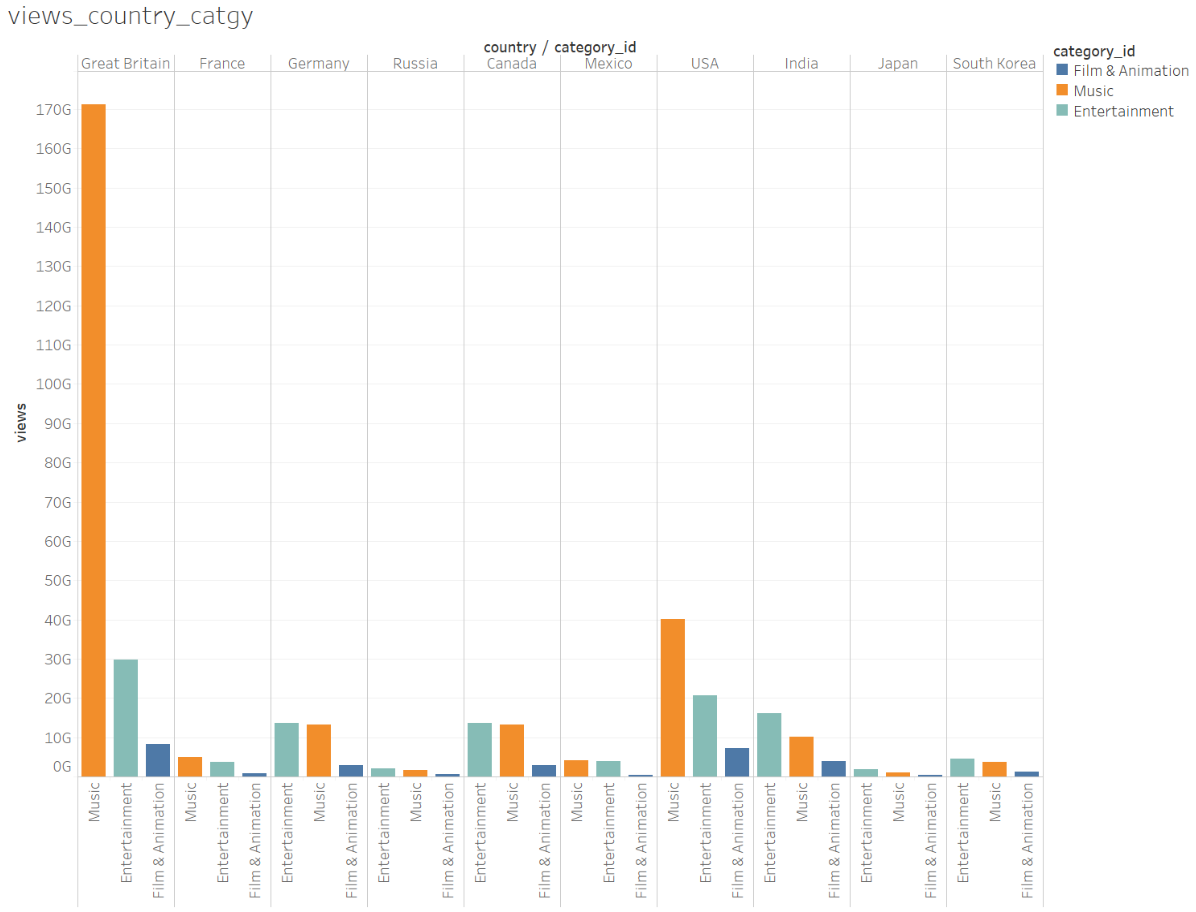
As for the relationship between number of views and likes (figure 2-3). It is a scatter plot, and each circle stand for a video. The colour shows different video categories and identify the trend line for every category. Due to that the YouTube dataset is trending videos that by views' preferences, the higher correlation coefficient means more repetitive views.

According to the visual analysis result, the categories of Movie, Shows and Trailers have the top 3 strongest positive correlation. It explains that viewers usually repeat playing the kinds of videos.

Task B

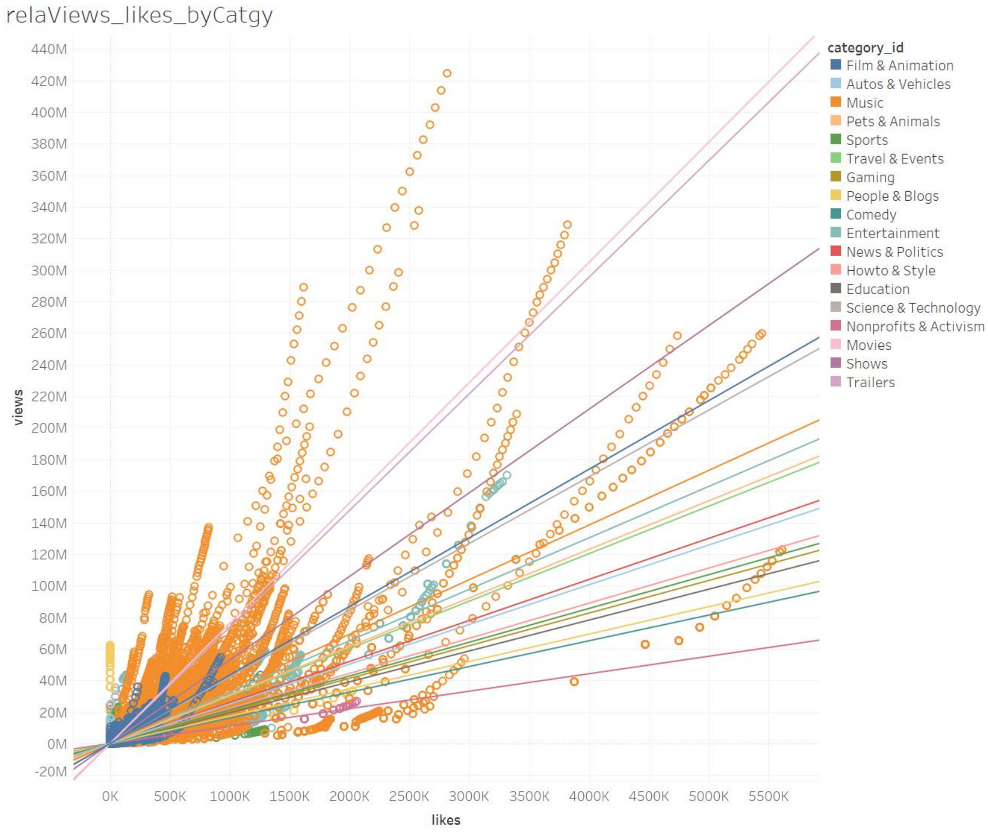
First, we create the visual graphs of US channel category, to make it easier to identify the channel information.

