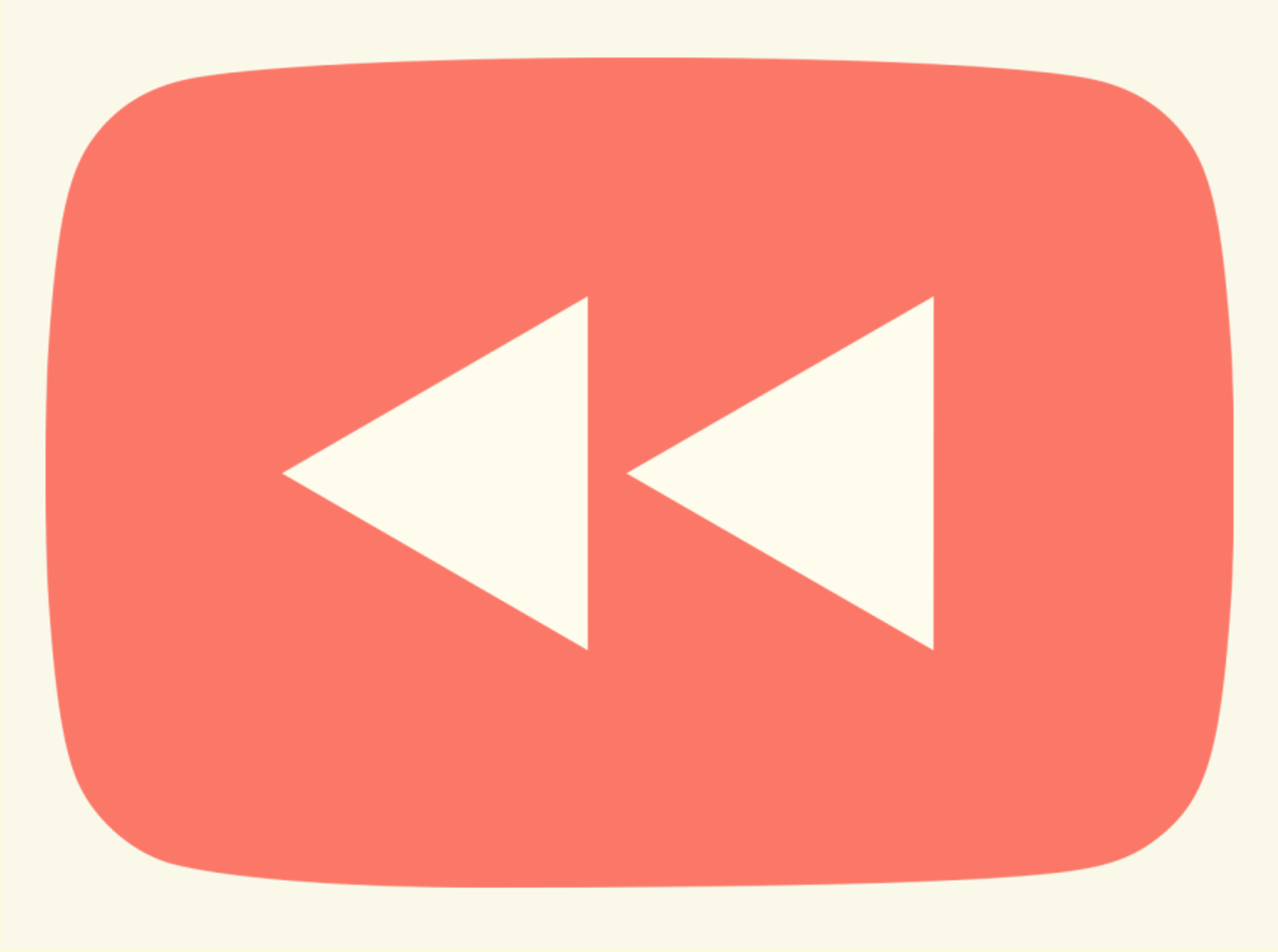
YOUTUBE ANALYZER (CPT_S 415 PROJECT Team- BitsN'Bytes)

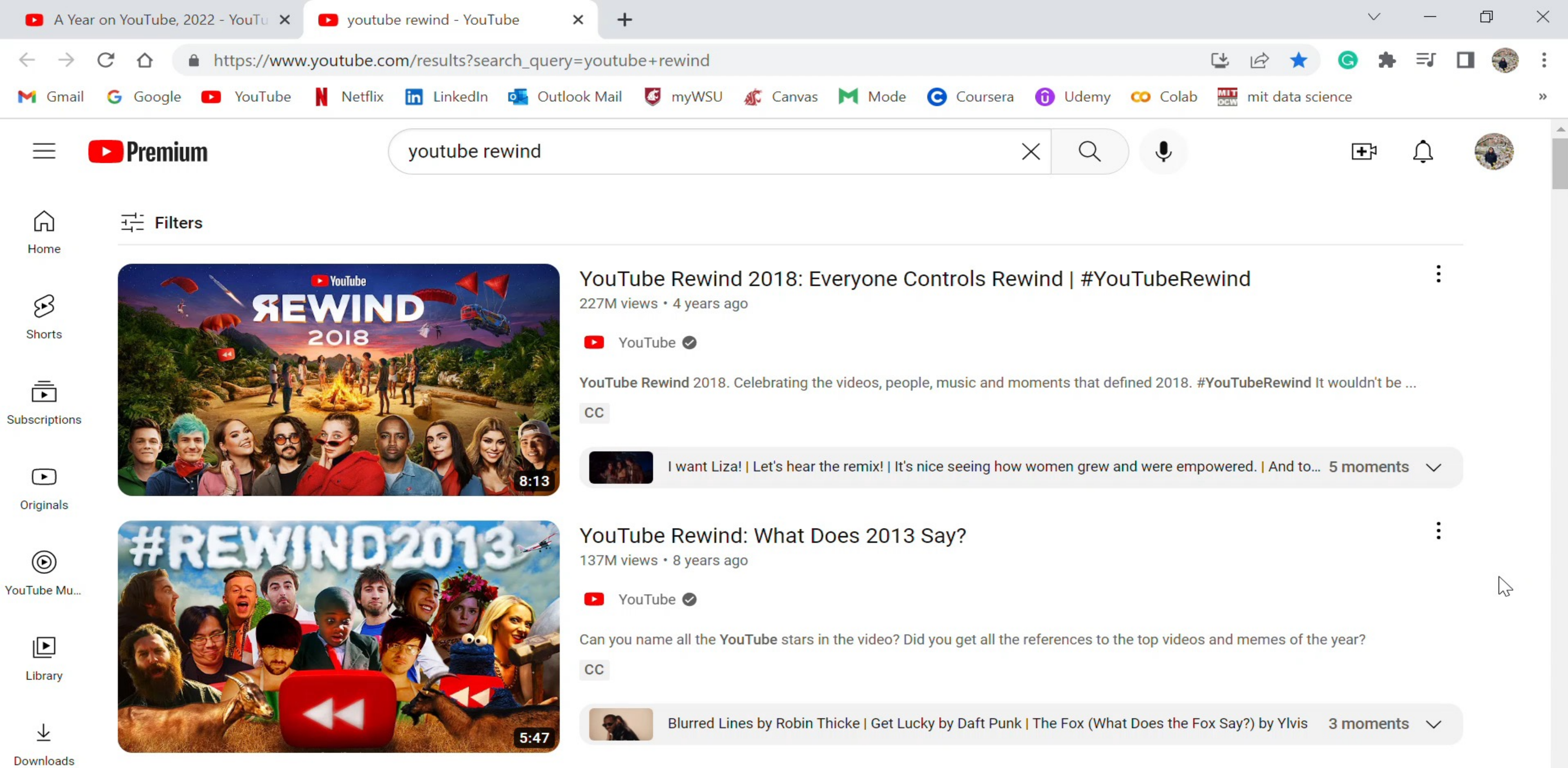
PRESENTED BY-
VISHNU PRIYA
PINAKI PRASAD
PRIYANKA GHOSH DASTIDAR
JENNIFER THOMSON

Back Story



Have you ever heard of **'YOUTUBE ЯEWIND?'** It was an annual video series created and produced by Youtube from 2010-2018

Youtube Rewind



YouTube Rewind 2018: Everyone Controls Rewind | #YouTubeRewind

227M views • 4 years ago

YouTube

YouTube Rewind 2018. Celebrating the videos, people, music and moments that defined 2018. #**YouTubeRewind** It wouldn't be ...

CC

I want Liza! | Let's hear the remix! | It's nice seeing how women grew and were empowered. | And to... 5 moments

YouTube Rewind: What Does 2013 Say?

