COGS108 Project Proposal

Yanyu Tao, Prasen Shakya, Abigail Diamse, Yuhua Xu, Weisi Luo, Fangyu Liu

TOTAL POINTS

8.75 / 10

QUESTION 1

1 Question & Hypothesis 1.5 / 2

✓ + **2 pts** **Question & Hypothesis section included**

✓ - **0.25 pts** **question not stated or is unclear**

- **0.25 pts** question not answerable with data

- **0.25 pts** hypothesis not stated or is unclear

✓ - **0.25 pts** **explanation of why hypothesis was chosen not stated**

- **1 pts** question missing

I'm confused as to what exactly your question is and why you're interested in it ....isn't a video trending dependent on how many times it's viewed? Like isn't that part of the definition? Doesn't this *have* to be true? I suggest revisiting your question to think about what you're actually interested in. Why do you care about view count? Or is it trending videos you want to better understand? If your goal is to understand what makes a trending video, you likely want to determine what influences view count or trending. Is it genre? What's in the video (cats, babies, etc.)? Is it length? Does it matter who shares it originally? Or what part of the world it is shared from? These questions and determining how to measure those and their relation to trending videos/view count is much more in line with what would make for an interesting COGS108 project. Right now it looks to me (and I'd be happy to be wrong if you explained further) like you're just measuring how many views it takes to be considered 'trending'.

QUESTION 2

2 Background 1 / 1

✓ + **1 pts** **Background Section Included**

- **0.25 pts** reference / link missing

- **0.5 pts** No background information about anything on the topic included

Would like to see, in your final report, a summary of the findings from each of those links (Rather than just the link to them. Otherwise, this is a good start.

QUESTION 3

3 Ethical Considerations 1.25 / 2

✓ + **2 pts** **Ethical Considerations Included**

✓ - **0.75 pts** **Details/ethical considerations lacking**

- **0.75 pts** Unclear or unthoughtful

Consider the questions raised here to add additional thought to this section: http://deon.drivendata.org/ ....explain how bias would be removed by focusing on a specific US region. The note about bias due to China and NK makes enough sense, but there are additional considerations to be made.

QUESTION 4

4 Data 3 / 3

✓ + **3 pts** **Data Section Included**

- **1 pts** Explanation of datasets missing or lacking in detail

- **1 pts** Source(s) not included

- **0.5 pts** Minor detail missing (size of dataset etc)

QUESTION 5

5 Team Expectations 1 / 1

✓ + **1 pts** **Team Expectations Included**

QUESTION 6

6 Timeline proposal **1 / 1**

- ✓ **+ 1 pts** **Timeline proposal included**
- **- 0.5 pts** Specific Dates missing
- **- 0.5 pts** table not customized for their group

I like that you have Everyone contributing, but be sure that everyone *is* contributing....and that all the work doesn't fall to one person over and over again.

Have You Seen This Cat Video?

TEAM NAME:
YAY
TEAM MEMBERS:
Abigail Diamse A14135992
Yanyu Tao A13961185
Weisi Luo A13629635
Yuhua Xu A13797321
Fangyu Liu A53255336
Prasen Shakya A14706492

DATA SCIENCE QUESTION(S) & HYPOTHESIS:

How does the trending type of videos affect the view count?
We predict the relationship between types or genres of the videos are directional to the popularity of the videos. The popularity of the videos will be measured by the number of views.

BACKGROUND:

Considering that Youtube is a constantly growing website with billions of videos being uploaded every day, many wonder what factors decide which videos end up viral, what leads to videos becoming number 1 on the website, etc. As Youtube has a vast audience of different age groups, cultures, backgrounds, etc., the number of views that video gets could be an indicator as to its popularity especially within their specific genres.

"Youtube and the Dynamics of the Participatory Culture"
http://eprints.qut.edu.au/18431/1/18431.pdf

How we plan to expand on this: We plan to explore different trends within several different genres in Youtube in an attempt to see a relationship between the popularity of a video different trends that occur. The concept of "Participatory Culture" is the idea that people are jumping onto trends that at popularity at the time, thus increasing the potential of those videos to become more popular. Thus, we plan to find and make visual this relationship.