The diagram (figure 2-4) shows the number of each category's channel number. Every small block in the picture represent a channel, the size



Sum of views for each category_id broken down by country. Colour shows details about category_id. The view is filtered on category_id, which keeps Film & Animation, Music and Entertainment.

Figure 2-2 the top 3 favourite video categories in different regions



Sum of likes vs. sum of views. Colour shows details about category_id. Details are shown for id. The view is filtered on sum of likes, which keeps non-Null values only.

Figure 2-3 Relationship between number of views and likes

of a block means how much video a channel had on the trending list, and by analysing the picture we can easily find entertainment and music are the most popular category in the US and has most contents and many channels were making videos.

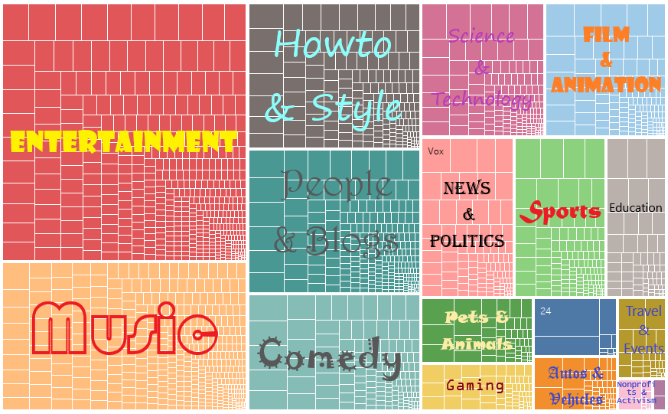


Figure 2-4 US channel category

After we collected the top 5 channel category of US, we then analyse the top 10 channels existing in them.

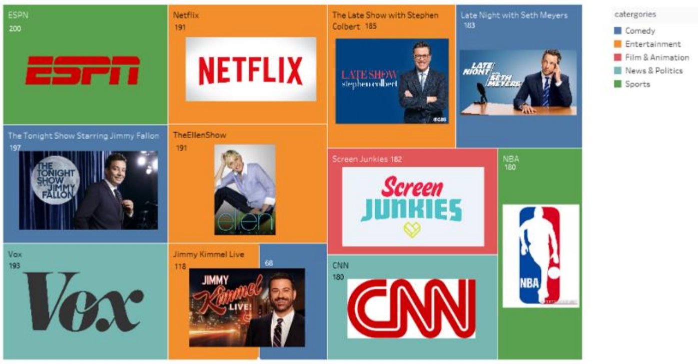
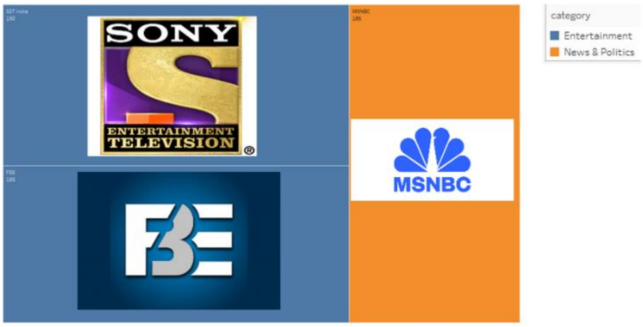
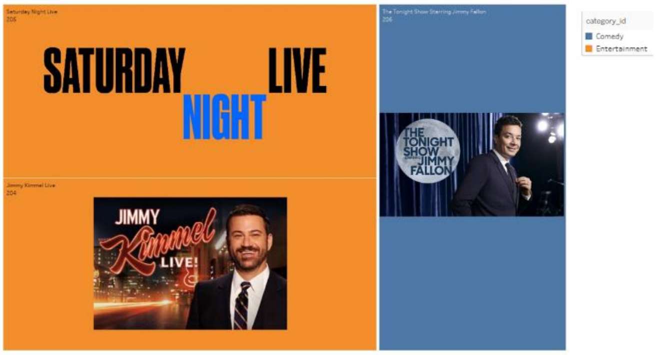


Figure 2-5 US top 10 channels

Figure 2-5 displays the top 10 channels has the most quantity of videos on the list in the US region. The color represents the what category the channel's videos and the size means the number of videos on the list. ESPN has 200 videos on the list and all sports videos, however, entertainment also has a huge proration in the top 10. And for all top 10 channels only Jimmy Kimmel Live has videos of 2 categories, entertainment and comedy, others channel only makes videos of one category.

After the analyse in US, we also did the collection in the CA, GB, and India, and the results are in the Figure 2-6.





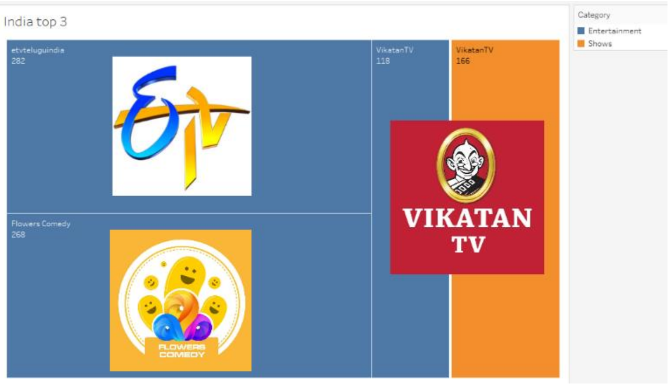


Figure 2-6 top 3 channels in CA, GB and India

As the diagram above shows, entertainment is the most popular category around the world and most top channels only make one category of videos.

Some channels had on many countries 'list and the graph below show the percentage of its videos on each country's list.

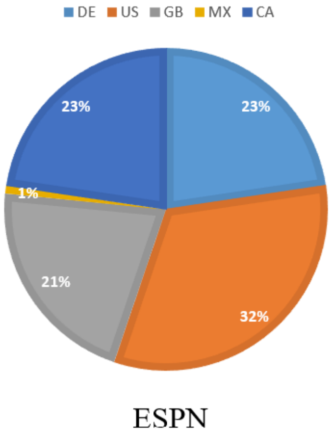


Figure 2-7 ESPN

Most ESPN videos were on US which is 32% and Mexico only form 1% of the graph.

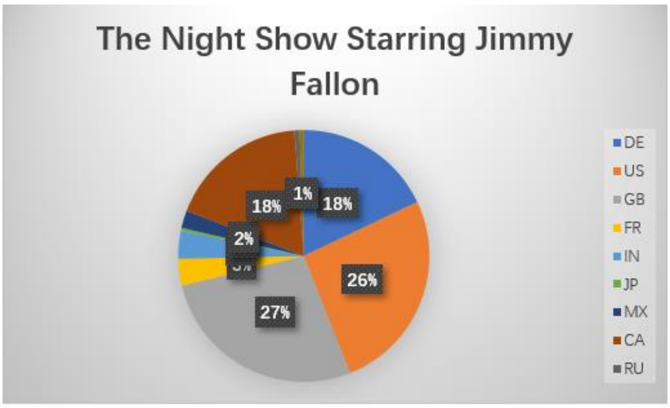


Figure 2-8 The Night Show Starring Jimmy Fallon

GB, US, CA and DE are the biggest fans of the show.