137M views • 8 years ago

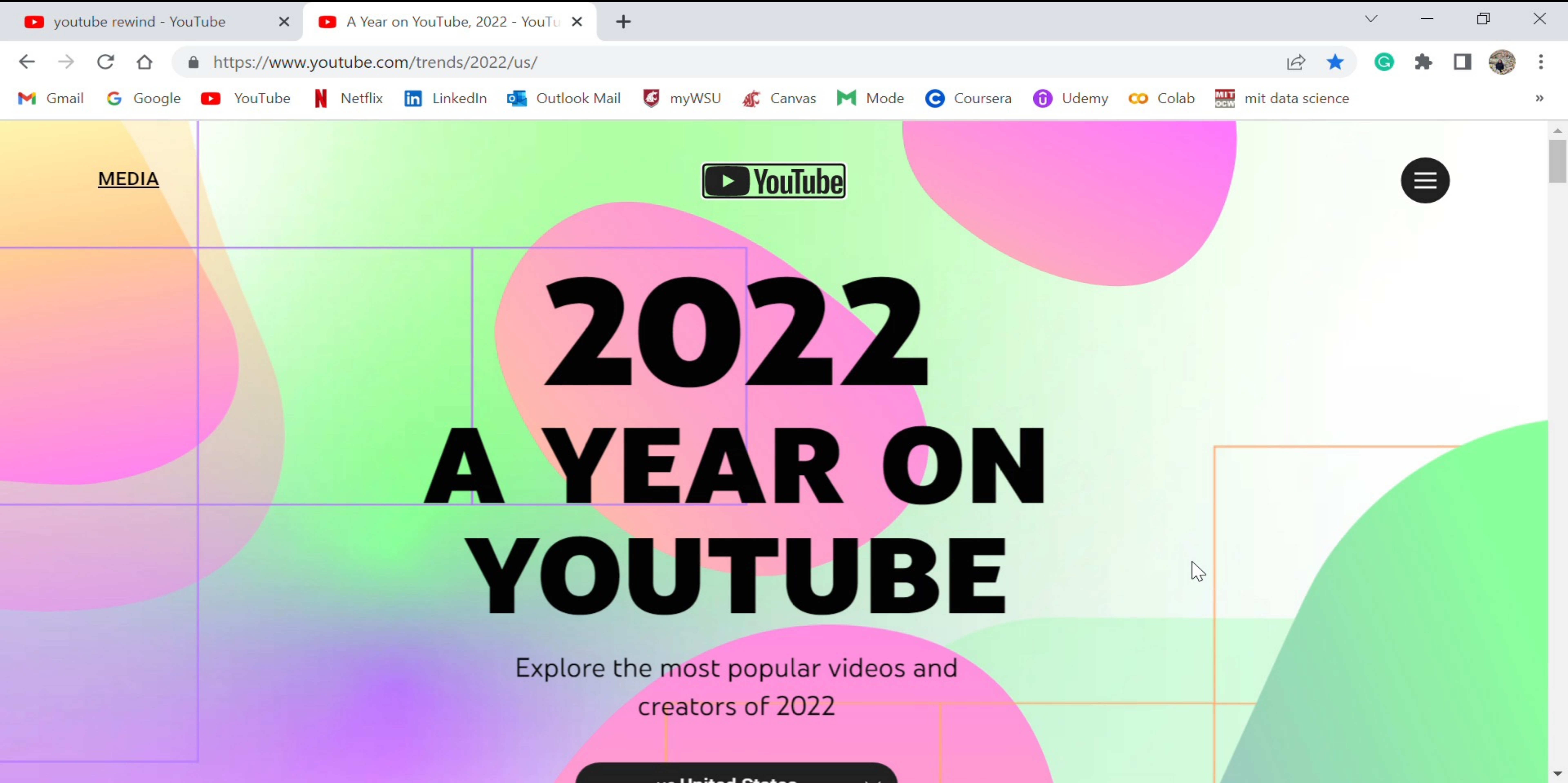
YouTube

Can you name all the **YouTube** stars in the video? Did you get all the references to the top videos and memes of the year?

CC

Blurred Lines by Robin Thicke | Get Lucky by Daft Punk | The Fox (What Does the Fox Say?) by Ylvis 3 moments

Youtube Trends 2022



2022

A YEAR ON YOUTUBE

Explore the most popular videos and creators of 2022

us United States

Points To Discuss:

- Top Categories
- Top Rated Videos
- Range queries
- Visualizations
- Network Aggregation
- Page rank