"On the prediction of popularity of trends and hits for user generated videos"
https://dl.acm.org/citation.cfm?id=2433489

How we plan to expand on this: We plan to present the performance of different genres of videos using prediction models. We will evaluate the prediction models on a random sample of Youtube videos

1 Question & Hypothesis 1.5 / 2

- ✓ **+ 2 pts** Question & Hypothesis section included
- ✓ **- 0.25 pts** question not stated or is unclear
- **- 0.25 pts** question not answerable with data
- **- 0.25 pts** hypothesis not stated or is unclear
- ✓ **- 0.25 pts** explanation of why hypothesis was chosen not stated
- **- 1 pts** question missing

I'm confused as to what exactly your question is and why you're interested in it ....isn't a video trending dependent on how many times it's viewed? Like isn't that part of the definition? Doesn't this *have* to be true? I suggest revisiting your question to think about what you're actually interested in. Why do you care about view count? Or is it trending videos you want to better understand? If your goal is to understand what makes a trending video, you likely want to determine what influences view count or trending. Is it genre? What's in the video (cats, babies, etc.)? Is it length? Does it matter who shares it originally? Or what part of the world it is shared from? These questions and determining how to measure those and their relation to trending videos/view count is much more in line with what would make for an interesting COGS108 project. Right now it looks to me (and I'd be happy to be wrong if you explained further) like you're just measuring how many views it takes to be considered 'trending'.

Have You Seen This Cat Video?

TEAM NAME:
YAY
TEAM MEMBERS:
Abigail Diamse A14135992
Yanyu Tao A13961185
Weisi Luo A13629635
Yuhua Xu A13797321
Fangyu Liu A53255336
Prasen Shakya A14706492

DATA SCIENCE QUESTION(S) & HYPOTHESIS:

How does the trending type of videos affect the view count?
We predict the relationship between types or genres of the videos are directional to the popularity of the videos. The popularity of the videos will be measured by the number of views.

BACKGROUND:

Considering that Youtube is a constantly growing website with billions of videos being uploaded every day, many wonder what factors decide which videos end up viral, what leads to videos becoming number 1 on the website, etc. As Youtube has a vast audience of different age groups, cultures, backgrounds, etc., the number of views that video gets could be an indicator as to its popularity especially within their specific genres.

"Youtube and the Dynamics of the Participatory Culture"
http://eprints.qut.edu.au/18431/1/18431.pdf

How we plan to expand on this: We plan to explore different trends within several different genres in Youtube in an attempt to see a relationship between the popularity of a video different trends that occur. The concept of "Participatory Culture" is the idea that people are jumping onto trends that at popularity at the time, thus increasing the potential of those videos to become more popular. Thus, we plan to find and make visual this relationship.

"On the prediction of popularity of trends and hits for user generated videos"
https://dl.acm.org/citation.cfm?id=2433489

How we plan to expand on this: We plan to present the performance of different genres of videos using prediction models. We will evaluate the prediction models on a random sample of Youtube videos

2 Background 1 / 1

✓ **+ 1 pts** **Background Section Included**

- 0.25 pts reference / link missing

- 0.5 pts No background information about anything on the topic included

Would like to see, in your final report, a summary of the findings from each of those links (Rather than just the link to them. Otherwise, this is a good start.

ETHICAL CONSIDERATIONS:
The dataset is online through kaggle that is an open community for data scientists and machine learners. Thus, we have the permission to use the whole dataset for exploring and studying. And there should not be privacy concerns regarding the dataset because there is no youtuber written in the datasets.
The potential bias in our dataset is some countries are excluded (such as China and North Korea) because youtube cannot be used in those countries. Moreover, there should not be a privacy issue since the youtube is public and youtuber should understand whatever they post is public.
In order to solve the potential bias in the dataset, we plan to narrow down our analysis in a specific region(US).