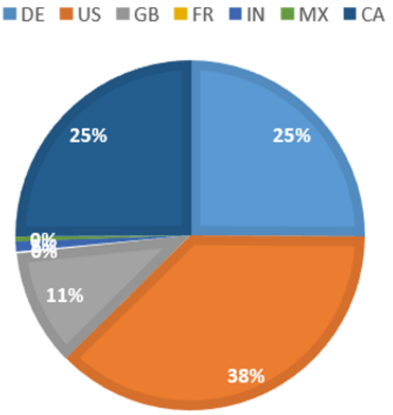


Figure 2-9 Vox

US, CA, DE like the channel most and other regions only form 12%.

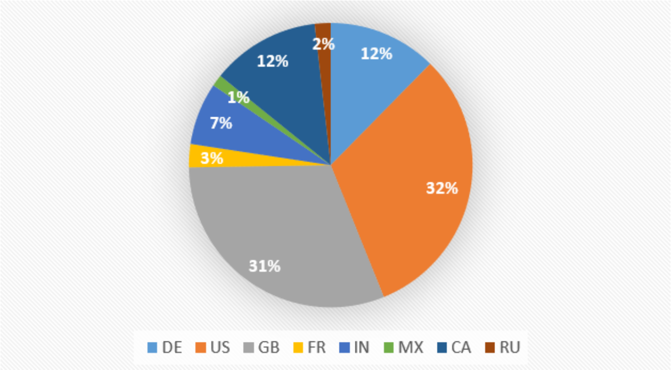


Figure 2-10 Netflix

GB, US form a huge part of 63% other region

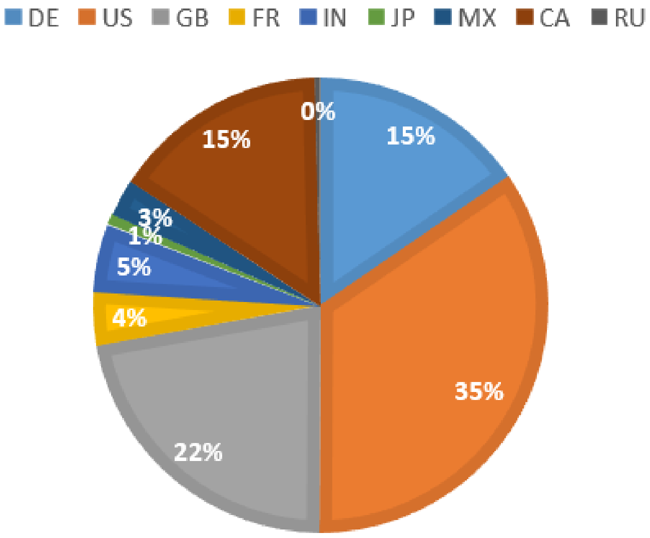


Figure 2-11 Screen Junkies

US and GB form the 57% of Screen Junkies' audience.

Task C

There will be two groups of visualization, each supporting the sub-topic depicted above. The first group of visualization will generate several figures for general statistics on incubation and trending duration based on different category. Histogram is drawn to show the mean, standard deviation and median for the incubation. Heat map is drawn to illustrate the intense and distribution of trending duration in different category.

In the second group of visualization, we draw area figure to show the relationship between incubation period, duration and DI-value. The x-axis should include hours in a day and days in a month. For easier and more directed visualization, we project the best days and hours of a category into calendar and clock. User are able to read the best reference time for a video.

Firstly, re-organize all of our data, create new excel for each category and record publishing date, trending date, and calculate Incubation Period of each video. Then, count the number of different Incubation Period value appears (frequency) and calculate their average, median, mode and standard deviation. It should be emphasized that the videos which take a long time to become prevalent (Incubation Period is more than 10 days) should be ignored when we calculate the duration since in this assignment, we only discuss general rules. Then use Tableau to complete histogram, set Frequency to row, and Incubation Period to column, and change the color and label for better visualization. The final summary visualization is shown in figure 2-12.

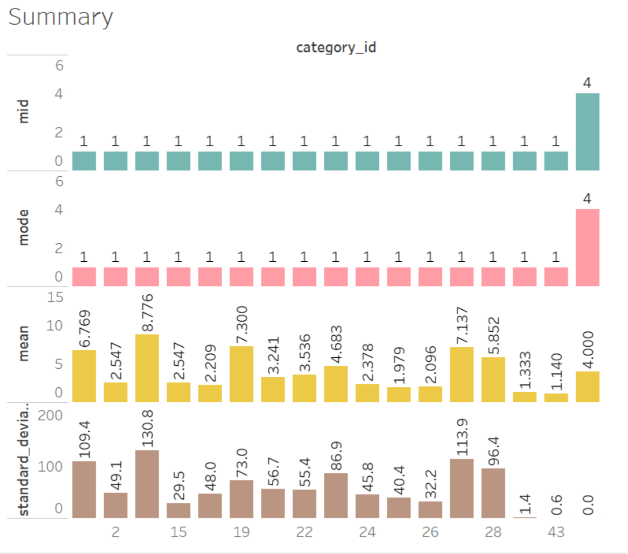


Figure 2-12 the statistic of IP in all categories

According to the figure 1, we can know that the median and mode of Incubation Period of all video categories are 1, which means that for most newly uploaded video, the next one day is the most important duration when the video is most likely to become prevalent. Hopefully this result can stand a chance to benefit both video bloggers and advertisement investors.

The heat map for trending duration is generated using the *matplot* and *seaborn* library in python. The density of the color in heat map shows the frequency of trending duration data. The trending duration is mainly retrieved from the trending date column. We clean the original data set to get the desired data, then we sort the data firstly by video ID, then by the trending date. The manipulated data set is prepared to cut to get the first trending date and the last trending date by drop the duplication and keep the first or the last record of a same video ID. The type of trending duration is datetime with the '%Y-%m-%dT%H:%M:%S.%fZ' format.

The python *seaborn* is used to draw the figure. By calling the heatmap () function, we can get the heat map result, shown in figure 2-13.

The deeper density means more counts in that pair of data. For example, as shown in this figure, we can conclude that entertainment, music, comedy and blogs are the four main popular categories that has longer trending duration while movies and trailers are extremely limited in the trending duration. This figure give insights to investors and YouTubers when they try to determine the video content they are going to film.