Project- Flow

STEP 1

Data Collection

STEP 2

Data Pre processing

STEP 3

Search Algorithms

STEP 4

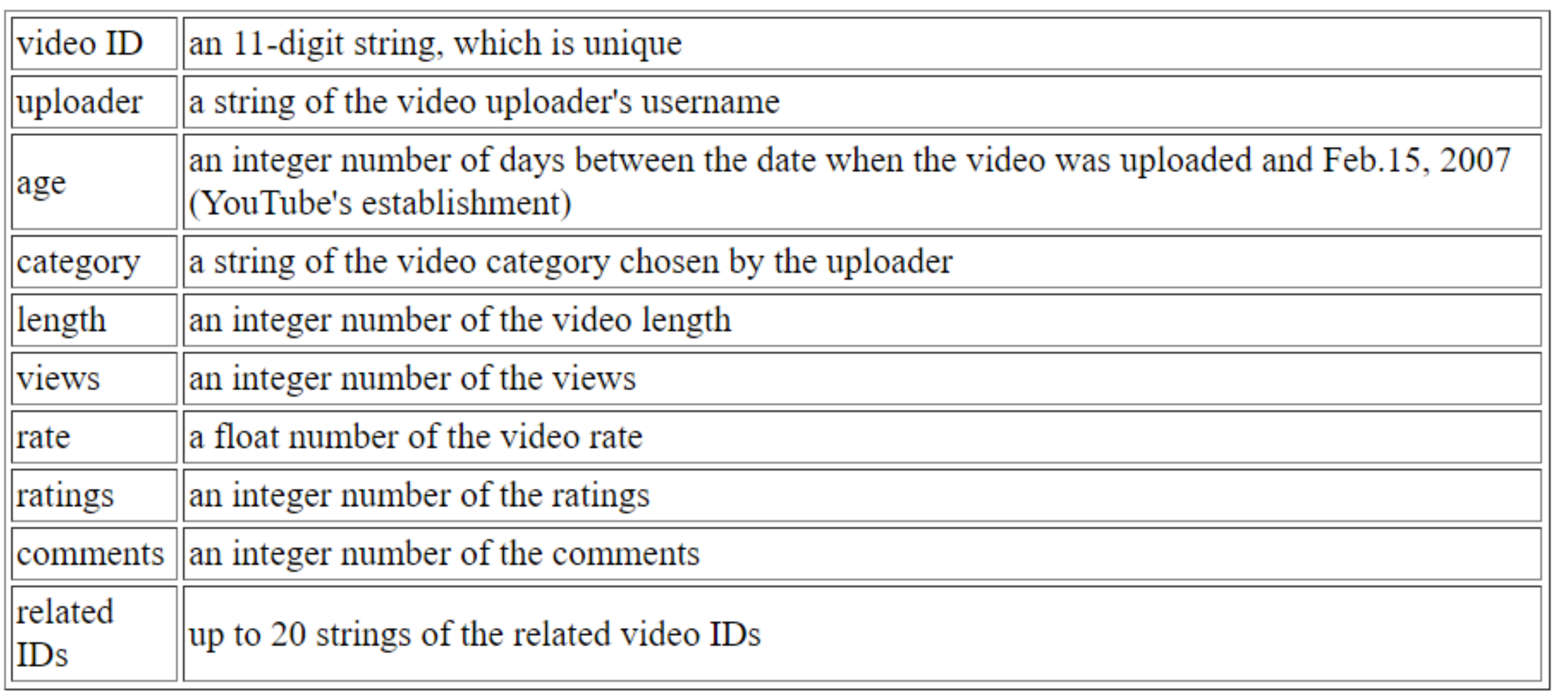
Visualizations

STEP 5

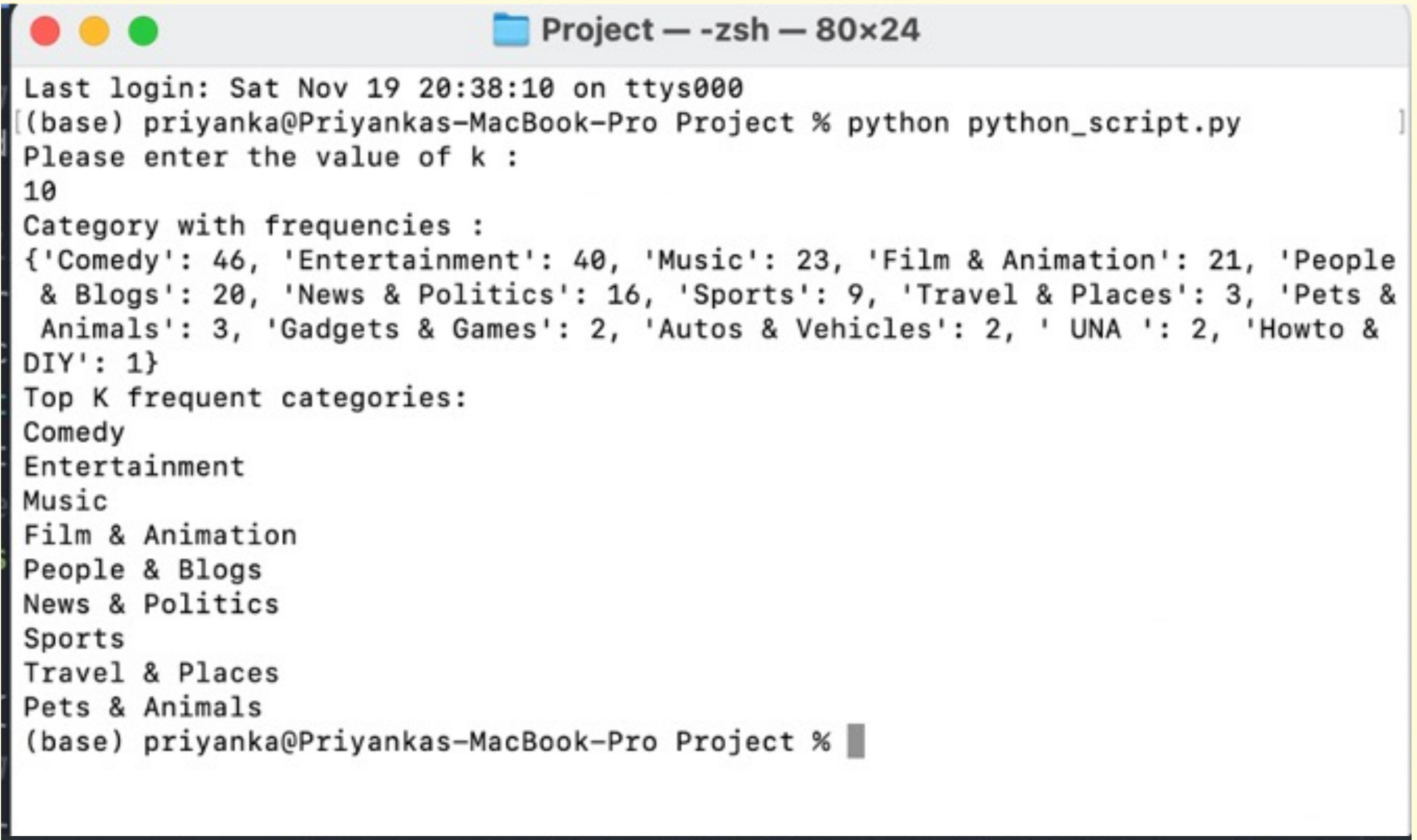
Network Aggregation& Pagerank

THE DATA

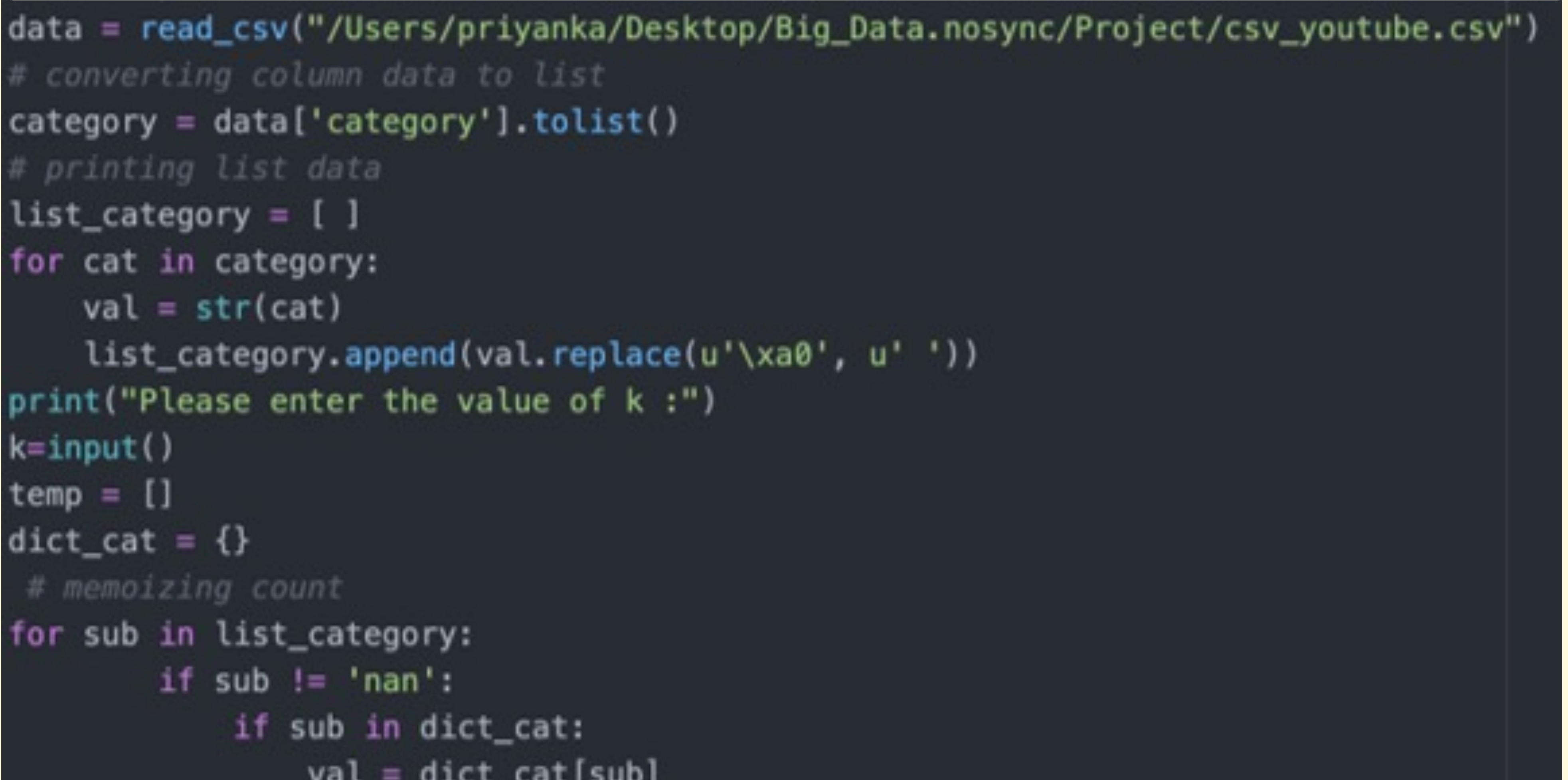
The Data is taken from https://netsg.cs.sfu.ca/youtubedata/



video ID	an 11-digit string, which is unique
uploader	a string of the video uploader's username
age	an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
category	a string of the video category chosen by the uploader
length	an integer number of the video length
views	an integer number of the views
rate	a float number of the video rate
ratings	an integer number of the ratings