DATA:
Dataset Name: Trending YouTube Video Statistics
Link to the dataset: https://www.kaggle.com/datasnaek/youtube-new

The datasets concludes the trending youtube video data, which contains 5 csv files and 5 json files(for 5 different regions). In our project, we decided to use one csv file called “USvideos.csv” as our dataset. The number of observations in our dataset is 40379. And this dataset includes video titles, channels, categories, publish time, number of views, number of likes and dislikes, description of the video,etc.

TEAM EXPECTATIONS AGREEMENT

Read over the COGS108 Team Policies individually. Then, include your group’s expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member’s name above and by adding their name to the Gradescope submission, you are indicating that you have read the COGS108 Team Policies, accept your team’s expectations below, and have every intention to fulfill them.

These expectations are for your team’s use and benefit—they won’t be graded for their details. Goals should be realistic: “No group member will never miss a meeting and everyone will always show up early” is probably unrealistic, but “Group members will attend almost every meeting and will communicate their absence at least a day in advance of the group meeting” and “When group members are unable to attend a meeting, they will submit their notes and progress ahead of the group meeting” are realistic expectations. Expectations for deadlines,

3 Ethical Considerations 1.25 / 2

✓ **+ 2 pts Ethical Considerations Included**

✓ **- 0.75 pts Details/ethical considerations lacking**

- 0.75 pts Unclear or unthoughtful

Consider the questions raised here to add additional thought to this section: http://deon.drivendata.org/ ....explain how bias would be removed by focusing on a specific US region. The note about bias due to China and NK makes enough sense, but there are additional considerations to be made.

ETHICAL CONSIDERATIONS:
The dataset is online through kaggle that is an open community for data scientists and machine learners. Thus, we have the permission to use the whole dataset for exploring and studying. And there should not be privacy concerns regarding the dataset because there is no youtuber written in the datasets.
The potential bias in our dataset is some countries are excluded (such as China and North Korea) because youtube cannot be used in those countries. Moreover, there should not be a privacy issue since the youtube is public and youtuber should understand whatever they post is public.
In order to solve the potential bias in the dataset, we plan to narrow down our analysis in a specific region(US).

DATA:
Dataset Name: Trending YouTube Video Statistics
Link to the dataset: https://www.kaggle.com/datasnaek/youtube-new

The datasets concludes the trending youtube video data, which contains 5 csv files and 5 json files(for 5 different regions). In our project, we decided to use one csv file called “USvideos.csv” as our dataset. The number of observations in our dataset is 40379. And this dataset includes video titles, channels, categories, publish time, number of views, number of likes and dislikes, description of the video,etc.

TEAM EXPECTATIONS AGREEMENT

Read over the COGS108 Team Policies individually. Then, include your group’s expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member’s name above and by adding their name to the Gradescope submission, you are indicating that you have read the COGS108 Team Policies, accept your team’s expectations below, and have every intention to fulfill them.

These expectations are for your team’s use and benefit—they won’t be graded for their details. Goals should be realistic: “No group member will never miss a meeting and everyone will always show up early” is probably unrealistic, but “Group members will attend almost every meeting and will communicate their absence at least a day in advance of the group meeting” and “When group members are unable to attend a meeting, they will submit their notes and progress ahead of the group meeting” are realistic expectations. Expectations for deadlines,

4 Data 3 / 3

- ✓ **+ 3 pts** **Data Section Included**
- **- 1 pts** Explanation of datasets missing or lacking in detail
- **- 1 pts** Source(s) not included
- **- 0.5 pts** Minor detail missing (size of dataset etc)

ETHICAL CONSIDERATIONS:
The dataset is online through kaggle that is an open community for data scientists and machine learners. Thus, we have the permission to use the whole dataset for exploring and studying. And there should not be privacy concerns regarding the dataset because there is no youtuber written in the datasets.
The potential bias in our dataset is some countries are excluded (such as China and North Korea) because youtube cannot be used in those countries. Moreover, there should not be a privacy issue since the youtube is public and youtuber should understand whatever they post is public.
In order to solve the potential bias in the dataset, we plan to narrow down our analysis in a specific region(US).