Figure 2-14 and 2-15 chronologically describe the DI-value for general videos in all categories. The edge differences between duration and incubation is proportion to the DI_value. If the DI_value is relatively larger at amount of time, for instance, 3am, 5am and 4pm GMT in figure 2 and 16th, 27th in a month, we say a video uploaded at that time is more likely to have more views. Or on the other side, we can draw a conclusion that, it would be wise if YouTubers avoid updating their videos during 10 pm to 12 pm GMT, and 10th, 29th in a month.

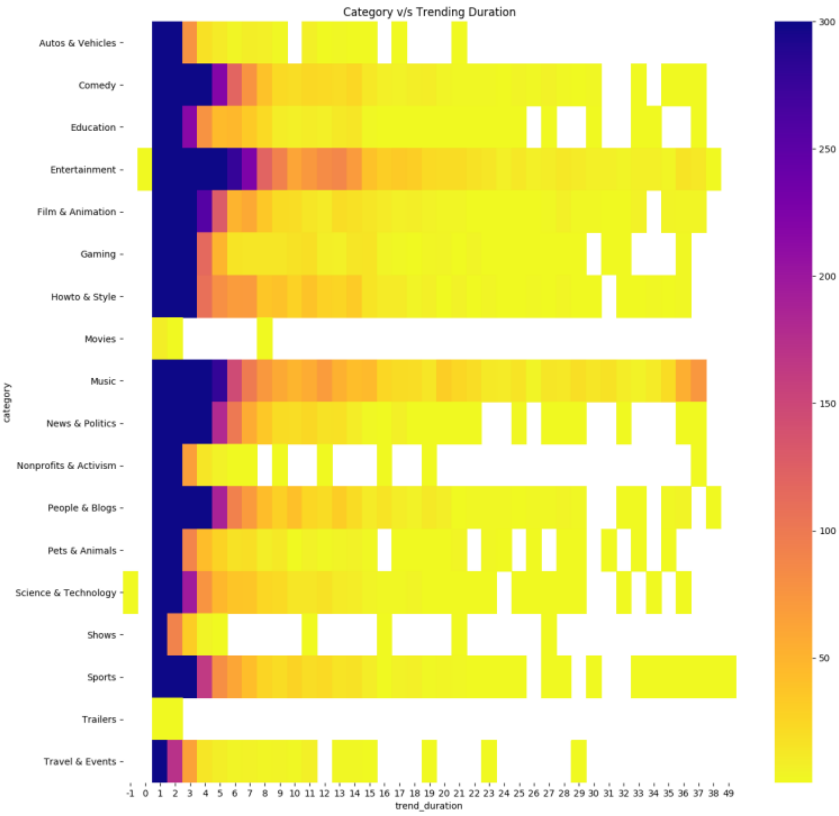


Figure 2-13 heat map for trending duration in different category

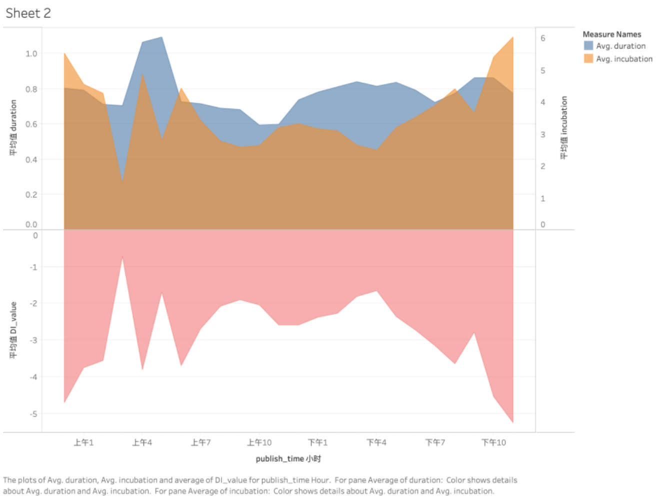


Figure 2-14 DI-value in hours: the plot of avg.duration, avg.incubation and average of DI-value. Color shows details about their value.

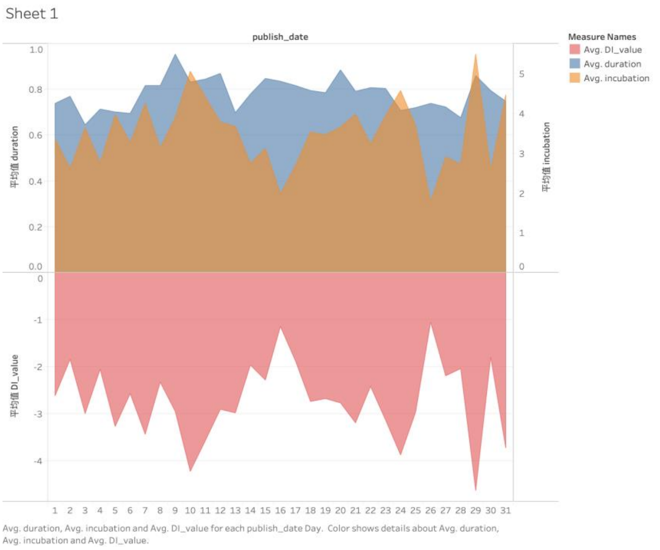
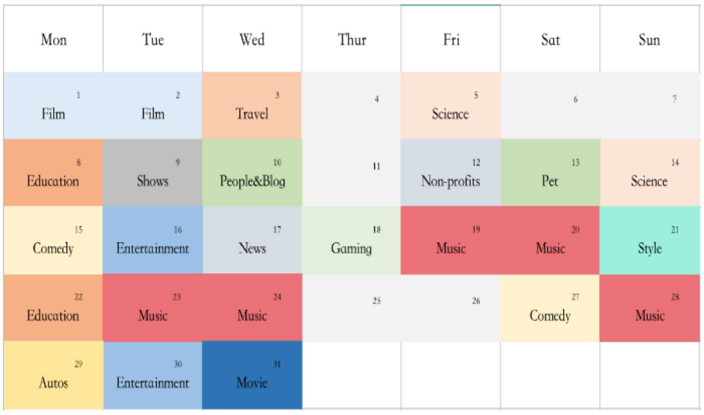


Figure 2-15: DI-value in days: the plot of avg.duration, avg.incubation and average of DI-value. Color shows details about their value.

And by finding the best time to upload videos for all category, we can combine them into two figures.



