YT project

March 2, 2024

1 Introduction

What do I watch on youtube? How has this content mix changed over time?

At what times of the day / year do I tend to watch youtube most?

2 Part 1: Data Cleaning and Category Analysis

I chose to download the watch history from google takeout as json rather than html so that we don't have to scrape it. In a different notebook, I do the html scraping method using beautiful soup.

Here we are preparing the dataset, which should include the video title, url, and timestamp.

The main task in this section is figuring out which distinct categories the videos in the dataframe belong to. Here I attempt unsupervised learning methods, as well as supervised regression methods.

[140]:
```
import json
import pandas as pd
with open('../Data Project/watch-history.json', 'r') as file:
    data = json.load(file)

#convert to pandas df
watch_df = pd.json_normalize(data)
```

Data Cleaning

[142]:
```
#removing ads, and unnecessary columns
watch_df = watch_df[watch_df['details'].isna()] #entries only have a␣
 ↪"description" and "details" if they were ad videos
watch_df = watch_df.drop(['header','description', 'details', 'products',␣
 ↪'activityControls'], axis = 1)
watch_df.rename(columns = {'titleUrl':'URL', 'subtitles':'Channel'}, inplace =␣
 ↪True)

#gets rid of the "watched" part of the string in title value
for idx in range(0, len(watch_df)):
    watch_df.iloc[idx]['title'] = watch_df.iloc[idx]['title'][8:]
```

```
#convert messy timestamp label into dates and time of day
from datetime import datetime
watch_df['time'] = pd.to_datetime(watch_df['time'])
#have 1 column for time of day, another for date
watch_df['time_24hr'] = watch_df['time'].dt.strftime('%H:%M:%S')
watch_df['date'] = watch_df['time'].dt.date
watch_df = watch_df.drop(['time'], axis = 1)
```

3 UNSUPERVISED LEARNING: K-means Clustering to try to separate data into different video categories

The issue is, at this point we can't really analyze the data since we don't have video category labels. API calls to google cloud would work but is a costly option. Let's attempt to use clustering to practice our skills and see if the method would achieve the result we want.

[4]:
```
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy
import matplotlib.pyplot as plt
import plotnine as p9
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.cluster import KMeans
import random
import re
import string
```

[147]:
```
#Download stopwords from NLTK
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

def processing(content):
    punctuation_pattern = re.escape(string.punctuation)  # Escapes all␣
 ↪punctuation characters
    emoji_pattern = "[\U0001F600-\U0001F64F]"  # Basic pattern to match some␣
 ↪common emojis
    combined_pattern = f'[{punctuation_pattern}{emoji_pattern}]'
    clean_content = re.sub(combined_pattern, '', content)

    #Tokenize and convert to lower case
    tokens = word_tokenize(clean_content.lower())
```

```
    #Remove stopwords
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [word for word in tokens if word not in stop_words]

    #Lemmatization
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in␣
 ↪filtered_tokens]

    #Important words only- certain length
    important_tokens = [token for token in lemmatized_tokens if not (len(token)␣
 ↪<= 2 or len(token) >= 21)]

    return ' '.join(important_tokens)
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/marcuschandra/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/marcuschandra/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/marcuschandra/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[148]:
```
#Dataset
titles = watch_df["title"].tolist()

#Processing texts
processed_texts = [processing(title) for title in titles]

#TFIDF vectorization
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(processed_texts)
```

[149]:
```
#Inspection to verify that processing works
processed_texts
```

[149]:
```
['run python file using another python file using subprocess easy',
 'push change visual studio code github tech project',
 'using git visual studio code official beginner tutorial',
 'extremely helpful guide merge conflict',
 'call one python file another pytthon file vice versa bash scripting using
subprocess',
 'unity call method/function another script',
 'exclusive class interval statistic math',
 'converting inclusive class interval exclusive math',
```

```
'marginal distribution strange support',
'statistic probability joint marginal distribution',
'easy minute chili oil noodle recipe',
'anonymously remaking street sign nyc artist design streetart',
'average redditor security airport cringe redditor',
'try laugh noise funny video lol credit kamera kastro link comment',
'ate hot dog …',
'know guess language stranger speaking lineup',
'asian buffet',
'double life modern family comedy central africa',
'tomorrowland movie scene',
'tomorrowland apocalypse speech',
'tomorrowland 2015 shocking scene',
'pork chive dumpling lunar new year lucky food',
'make pork dumpling asianfood recipe',
'people night eat 17-course omakase line bon appétit',
'modern family phil get revenge family ft. back-up luke',
'ran marathon heelys',
'power fortune cooky',
'believe',
'british footballer try chimaek korean fried chicken beer',
'costco wagyu pan sear vs. sou vide win wagyu steak costco',
'asian meal girl meal',
"expensive ingredient 've cooked",
"n't mess phil …",
'little brother cartwheeled said better short',
'really smart… movie fyp',
'german try american coke',
'twisting piece metal cow hoof hooftrimming hooftrimmer asmr farming hoofcare',
'lie job pick chick short',
'much spent destination wedding lake como italy short',
'applied 415 quant job learn mistake',
'influential economist minute',
'asian food ordering mindset',
'rare steak cheese sandwich   ',
'gordon ramsay show caring side hellskitchen gordonramsay',
'rum ball short',
'restaurant world insane meal',
"succession 3x03 government n't fcking pez dispenser",
'day life personal chef surprise birthday dinner',
'florida stanley forever always office',
'suit clip jenny save harvey name',
'thank fish familyguy short',
'chef tell difference steak tuna hellskitchen',
'japan think westerner eat',
'authentic lebanese shawarma brooklyn   ',
'ordering expensive food inexpensive restaurant short',
```

```
 'pronounce mailys pronouncenames.com',
 'open combination lock',
 'open umpire caught cheating coco gauff',
 'sharapova reacts djokovic priceless impression djokovic tennis tennisplayer',
 'super sinner fire',
 'shipping route one world dangerous wsj',
 'mila kunis meg family guy one makeover',
 'incredible point',
 'pronounce powhatan way uk/british us/american english pronunciation',
 'pronounce haudenosaunee correctly',
 'probability exponential distribution finding cumulative distribution
function',
 'uniform distribution explained example',
 'book habakkuk summary complete animated overview',
 'find find_all web scraping python',
 'scraping data real website web scraping python',
 'extract data nested html tag explained web scraping tutorial english',
 'html exclude unwanted tag beautifulsoup python',
 'uninstall homebrew 2022 macos within min catalina big sur',
 'beautifulsoup request web scraping python',
 'pyscript run python browser end javascript',
 'senior programmer junior developer short',
 'project add data portfolio',
 'web scraping pro par html python',
 'parse html content using beautifulsoup complete guide web scraping tutorial',
 'web scraping python part parsing html beautiful soup',
 'python project scrape country population data html table csv excel using
python',
 'learn easily export html table excel csv pdf code mark',
 'convert download html table csv file using javascript',
 'master sentiment analysis using python nltk uncover text insight data
science',
 'python project scrape youtube using youtube data api analyze visualize youtube
data',
 'create youtube analytics report free looker studio template',
 'build youtube analytics dashboard data studio template cyad ep1',
 'algorithmic trading using python full course',
 'get evaluate startup idea startup school',
 'succession matsson interview tom s04ep10 series finale',
 'succession greg salary revealed s04ep10 series finale',
 'modern family sal threatens lily',
 'went business school',
 'return scale cobb douglas production function',
 'isoquant isocost cost minimization',
 'project add data portfolio',
 'data analyst portfolio great example',
 'korean take leftover home everythingkimchi short',
```

```
 'half japanese japan',
 'thing china dont make sense',
 'biggest problem adoption wsj',
 'foreign policy second',
 'chance',
 'mark zuckerberg demand recognition harvard social network 2010',
 "look n't define them.2 short viral",
 'elevator scene person interest',
 'look pizzashop vega donutshop secretpiza vegasfood italian businesspartner
funny',
 'server math',
 'modern family phil basketballer',
 'vietnamese wingman',
 'japanese salaryman food hiroshima anyone else hooked okonomiyaki',
 'clever policeman',
 'bought house 20,000',
 'find absolute bargain january sale short',
 'sensitivity training workplace utopia short',
 'pov date cheap',
 'entrepreneur reject peter jones million offer| dragon den',
 'asian squid game ft. danielthrasher',
 'must protect nyc pizza slice cost',
 'fast food japan',
 "'escape velocity contribute idea black hole",
 'nūes grave dia los muertos oaxaca city méxico official music video',
 'arrival trailer 2016 paramount picture',
 'combination formula probability combinatorics probability statistic khan
academy',
 'permutation formula probability combinatorics probability statistic khan
academy',
 'algorithmic trading using python full course',
 'tozo nc7 tozo nc9 active noise cancelling anc wireless earbuds review must
see',
 'unboxing tozo nc9 earbuds full review',
 'recruiter catch candidate cheating interview |real frontend developer role
interview',
 'bcg pymetrics test everything need pas',
 'balloon game assessment pymetrics arctic shore game pwc morgan use hire',
 'learn 2024 could start',
 'citadel method secret sauce quantitative investing martin shkreli',
 'young people pro-palestine',
 'matthew mcconaughey pick girl second',
 'free violate law physic sean carroll',
 '22km… count running run gym relatable comedy funny viral',
 'beat tube next station',
 'part brucewayne gotham wayne batman series',
 'make korean sojubombs soju beer cocktail sojucocktail koreanfood koreatown',
```

```
 'getting drunk italy',
 'gang leader day sudhir venkatesh',
 'japan cheap japanese interview',
 'modern family haley virgin sts',
 'british actor american actor key peele',
 'future queen spain embarks 4-year military bootcamp',
 'romantic way meet partner freshoffttheboat short',
 'power couple make 2500/day dog daycare',
 'hasbro game taboo board game',
 'fetch apis python api fetching python',
 'dare ask separate check americanhigh restaurant restaurantlife server brunch',
 'day year japan heading work',
 'race understand corruption sport movie afrovibes africa hollywoodmovie',
 'haley meet boyfriend parent part modern family short',
 'gordon daughter megan sends back birthday meal hell kitchen',
 'week life study abroad barcelona',
 'truth uchicago fun die uchicago',
 'mr. bean disliked',
 'modern family luke geography teacher sts',
 'guess footballer trent alexander-arnold',
 'absolutely jaw-dropping rally',
 'switched accent end …',
 'louis reminds teacher amazonprimevideo harvey louislitt suit viral series
netflix',
 'getting kicked college lecture',
 'thickest boston accent ever',
 'modern family gloria phill kiss-cam final photography sts',
 'different… movie fyp',
 'airbnb guest leaf home trashed',
 'pov dating girl nyc dating nyc date',
 'beef gordon ramsay',
 'step aside haley alex time shine short modernfamily alexdunphy',
 'modern family haley andy get caught',
 'freakonomics steven levitt secret making tough choice',
 'made million attack israel',
 'take white friend black barbershop ',
 'rookie … therookie gutfeeling',
 'asian bargain',
 'chunkz genius',
 'math professor actual field medalist… oxbridge short',
 'phil caught flirting hospital modern family s01e11 comedy clip',
 'pickpocketed pickpocket',
 'catch keep',
 'warren buffett bill ackman asked warren question become billionaire investing
stock',
 'yardman science  \u200d ',
 'playing pickleball like tennis player pickleball',
```