DATA:
Dataset Name: Trending YouTube Video Statistics
Link to the dataset: https://www.kaggle.com/datasnaek/youtube-new

The datasets concludes the trending youtube video data, which contains 5 csv files and 5 json files(for 5 different regions). In our project, we decided to use one csv file called "USvideos.csv" as our dataset. The number of observations in our dataset is 40379. And this dataset includes video titles, channels, categories, publish time, number of views, number of likes and dislikes, description of the video,etc.

Team Expectations Agreement

Read over the COGS108 Team Policies individually. Then, include your group's expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member's name above and by adding their name to the Gradescope submission, you are indicating that you have read the COGS108 Team Policies, accept your team's expectations below, and have every intention to fulfill them.

These expectations are for your team's use and benefit—they won't be graded for their details. Goals should be realistic: "No group member will never miss a meeting and everyone will always show up early" is probably unrealistic, but "Group members will attend almost every meeting and will communicate their absence at least a day in advance of the group meeting" and "When group members are unable to attend a meeting, they will submit their notes and progress ahead of the group meeting" are realistic expectations. Expectations for deadlines, how you'll work together, meeting attendance and participation, and project completion should all be considered and details included below.

INCLUDE YOUR TEAM'S EXPECTATIONS HERE

- Members will meet at least once a week.
- Coordinate and communicate through Slack and GMail.
- Set deadline for given task on day before weekly meeting. Otherwise, decide collectively on a change of date.
- Work on shared Google Docs for ideas and minute meetings.
- When group members are unable to attend they can share their ideas and progress on shared minute-meeting Google Doc and through Slack.
- Finish the codes and texts by due day

5 Team Expectations 1 / 1

✓ + 1 pts **Team Expectations Included**

Project Timeline Proposal

Include actual dates and times for due dates and meetings below, not just what week they'll be completed

	Draft Text?	**Write Code?**	**Proposed due date**	**Discuss at team meeting**	**Edit?**
Initial team meeting	NA	NA	NA	week 2	NA
Background Research	Yuhua Abigail Prasen	NA	4/19/2019 week 3	4/23/2019 week 4	Everyone
Question & Hypothesis	Yuhua	NA	4/19/2019 week 3	4/23/2019 week 4	Everyone
Ethical Considerations	Weisi Yanyu Fangyu	NA	4/19/2019 week 3	4/23/2019 week 4	Everyone
Dataset	Weisi Fangyu	Everyone	4/23/2019 week 4	4/23/2019 week 4	Everyone
Data Wrangling	Everyone	Everyone	4/23/2019 week 4	4/23/2019 week 4	Everyone
Descriptive	Everyone	Everyone	5/05/2019 week 5	5/07/2019 week 6	Everyone
Exploratory	Everyone	Everyone	5/05/2019 week 5	5/07/2019 week 6	Everyone
Analysis - Part I	Everyone	Everyone	5/11/2019 week 6	5/14/2019 week 7	Everyone
Analysis - Part II	Everyone	Everyone	5/11/2019 week 6	5/14/2019 week 7	Everyone
Analysis - Part III	Everyone	Everyone	5/11/2019 week 6	5/14/2019 week 7	Everyone

Summarize Results	Everyone	NA	5/18/2019 week 7	5/21/2019 week 8	Everyone
Conclusions	Everyone	NA	5/18/2019 week 7	5/21/2019 week 8	Everyone

Once completed, save this document as a PDF & submit on Gradescope. <u>Be sure to add each team member’s name to the Gradescope submission.</u>

6 Timeline proposal **1 / 1**

✓ **+ 1 pts Timeline proposal included**

- 0.5 pts Specific Dates missing

- 0.5 pts table not customized for their group

I like that you have Everyone contributing, but be sure that everyone *is* contributing....and that all the work doesn't fall to one person over and over again.