comments	an integer number of the comments
related IDs	up to 20 strings of the related video IDs



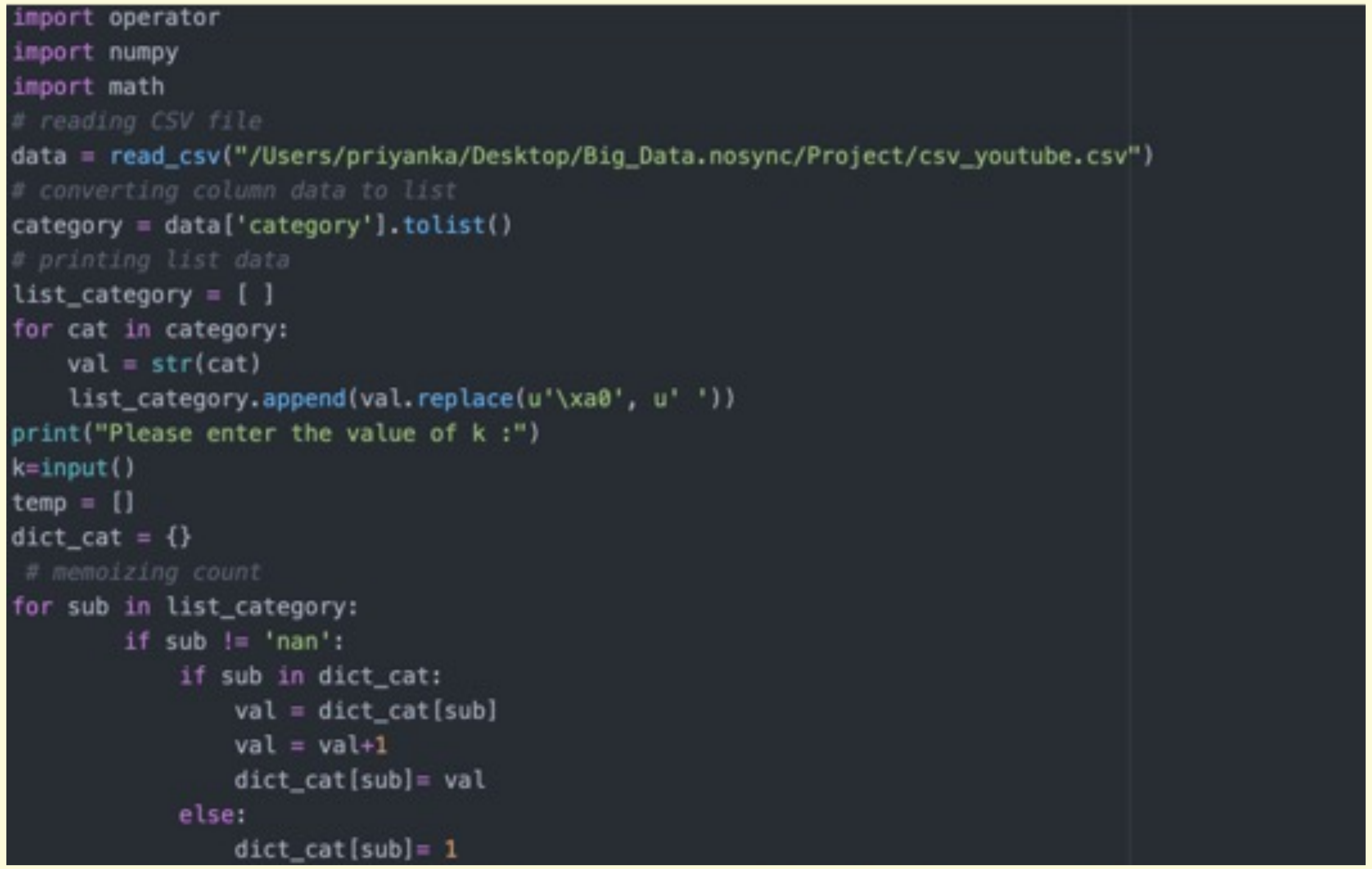
```
Last login: Sat Nov 19 20:38:10 on ttys000
(base) priyanka@Priyankas-MacBook-Pro Project % python python_script.py
Please enter the value of k :
10
Category with frequencies :
{'Comedy': 46, 'Entertainment': 40, 'Music': 23, 'Film & Animation': 21, 'People
 & Blogs': 20, 'News & Politics': 16, 'Sports': 9, 'Travel & Places': 3, 'Pets &
 Animals': 3, 'Gadgets & Games': 2, 'Autos & Vehicles': 2, ' UNA ': 2, 'Howto &
DIY': 1}
Top K frequent categories:
Comedy
Entertainment
Music
Film & Animation
People & Blogs
News & Politics
Sports
Travel & Places
Pets & Animals
(base) priyanka@Priyankas-MacBook-Pro Project %
```