```
 'sampras agassi epic point',
 'meet kind people stealing money kelly richmond pope',
 'making expensive bowl froyo short',
 'successions4e7 greg fire dozen people zoom',
 'gordon ramsay cook chef hell kitchen',
 'first-ever perfect score blind taste test impress chef ramsay hell kitchen',
 'finally tried big-mac',
 'harvey absolute savage minute straight short suit',
 'finally reviewing hardest get reservation nyc carbone review foodreview
carbone nyc',
 '100 sack merry christmas',
 '2023 national spelling-bee champion spellingbee short',
 'call millman',
 'iconic goal hockey history short',
 'novak djokovic control',
 'gordon ramsay finished plate rare moment',
 'notice jake amy came full circle short brooklyn nine-nine',
 'mom get huge hotel bill freshoffttheboat short',
 'well pick every spot bobby axelrod-billions',
 'lamborghini butt-kick ferrari disrespect lamborghini ferrari revenge
shortsfeed',
 'elite lifter return first gym… ',
 'heart warming shopping spree shakepaid',
 'louis litt know make introduction suit louislitt rickhoffman short',
 'asian bad math',
 'gordon ramsay reacts lazy pot noodle beef wellington gordonramsay short',
 'gordon ramsey shocked',
 'gordon ramsay using snoop dogg technique extra fragrant mint',
 'laurie bream awkward moment',
 'modern family haley cheat arvin dylan',
 'stop jumping control shot tennis tennistip',
 'make learning addictive social medium luis von ahn ted',
 'get caught unproductive vs. way thinking',
 'pete sampras built different',
 'margot robbie husband pinch butt red carpet short',
 'mcenroe destroys berrettini',
 'got sick end life dr. morgan levine short longevity aging',
 'one nastiest napoleonic weapon',
 "'sunk cost fallacy poker liv boeree poker short",
 'psychopath rise power brian klaas',
 'coke reveal',
 'guess job finally short',
 'survive hood',
 'gordon mum still know best short',
 'http v=w8ft8vubycc',
 'calculation helped save daughter sankalpbharat neet movie physic math
ytshorts',
```

```
'mcdonald first outlet.  thefounder',
'creepiest b99 character brooklyn nine-nine',
'mike told jack spent night making email',
'secret michelin street food insanely good',
'2023 biggest breakthrough computer science',
'worst year bank since 2008 film',
'easiest way guarantee tip ubereats doordash fooddelivery short',
'delivering rich london london wealthy expensive money delivery satisfying',
'novak djokovic pretend point',
'novak djokovic tagged roger federer',
'cant control jamaican',
'pedro pascal good mexican food nyc',
'society allocate mathematician grant sanderson 3blue1brown',
'haley new job part modern family short',
'classic federer nadal point',
'gambler beat roulette',
'markov chain clearly explained part',
'jim simon trading secret 1.1 markov process',
"'mistake changed nhl hockey forever",
'harry met sally 1989 whole car talk',
'transformed canada goose billion-dollar brand',
'oshie olympic shootout nbc english',
'jim ross call t.j. oshie shootout winning goal vs. russia',
'shootout created hockey legend…',
"money n't buy happiness",
'think right masculinity motivation motivationalquotes trump donaldtrump',
'one hardest goodbye',
"could n't afford surgery happened foryou",
'craziest tennis point ever',
'crazy djokovic championship point',
'winklevoss twin crypto program failed',
'novak djokovic forehand smooth practice davis cup final',
'claire gold digger comment come back haunt clip modern family',
'night agent got watch show nightagent movie shinebykingsleyking afrovibes',
'alcaraz tiafoe hilarious point',
'http v=isztov7gc0s',
'2023 biggest breakthrough math',
'japanese economic miracle explosive growth',
'100 cheese burger philippine',
'cop jedi',
'chunkz genius short chunkz viral trending',
'guessing salary london',
'breathalyzing passenger uberdriver',
'pizza delivery scene minute',
'know right',
'catch win 1,000',
'law student know right refusal like bos    copdismissed 4thamendment
```

```
idrefusal',
 'hot pot hack … kinda lifehacks',
 "'you subscription harvey specter monthly know suit",
 'australian british person',
 'always try human beings ',
 'challenge make sushi quickest  sushi challenge',
 'easy way make better ramen',
 "harvey could n't help suit",
 'harvey want bullsh translated english s05 e05 suit',
 'simple way get free coffee need bean short',
 'stick',
 'chef decides cook expensive dinner hellskitchen gordonramsay',
 'kindness long way help',
 'plan get cancelled',
 'uncle roger review biggest fried rice myth made lau',
 'harvey specter name law game s06 e03 suit',
 'every eliza hamilton',
 'hamilton official trailer disney+',
 'dark side credit card reward nyt opinion',
 'e158 global trade disrupted adobe/figma canceled realtor sued trump blocked',
 'first ever table tennis pump fake short',
 'risky little game friend friend',
 'build sale skill dubai dubai alexdebare',
 'modern family phil realising claire cheated short',
 'tall people short person',
 'harvey impressed client harvey louislitt suit viral series amazonprimevideo',
 'samsnextgen stacked signaturedetails giftwrapped 4dfit roshanmelwani suit',
 'trevor noah simple question ended debate',
 "people n't",
 'louis stop mudding louislitt harvey suit viral series netflix
amazonprimevideo',
 'haley meet boyfriend parent part modern family short',
 'man knock self fight hockey game',
 'wait body slam hockey fight fyp preds nhl short',
 'huge hit fight saved peterborough petes season',
 'hockey fight brutal  hockey fight trending',
 'slafkovsky 1st victim',
 'http v=csefqo_cfwq',
 'still thinking goalie fight battle alberta',
 'hockey fight best',
 'hockey fight getting crazier via luke_hohlt nhl',
 'stood table',
 'vega guinness challenge',
 'helping small business mexico',
 'http v=ufbzbfndnba',
 'alex clothes shopping part modern family short',
 'alex strange grieving process part modern family short',
```

```
'behind apple split goldman sachs wsj tech news briefing',
'warren buffett private equity firm typically dishonest',
'jensen huang nvidia future a.i dealbook summit 2023',
"tesla ceo elon musk 'll say want say lose money",
"'you lied elon musk slaughter bbc reporter live interview",
'people ask',
'love richard dualipa dualipahoudini houdini dualipavideos',
'adam sandler',
'haley job interview part modern family short',
'drunk shakespeare chicago review',
'buzzing drunk shakespeare downtown houston',
'chipotle chicken bowl reaction',
'beta squad expect',
'beta squad insane match girlfriend boyfriend',
'chicken shop date chunkz filly',
'filly expect ksi',
'niko get away',
'chunkz hold back',
'darkest violates nella rose',
'chunkz harry pinero chefasylum',
'beta jackson',
'chunkz emotional',
'filly hold back',
'chunkz uncomfortable',
'beta squad believe',
'chunkz get emotional',
'name country beginning',
'kenny violated',
'chunkz natural',
'chunkz expect',
'beta squad believe',
'beta squad violated',
'beta squad play guess musician dave',
'filly smartest',
'bradley jamaican',
'filly hyped',
'niko violates darkest',
'beta squad ruthless …',
'sancho sterling react chunkz goal',
'ishowspeed admits messi better',
'gangster rapper crazy',
'beta squad believe',
'ricky move',
'sound wrong',
'niko loses 2.5 million mrbeast private jet',
'filly could complete challenge',
'name word rhyme trigger',
```

```
'beta squad serious',
'chunkz pro',
'yung filly smartest',
'filly know',
'chunkz emotional',
'yung filly violates niko',
'yung filly could mad …',
'beta squad setup niko',
'jidion expect',
'beta squad act like kai cenat',
'beta squad expect',
'switched accent end …',
'yung filly happy …',
'beta squad expect',
'chunkz expect …',
'harry pinero smooth',
'harry pinero moving mad',
'harry pinero violates beta squad',
'short people tall person',
'ksi give 1000 say',
'harry violates beta squad',
'niko believe',
'ball never debatable',
'brought friend beer war ',
'awkward phone call',
'viral watermelon hack',
'glimspe dinner party apartment',
'silicon valley s02e10 server overload epic funny scene',
'saying kanye lyric stranger',
"telling people thing n't need know",
'oreo mcflurry',
'confessing stranger',
'steal princepine toplodge',
'jamaican give lady advice',
'maurice',
'telling jamaican joke make sense',
'giving side quest stranger',
'oversharing customer',
'asking security defend honor',
'cheating girl',
'wagwan bossy',
'giving bad direction',
'http v=jppjmypygm8',
'http v=w6lq1i5iuuo',
'finally meet work wife',
'built private jet office',
'panda express gourmet',
```

```
 'gotham',
 "'ve eaten 10,000 time",
 'need stop using recipe',
 'inside pantone company turn color money wsj economics',
 'kim jong outfit reveal north korea wsj',
 'ranking every restaurant legendary japanese shopping street',
 'got nice tip craigferguson latelateshow standupcomedy',
 'louis want sue harvey',
 'give presentation gavin belson',
 'gavin goon',
 'http v=8zmoxx4t4yw',
 'smith technique',
 'student juggle prove sober',
 'osmo walmart',
 'ranking every death row meal philip ray workman',
 'scottie got litt',
 'chunkz pro',
 'algorithmic trading using python full course',
 'fit quant finance',
 'algorithmic trading using python full course',
 'yung filly violates niko',
 'american british person',
 'kai cenat',
 'niko genius',
 'filly smart',
 'niko clutch guess',
 'short people tall person',
 'female secret male',
 'lost mind',
 'female secret male',
 'female secret male',
 'chunkz lost mind',
 'female secret male',
 'funny',
 'female secret male',
 'sharky genius',
 'got job google ',
 'military female tighten belt tiktok paulasamira',
 'consultant vs. product manager  ft bykchoi',
 'google engineer salary level short',
 'jeff bezos really approachable wealth jeffbezos celebrity entrepreneur ceo',
 'completely called',
 'investment banking fit question quick prep',
 'conjoined triangle success',
 'math     short bryancranston lottery viral short shortsvideo trending
lucky',
 'rich guy disrespect server',
```

```
 'part',
 'house smell rat short fyp',
 'haley job interview part modern family short',
 'modernfamily phildunphy clairedunphy sarahhyland series disneyplus',
 'best prank b99 history short brooklyn nine-nine',
 'louis save day viral short series suit harvey louislitt',
 'caught fish chip river thames',
 'costco tuna sushi danger',
 'lost found oscar shortlisted stop-motion animation short week',
 'fold pack suit right way',
 'e156 ivy league antisemitism macro saas recovery gemini figma deal delay big
friedberg update',
 'fold blazer',
 'fold jacket travel luggage wrinkle howtofold nowrinkles suitcafe',
 'in-n-out work',
 'costco salmon mre sort',
 'foodie breakfast short',
 'sliced brisket order prep',
 'elle wood girl legally blonde 2001',
 'effective way offset depreciation jet',
 'man hanged laughed. short viral recap',
 'chess make football billion',
 'undercover cop walk italian deli boston short boston italianfood italian
comedyvideo',
 'costco salmon nobu sashimi sushi',
 'generation gap smoking vaping old dad 2023',
 'favorite party trick explained',
 'boarding airplane',
 'real authentic mexican littleitalyla.com',
 "remade first viral video still best dish 've ever made",
 'japanese morning ramen',
 'japanese wagyu hotel room short',
 'japan cheapest meal short',
 'money ethic mike find professor gerard lying suit',
 'harvard ethic professor seek harvey help suit',
 'mitchell paris part modern family short',
 'louis onto mike',
 'nguyen pronounce vietnam popular name vietnamese learnvietnamese vietnam',
 'pronounce nguyen according people named nguyen',
 'cybertruck beat porsche 911 towing 911',
 'tesla cybertruck hummer drag race',
 'http v=5atzeijeu98',
 'barbarian gate story ross johnson nabisco takeover',
 'david rubenstein explains lack private equity deal',
 '240 day request made cantonese family style dinner',
 'shooter bob lee sniper set search assassin plan kill president. movie',
 'bonding girlfriend dad',
```

```
'trader joe salmon vs. sushi-grade',
'engel aggregation',
'envelope theorem constrained optimization roy identity',
'm5e12 microeconomics certainty equivalence risk premium',
'leg wrestling',
'axe taylor mentorship two',
'begini rasanya steak keripik kaca',
'man grow billion',
'well bobby axelrod-billions',
'shoot sonny',
'reminds best day life bobby axelrod-billions',
'become billionaire bobby axelrod-billions',
'get one life  bobby axelrod billion',
'busukin daging pake jasuke',
'ngelem daging biar jadi steak emang bisa',
'masak daging busuk pake saus korea',
'kasih coba orang yang udah nungguin daging busuk ini',
'hasil ngelem daging seharian',
'bikin jus wasabi buat daging pes banget',
'arghh jus wasabi gini banget rasanya',
'borong gerobak nasi goreng',
"excel save hour time people n't know",
'alternating series test',
'axe wag loyalty billion',
'axe bos straight killer ',
'japan unhealthy noodle short',
'asynchronous video interview prepare usually asked',
'global internship interview tip',
'mckinsey last week tonight john oliver hbo',
'first five minute every season suit',
'lehman trilogy broadway show clip',
'ntl lehman trilogy official trailer',
'      marx met confucius translated clip',
'',
'                  ',
'october 2023 yet singing pain',
'october 2023 jesus asks faith',
'update bold faith campaign 210 java road',
'june 2023 spoken word receiving feedback',
'november 2023 cultivating city one 7.4 million',
'dim-summary november 2023',
'november 2023 cultivating city problem work',
'bold move brilliant player  axe_billions',
'harvey save kid career suit short',
'wagyu beef one world faked food wagyu beef japan',
'all-in summit ray dalio rise fall nation changing world order',
'all-in summit stephen wolfram computation nature universe',
```