Mon	Tue	Wed	Thur	Fri	Sat	Sun
1 Film	2 Film	3 Travel	4	5 Science	6	7
8 Education	9 Shows	10 People&Blog	11	12 Non-profits	13 Pet	14 Science
15 Comedy	16 Entertainment	17 News	18 Gaming	19 Music	20 Music	21 Style
22 Education	23 Music	24 Music	25	26	27 Comedy	28 Music
29 Autos	30 Entertainment	31 Movie				

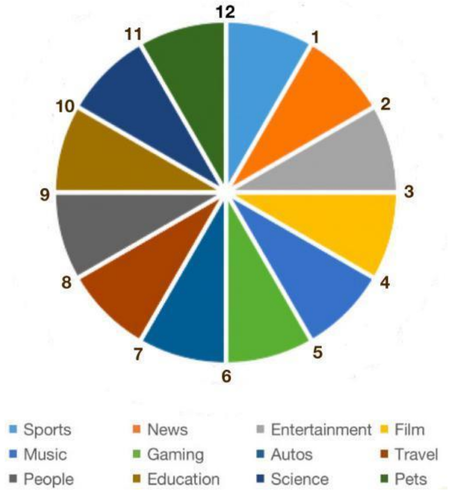


Figure 2-16 best time among different categories

These two figures are high conclusion visualization revealing the best day and the best hour for a specific categorized video. For example, a science video is suggested to be post during 14th at 10am GMT. This figure is of decent importance and has mammoth value for both the YouTubers and investors.

3.Implementation

Task A

For the data processing, the dataset includes the daily trending videos of ten countries, and the dataset is in separate files for each county. It includes titles, publish times, tags, description, comments and numbers of views, likes and dislikes, comment count. it is not necessary for use all of them. The raw dataset includes literal data in different languages, such as comments and descriptions columns, it leads to garbled data because of commas and semicolons. Thus, this step cleans the useless data in order to align all columns. Furthermore, we combined the dataset of the ten countries into one file.

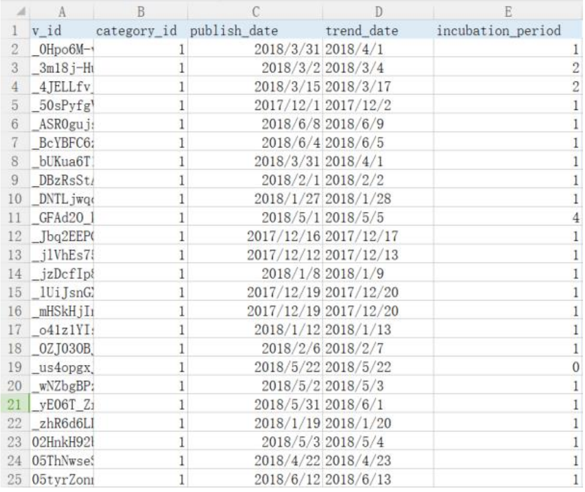
Tableau is our main tool for handling this. We choose the most reasonable visual format to demonstrate the result based on the business strategy. Attempt to visualize the dataset in different forms and estimate them by standards to obtain the most reasonable graph.

Task B

In this task, we use the Tableau combine with the paint tools to demonstrate the best performance which have the data and the beautiful photos to show the results.

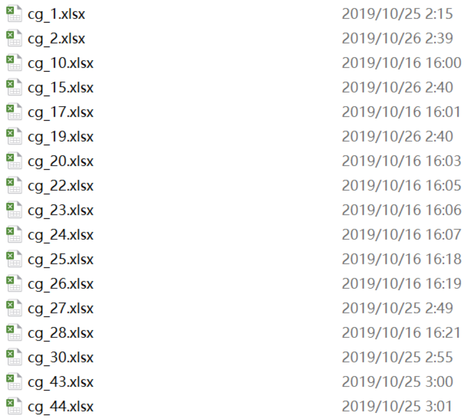
Task C

Task C mainly include two stages. Firstly, re-organize all our data, create new excel for each category and record publishing date, trending date, and calculate Incubation Period of each video.



	A	B	C	D	E
1	v_id	category_id	publish_date	trend_date	incubation_period
2	_0Hpo6M-'	1	2018/3/31	2018/4/1	1
3	_3m18j-Hı	1	2018/3/2	2018/3/4	2
4	_4JELLfv_	1	2018/3/15	2018/3/17	2
5	_50sPyfg'	1	2017/12/1	2017/12/2	1
6	_ASR0guj:	1	2018/6/8	2018/6/9	1
7	_BcYBFC6:	1	2018/6/4	2018/6/5	1
8	_bUKua6T	1	2018/3/31	2018/4/1	1
9	_DBzRsSt/	1	2018/2/1	2018/2/2	1
10	_DNTLjwq	1	2018/1/27	2018/1/28	1
11	_GFAd20_I	1	2018/5/1	2018/5/5	4
12	_Jbq2EEP(	1	2017/12/16	2017/12/17	1
13	_j1VhEs7!	1	2017/12/12	2017/12/13	1
14	_jzDcfIp{	1	2018/1/8	2018/1/9	1
15	_1UiJsnG]	1	2017/12/19	2017/12/20	1
16	_mHSkHjIı	1	2017/12/19	2017/12/20	1
17	_o41z1YI:	1	2018/1/12	2018/1/13	1
18	_0ZJ030B_	1	2018/2/6	2018/2/7	1
19	_us4opgx_	1	2018/5/22	2018/5/22	0
20	_wNZbgBP:	1	2018/5/2	2018/5/3	1
21	_yE06T_Zı	1	2018/5/31	2018/6/1	1
22	_zhR6d6LI	1	2018/1/19	2018/1/20	1
23	02HnkH92I	1	2018/5/3	2018/5/4	1
24	05ThNwse!	1	2018/4/22	2018/4/23	1
25	05tyrZonı	1	2018/6/12	2018/6/13	1

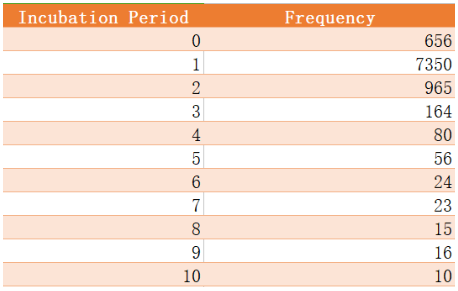
Figure 3-1 example - new excel of category 1



cg_1.xlsx	2019/10/25 2:15
cg_2.xlsx	2019/10/26 2:39
cg_10.xlsx	2019/10/16 16:00
cg_15.xlsx	2019/10/26 2:40
cg_17.xlsx	2019/10/16 16:01
cg_19.xlsx	2019/10/26 2:40
cg_20.xlsx	2019/10/16 16:03
cg_22.xlsx	2019/10/16 16:05
cg_23.xlsx	2019/10/16 16:06
cg_24.xlsx	2019/10/16 16:07
cg_25.xlsx	2019/10/16 16:18
cg_26.xlsx	2019/10/16 16:19
cg_27.xlsx	2019/10/25 2:49
cg_28.xlsx	2019/10/16 16:21
cg_30.xlsx	2019/10/25 2:55
cg_43.xlsx	2019/10/25 3:00
cg_44.xlsx	2019/10/25 3:01

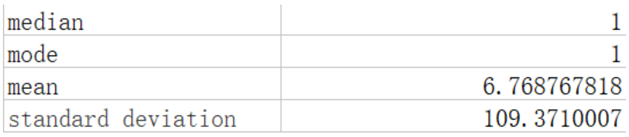
Figure 3-2 re-organize all the data

Then, count the number of different Incubation Period value appears (frequency) and calculate their average, median, mode and standard deviation. It should be emphasized that the videos which take a long time to become prevalent (Incubation Period is more than 10 days) should be ignored when we calculate the duration since in this assignment ,we only discuss general rules.