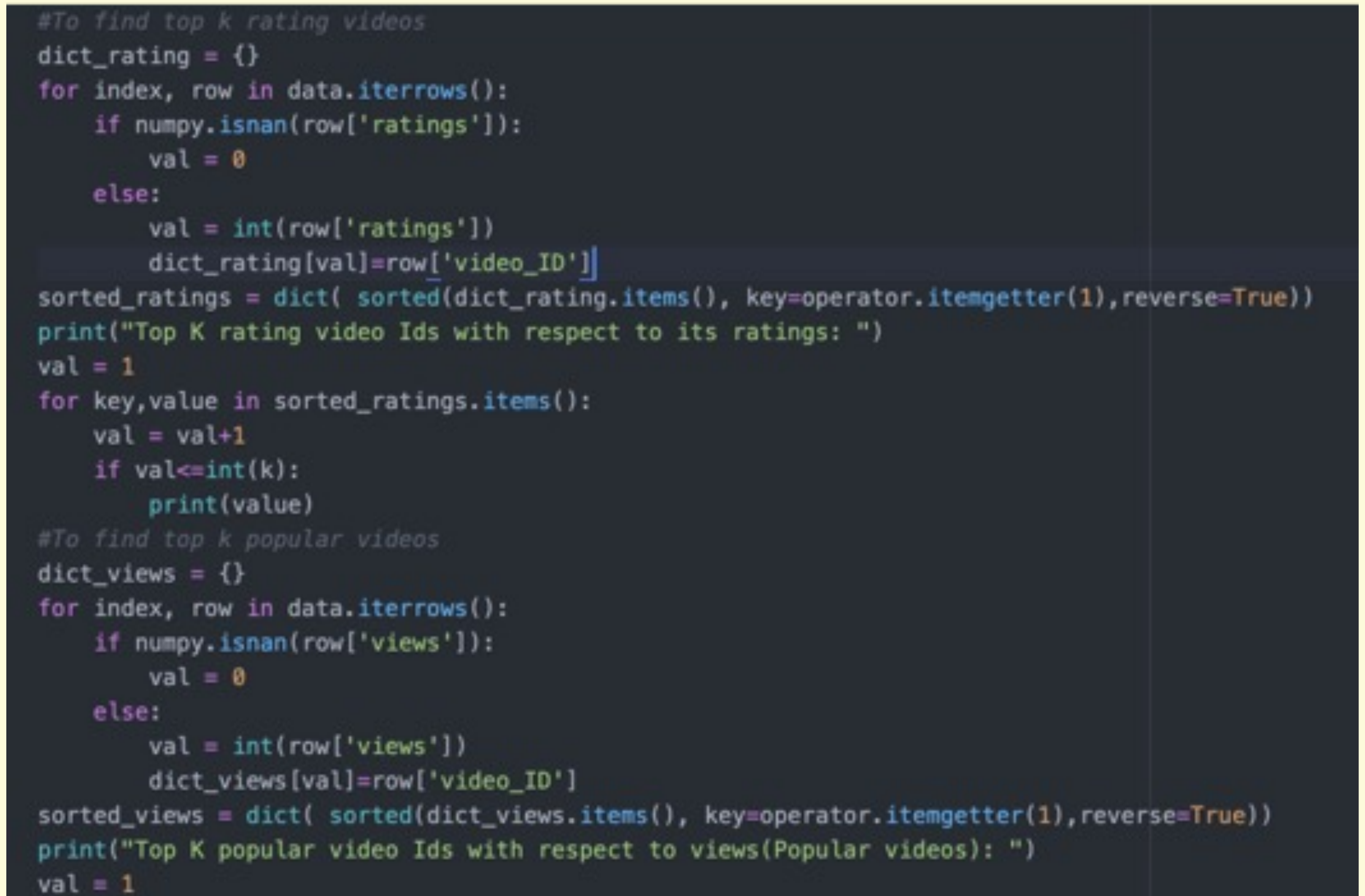
```
data = read_csv("/Users/priyanka/Desktop/Big_Data.nosync/Project/csv_youtube.csv")
# converting column data to list
category = data['category'].tolist()
# printing list data
list_category = [ ]
for cat in category:
    val = str(cat)
    list_category.append(val.replace(u'\xa0', u' '))
print("Please enter the value of k :")
k=input()
temp = []
dict_cat = {}
 # memoizing count
for sub in list_category:
        if sub != 'nan':
            if sub in dict_cat:
                val = dict_cat[sub]
```

Find Top k categories in which most number of videos are uploaded

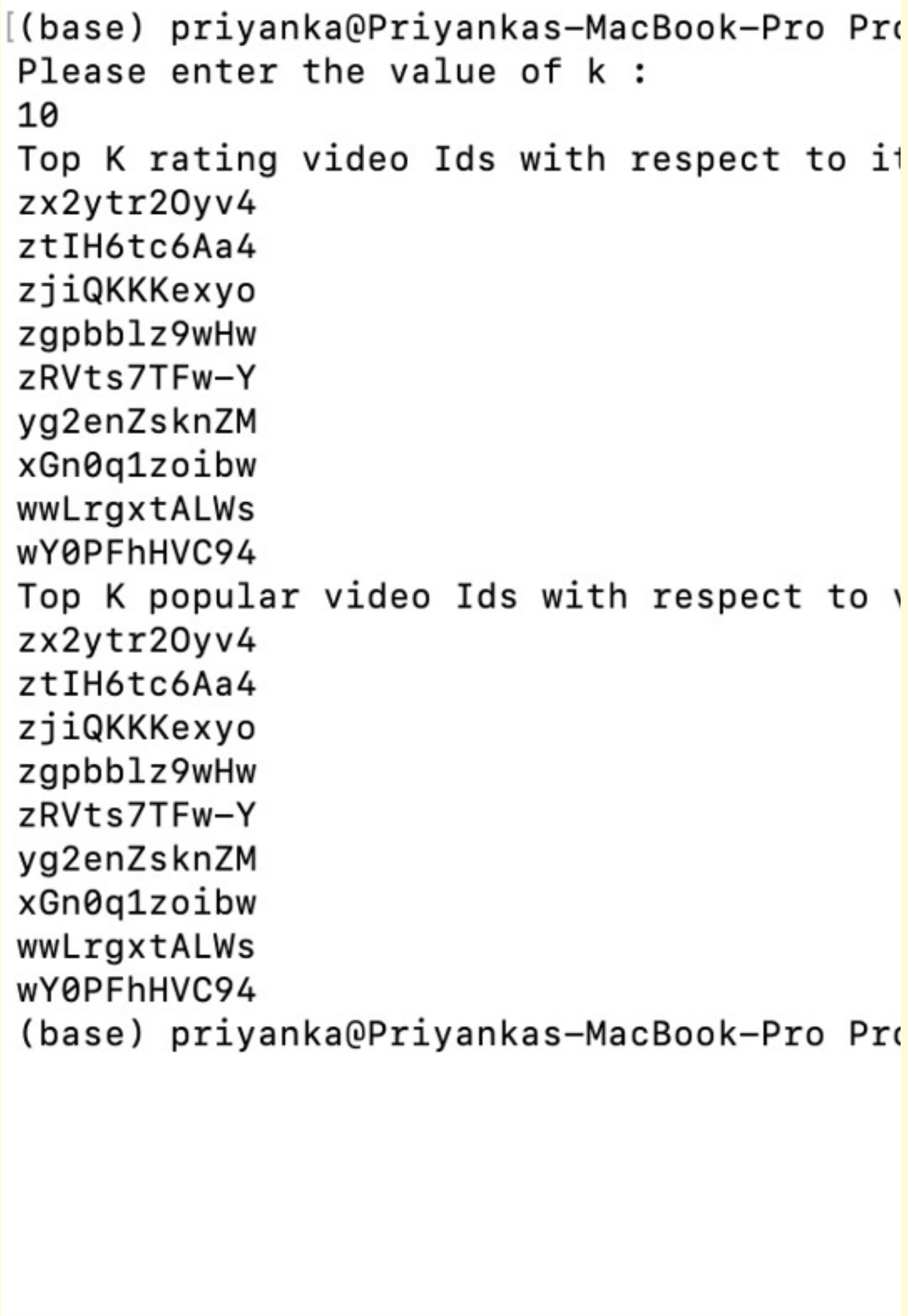
Find Top k rated videos; top k most popular videos



```
import operator
import numpy
import math
# reading CSV file
data = read_csv("/Users/priyanka/Desktop/Big_Data.nosync/Project/csv_youtube.csv")
# converting column data to list
category = data['category'].tolist()
# printing list data
list_category = [ ]
for cat in category:
    val = str(cat)
    list_category.append(val.replace(u'\xa0', u' '))
print("Please enter the value of k :")
k=input()
temp = []
dict_cat = {}
 # memoizing count
for sub in list_category:
        if sub != 'nan':
            if sub in dict_cat:
                val = dict_cat[sub]
                val = val+1
                dict_cat[sub]= val
            else:
                dict_cat[sub]= 1
```