```
 'minute chuck sr. unhinged billion showtime',
 'hire',
 'explaining 7500 purchase wife',
 'people correctly',
 'peace offering …',
 'day life',
 'restaurant steak better',
 'insanely delicious sushi 6.97 grocery salmon',
 'chunkz show bollywood vocal',
 'chunkz clutch guess',
 'speed chunkz lose 2.5 million private jet',
 'sharky get pressed',
 'ksi hold back',
 'ksi give 1000 say',
 'way said',
 'niko australian',
 'switched accent end …',
 'wrong choice',
 'making pink pasta parishilton',
 'religious education god',
 'really thought',
 '0dte inside explosion ultra-risky option trading wsj',
 'matthew perry went head-to-head journalist peter hitchens drug addiction 2013
bbcnews',
 'hour steak',
 'batman',
 'stay strong king bench press short',
 'http v=bgxx8hij_li',
 'made monster… movie fyp',
 'gangster rapper crazy',
 'discharging patient emergency room',
 'man named shelby peakyblinders short',
 'whole food salmon sushi danger salmon sushi',
 'mark wahlberg right mark short',
 'donna reading jessica',
 'tried bangkok legendary crab glass noodle street eats bon appétit',
 "best thing 've ever eaten short",
 'bought airline',
 'episode teaching next date cook cedriklorenzen food cooking',
 'wagyu tallow béarnaise sauce',
 'duality utility maximization expenditure minimization',
 'duality lagrangian dual problem',
 'a.8 consumption duality consumption microeconomics',
 'guy trying smash',
 'harvey congratulating mike making junior partner',
 'snoop dogg shaq fried chicken',
 'best beginner guide hedge fund private equity venture capital',
```

```
 'non-investment banking finance job high pay great work life balance',
 'know point academy … episode 157 highlight',
 'mission impossible dead reckoning part one',
 'hard feeling',
 'college endowment massive',
 'gretchen showing loyalty harvey',
 'introducing donna replacement amazing experienced gretchen suit',
 'harvey hiring new secretary',
 'old barbie movie classical playlist',
 'sharky get stabbed back',
 'became member british royal family',
 'found missing 100m plane …',
 'gaza strip',
 'http v=f8ujyswd-eo',
 'robbing bank… movie fyp',
 'louis wanted surprise donna',
 'cocktail bar',
 'harvey mike take spontaneous trip atlantic city short suit',
 'salt steak',
 'part hospital short recommended',
 'mike take witness suit short',
 'flirting work real estate agent short',
 'gretchen showing loyalty harvey',
 'esas son reebok son nike truly hilarious audio clip',
 'business opportunity ozark short',
 'http v=zw7k8uahgqu',
 'michelin starred ramen',
 'giving side quest stranger',
 'seasonal breakfast japan',
 "ramen 've never heard short",
 'lasagna olivia tiedemann',
 'illegal grilled cheese',
 'foreign market',
 'much bonus harvey louislitt suit viral series netflix amazonprimevideo',
 'basecamp antarctica',
 'modelo used data become america top beer wsj economics',
 'men changing room',
 'russia belong europe asks singapore foreign minister',
 'massively regret threatening gangster replica gun    snatch desert eagle
scene',
 'openai ceo kid studying',
 'cooking top empire state building',
 'licence',
 'cocktail bar',
 'tourist friendly two michelin star sushi omakase tokyo 150 sushi umi',
 'billion season episode series finale promo showtime',
 'tragedy old face andre thewire hbo sceneremix bankofenglandglass zorrohuh full
```

```
version',
 'wealth nation summary adam smith',
 'roasting pitch deck used raise 5,000,000',
 'israel palestinian century conflict',
 'exclusive alvin leung unveils secret dubai demon duck restaurant',
 'essential adam smith moral sentiment',
 'friend friend',
 'http v=sticfbfaoog',
 'could fix short',
 'working home',
 'street thug stunned real gangster gentleman best scene',
 'asking girl rate vs. chris hemsworth',
 'diversification holy grail investing',
 'ranking every restaurant legendary japanese shopping street',
 'steak ranked worst best short',
 'non-glasses wearer',
 'one thing menu … short',
 'turning costco trout 300 sushi epic transformation sushi',
 'frozen costco scallop sushi danger sushi costco',
 'job interview',
 'get good small talk even enjoy',
 'israel declares war hamas following unprecedented wide-scale attack wsj',
 'chinese indian leaving india ',
 'cooking sister vegan',
 'http v=xhvp5otsvys',
 'hedge fund',
 "n't help seeitoff podcast maxfosh",
 'macaroni shoe',
 'maurice',
 'jamaican accent switch',
 'neil degrasse tyson flaw eyewitness testimony',
 'elevate costco salmon korean raw soy marinated delight koreanfood costco
salmon',
 'inside tech fight climate change exponentially azeem azhar',
 'fried chicken war curse popeyes',
 'costco sushi hack make sushi frozen scallop',
 'walmart tuna sushi danger sushi tuna sashimi',
 'trade bouncy stock quant interview question',
 'solve quickly investment bank interview question',
 'tiktok e-commerce erode alibaba sea margin',
 'rise fall possible rise san francisco downtown wsj',
 'far car zero fuel',
 'beat nerd muscle ups win 100',
 'arcade worker… movie fyp',
 'never happened',
 'pull hair prank omegle funny',
 'ray dalio country invest',
```

```
 'babe wanted egg bacon cheese',
 'undetectable',
 'all-in summit ray dalio rise fall nation changing world order',
 'would happen everyone stopped eating meat tomorrow carolyn bean',
 'nothing except everything',
 'modern family manny real dad',
 'liberal parent react christian daughter',
 'fairground game',
 'vs. dente compilation',
 'toothpaste dentist recommend',
 "n't funny",
 'dad airport',
 'ben shapiro raised 4.8 million daily wire',
 'kid crazy try punch',
 'running marathon without training',
 'two double bed country old men 2007 short nocountryforoldmen moviescene',
 'james maddison best response bournemouth fan …',
 "'sorry speaking english giroud funny exchange reporter short",
 'goalkeeper receive red card within minute',
 'christian pulisic olivier giroud test friendship',
 'olivier giroud bicycle kick poland olivergiroud giroud bicyclekick fifa22
france',
 'assist goal ozil giroud',
 'olivier giroud becomes milan goalkeeper   football acmilan short',
 'pulisic goal giroud save genoa 0-1 milan highlight serie',
 'cobb-douglas utility function',
 'utility maximization cobb-douglas utility function',
 'http v=noeiu3nwydc',
 'hustle big white men jump scene jack harlow',
 'brought nerf gun shooting range',
 'know highschool… movie fyp',
 'pt1 hospital recommended',
 'caucasian filled rank army get 4dfit cap plentyofconsent.com',
 'least upper bound proof',
 'change numeraire',
 'jaylen arthur yang snowflake 2023 contemporary music competition',
 'welcome autumn quarter message dean hale',
 'tvb drama    60fps 1/25             tvb   ',
 'airport security',
 'samsnextgen roshanmelwani samstailor tiktoktailor celebritytailor
suitwhisperer bespokesuits',
 'turning costco trout 300 sushi epic transformation sushi',
 'all-in summit bill gurley present 2,851 mile',
 'private equity everything',
 'private equity everything',
 'chick-fil-a milking 1000 k8s cluster',
 'learn speak indonesian day',
```

```
 'forget follow humour fat loss fact alongside health fitness exercise',
 'many coat hanger hold weight',
 'sheldon school leader',
 'translated hogwarts french',
 'round friend',
 'trying break glass voice',
 'ton rubber band ball drop 2000ft',
 'story behind uno reverse card',
 'inappropriate kiss … whoseline scenesfromahat',
 'new honey pepper pimento sandwich',
 'mike turn junior partner',
 'school shooter  fallout short',
 'asking girl chrisbumstead',
 'free sample short',
 'andrew tate edit viral short islam islamicshorts fyp ',
 'japanese style fried rice short',
 'miracle',
 'chinese waiter shocked interpret chinese friend',
 'mha colour fyp  edit colour colourtrends mha myheroacademia anime',
 'guess business buy something blindfolded short',
 'get dream job',
 'alisha lehmann revenge short',
 'broke ankles ',
 'last throw',
 'fix hole wall',
 'bouncer',
 'riding queen year \U0001f979  equestrianrider equestrian horse
equestrianlife',
 'bottle flip level level 100 part',
 'absolutely terrifying kristianlandgren',
 'match rhythm   dance',
 'president 9/11 attack klemfamily',
 'video removed',
 'got presidential pardon naughty school',
 'cancelled ride…but still deserves star uberdriver',
 'cappuccino prank',
 'undercover cop boston',
 'coldest table tennis player ever',
 'transgender dophile confronted walmart meeting 12-year-old',
 'making dictator favorite food vladimir putin',
 'http v=h-dudjlgxag',
 'graffiti legally',
 'using fake weight public gym',
 'way said',
 'broke international security convention',
 'see always true motivation inspiration sigma short',
 'recreated primary school sport day',
```

```
"lily mark '19 speaking shanghainese grandma",
'foreigner speaking shanghainese',
'parent like',
'all-in summit ray dalio rise fall nation changing world order',
'3-year life update marriage phd new job',
'glad rejected oxford thewaythingsgo trendingonshorts',
'shanghainese learn basic phrase angela',
'   ',
'       ',
"lily mark '19 speaking shanghainese grandma",
'aussie wake coma speaking mandarin feed',
'foreigner speaking shanghainese',
"'succession character finance analyzed money expert wsj spent",
"s04e02 one email n't exhaust logan roy succession season episode brian cox",
'best roman roy succession hbo kieran culkin',
'one roman roy scene',
'harry pinero violates beta squad',
'speed chunkz lose 2.5 million private jet',
'sharky get pressed',
'way said',
'chunkz clutch guess',
'pressure washing entire subdivision',
'restaurant cost 500k',
'http v=8utkf_e9ame',
'finishing high school',
'still burnt',
'claire discovered truth',
'soap like',
'feel wrong making sign',
'bradford meet guy civilian arrest',
'super easy quick beef taco cooking asmr food short',
'started religion',
'video removed',
'prime new hydration sponsor barcelona short',
'bouncer feat stephen try jack joseph',
'cartel delivers good',
'fastest way learn code codingtips startup short',
'behind scene met top firm lightspeed venture partner',
'student get caught plagiarism',
'dinit',
'shot year',
'learnt hokkien week',
'making grilled cheese sandwich using cheese bread',
'gordon ramsay fried chicken',
'video removed',
'debate anyone',
"'ve never seen someone jump higher life",
```

```
'got pizza delivered moving boat',
'calling queen government name part stephanie prince short dragqueen',
'tipping america',
'jessica absolute bos suit',
'answering phone front friend',
'winning long war ukraine',
'inside world largest cargo shipping bottleneck today wsj',
'duality indian couple',
'humanely kill squid food/bait',
'maui resident put brush fire saving home possibly neighborhood',
'harvey deposing woodall part',
'bouncer',
'muslim interrogation comedyskit 60seconds funnyclips muslim funny',
'dad restaurant',
'',
'yukhoe short',
'deported america',
"trevor n't mike wedding",
'biggest restaurant scam',
'ice cream store stopped using spoon sciencefacts amazonfinds',
'tennis player struggle make living',
'flying iphone trick shot',
'mountain lion stalk elk hunter idaho saved glock27 warning shot',
'delivery',
'http v=gydfj7yrhdw',
'best reaction fails gymnastics sport flip flipfail trampoline',
'meat loaf surprise ingredient short meatloaf thefword',
'3000 year ago… shocking movie fyp',
'pov time month … shortsfeed short',
'error sigmund freud life lesson rule  motivation quote aphorism',
'sheldon bar',
'three charging coyote rolled run dropped running one man',
'bought roundabout',
'vandalizing car ft. zachcray',
'amazing',
'http v=zfy-fhqect4',
'bridge needed learn trick skimboarding',
'never djokovic',
'samsnextgen flyingdutchman stacked signaturedetails giftwrapped 4dfit roshan',
'100 caviar grilled cheese',
'ultimate lookup guide xlookup vlookup hlookup',
'excel inserting column keyboard shortcut mac',
'excel inserting column keyboard shortcut mac',
'looking back year achievement mediacorp year-end special',
"n't indonesia speak dutch documentary",
'faked food world big business insider business',
'highsnobiety visit story behind hong kong hour suit',
```