Incubation Period	Frequency
0	656
1	7350
2	965
3	164
4	80
5	56
6	24
7	23
8	15
9	16
10	10

Fig3-3: example - count the frequency



median	1
mode	1
mean	6.768767818
standard deviation	109.3710007

Figure 3-4 statistics

And summarize all statistical values.

Then use Tableau to complete histogram, set Frequency to row, and Incubation Period to column, and change the color and label for better visualization.

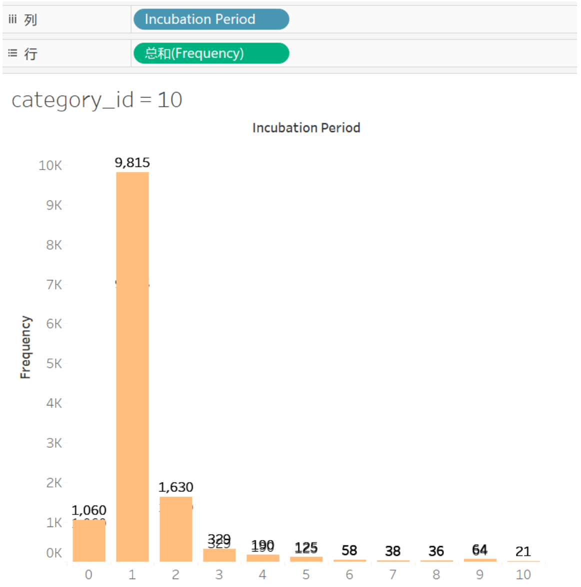


Figure 3-5: example of category 10

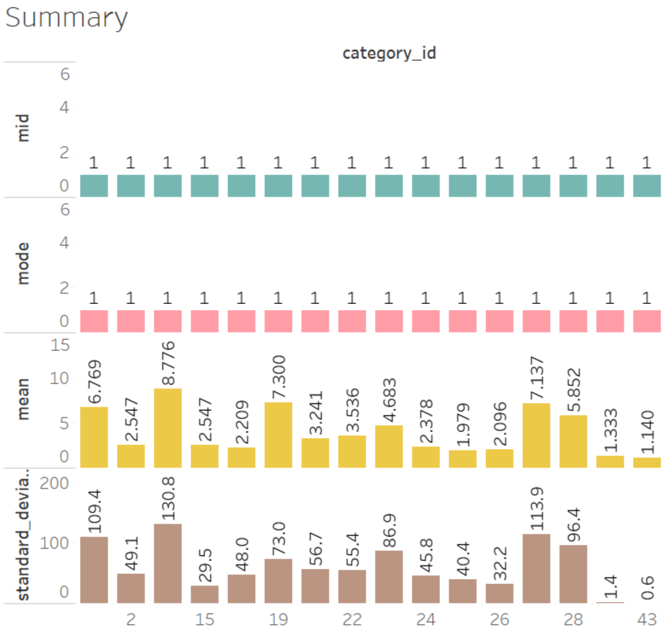


Figure 3-6 summary

This is the summary of the implementation of the first stage. For the second stage, we create our own calculation methods to find the best time of uploading combined with some python codes and use of Tableau. Final results have been demonstrated In figure 2-16.

4.Evaluation

4.1 Results

As the figure 4-1 demonstrates, evaluation is the one of the most significant part within the plan to promote the development of the project, and we have done the evaluation part several times for acquiring better result. To collect more clearly and precisely results, we have many iterations of the process of system development cycle. For the evaluation part, we use the Cognitive Walkthrough method to evaluate our visualizations and to check whether the final visualizations are really performed well to support the tasks. We will analyse the results of them as follow.

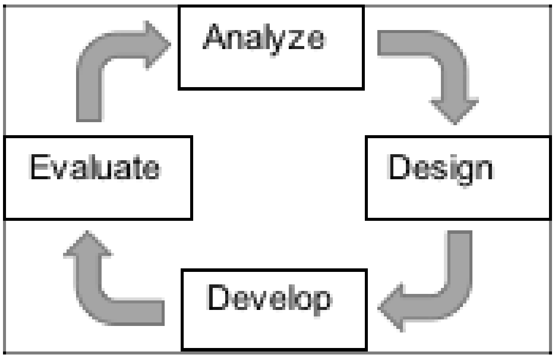


Figure 4-1 Process of the project

Task A

This task aims to analyse the market size of the YouTube trending videos in the global world. According to our visualizations which is shown before, we use a geographical map to show the total number of views of different countries with the different colour. We can conclude that the Great Britain has the largest market size and Japan is the smallest. Then, we analyse each region in-depth to identify what are their favourite YouTube videos categories. It can be seen from our visualization that, the top 3 influential categories of YouTube videos are Music, Entertainment, and Film & Animation. Then we analyse the relationship between numbers of views and likes among multiple dimensions which can show whether there is a large amount of repeating viewing by individuals. Based on our implementation, Movie, Shows and Trailers are the top 3 categories whose videos are repeat viewing in multiple times.

Task B

This task is to identify which channel of videos is the most popular and attractive. We analyse it by each country. As the visualization of US channel-category, Entertainment occupy the largest area of the graph which means it has the maximum number of videos which on the trending list, so Entertainment is the most popular channel. Then we focus on the specific channel of US. The channel of ESPN has 200 videos which belong to sport category on the list. We also analyze the top 3 channels of other countries, such as CA, GB, India and so on. Entertainment is the most popular category in all of these countries. Most of these top channels make only one category of videos. Because there are some channels are on the list of many countries, we also analyse the detail of it. For ESPN, it is distributed in five countries, US accounts for 32% which is the largest. DE, GB, CA have the similar proportion. For the Night Show Starring Jimmy Fallon, it is distributed in nine countries, GB is the largest one which is 27% and US is followed by it with 26%. For Vox, it is distributed in seven countries with US is 38% and DE and CA are 25% separately. For Netflix, it is distributed in eight countries. US is the largest one and GB almost the same with it. For Screen Junkies, it is distributed in nine countries, US is the largest one again.