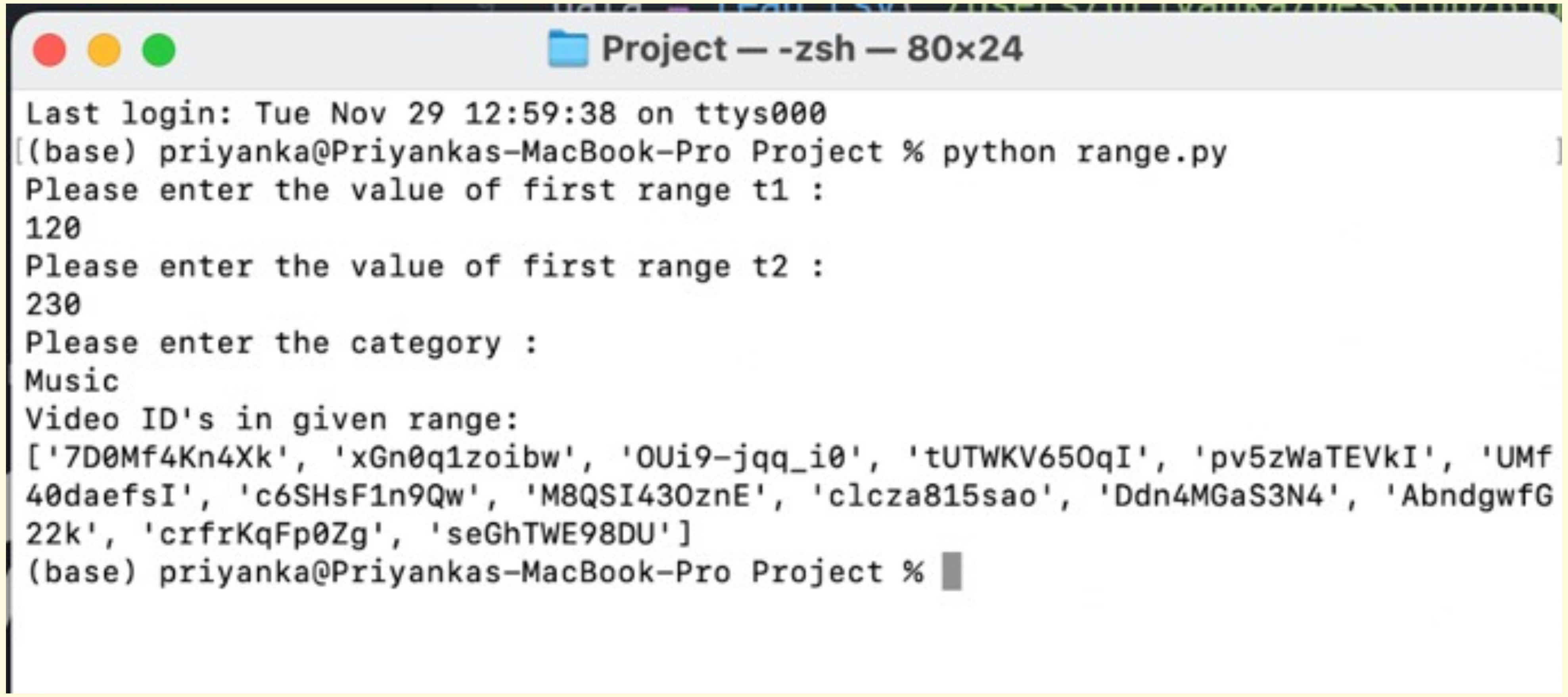


```
#To find top k rating videos
dict_rating = {}
for index, row in data.iterrows():
    if numpy.isnan(row['ratings']):
        val = 0
    else:
        val = int(row['ratings'])
        dict_rating[val]=row['video_ID']
sorted_ratings = dict( sorted(dict_rating.items(), key=operator.itemgetter(1),reverse=True))
print("Top K rating video Ids with respect to its ratings: ")
val = 1
for key,value in sorted_ratings.items():
    val = val+1
    if val<=int(k):
        print(value)
#To find top k popular videos
dict_views = {}
for index, row in data.iterrows():
    if numpy.isnan(row['views']):
        val = 0
    else:
        val = int(row['views'])
        dict_views[val]=row['video_ID']
sorted_views = dict( sorted(dict_views.items(), key=operator.itemgetter(1),reverse=True))
print("Top K popular video Ids with respect to views(Popular videos): ")
val = 1
```

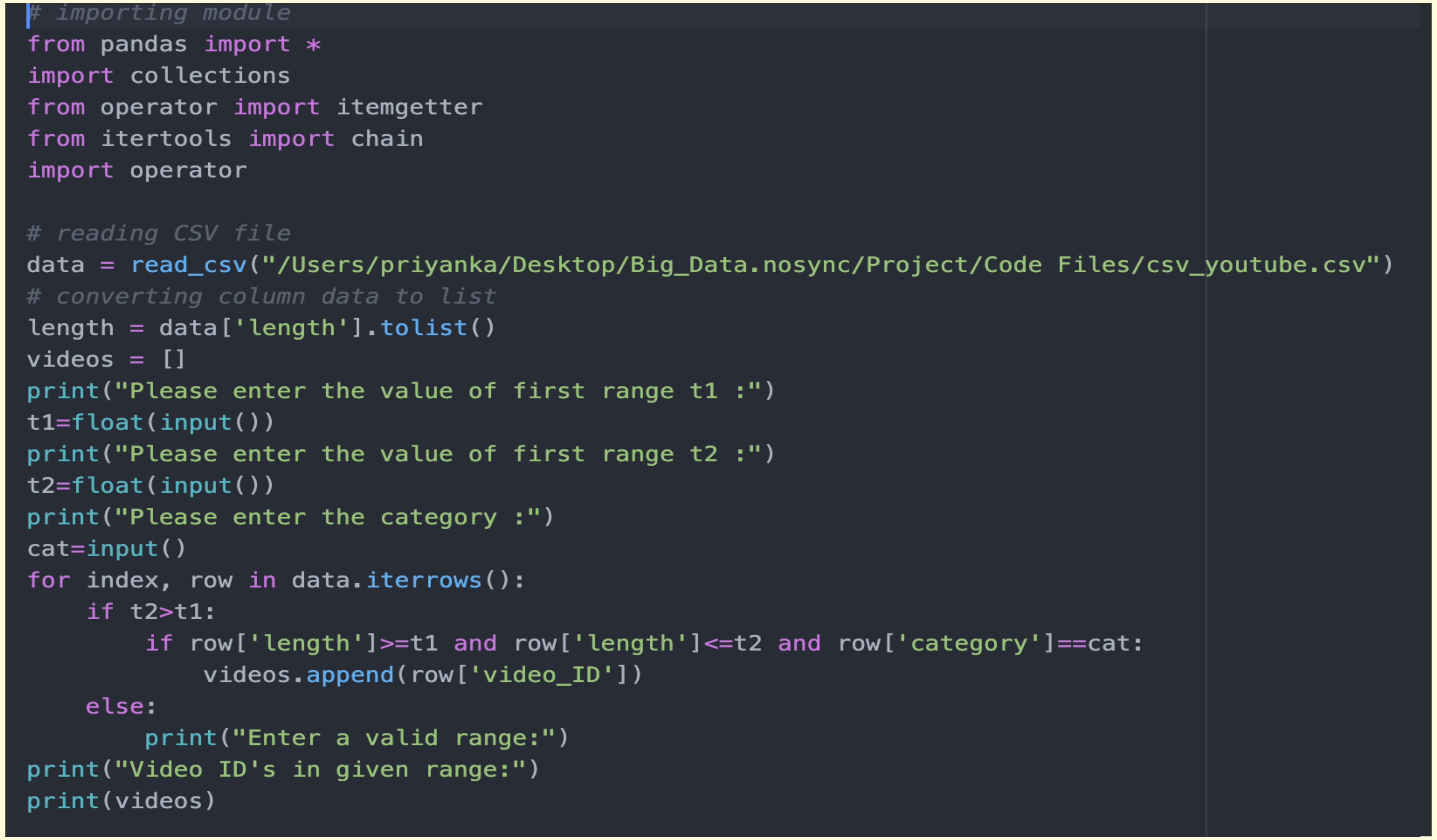


```
(base) priyanka@Priyankas-MacBook-Pro Pr
Please enter the value of k :
10
Top K rating video Ids with respect to i
zx2ytr2Oyv4
ztIH6tc6Aa4
zjiQKKexyo
zgpbblz9wHw
zRVts7TFw-Y
yg2enZsknZM
xGn0q1zoibw
wwLrgxtALWs
wY0PFhHVC94
Top K popular video Ids with respect to
zx2ytr2Oyv4
ztIH6tc6Aa4
zjiQKKexyo
zgpbblz9wHw
zRVts7TFw-Y
yg2enZsknZM
xGn0q1zoibw
wwLrgxtALWs
wY0PFhHVC94
(base) priyanka@Priyankas-MacBook-Pro Pr
```

Range queries: Find all videos in categories X with duration within a range [t1, t2]



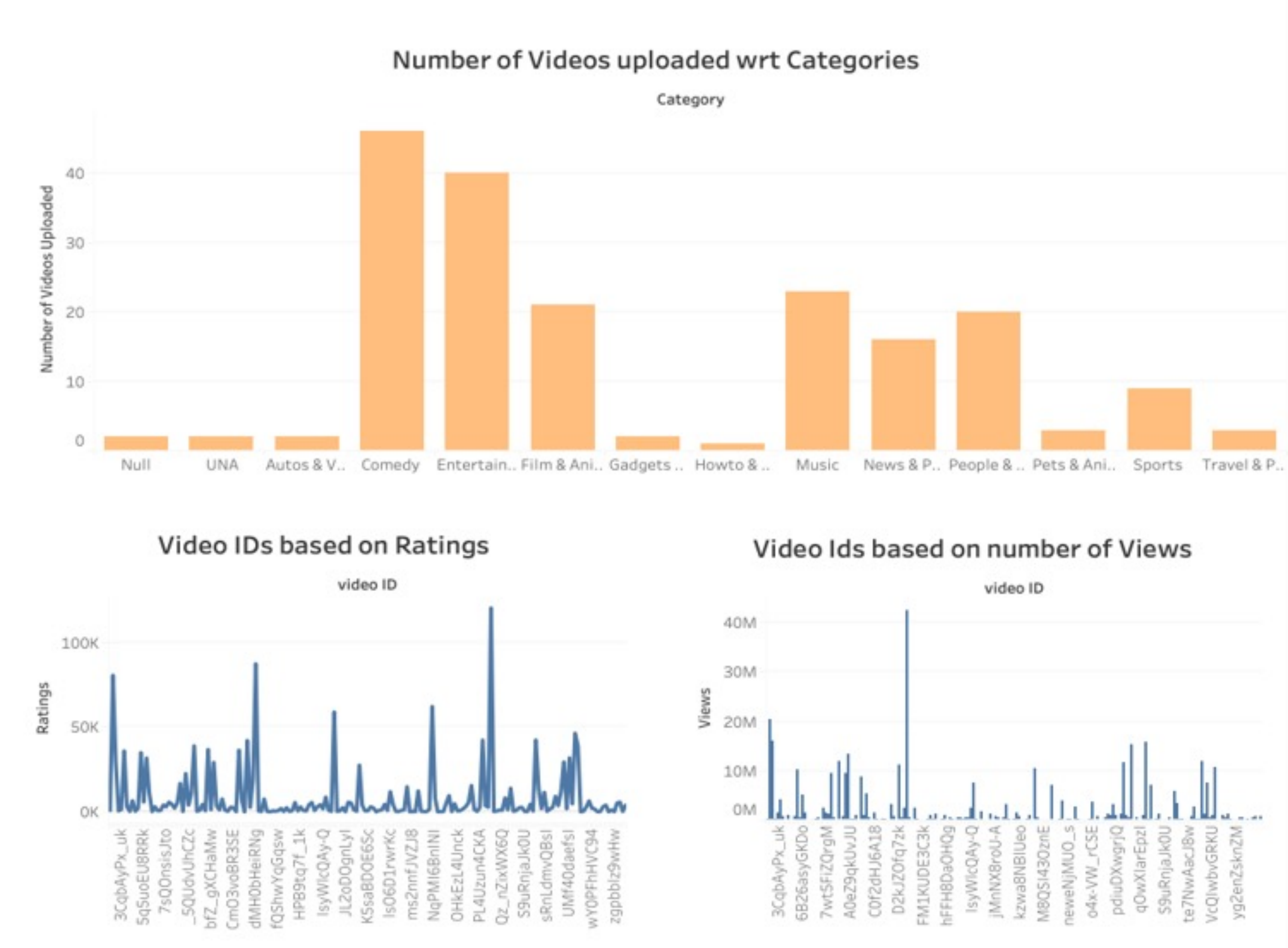
```
Last login: Tue Nov 29 12:59:38 on ttys000
(base) priyanka@Priyankas-MacBook-Pro Project % python range.py
Please enter the value of first range t1 :
120
Please enter the value of first range t2 :
230
Please enter the category :
Music
Video ID's in given range:
['7D0Mf4Kn4Xk', 'xGn0q1zoibw', 'OUi9-jqq_i0', 'tUTWKV65OqI', 'pv5zWaTEVkI', 'UMf
40daefsI', 'c6SHsF1n9Qw', 'M8QSI43OznE', 'clcza815sao', 'Ddn4MGaS3N4', 'AbndgwfG
22k', 'crfrKqFp0Zg', 'seGhTWE98DU']
(base) priyanka@Priyankas-MacBook-Pro Project %
```



```python
# importing module
from pandas import *
import collections
from operator import itemgetter
from itertools import chain
import operator

# reading CSV file
data = read_csv("/Users/priyanka/Desktop/Big_Data.nosync/Project/Code Files/csv_youtube.csv")
# converting column data to list
length = data['length'].tolist()
videos = []
print("Please enter the value of first range t1 :")
t1=float(input())
print("Please enter the value of first range t2 :")
t2=float(input())
print("Please enter the category :")
cat=input()
for index, row in data.iterrows():
    if t2>t1:
        if row['length']>=t1 and row['length']<=t2 and row['category']==cat:
            videos.append(row['video_ID'])
    else:
        print("Enter a valid range:")
print("Video ID's in given range:")
print(videos)
```