```
 'awkward first date prank',
 'asking girl girlfriend powerpoint presentation',
 'earned it  basketball couple foodislife comedy',
 'tequila take oscar red carpet',
 'hawk-eye transformed u.s. open sport',
 'type fresh frozen salmon eat raw walmart whole food',
 'ukraine soldier get shot russian sniper short ukraine',
 'escape robber …wait stevenhe',
 'legally became parent favourite child',
 'sam tailor hong kong',
 'join army get 4dfit plentyofconsent.com roshanmelwani samstailor
tiktoktailor',
 'army continues grow get 4dfit cap plentyofconsent.com roshanmelwani tailor',
 'looking back year achievement mediacorp year-end special',
 'wagyu slice twist unleash taste bud dice epic topping sushi wagyu',
 'pro eater vs. endless pasta olive garden',
 'jamaican story time',
 'giving gta mission stranger',
 'mocking people',
 'confessing stranger',
 'little smoothie action',
 'jamaican drive thru song',
 'telling jamaican joke make sense',
 'skimboarders dream wave',
 'jessica never breakast bar life',
 'jesse walt flashback camino short',
 'gym girl gold bar challenge',
 'wild short',
 'treat champion',
 'thing escalated short short viral edit prank gonewrongprank',
 'tim well world record',
 'make pasta living',
 'http v=ari-hrtktvm',
 'wooden cutting board care short',
 'another classic recipe',
 'moroccan eat',
 'http v=ly2th4b8m4o',
 'town save tax moving front door',
 'singapore local teach perfect hainanese chicken rice origin',
 'crimewatch 2023 ep6 slashing case two young suspect',
 'canada territory explained',
 'samsmodel plentyofconsent stacked signaturedetails giftwrapped 4dfit roshan',
 'get 4dfit cap plentyofconsent.com roshanmelwani plentyofconsent samstailor
tailor',
 'benjamin mike louislitt harvey suit viral series netflix amazonprimevideo',
 'joe gibbs racing crew quick',
 'daily meme pt29',
```

```
 'marta kostyuk happy last backhand tennis',
 'mike tyson living nyc short viral interview boxing miketyson celeb',
 'two londoner one got somesing beddar    ',
 'luke damant eats 0.20 potato swirl bangladesh short',
 'way tried beachprotour',
 'deserved beachprotour',
 'okaaayyy beachprotour',
 'rally everything beachprotour',
 'opponent want score point hughes everywhere   ',
 'difference novice veteran using excavator',
 'say n-word 100 ft. zople',
 'extreme capital city world',
 'plane',
 'put glove surgery short',
 'nolan help blackmail victim',
 'love girl follows rule',
 'spinning wheel defy gravity obeys physic funny fyp reel short shortsvideo',
 'old man ability unique world  ',
 'always take receipt end ing world short',
 'modern family haley new boyfriend',
 'samsmodel stacked signaturedetails giftwrapped 4dfit roshanmelwani suit',
 'pablo escobar coke boat cartel smuggling technique  ',
 'manchester united fan kick ball onto barcelona resident balcony viralhog',
 'cashier cheat customer money short gasstation',
 'lukaku said micah richards',
 'milan friend cat traumatized waiter',
 'heroic mom save terrified five-year-old raccoon attack usa today short',
 'classic kimboslice ufc streetfighter mma mmafighter ufc fightnight boxing
ufcfighter',
 'modern family gloria meet ronaldo',
 'medieval surgical device invented remove arrow king henry face',
 'way  bobby axelrod_billions link comment',
 'sell war american people',
 'dish killing indian restaurant',
 'world  friendliest lumberjack short gasstation',
 'celosextremos jajaja novio fpy',
 'modern family thanksgiving dinner',
 "buying alcohol 're",
 'convenient form transportation short',
 'hero tackle fleeing drunk driver killed texas cop crash',
 'man best job world',
 'dad pass vibe check',
 'mali removed french national language',
 'america got stupid',
 'excel negative number bracket',
 'hooded youth',
 'legit checking nike dunk low legitcheck ebay ebaysneakers shoe sneaker
```

```
nikedunk',
 'tricking entire country ft. maxfosh',
 'fell prank funny comedy omegle comedyprank',
 'got jeff bezos speeding ticket bicycle',
 'japan better bbq us…  ',
 'system worked better mine shameless short viral',
 'beta squad setup chunkz',
 'teacher toilet',
 'lasagna sandwich cookingwithkian',
 'taiwanese man pairing singaporean men vietnamese woman',
 'computer write proof point mathematician',
 'legendary recipe',
 'lawyer win first ever case suit',
 'reese witherspoon billion-dollar secret codie sanchez expose book-to-movie
strategy',
 'gymbro catch slacking … shortsfeed short',
 'pov gymbros see filming … shortsfeed short',
 'five guy mcdonald shook',
 'googling symptom',
 'bought sport car',
 'barber',
 'tell mate look',
 'picked last',
 'fish chip shop',
 'english teacher',
 'amy meet sheldon sibling',
 'rest steak short',
 'favorite pasta',
 'ranking every death row meal ricky ray rector',
 'http v=gjdolhzdu7q',
 'interstate trailer',
 'bet eat double cheeseburger hour interstate',
 'video removed',
 'costco salmon sushi hack',
 'beta squad snake harry pinero',
 'harry pinero love gyat',
 …]
```

Testing a range of values of k, using an elbow plot to plot inertia against value of K

[ ]:
```
random.seed(42)
inertias = []
k_values = []
for i in range(1,21): #chose range up till 21 because on previous attempt␣
 ↪gradient didn't seem to level off
    kmeans = KMeans(n_clusters=i, n_init = 20, init = 'k-means++')
    kmeans.fit(X)
```

```
    inertias.append(kmeans.inertia_)
    k_values.append(i)

inertias_df = pd.DataFrame({
    "K-value": k_values,
    "Within-cluster variance sum": inertias
})
print(inertias_df)
```

[ ]:
```
print(p9.ggplot(inertias_df, p9.aes(x = 'K-value', y = 'Within-cluster variance␣
 ↪sum')) +
p9.geom_vline(xintercept = 2, color = "red") +
p9.geom_vline(xintercept = 6, color = "red") +
p9.geom_vline(xintercept = 9, color = "red") +
p9.geom_vline(xintercept = 13, color = "red") +
p9.geom_vline(xintercept = 18, color = "red") +
p9.geom_line() +
p9.scale_x_continuous(name = "$K$") +
p9.scale_y_continuous(name = "Inertia") +
p9.theme(legend_position = "none", figure_size = [6, 3.5]))
```

The red line indicate noticeable inflection points that would make for good K value candidates. Let's choose 13 since we are expecting a relatively large number of categories.

[150]:
```
kmeans = KMeans(n_clusters=13, n_init = 20, init = 'k-means++')
kmeans.fit(X)
```

[150]:
```
KMeans(n_clusters=13, n_init=20)
```

[151]:
```
cluster_groups = kmeans.labels_
watch_df["13 clusters"] = cluster_groups
```

[ ]:
```
#Checking to see the distribution of cluster sizes
from collections import Counter
counts = Counter(cluster_groups)
counts_df = pd.DataFrame(counts.items(), columns=['Item', 'Count'])
print(counts_df)
```

[152]:
```
group_0 = watch_df[watch_df["13 clusters"] == 0]
group_1 = watch_df[watch_df["13 clusters"] == 1]
group_2 = watch_df[watch_df["13 clusters"] == 2]
group_3 = watch_df[watch_df["13 clusters"] == 3]
group_4 = watch_df[watch_df["13 clusters"] == 4]
group_5 = watch_df[watch_df["13 clusters"] == 5]
group_6 = watch_df[watch_df["13 clusters"] == 6]
group_7 = watch_df[watch_df["13 clusters"] == 7]
group_8 = watch_df[watch_df["13 clusters"] == 8]
group_9 = watch_df[watch_df["13 clusters"] == 9]
```

```
group_10 = watch_df[watch_df["13 clusters"] == 10]
group_11 = watch_df[watch_df["13 clusters"] == 11]
group_12 = watch_df[watch_df["13 clusters"] == 12]
```

[43]:
```
group_11.sample(20)
```

[43]:

	title	URL
6608	Everything Is Stupid - The Metaverse | The Dai…	https://www.youtube.com/watch?v=XVNOUtMVxyw
11927	Remember Traveling? | The Daily Show	https://www.youtube.com/watch?v=2XBTeqo7U0Q
11986	Joe Biden's Path to the Presidency | The Daily…	https://www.youtube.com/watch?v=RP8rCUfB8M8
14973	2020 December Democratic Debate in Los Angeles…	https://www.youtube.com/watch?v=TtUld2FZoAg
12026	The Most Scandal-Plagued Presidency Ever - A L…	https://www.youtube.com/watch?v=bb-4LGi5bXU
12119	What the Hell Happened This Week? - Week Of 11…	https://www.youtube.com/watch?v=q_hnCQBNYsc
12094	Trump's Impeachment Acquittal, State of the Un…	https://www.youtube.com/watch?v=_qZNksTX4zk
11686	The Best of Ronny Chieng - Wrestling, Bitcoin …	https://www.youtube.com/watch?v=g0MkYazCxyA
12560	Edward Snowden - "Permanent Record" & Life as …	https://www.youtube.com/watch?v=PArFP7ZJrtg
14256	Coronavirus Misinformation & Toilet Paper Pani…	https://www.youtube.com/watch?v=Ls0ZormAOhU
12535	Trump's Sister Caught Talking S**t & Trump Bul…	https://www.youtube.com/watch?v=jNjYqw8aX5A
4746	Not Everyone Is Mourning The Queen's Death & N…	https://www.youtube.com/watch?v=u7Wja0rdc-U
11790	Black Home Ownership - If You Don't Know, Now …	https://www.youtube.com/watch?v=NEKYqsIZMn8
11307	Trump's Travel Ban: A Disaster in Four Acts | …	https://www.youtube.com/watch?v=A2tyXZmj2Ew
14955	Trump Orders Assassination of Top Iranian Gene…	https://www.youtube.com/watch?v=SXARDOeSDyg
11409	Jordan Klepper vs. Trump Supporters: The Compl…	https://www.youtube.com/watch?v=7OeeHz0uNdM
12103	What the Hell Happened This Week? Week of 9/14…	https://www.youtube.com/watch?v=9mXtS3-ffDE
15042	2019 Was Stupid | The Daily Show	https://www.youtube.com/watch?v=cof3YrF-_K0
12089	Crazy Rich Nation | The Daily Show	
14588	The Oscars, Coronavirus Updates & United's Pla…	

```
12089  https://www.youtube.com/watch?v=TtziF8sgZ0I
14588  https://www.youtube.com/watch?v=7UcJtY-aFsg