Task C

This task is to identify the best video launching time and evaluate the relationship between category and how long it takes for the related video to become prevalent. The figures in Task C visualization part support well to solve the two topics. In the first topic, we draw two general statistic pictures about incubation and trending duration. Then we further our analysis on the calculation among them and get the DI-value area figure. DI-value is strongly referenced to find the best time to upload a video.

4.2 Discussion

Our tasks are business-oriented which focus on offer the business value for both markets and users. We have achieved some practical and useful results to help analyse the related factors which may affect the trending of videos. The detail information which is provided by our tasks can help understand the trending of popular, what the market size, what is the most popular category, and when is the best time to upload a video among multiple countries. We will analyse more in-depth in the future to get more meaningful implementation.

5. Conclusion

Our major target is to use the YouTube trending dataset to find the business value behind them, and this have been done successfully since we can justify them based on three main factors:

1)The market size in of different area. In Task A we have performed detailed trending of different areas, we use different diagrams also with the data after pre-processing to handle the factors that may affect the results.

2)The channel information. From the Task B we support the information of the relationship between channel and category and based on this we could provide some advertisement guidance to those people who want to invest.

3)The correct time for the video to become ‘trending video’. Task C we have finished some analysis based on the time factor and put some new ideas and calculation methods to find the better results, then we collected the time for the video to become ‘trending’, and this means it could put more provide to those youtubers for improving there efficiency or also attracting the investors to invest some money into those ‘potential’ videos.

In conclusion, in this project we have analysed the data and transferred them into the visual diagrams to practice our visualization skills. Furthermore, we accomplish some meaning things to the real world like the business factors. Also, we may have some insufficient things that we did not concern, we have satisfied basic requirements. In future, we may improve our developing methods or put more new research questions based on the dataset.

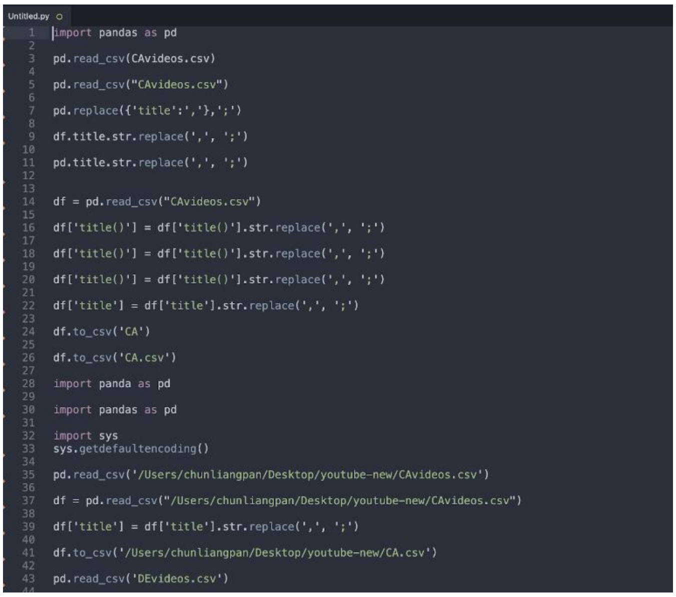
References

Steed, C. A., Potok, T. E., Patton, R. M., Goodall, J. R., Maness, C., Senter, J. (2012, October). Interactive visual analysis of high throughput text streams. In International visual text analytics workshop (pp. 1-4).

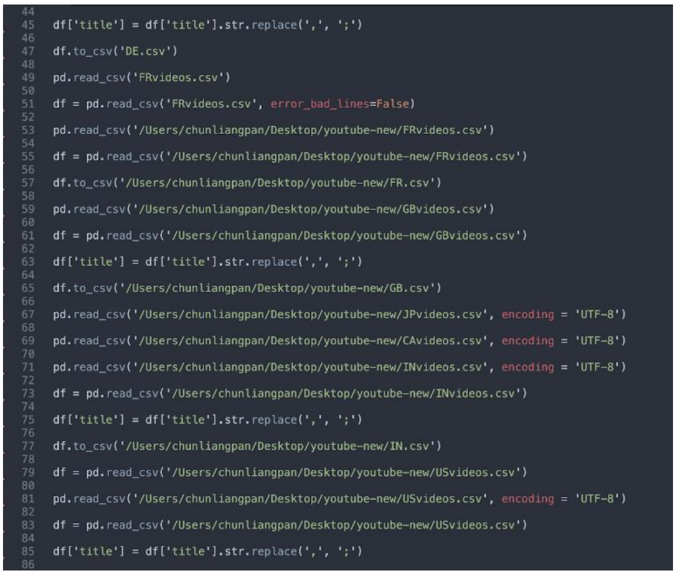
Taylor Loren, The Best Time to Post on Instagram in 2019, According to 12 Million Posts.

Appendix: Code

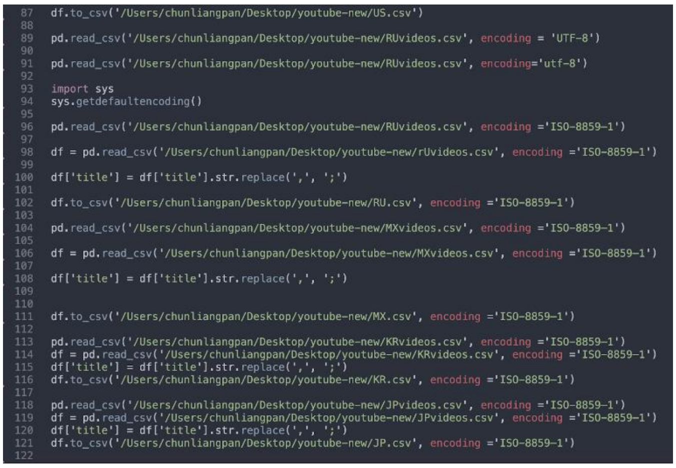
In order to fix these problems, the cleanup team cleaned the dataset with the programming language, Python and the library Pandas because there are certain invalid values in the dataset and some data structures have bugs. The specific cleaning code is shown below.



```
import pandas as pd

pd.read_csv(CAvideos.csv)

pd.read_csv("CAvideos.csv")

pd.replace({'title':','},';')

df.title.str.replace(',', ';')

pd.title.str.replace(',', ';')


df = pd.read_csv("CAvideos.csv")

df['title()'] = df['title()'].str.replace(',', ';')

df['title()'] = df['title()'].str.replace(',', ';')

df['title()'] = df['title()'].str.replace(',', ';')

df['title'] = df['title'].str.replace(',', ';')

df.to_csv('CA')

df.to_csv('CA.csv')

import panda as pd

import pandas as pd

import sys
sys.getdefaultencoding()

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/CAvideos.csv')

df = pd.read_csv("/Users/chunliangpan/Desktop/youtube-new/CAvideos.csv")

df['title'] = df['title'].str.replace(',', ';')

df.to_csv('/Users/chunliangpan/Desktop/youtube-new/CA.csv')

pd.read_csv('DEvideos.csv')
```



```

df['title'] = df['title'].str.replace(',', ';')

df.to_csv('DE.csv')

pd.read_csv('FRvideos.csv')

df = pd.read_csv('FRvideos.csv', error_bad_lines=False)

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/FRvideos.csv')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/FRvideos.csv')

df.to_csv('/Users/chunliangpan/Desktop/youtube-new/FR.csv')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/GBvideos.csv')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/GBvideos.csv')

df['title'] = df['title'].str.replace(',', ';')

df.to_csv('/Users/chunliangpan/Desktop/youtube-new/GB.csv')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/JPvideos.csv', encoding = 'UTF-8')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/CAvideos.csv', encoding = 'UTF-8')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/INvideos.csv', encoding = 'UTF-8')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/INvideos.csv')

df['title'] = df['title'].str.replace(',', ';')

df.to_csv('/Users/chunliangpan/Desktop/youtube-new/IN.csv')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/USvideos.csv')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/USvideos.csv', encoding = 'UTF-8')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/USvideos.csv')

df['title'] = df['title'].str.replace(',', ';')