Visualizations

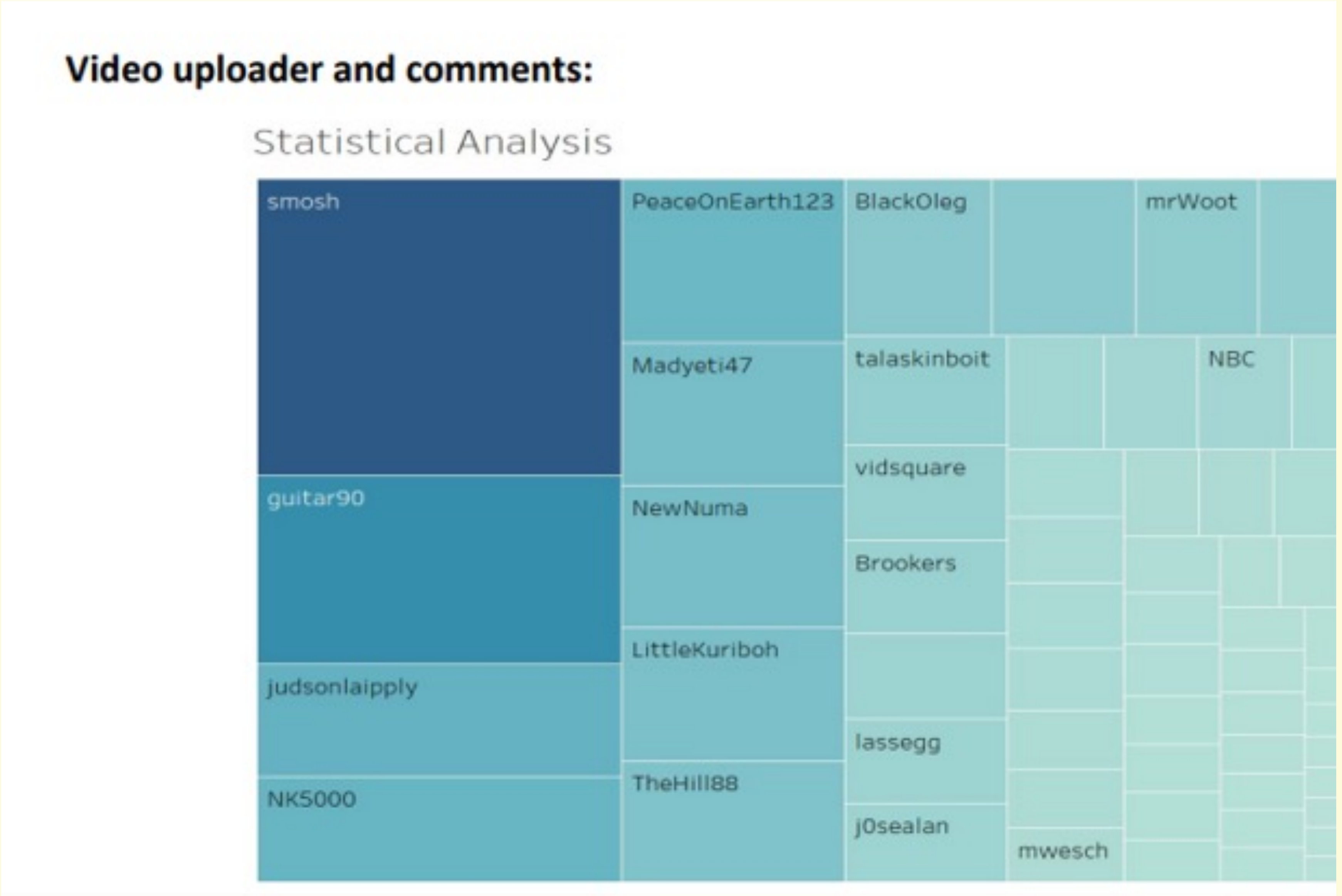


Number of Videos uploaded wrt Categories

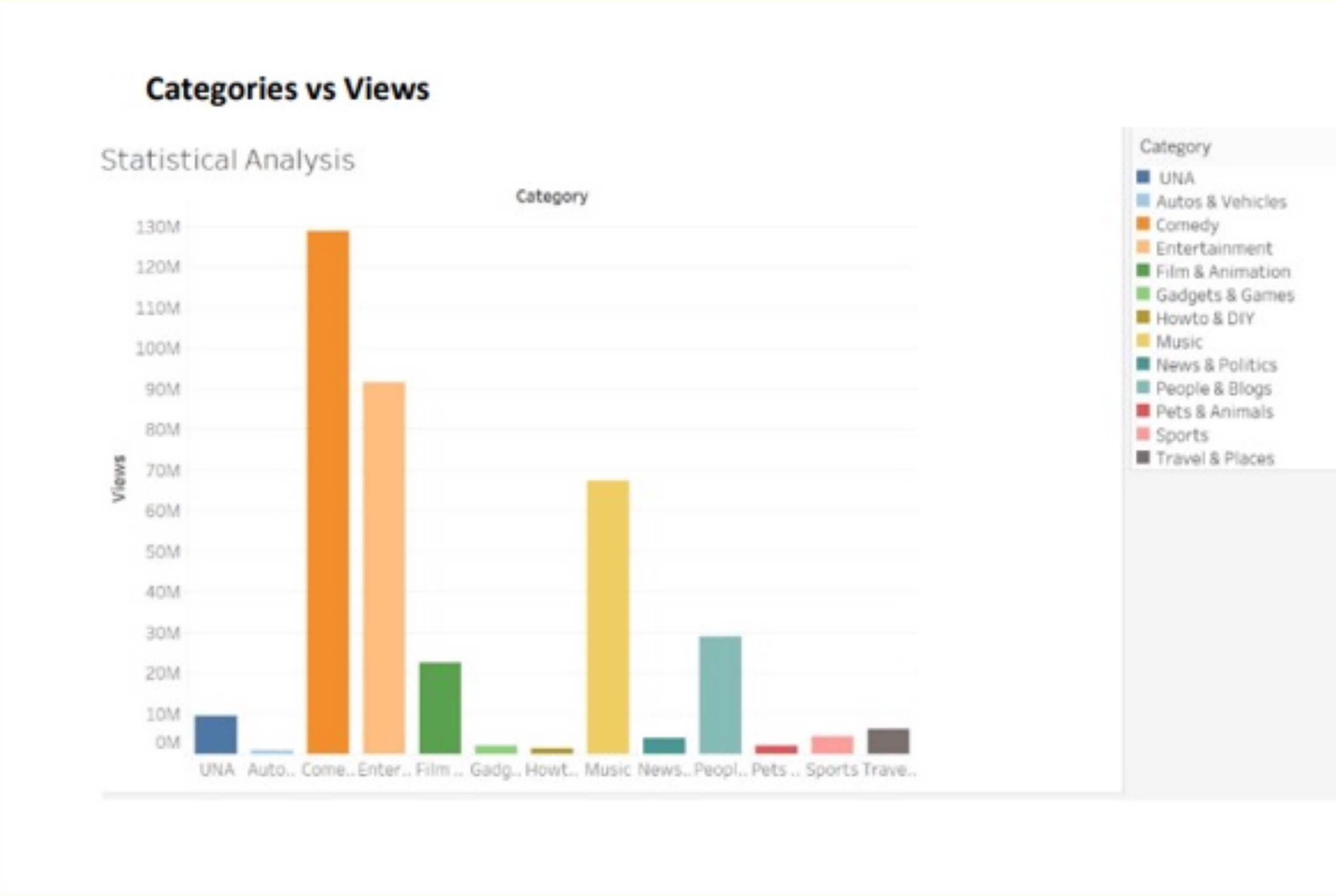
Video IDs based on Ratings

Video Ids based on number of Views

Visualizations

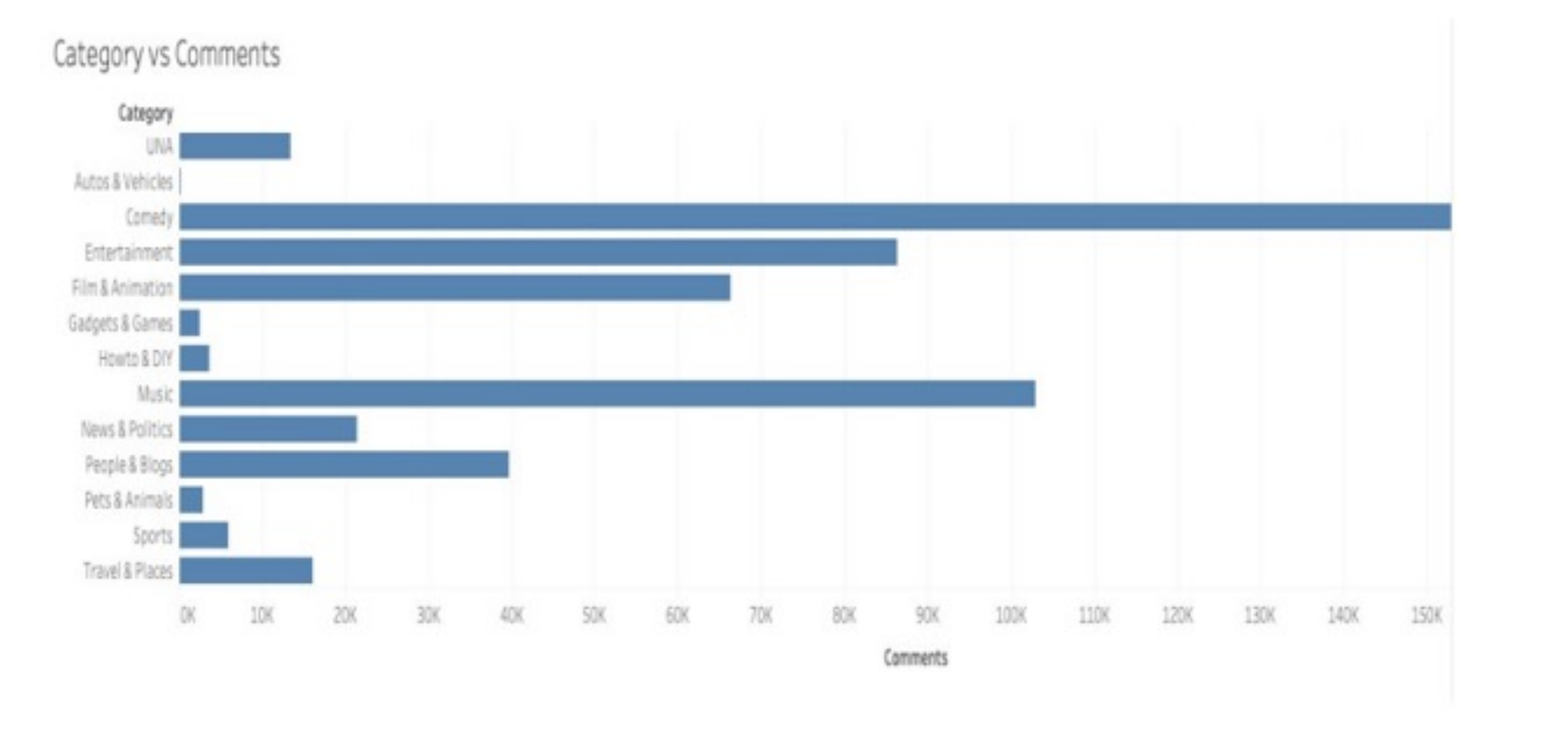


Video uploader and comments:

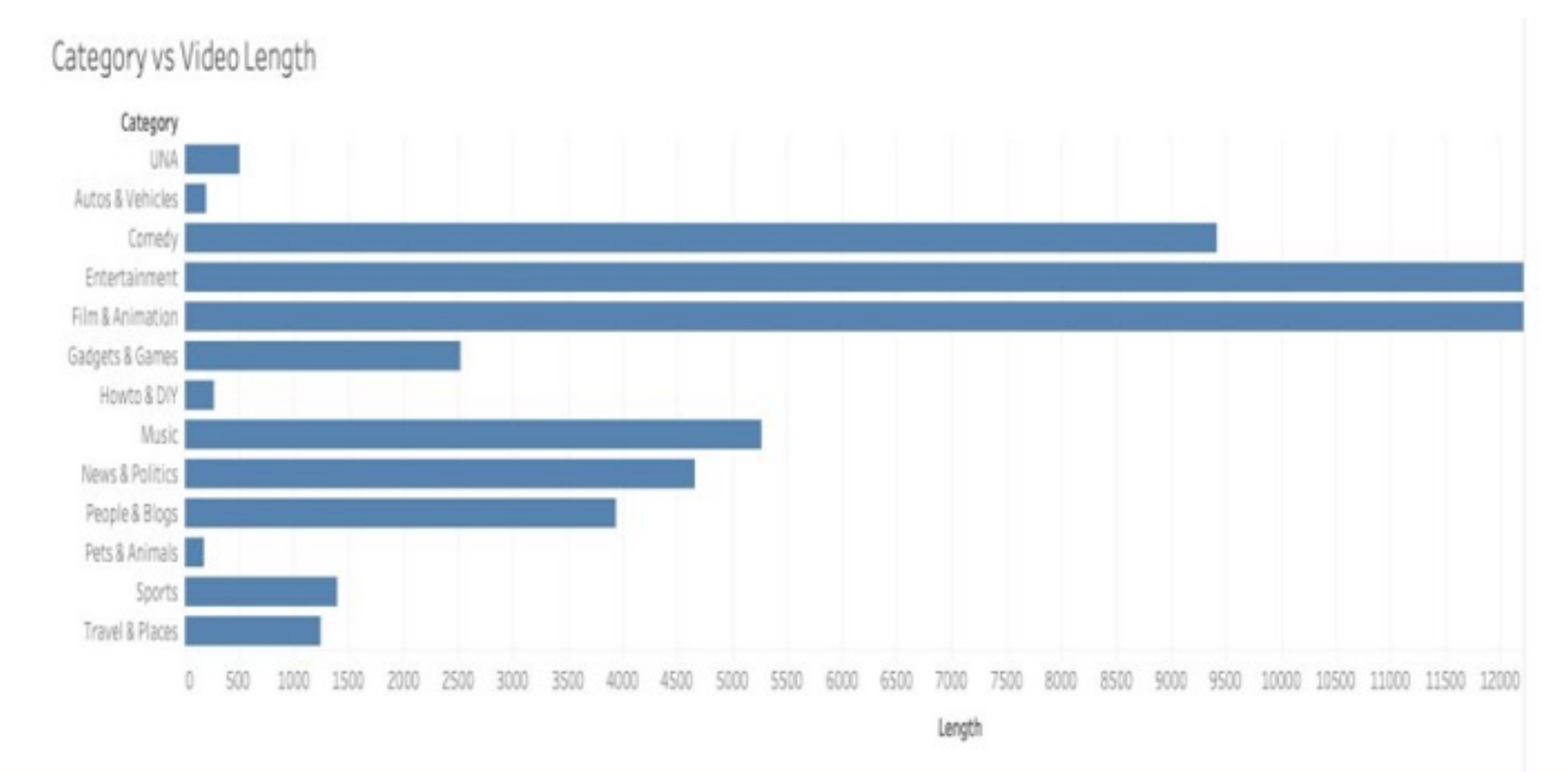


Categories vs Views

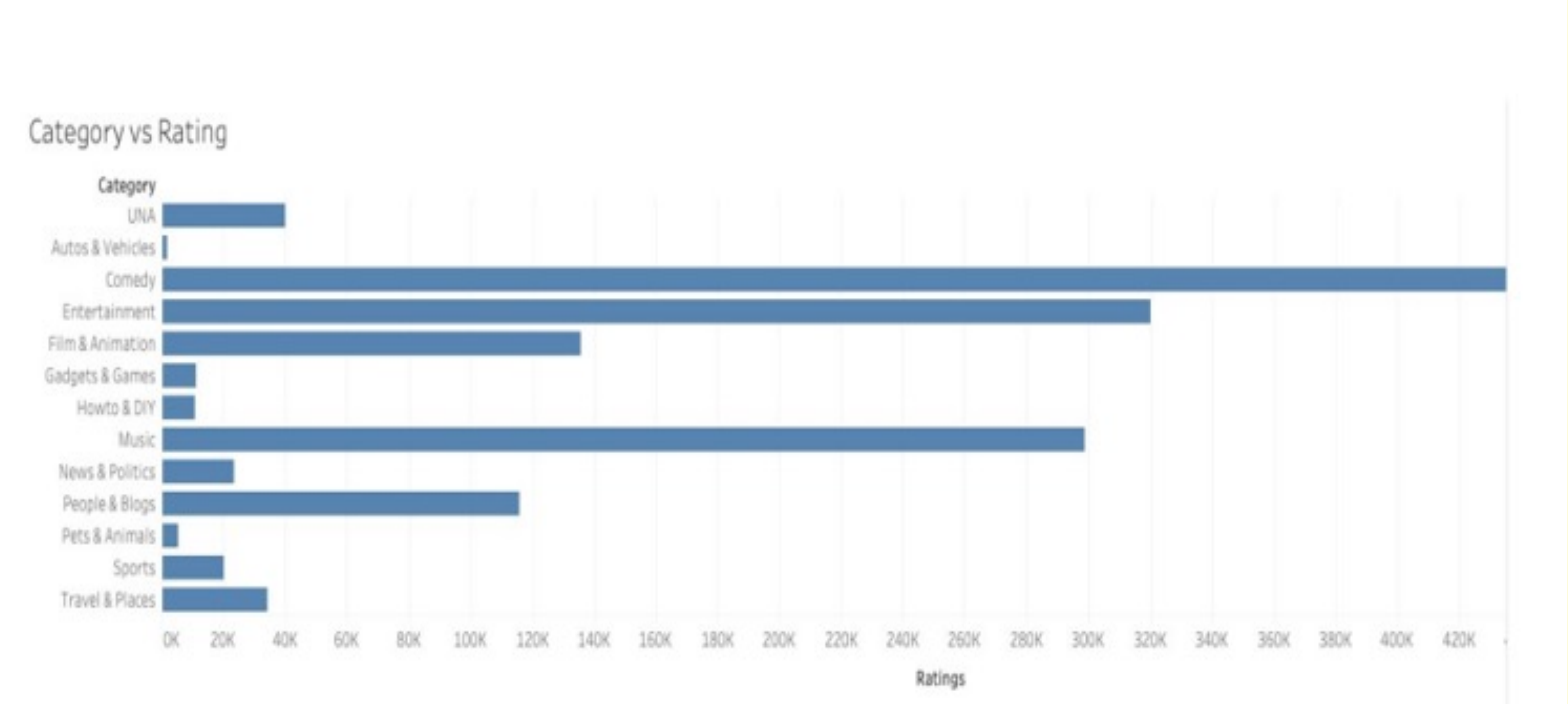
Visualizations



Visualizations



Visualizations



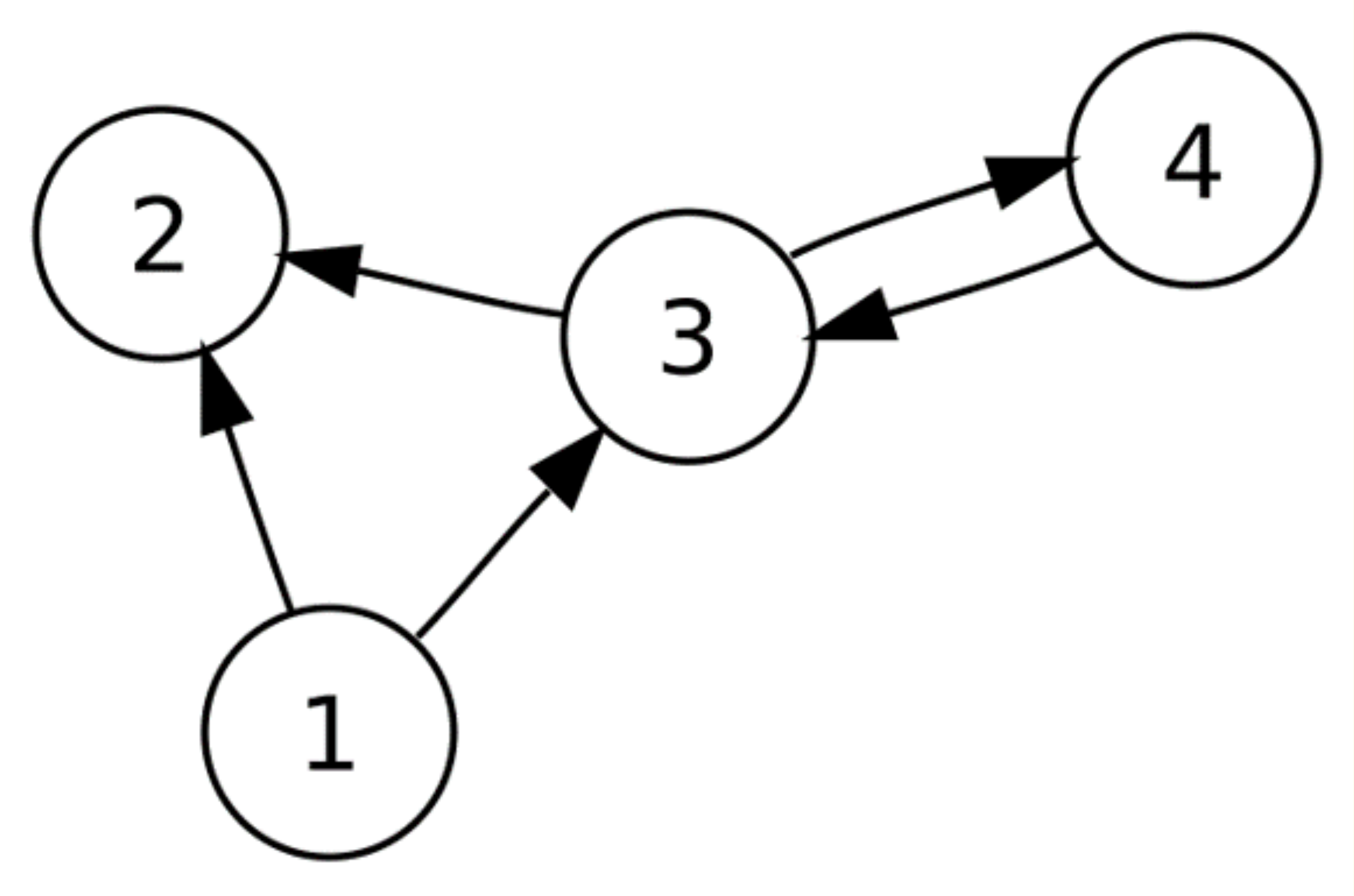
Why Spark?

- Fast
- Developer friendly
- Multiple workloads

Few Terms to introduce :

- Degrees (indegrees and outdegrees)
- PageRank Algorithm

Graph Nodes and Degrees



Node 1:

Degree = 2

Indegree = 0

Outdegree =

2

Node 2:

Degree = 2

Indegree = 2

Outdegree = 0

Node 3:

Degree = 4

Indegree = 2

Outdegree =

2

Node 4:

Degree = 2

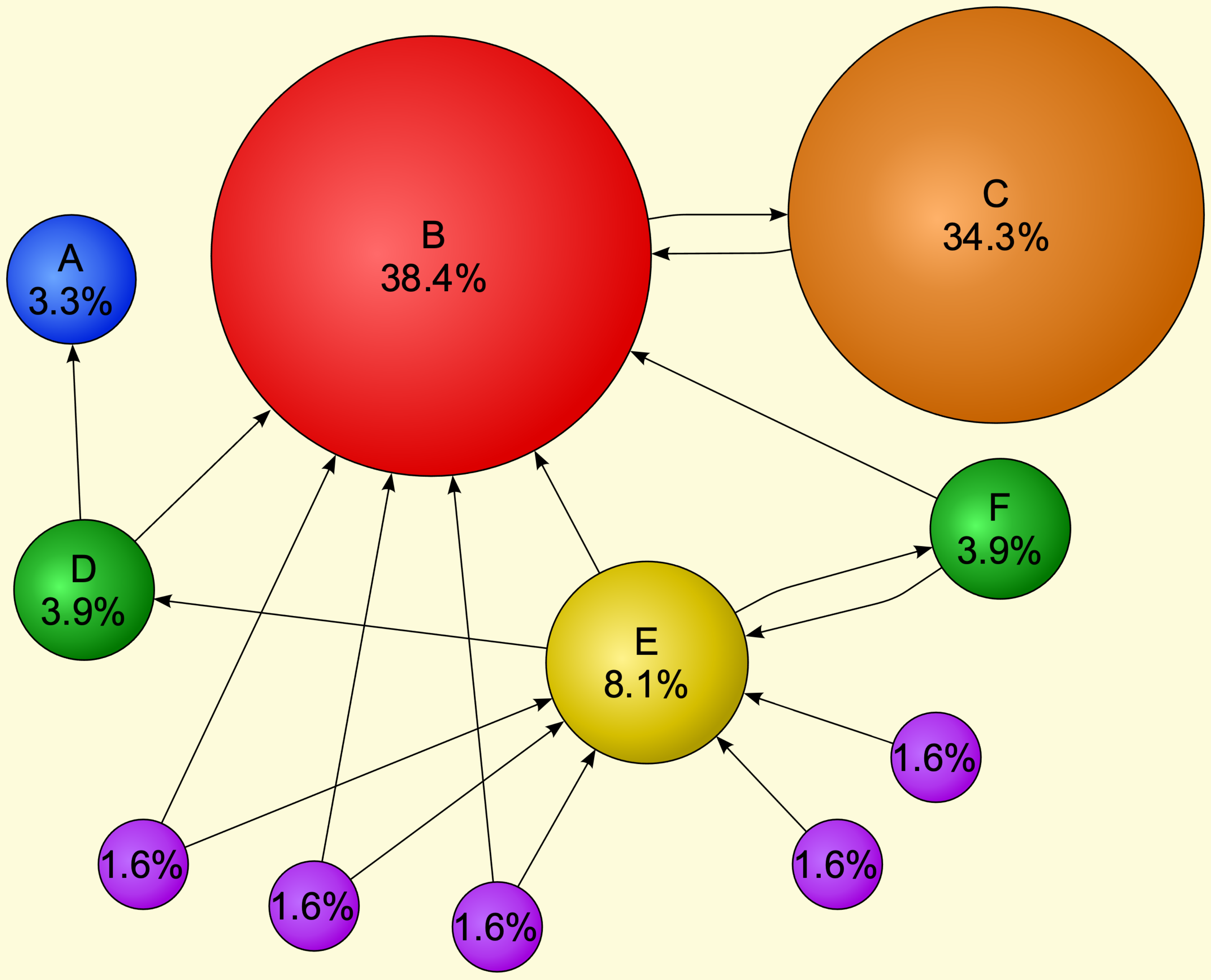
Indegree = 1

Outdegree = 1

Page Rank Algorithm

1. Originally developed by Google Search to rank web pages in their search engine results.

2. Measures importance of a node based on quantity as well as quality of other nodes pointing towards it.



DEMO

THANK YOU