                                                  Channel time_24hr  \
6608   [{'name': 'The Daily Show', 'url': 'https://ww…  15:43:43
11927  [{'name': 'The Daily Show', 'url': 'https://ww…  16:04:35
11986  [{'name': 'The Daily Show', 'url': 'https://ww…  13:56:22
14973  [{'name': 'The Daily Show', 'url': 'https://ww…  16:56:44
12026  [{'name': 'The Daily Show', 'url': 'https://ww…  16:30:11
12119  [{'name': 'The Daily Show', 'url': 'https://ww…  15:14:57
12094  [{'name': 'The Daily Show', 'url': 'https://ww…  20:08:40
11686  [{'name': 'The Daily Show', 'url': 'https://ww…  15:41:09
12560  [{'name': 'The Daily Show', 'url': 'https://ww…  02:43:30
14256  [{'name': 'The Daily Show', 'url': 'https://ww…  05:08:15
12535  [{'name': 'The Daily Show', 'url': 'https://ww…  04:16:45
4746   [{'name': 'The Daily Show', 'url': 'https://ww…  17:58:32
11790  [{'name': 'The Daily Show', 'url': 'https://ww…  15:05:15
11307  [{'name': 'The Daily Show', 'url': 'https://ww…  09:06:49
14955  [{'name': 'The Daily Show', 'url': 'https://ww…  12:16:02
11409  [{'name': 'The Daily Show', 'url': 'https://ww…  17:55:29
12103  [{'name': 'The Daily Show', 'url': 'https://ww…  17:20:21
15042  [{'name': 'The Daily Show', 'url': 'https://ww…  05:09:09
12089  [{'name': 'The Daily Show', 'url': 'https://ww…  10:21:26
14588  [{'name': 'The Daily Show', 'url': 'https://ww…  04:06:07

            date  Category Prediction  13 clusters
6608   2022-04-24                   23           11
11927  2020-11-27                   23           11
11986  2020-11-22                   23           11
14973  2020-01-05                   23           11
12026  2020-11-18                   23           11
12119  2020-11-14                   23           11
12094  2020-11-14                   23           11
11686  2020-12-07                   23           11
12560  2020-08-25                   23           11
14256  2020-03-12                   23           11
12535  2020-08-31                   25           11
4746   2022-09-17                   24           11
11790  2020-12-04                   23           11
11307  2021-01-09                   23           11
14955  2020-01-07                   23           11
11409  2020-12-29                   25           11
12103  2020-11-14                   23           11
15042  2020-01-02                   23           11
12089  2020-11-15                   23           11
14588  2020-02-14                   23           11
```

Group 11 seems to be about news/politics

We end up with the following categories

[15]:
```
cluster_dict = {
"Cluster 0":"Comedy, funny moments, sit-coms, satire",
"Cluster 1":"Unclear",
"Cluster 2":"Movie/TV trailers",
"Cluster 4":"Movie Scenes",
"Cluster 5":"Larry david 'Curb Your Enthusiasm'",
"Cluster 6":"Videos that were removed due to copyright",
"CLuster 7":"Short form videos, assorted topics",
"Cluster 8":"Fortnite",
"Cluster 9":"Removed videos",
"Cluster 10":"Guitar Lessons",
"Cluster 11":"News/Politics",
"Cluster 12":"Brooklyn Nine-Nine"}
```

[153]:
```
#since we flagged removed videos, we can update our watch_df
watch_df = pd.concat([group_0, group_1, group_2, group_3, group_4, group_5,␣
 ↪group_7, group_8, group_10, group_11, group_12], ignore_index=True)
```

Reclustering the unclear videos

Trying once more to see if we can get good clusters when rerunning the process on "unclear" videos. Then later on we can think about combining small clusters into larger clusters of a similar kind.

[33]:
```
cluster_1_titles = group_6["title"].tolist()

# Processing texts
cluster_1_processed = [processing(title) for title in cluster_1_titles]

# Convert to TF-IDF vectors
vectorizer = TfidfVectorizer()
Y = vectorizer.fit_transform(cluster_1_processed)
```

[18]:
```
random.seed(42)
inertias = []
k_values = []
for i in range(1,21): #chose range up till 21 because on previous attempt␣
 ↪gradient didn't seem to level off
    kmeans = KMeans(n_clusters=i, n_init = 25, init = 'k-means++')
    kmeans.fit(Y)
    inertias.append(kmeans.inertia_)
    k_values.append(i)

inertias_df = pd.DataFrame({
    "K-value": k_values,
    "Within-cluster variance sum": inertias
})
```

```
print(inertias_df)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/var/folders/8j/bp7vb7mn3nz6lvwvkhglmg2c0000gn/T/ipykernel_8585/474640288.py in
 ↪<module>
      4 for i in range(1,21): #chose range up till 21 because on previous␣
 ↪attempt gradient didn't seem to level off
      5     kmeans = KMeans(n_clusters=i, n_init = 25, init = 'k-means++')
----> 6     kmeans.fit(Y)
      7     inertias.append(kmeans.inertia_)
      8     k_values.append(i)

NameError: name 'Y' is not defined
```

[34]:
```
print(p9.ggplot(inertias_df, p9.aes(x = 'K-value', y = 'Within-cluster variance␣
 ↪sum')) +
p9.geom_vline(xintercept = 2, color = "red") +
p9.geom_vline(xintercept = 5, color = "red") +
p9.geom_vline(xintercept = 7, color = "red") +
p9.geom_vline(xintercept = 11, color = "red") +
p9.geom_vline(xintercept = 16, color = "red") +
p9.geom_line() +
p9.scale_x_continuous(name = "$K$") +
p9.scale_y_continuous(name = "Inertia") +
p9.theme(legend_position = "none", figure_size = [6, 3.5]))
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/var/folders/8j/bp7vb7mn3nz6lvwvkhglmg2c0000gn/T/ipykernel_8585/883780635.py in
 ↪<module>
----> 1 print(p9.ggplot(inertias_df, p9.aes(x = 'K-value', y = 'Within-cluster␣
 ↪variance sum')) +
      2 p9.geom_vline(xintercept = 2, color = "red") +
      3 p9.geom_vline(xintercept = 5, color = "red") +
      4 p9.geom_vline(xintercept = 7, color = "red") +
      5 p9.geom_vline(xintercept = 11, color = "red") +

NameError: name 'inertias_df' is not defined
```

Trying k=5

[35]:
```
kmeans_round2 = KMeans(n_clusters=11, n_init = 25, init = 'k-means++')
kmeans_round2.fit(Y)
cluster_groups_2 = kmeans_round2.labels_
```

```
group_1["11 clusters"] = cluster_groups_2
```

```
/var/folders/8j/bp7vb7mn3nz6lvwvkhglmg2c0000gn/T/ipykernel_8585/3661136591.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

[36]:
```
#Checking to see the cluster distribution
from collections import Counter
counts = Counter(cluster_groups_2)
counts_df = pd.DataFrame(counts.items(), columns=['Item', 'Count'])
print(counts_df)
```

```
    Item  Count
0      6  14134
1      3   1515
2     10    226
3      5    378
4      2    446
5      9    204
6      0    393
7      7    368
8      4    216
9      8    182
10     1    105
```

[151]:
```
group_1_0 = group_1[group_1["11 clusters"] == 0]
group_1_1 = group_1[group_1["11 clusters"] == 1]
group_1_2 = group_1[group_1["5 clusters"] == 2]
group_1_3 = group_1[group_1["5 clusters"] == 3]
group_1_4 = group_1[group_1["5 clusters"] == 4]
group_1_5 = group_1[group_1["5 clusters"] == 5]
group_1_6 = group_1[group_1["5 clusters"] == 6]
group_1_7 = group_1[group_1["5 clusters"] == 7]
group_1_8 = group_1[group_1["5 clusters"] == 8]
group_1_9 = group_1[group_1["5 clusters"] == 9]
group_1_10 = group_1[group_1["5 clusters"] == 10]
```

[151]:
```
                                                   title  \
26099                                      Must Love Dogs
23297  Orange Juice VS Nickatnyte | Clash Royale King…
16638            Pink Floyd - Time solo (Cover by Chloé)
2913   Disney Vs. DeSantis: Why Florida's Governor To…
10846               My FAVORITE Chess Openings by Rating
20428                              love's so hard to find
```

```
2873   Make an Awesome Excel Dashboard in Just 15 Min…
16837                  RICH VS POOR KIDS 3 - DINNER TIME
19283                      THE COOKIE CHALLENGE PRANK…
25606  Daniel Bryan's final match: Daniel Bryan & Joh…
21327                                WILLIAM LOVES DOGS
23129                          $1 Bagel vs. $1,000 Bagel
21186  Maroon 5 surprise a teen for the party of the …
12246    How Popeyes' Chicken Sandwich Changed Fast Food
10578       Ludwig reacts to xQc vs Pokimane chess match
13692             Real Life Trick Shots 3 | Dude Perfect
20759                                   SCHOOL vs WORK
22195  Thanos VS. Deadpool   Josh Brolin And Ryan Rey…
1361           COULD YOU CATCH THIS FOR 10,000 DOLLARS …
22148  Bloopers That Make Us Love Ryan Reynolds Even …
12327                     How to Remain Calm With People
21721            Avicii - Waiting For Love (Lyric Video)
4429   How rich people vs really rich people board an…
7207              War in Ukraine: are sanctions working?
7349   The Most Brutal Attacking Tennis by Roger Federer
15515              SURPRISING PEOPLE WITH KYLIE JENNER!!
1303                         The $1,000,000 Fruit Salad.
25774  Shane McMahon and Daniel Bryan announce huge t…
6202    Going to a dentist for the first time in the USA
25865  Daniel Bryan vs. Randy Orton -- No Disqualific…
21974                             Homer Simpson play bass
11400  $1.5M Champions Chess Tour: Airthings Masters …
10575     How do governments make money? | CNBC Explains
26       How to make pork dumplings   #asianfood #recipe
17711  Why incompetent people think they're amazing -…
19396  Hardcore Pawn - Rule Breaker Tries To Score Ca…
16132         WWII enemies reunited in D-Day anniversary
7458        College Admissions: Inside the Decision Room
2802                 Solo 2 DAYS Eating ONLY What I Catch
22810  LFC's epic Rock, Paper, Scissors challenge | M…
12507   6 Sep 2020 | Love Letters:  We Are Loved to Love
4475              Who's the hottest girl at this college?
22540             I Covered Myself With Tattoos For A Day
1125                                 JAMAICAN STORY TIME
15796              London Hacks - Living on £1 a Day | #4
20372                           2,500,000 SUBSCRIBERS…
23669     Inside The Lives Of The Rich Kids Of Singapore
21397  Mom makes dress for daughter out of husband's …
17113  Roger Federer Practice 2014 BNP Paribas Open P…
20636  KEVIN HART VS MICHAEL DAPAAH - UK VS USA SLANG…

                                                   URL  \
26099  https://www.youtube.com/watch?v=51lKPhLMvi8
```