```



```
df.to_csv('/Users/chunliangpan/Desktop/youtube-new/US.csv')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/RUvideos.csv', encoding = 'UTF-8')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/RUvideos.csv', encoding='utf-8')

import sys
sys.getdefaultencoding()

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/RUvideos.csv', encoding ='ISO-8859-1')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/rUvideos.csv', encoding ='ISO-8859-1')

df['title'] = df['title'].str.replace(',', ';')

df.to_csv('/Users/chunliangpan/Desktop/youtube-new/RU.csv', encoding ='ISO-8859-1')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/MXvideos.csv', encoding ='ISO-8859-1')

df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/MXvideos.csv', encoding ='ISO-8859-1')

df['title'] = df['title'].str.replace(',', ';')


df.to_csv('/Users/chunliangpan/Desktop/youtube-new/MX.csv', encoding ='ISO-8859-1')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/KRvideos.csv', encoding ='ISO-8859-1')
df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/KRvideos.csv', encoding ='ISO-8859-1')
df['title'] = df['title'].str.replace(',', ';')
df.to_csv('/Users/chunliangpan/Desktop/youtube-new/KR.csv', encoding ='ISO-8859-1')

pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/JPvideos.csv', encoding ='ISO-8859-1')
df = pd.read_csv('/Users/chunliangpan/Desktop/youtube-new/JPvideos.csv', encoding ='ISO-8859-1')
df['title'] = df['title'].str.replace(',', ';')
df.to_csv('/Users/chunliangpan/Desktop/youtube-new/JP.csv', encoding ='ISO-8859-1')

```

12th September 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Choose dataset + identify tasks

Discussion:

Before meeting, each number of our group has already analyzed the datasets which assigned by lecturer. Therefore, we have more time to discuss the analysis and understanding of each dataset and choose the most worthy dataset——YouTube. In addition, we share the experience of the visualization tools which we familiar with. After that, we try to identify the basic tasks about this dataset and determine to use which tools is better. Last, we determine to think about more detailed tasks and discuss at next meeting.

19th September 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Initial report draft

Discussion:

This week, we discuss more detailed tasks needed to visualization. There are two main categories of tasks about the dataset. One is focus on the relationship which are shown obviously among different dimensions, the other is trying to develop and speculate potential valuable factors that may be relevant to the real world. We finally identify two abstract tasks which contains several concrete tasks separately. After that, we discuss the contents should be involved in the initial report like analysis, visualization, implementation, evaluation, and so on. Then we make a draft of the initial draft and assign tasks for each member.

26th September 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Complete initial report

Discussion:

We mainly focus on completing the initial report during this week. We firstly discuss the progress about the tasks from the last week to now separately. Then we share the difficult and uncertain contents and analyze the solution of it together. In terms of the visualization part, we try to determine the way to visualize our tasks and which kinds of graph are the best to show visual representation of the results apparently. After that, we continue working on improving the final version of the initial report. We appoint upload ourselves works to Google Drive before a set time and collate them to the final version which has a formal format.

10th October 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Project presentation

Discussion:

The analysis team figured out what requirements of the presentation are. Based on the initial report, we discussed about the structure and design template related to the project presentation. After discussing about the structure and design of presentation, then the team move to the task assigning.

Because the analysis team has 7 team members, there are 4 members mainly responsible for providing context and graphics of the presentation. The remaining three people are responsible for the speech and make detailed adjustments to the slide according to their own time.

After clarifying respective responsibilities, our analysis team set the Sunday of week 9 as the deadline for submitting reports within the group, and agreed to upload their content to google drive.

17th October 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Project presentation

Discussion:

This week, the main discussion direction of the analysis team was to simulate the presentation because our team was assigned to Week10 for presentation. In addition, since we are merging the main tasks on the basis of the initial report, and discussing them by a hypothetical way, this group discussion is the last time we unified the idea before the presentation.

The main content of the discussion is that the four content providers of the group will give a detailed introduction to their respective parts and will be repeated by the speaker responsible for his module. If the retelling is consistent with the introduction, it is passed, and vice versa, the discussion is continued to reach an agreement.

After all of team members agreed with comments, we did a drill and successfully completed the presentation in a limited time.

24th October 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Project final report

Discussion:

The analysis team reviewed the work of the other groups in the week 10 presentation and sorted out some of the shortcomings and improvements. In the framework of the presentation idea and the initial report, our team made the corresponding analysis of final report and task assignment for the final report.

Similar to the task assignment of presentation, our four content providers are primarily responsible for the part of task analysis and visualization. The remaining three are responsible for completing the introduction, implementation and evaluation sections of the final report.

Finally, our team agreed to meet at week 13 and check the quality of the paper to ensure that the work was submitted on time and in good quality.

1st November 2019
3:00-5:00 PM
School of Computer Science

Weekly Meeting

Attendees: All group members doing this project

Agenda item: Complete final report

Discussion:

The main task of this week is to complete the final report. During the meeting, we discuss the content of each subtask of the final report and assign the subtasks to each member. Because we have determined the final tasks which is needed to be visualized before, we just talk about whether the visualization can be improved to a better representation. In addition, we decide to share our work and progress through Google Drive to better cooperation. After every group member finish their work, we collate them together. However, the report is longer than the page limitation, we discuss each part of the report to make them concise.