23297 https://www.youtube.com/watch?v=2zGa-edq9nk
16638 https://www.youtube.com/watch?v=VVCFVoqo1Tw
2913 https://www.youtube.com/watch?v=Ji_lt9WagMA
10846 https://www.youtube.com/watch?v=NFod-ozimmM
20428 https://www.youtube.com/watch?v=DKntq6YMnVg
2873 https://www.youtube.com/watch?v=jeYjtEX3RAE
16837 https://www.youtube.com/watch?v=4tD_SMikPI4
19283 https://www.youtube.com/watch?v=wKWi2ZCd-yQ
25606 https://www.youtube.com/watch?v=yzMDOSlsmpI
21327 https://www.youtube.com/watch?v=48TD9Babigw
23129 https://www.youtube.com/watch?v=rdeQT7KkqM8
21186 https://www.youtube.com/watch?v=UEXpPJSLP4A
12246 https://www.youtube.com/watch?v=fM_Gb9YBxd0
10578 https://www.youtube.com/watch?v=_3Wt9mRy014
13692 https://www.youtube.com/watch?v=qlzVPauUgw8
20759 https://www.youtube.com/watch?v=95-YCaj5HJQ
22195 https://www.youtube.com/watch?v=5RXtdSMX8ec
1361 https://www.youtube.com/watch?v=ljThMgchbPw
22148 https://www.youtube.com/watch?v=Rv9VC7ZPrH0
12327 https://www.youtube.com/watch?v=du035tg-SwY
21721 https://www.youtube.com/watch?v=-ncIVUXZla8
4429 https://www.youtube.com/watch?v=g2UYileH-P4
7207 https://www.youtube.com/watch?v=zOfGS7Rinvs
7349 https://www.youtube.com/watch?v=h61F2j_2GTE
15515 https://www.youtube.com/watch?v=hTMt0wpPfyY
1303 https://www.youtube.com/watch?v=csPWSyBI4XM
25774 https://www.youtube.com/watch?v=Egi0wL9jTlw
6202 https://www.youtube.com/watch?v=C2Kus5Gph4g
25865 https://www.youtube.com/watch?v=3hZPQSdlPEY
21974 https://www.youtube.com/watch?v=YPhO5g39vpg
11400 https://www.youtube.com/watch?v=SPdbVKiuyAM
10575 https://www.youtube.com/watch?v=ekLH2OaSefI
26 https://www.youtube.com/watch?v=D1k5w1fC4us
17711 https://www.youtube.com/watch?v=pOLmD_WVY-E
19396 https://www.youtube.com/watch?v=c5xQ2s428FM
16132 https://www.youtube.com/watch?v=X2JHgZLNuPw
7458 https://www.youtube.com/watch?v=Y-OLlJUXwKU
2802 https://www.youtube.com/watch?v=77ra2oxMfJ0
22810 https://www.youtube.com/watch?v=FnLzNpDZO_M
12507 https://www.youtube.com/watch?v=s9cABawWV7Q
4475 https://www.youtube.com/watch?v=B4KY0TYOyCE
22540 https://www.youtube.com/watch?v=TV3iMgQAekM
1125 https://www.youtube.com/watch?v=nMBml1iNvtk
15796 https://www.youtube.com/watch?v=pOM9A-kqUwA
20372 https://www.youtube.com/watch?v=TqKKjvq4o_U
23669 https://www.youtube.com/watch?v=KmwxpFzEx6g
21397 https://www.youtube.com/watch?v=iGgM4TfB3HQ

```
17113  https://www.youtube.com/watch?v=LE15sEsJZso
20636  https://www.youtube.com/watch?v=fh2zSxwTU44

                                                  Channel time_24hr  \
26099  [{'name': 'TheEllenShow', 'url': 'https://www…  01:58:35
23297  [{'name': 'Clash Tournament', 'url': 'https://…  09:04:15
16638  [{'name': 'Rockloe', 'url': 'https://www.youtu…  03:46:10
2913   [{'name': 'CNBC', 'url': 'https://www.youtube…  21:04:13
10846  [{'name': 'GothamChess', 'url': 'https://www.y…  18:57:09
20428  [{'name': 'Furia G', 'url': 'https://www.youtu…  03:23:02
2873   [{'name': 'Kenji Explains', 'url': 'https://ww…  23:56:01
16837  [{'name': 'Jimi Jackson', 'url': 'https://www…  12:24:33
19283  [{'name': 'Woody & Kleiny', 'url': 'https://ww…  02:25:45
25606  [{'name': 'WWE', 'url': 'https://www.youtube.c…  15:39:53
21327  [{'name': 'william coach en seduction humorist…  16:30:39
23129  [{'name': 'BuzzFeedVideo', 'url': 'https://www…  08:22:38
21186  [{'name': 'Maroon 5', 'url': 'https://www.yout…  05:31:31
12246  [{'name': 'CNBC', 'url': 'https://www.youtube…  11:05:47
10578  [{'name': 'Ludwig Reacts', 'url': 'https://www…  18:04:06
13692  [{'name': 'Dude Perfect', 'url': 'https://www…  08:31:36
20759  [{'name': 'Smosh', 'url': 'https://www.youtube…  23:15:49
22195  [{'name': 'EpicMashups', 'url': 'https://www.y…  06:30:16
1361   [{'name': 'Asa Lachesa', 'url': 'https://www.y…  07:19:17
22148  [{'name': 'Looper', 'url': 'https://www.youtub…  08:35:02
12327  [{'name': 'The School of Life', 'url': 'https:…  13:42:52
21721  [{'name': 'AviciiOfficialVEVO', 'url': 'https:…  16:04:58
4429   [{'name': 'Nicholas Crown', 'url': 'https://ww…  03:56:55
7207   [{'name': 'The Economist', 'url': 'https://www…  22:33:13
7349   [{'name': 'Tennis is Life', 'url': 'https://ww…  17:05:23
15515  [{'name': 'David Dobrik', 'url': 'https://www…  03:42:14
1303   [{'name': 'Iron Chef Dad', 'url': 'https://www…  11:17:30
25774  [{'name': 'WWE', 'url': 'https://www.youtube.c…  09:39:47
6202   [{'name': 'Silicon Valley Girl', 'url': 'https…  07:42:48
25865  [{'name': 'WWE', 'url': 'https://www.youtube.c…  09:22:25
21974  [{'name': 'Artie Cortés', 'url': 'https://www…  16:34:27
11400  [{'name': 'Magnus Carlsen', 'url': 'https://ww…  03:27:28
10575  [{'name': 'CNBC International', 'url': 'https:…  16:59:21
26     [{'name': 'FeedMi  ', 'url': 'https://www.yout…  01:12:25
17711  [{'name': 'TED-Ed', 'url': 'https://www.youtub…  06:15:17
19396  [{'name': 'truTV', 'url': 'https://www.youtube…  06:30:49
16132  [{'name': 'Channel 4 News', 'url': 'https://ww…  03:11:38
7458   [{'name': 'Bloomberg Originals', 'url': 'https…  22:58:21
2802   [{'name': 'Wade Papenfus - Offshore Tales', 'u…  17:21:03
22810  [{'name': 'Liverpool FC', 'url': 'https://www…  12:38:23
12507  [{'name': 'Island ECC', 'url': 'https://www.yo…  07:40:20
4475   [{'name': 'James Seo', 'url': 'https://www.you…  20:59:00
22540  [{'name': 'BuzzFeedVideo', 'url': 'https://www…  10:47:13
```

1125	[{'name': 'sidequestz', 'url': 'https://www.yo…	17:56:08
15796	[{'name': 'THE HACK', 'url': 'https://www.yout…	11:33:01
20372	[{'name': 'Memeulous', 'url': 'https://www.you…	08:52:05
23669	[{'name': 'TheThings Celebrity', 'url': 'https…	08:42:02
21397	[{'name': 'Good Morning America', 'url': 'http…	05:49:58
17113	[{'name': 'TheUSTATennisPlayer', 'url': 'https…	16:54:46
20636	[{'name': 'Michael Dapaah', 'url': 'https://ww…	03:44:38

	date	13 clusters	5 clusters	11 clusters
26099	2016-01-10	1	0	0
23297	2018-03-04	1	0	0
16638	2019-08-22	1	0	0
2913	2023-06-09	1	2	0
10846	2021-02-24	1	0	0
20428	2018-10-27	1	0	0
2873	2023-06-12	1	0	0
16837	2019-08-09	1	0	0
19283	2019-03-01	1	0	0
25606	2016-08-01	1	2	0
21327	2018-08-13	1	0	0
23129	2018-03-26	1	2	0
21186	2018-08-17	1	0	0
12246	2020-10-28	1	0	0
10578	2021-03-15	1	0	0
13692	2020-05-06	1	0	0
20759	2018-09-19	1	1	0
22195	2018-06-02	1	2	0
1361	2023-09-04	1	0	0
22148	2018-06-11	1	0	0
12327	2020-10-11	1	0	0
21721	2018-07-26	1	0	0
4429	2022-11-10	1	0	0
7207	2022-03-05	1	0	0
7349	2022-03-03	1	0	0
15515	2019-11-18	1	0	0
1303	2023-09-05	1	0	0
25774	2016-07-27	1	2	0
6202	2022-05-19	1	0	0
25865	2016-07-24	1	2	0
21974	2018-07-03	1	2	0
11400	2020-12-30	1	0	0
10575	2021-03-16	1	0	0
26	2024-01-22	1	0	0
17711	2019-06-02	1	0	0
19396	2019-02-24	1	0	0
16132	2019-10-05	1	0	0
7458	2022-02-25	1	0	0

2802	2023-06-18	1	0	0
22810	2018-04-23	1	0	0
12507	2020-09-07	1	0	0
4475	2022-11-05	1	0	0
22540	2018-05-10	1	0	0
1125	2023-09-08	1	0	0
15796	2019-10-31	1	0	0
20372	2018-10-30	1	0	0
23669	2018-01-07	1	0	0
21397	2018-08-10	1	0	0
17113	2019-07-28	1	0	0
20636	2018-09-29	1	0	0

The clusters we're getting now aren't very good, after having tried different numbers of clusters.

Perhaps we need to try a better approach for labeling video categories

4 Training on existing dataset

I found a Kaggle dataset with video titles and categories labeled. It also has video descriptions, but we'll try to train on video titles only for now.

Credits to dataset owner:

Title: Youtube-video-dataset

Author: Rahul Anand

URL: https://www.kaggle.com/datasets/rahulanand0070/youtubevideodataset?resource=download

Logistic Regression

Since we're dealing with categories, logistic regression is a robust candidate for a dataset of this size

[8]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import make_pipeline
from sklearn.metrics import classification_report
import statsmodels.api as sm

kaggle_data = pd.read_csv("../Data Project/Youtube Video Dataset.csv")
```

[53]:
```
X = kaggle_data['Title'].tolist()
processed_texts = [processing(title) for title in X]
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(processed_texts)

y = kaggle_data['Category']
```

```
category_encoder = LabelEncoder()
y_encoded = category_encoder.fit_transform(y)

#Training/Test split
random.seed(42)
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.
 ↪35)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

# Print a classification report
print(classification_report(y_test, y_pred, target_names=label_encoder.
 ↪classes_))
```

```
                    precision    recall  f1-score   support

         Art&Music       0.99      0.97      0.98       344
              Food       0.95      0.93      0.94       358
           History       0.99      0.96      0.98       342
Science&Technology       0.84      0.98      0.90       394
     manufacturing       0.99      0.90      0.94       357
       travel blog       0.96      0.94      0.95       448

          accuracy                           0.95      2243
         macro avg       0.95      0.95      0.95      2243
      weighted avg       0.95      0.95      0.95      2243
```

[54]:
```
real_data = watch_df["title"].tolist()
real_processed = [processing(title) for title in real_data]
realX = vectorizer.transform(real_processed)


encoded_preds = model.predict(realX)
predicted_categories = label_encoder.inverse_transform(encoded_preds)


watch_df['Category Prediction'] = predicted_categories
```

The model is OK, but predictions are quite mediocre. We should train on a different dataset, larger with more category labels.

DATASET 2

Title: Trending YouTube Video Statistics

Author: Mitchell J, username "datasnaek"

URL: https://www.kaggle.com/datasets/datasnaek/youtube-new?select=USvideos.csv

We can combine the US, GB, CA video csv files to make a larger dataset which will be better for model training. To avoid overtraining on repeated video titles, we will remove titles that occur more than once

We can also combine it with this following similar dataset with different data entries

Title: Trending YouTube Video Statistics and Comments

Author: Mitchell J, username "datasnaek"

URL: https://www.kaggle.com/datasets/datasnaek/youtube?resource=download&select=GBvideos.csv

[194]:
```
with open('../Data Project/US_category_id.json', 'r') as file:
    Category_IDs = json.load(file)
Categories_dict = {block['id']: block['snippet']['title'] for block in␣
 ↪Category_IDs['items']}
```

[195]:
```
US_data = pd.read_csv("../Data Project/USvideos.csv")
CA_data = pd.read_csv("../Data Project/CAvideos.csv")
GB_data = pd.read_csv("../Data Project/GBvideos.csv")
combined_kaggle_1 = pd.concat([US_data, CA_data, GB_data], ignore_index=True)
combined_kaggle_1 = combined_kaggle_1[['video_id', 'title', 'channel_title',␣
 ↪'category_id']]
```

[196]:
```
US2_data = pd.read_csv("../Data Project/USvideos2.csv")
GB2_data = pd.read_csv("../Data Project/GBvideos2.csv")
US2_data = US2_data[['video_id', 'title', 'channel_title', 'category_id']]
GB2_data = GB2_data[['video_id', 'title', 'channel_title', 'category_id']]
```

[197]:
```
combined_kaggle_2 = pd.concat([combined_kaggle_1, US2_data, GB2_data],␣
 ↪ignore_index=True)
combined_kaggle_2 = combined_kaggle_2.drop_duplicates(subset=['video_id']) #␣
 ↪don't want to overtrain on the same videos
```

[198]:
```
#We want to reduce the weight of "family" category. We might consider removing␣
 ↪it altogether
heaviest = combined_kaggle_2[combined_kaggle_2['category_id'] == 24]
sampled_heaviest = heaviest.sample(n= 5000, random_state=42)
rest_df = combined_kaggle_2[combined_kaggle_2['category_id'] != 24]
combined_kaggle_2 = pd.concat([sampled_heaviest, rest_df])
```

Too many categories, making it too difficult to train our algorithm

[199]:
```
broader_categories = {
    'Entertainment': [1, 18, 23, 24, 30, 31, 32, 33, 34, 36, 37, 39, 40, 41,␣
 ↪42, 43, 44],
    'Lifestyle & How-to': [15, 21, 22, 26],
    'Music & Performing Arts': [10],
```

```
    'Science & Education': [28, 27, 35],
    'Sports & Gaming': [17, 20],
    'Vehicles & Technology': [2],
    'Travel & International': [19, 38],
    'News & Social Commentary': [8, 25, 29]
}

def map_to_group(cat_id):
    for group, ids in broader_categories.items():
        if cat_id in ids:
            return group
    return "unassigned" #else case

combined_kaggle_2['broad_category'] = combined_kaggle_2['category_id'].
 ↪apply(map_to_group)
encoder = LabelEncoder()
combined_kaggle_2['broad_category_encoded'] = encoder.
 ↪fit_transform(combined_kaggle_2['broad_category'])
combined_kaggle_2
```

[199]:

```
           video_id                                                title  \
77088   QVklmYejCVk                Denya Okhra S03 Episode 05 Partie 02
22566   vjnqABgxfO0                  The Grinch - Official Trailer (HD)
10418   45dj8U3xSFU  Volatile Owner Tears into Customer Over Microw…
50170   LjNhW2mR1lM            Black Mirror | Happy New Year | Netflix
51599   jQkZRWoE-wM                  Deivamagal Episode 1432, 05/01/18
…               …                                                   …
136623  V3jY8-VzSEA        Gucci Mane - Back On [Official Music Video]
136626  wOcRH5GbqTQ         Knuckle Puck - Gone (Official Music Video)
136642  NRjDqw1w7k8    ZHU - Waters of Monaco (Adidas China Pure Edit)
136649  K2SCpgCurVQ  Destiny 2: Wardcliff Coil Glitch! How To Get U…
136669  XQFeShp6UIY  World's First Prestige Leviathan Raid Completi…

             channel_title  category_id           broad_category  \
77088     Elhiwar Ettounsi           24            Entertainment
22566          Illumination          24            Entertainment
10418   Kitchen Nightmares           24            Entertainment
50170              Netflix           24            Entertainment
51599            VikatanTV           24            Entertainment
…                        …          …                        …
136623   OfficialGucciMane           10  Music & Performing Arts
136626         riserecords           10  Music & Performing Arts
136642             ZHUVEVO           10  Music & Performing Arts
136649                Kota           20          Sports & Gaming
136669               Gladd           20          Sports & Gaming

        broad_category_encoded
```

```
77088                        0
22566                        0
10418                        0
50170                        0
51599                        0
…                          …
136623                       2
136626                       2
136642                       2
136649                       5
136669                       5

[28218 rows x 6 columns]
```

[200]:
```
X = combined_kaggle_2['title'].tolist()
processed_texts = [processing(title) for title in X]
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(processed_texts)

y = combined_kaggle_2['broad_category_encoded']

#Training/Test split
random.seed(42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.58      0.83      0.68      1833
           1       0.63      0.59      0.61      1297
           2       0.84      0.64      0.73       568
           3       0.73      0.60      0.66       766
           4       0.69      0.27      0.39       368
           5       0.90      0.67      0.77       753
           6       0.84      0.27      0.41        59

    accuracy                           0.66      5644
   macro avg       0.74      0.55      0.61      5644
weighted avg       0.69      0.66      0.66      5644
```

[201]:
```
real_data = watch_df["title"].tolist()
real_processed = [processing(title) for title in real_data]
realX = vectorizer.transform(real_processed)
preds = model.predict(realX)
watch_df['Category Predicted'] = encoder.inverse_transform(preds)
```

[202]:
```
watch_df
```

[202]:

```
                                                   title  \
0          British Actors Vs American Actors | Key & Peele
1                             CORVETTE DRIVER KEYS A TESLA
2                                 Key & Peele - Das Negros
3                            Key & Peele - Alien Imposters
4                  Key & Peele’s Not-So-Clever Criminals
…                                                      …
23926  Enzo Amore & Big Cass vs. The Shining Stars: R…
23927  Who wants to be the guest ref for Big Show vs…
23928       WWE: "Crank It Up"   Big Show 9th Theme Song
23929  Cena vs. Orton vs. Triple H vs. Big Show - Fat…
23930  Annoying Orange - Big Top Orange (Ft. Madagasc…

                                                    URL  \
0      https://www.youtube.com/watch?v=gOvn7r-GYC0
1      https://www.youtube.com/watch?v=RB-ARvWD3mk
2      https://www.youtube.com/watch?v=m1bLXk6UVts
3      https://www.youtube.com/watch?v=DWO1pkHgrBM
4      https://www.youtube.com/watch?v=qLvxc83GrM4
…                                                 …
23926  https://www.youtube.com/watch?v=nPoNdKKjZOQ
23927  https://www.youtube.com/watch?v=_sqLYLuAXxM
23928  https://www.youtube.com/watch?v=ZLazJKggCd0
23929  https://www.youtube.com/watch?v=L2uneW6tcyI
23930  https://www.youtube.com/watch?v=UU2AZyf6EGQ

                                                 Channel time_24hr  \
0      [{'name': 'Comedy Central UK', 'url': 'https:/…  15:45:29
1      [{'name': 'Wham Baam Teslacam', 'url': 'https:…  05:10:52
2      [{'name': 'Comedy Central', 'url': 'https://ww…  05:57:16
3      [{'name': 'Comedy Central', 'url': 'https://ww…  05:55:18
4      [{'name': 'Key & Peele', 'url': 'https://www.y…  05:45:08
…                                                    …         …
23926  [{'name': 'WWE', 'url': 'https://www.youtube.c…  10:25:25
23927  [{'name': 'WWE', 'url': 'https://www.youtube.c…  02:39:23
23928  [{'name': 'Saint', 'url': 'https://www.youtube…  05:00:58
23929  [{'name': 'WWE', 'url': 'https://www.youtube.c…  13:50:05
23930  [{'name': 'Annoying Orange', 'url': 'https://w…  10:04:35
```

```
             date  13 clusters  Category Predicted
0      2024-01-01            0       Entertainment
1      2023-02-03            0  Lifestyle & How-to
2      2023-01-17            0       Entertainment
3      2023-01-17            0       Entertainment
4      2023-01-17            0       Entertainment
…           …          …                 …
23926  2016-07-26           12     Sports & Gaming
23927  2016-07-24           12       Entertainment
23928  2016-07-23           12       Entertainment
23929  2016-07-22           12     Sports & Gaming
23930  2013-04-27           12       Entertainment

[23931 rows x 7 columns]
```

[203]:
```
watch_df[watch_df["Category Predicted"] == "News & Social Commentary"].
↪sample(20)
```

[203]:
```
                                                  title  \
1783                School Shooter  // #fallout #shorts
11521     What is the circular economy? | CNBC Explains
3984   G20: The Economist interviews Indonesia's pres…
11413  The Difference Between The Stock Market And Th…
10418    Why is the dollar so powerful? | CNBC Explains
1589                   Global Internship Interview Tips
8417                        Why ISIS would attack Paris
2539   Why China's Economy Doesn't Want American Corn…
15916                       She speaks fluent American
10744        2019 BuffCo May Time Trials 1500m - Heat 1
6095   How Erdogan's Strategy Backfired: Turkey's Eco…
11398    What is currency manipulation? | CNBC Explains
15718       Revolution 360: Preparations for the strike
9465   Why These Headhunters Converted to Christianit…
10377             How Airlines Park Thousands Of Planes
7305   The Barclays Trading Strategy that Outperforms…
7281                 Prison Viral- @thealexanderdenning
19599                        In the Unlikely Event…
1341            TJ Oshie Olympic Shootout (NBC English)
2683   "Are You Destined to Deal?" With Goldman Sachs…

                                                    URL  \
1783   https://www.youtube.com/watch?v=eBVpp-q7Y74
11521  https://www.youtube.com/watch?v=__0Spwj8DkM
3984   https://www.youtube.com/watch?v=a7vMDrm8jLo
11413  https://www.youtube.com/watch?v=59im9CtR9YI
10418  https://www.youtube.com/watch?v=kkkH-0khoQw
1589   https://www.youtube.com/watch?v=SmjL7ZVaxz0
```

```
8417   https://www.youtube.com/watch?v=3bIvqS7gnQo
2539   https://www.youtube.com/watch?v=YPUj0sFEfxU
15916  https://www.youtube.com/watch?v=K9iR6sLwDKY
10744  https://www.youtube.com/watch?v=46GepIa5ypM
6095   https://www.youtube.com/watch?v=dkyudz4w5Gc
11398  https://www.youtube.com/watch?v=wEbrdxWw7ew
15718  https://www.youtube.com/watch?v=-8PNPAfkY8k
9465   https://www.youtube.com/watch?v=oLs-UoqzLlU
10377  https://www.youtube.com/watch?v=xpIs8Y9vgSs
7305   https://www.youtube.com/watch?v=8pYgz4YlQnE
7281   https://www.youtube.com/watch?v=pxw5B9DLMW8
19599  https://www.youtube.com/watch?v=4BuOlKZ_C2k
1341   https://www.youtube.com/watch?v=MUxJXzKY4LE
2683   https://www.youtube.com/watch?v=RpUJfW4WTKw

                                                  Channel time_24hr  \
1783   [{'name': 'PB The Prince', 'url': 'https://www…  10:20:27
11521  [{'name': 'CNBC International', 'url': 'https:…  10:36:44
3984   [{'name': 'The Economist', 'url': 'https://www…  17:43:29
11413  [{'name': 'CNBC', 'url': 'https://www.youtube…  12:39:48
10418  [{'name': 'CNBC International', 'url': 'https:…  06:44:13
1589   [{'name': 'CIEE Study Abroad', 'url': 'https:/…  03:05:10
8417   [{'name': 'Vox', 'url': 'https://www.youtube.c…  19:20:12
2539   [{'name': 'The Wall Street Journal', 'url': 'h…  19:09:39
15916  [{'name': 'Sean Walsh', 'url': 'https://www.yo…  10:33:50
10744  [{'name': 'Hideo Harry Loasby', 'url': 'https:…  09:55:53
6095   [{'name': 'TLDR News Global', 'url': 'https://…  20:41:16
11398  [{'name': 'CNBC International', 'url': 'https:…  17:33:23
15718  [{'name': 'RT', 'url': 'https://www.youtube.co…  01:11:27
9465   [{'name': 'National Geographic', 'url': 'https…  11:27:12
10377  [{'name': 'CNBC', 'url': 'https://www.youtube…  11:27:45
7305   [{'name': 'Benjamin', 'url': 'https://www.yout…  00:23:23
7281   [{'name': 'Alexander the Great TV', 'url': 'ht…  23:57:53
19599  [{'name': 'Dropout', 'url': 'https://www.youtu…  05:54:12
1341   [{'name': 'Kevin B', 'url': 'https://www.youtu…  02:51:49
2683   [{'name': 'University of Virginia School of La…  08:44:03

            date  13 clusters        Category Predicted
1783   2023-09-20            2  News & Social Commentary
11521  2019-12-16            2  News & Social Commentary
3984   2022-11-28            2  News & Social Commentary
11413  2019-12-29            2  News & Social Commentary
10418  2020-04-25            2  News & Social Commentary
1589   2023-11-17            2  News & Social Commentary
8417   2021-02-03            2  News & Social Commentary
2539   2023-07-24            2  News & Social Commentary
15916  2018-08-24            2  News & Social Commentary
```

```
10744  2020-03-15            2  News & Social Commentary
6095   2022-01-27            2  News & Social Commentary
11398  2019-12-29            2  News & Social Commentary
15718  2018-09-21            2  News & Social Commentary
9465   2020-08-31            2  News & Social Commentary
10377  2020-05-04            2  News & Social Commentary
7305   2021-07-08            2  News & Social Commentary
7281   2021-07-08            2  News & Social Commentary
19599  2016-01-01            2  News & Social Commentary
1341   2023-12-25            2  News & Social Commentary
2683   2023-07-15            2  News & Social Commentary
```

Finally we have a working model! Our videos are now well categorized and our dataframe is clean (we did our best).

5 Part 2: Data Analysis and Visualization

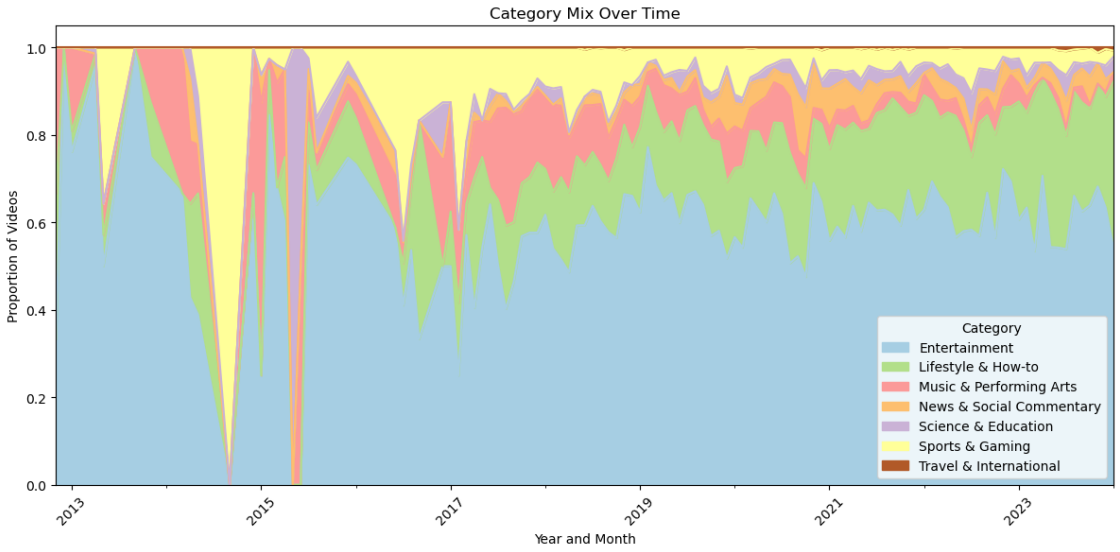
Category Mix over time

[212]:
```
watch_df['year_month'] = pd.to_datetime(watch_df['date'], format='%Y-%m-%d').dt.
↪to_period('M')
```

[212]:
```
                                                  title  \
0        British Actors Vs American Actors | Key & Peele
1                           CORVETTE DRIVER KEYS A TESLA
2                               Key & Peele - Das Negros
3                           Key & Peele - Alien Imposters
4                  Key & Peele's Not-So-Clever Criminals
…                                                    …
23926  Enzo Amore & Big Cass vs. The Shining Stars: R…
23927  Who wants to be the guest ref for Big Show vs…
23928       WWE: "Crank It Up"   Big Show 9th Theme Song
23929  Cena vs. Orton vs. Triple H vs. Big Show - Fat…
23930  Annoying Orange - Big Top Orange (Ft. Madagasc…

                                                    URL  \
0      https://www.youtube.com/watch?v=gOvn7r-GYC0
1      https://www.youtube.com/watch?v=RB-ARvWD3mk
2      https://www.youtube.com/watch?v=m1bLXk6UVts
3      https://www.youtube.com/watch?v=DWO1pkHgrBM
4      https://www.youtube.com/watch?v=qLvxc83GrM4
…                                                  …
23926  https://www.youtube.com/watch?v=nPoNdKKjZOQ
23927  https://www.youtube.com/watch?v=_sqLYLuAXxM
23928  https://www.youtube.com/watch?v=ZLazJKggCd0
23929  https://www.youtube.com/watch?v=L2uneW6tcyI
23930  https://www.youtube.com/watch?v=UU2AZyf6EGQ
```

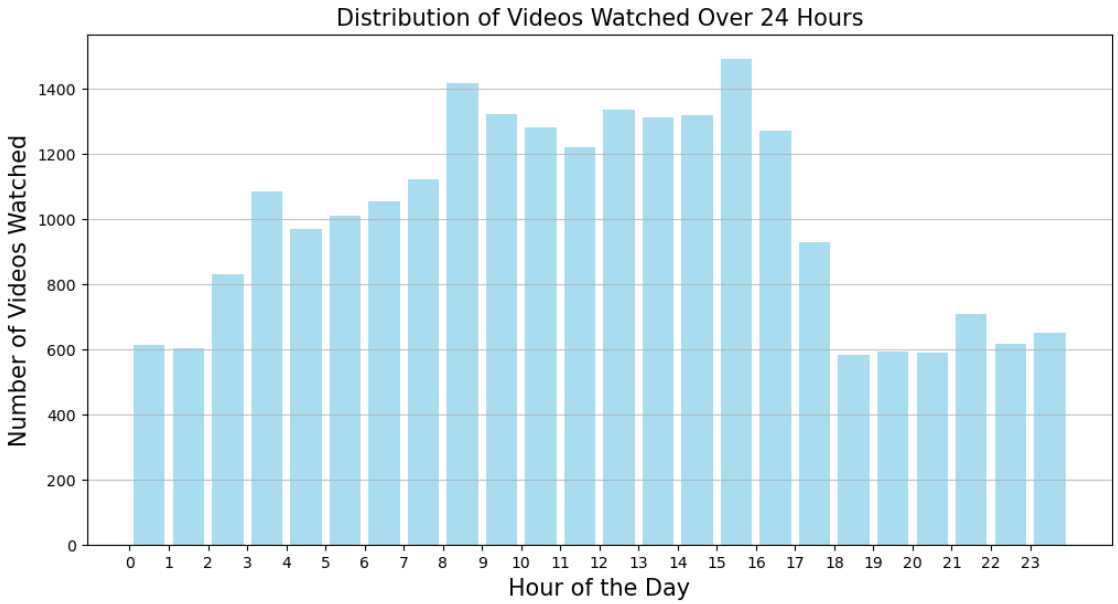
```
                                                  Channel time_24hr  \
0       [{'name': 'Comedy Central UK', 'url': 'https:/…  15:45:29
1       [{'name': 'Wham Baam Teslacam', 'url': 'https:…  05:10:52
2       [{'name': 'Comedy Central', 'url': 'https://ww…  05:57:16
3       [{'name': 'Comedy Central', 'url': 'https://ww…  05:55:18
4       [{'name': 'Key & Peele', 'url': 'https://www.y…  05:45:08
…                                                     …       …
23926   [{'name': 'WWE', 'url': 'https://www.youtube.c…  10:25:25
23927   [{'name': 'WWE', 'url': 'https://www.youtube.c…  02:39:23
23928   [{'name': 'Saint', 'url': 'https://www.youtube…  05:00:58
23929   [{'name': 'WWE', 'url': 'https://www.youtube.c…  13:50:05
23930   [{'name': 'Annoying Orange', 'url': 'https://w…  10:04:35

             date  13 clusters  Category Predicted date_column year_month
0      2024-01-01            0       Entertainment  2024-01-01    2024-01
1      2023-02-03            0  Lifestyle & How-to  2023-02-03    2023-02
2      2023-01-17            0       Entertainment  2023-01-17    2023-01
3      2023-01-17            0       Entertainment  2023-01-17    2023-01
4      2023-01-17            0       Entertainment  2023-01-17    2023-01
…             …            …                   …           …         …
23926  2016-07-26           12     Sports & Gaming  2016-07-26    2016-07
23927  2016-07-24           12       Entertainment  2016-07-24    2016-07
23928  2016-07-23           12       Entertainment  2016-07-23    2016-07
23929  2016-07-22           12     Sports & Gaming  2016-07-22    2016-07
23930  2013-04-27           12       Entertainment  2013-04-27    2013-04

[23931 rows x 9 columns]
```

[235]:
```
category_time_data = watch_df.groupby(['year_month', 'Category Predicted']).
 ↪size().unstack(fill_value=0)
category_time_data_normalized = category_time_data.div(category_time_data.
 ↪sum(axis=1), axis=0)
fig, ax = plt.subplots(figsize=(12, 6))
category_time_data_normalized.plot(kind='area', stacked=True, colormap=␣
 ↪'Paired', ax=ax)
ax.set_title('Category Mix Over Time')
ax.set_xlabel('Year and Month')
ax.set_ylabel('Proportion of Videos' if 'category_time_data_normalized' in␣
 ↪locals() else 'Number of Videos')
ax.legend(title='Category')

plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Youtube Activity hour of day distribution

[234]:
```
watch_df['hour'] = pd.to_datetime(watch_df['time_24hr'], format='%H:%M:%S').dt.
↪hour
plt.figure(figsize=(12, 6))
plt.hist(watch_df['hour'], bins=24, range=(0, 24), rwidth=0.8, color='skyblue',␣
↪alpha=0.7)
plt.xticks(range(0, 24))
plt.grid(axis='y', alpha=0.75)
plt.xlabel('Hour of the Day', fontsize=15)
plt.ylabel('Number of Videos Watched', fontsize=15)
plt.title('Distribution of Videos Watched Over 24 Hours', fontsize=15)
plt.show()
```



Year to year activity

[241]:
```
watch_df['year'] = pd.to_datetime(watch_df['date'], format='%Y-%m-%d').dt.year
watch_df['month'] = pd.to_datetime(watch_df['date'], format='%Y-%m-%d').dt.month
```

[242]:
```
watch_df
```

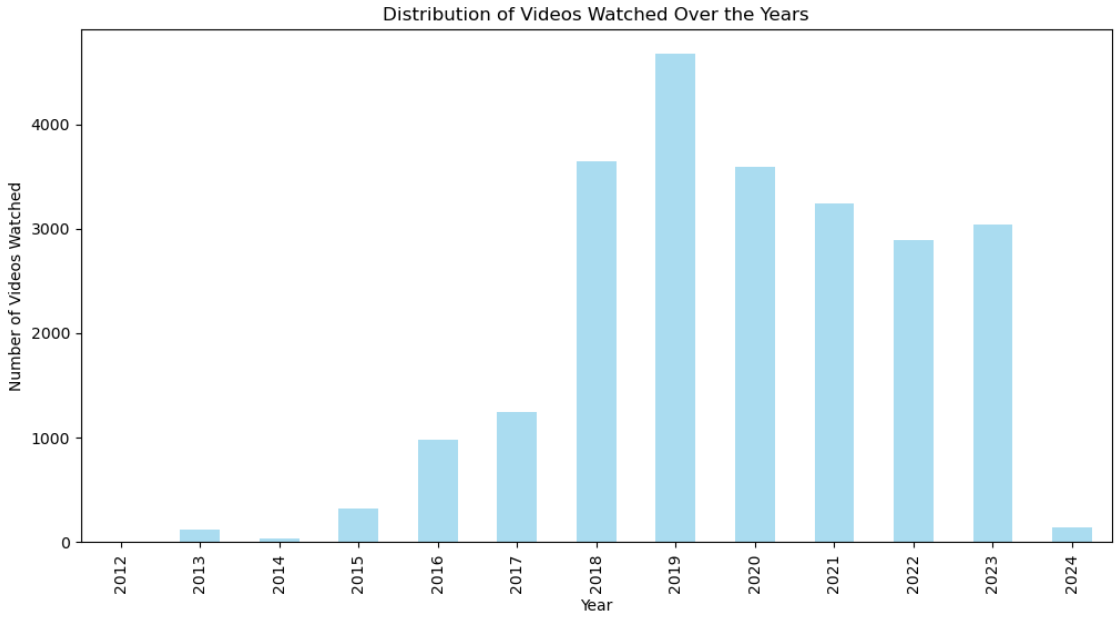
[242]:

```
                                                   title  \
0        British Actors Vs American Actors | Key & Peele
1                           CORVETTE DRIVER KEYS A TESLA
2                               Key & Peele - Das Negros
3                           Key & Peele - Alien Imposters
4                  Key & Peele's Not-So-Clever Criminals
…                                                    …
23926  Enzo Amore & Big Cass vs. The Shining Stars: R…
23927  Who wants to be the guest ref for Big Show vs…
23928       WWE: "Crank It Up"   Big Show 9th Theme Song
23929  Cena vs. Orton vs. Triple H vs. Big Show - Fat…
23930  Annoying Orange - Big Top Orange (Ft. Madagasc…

                                                   URL  \
0       https://www.youtube.com/watch?v=gOvn7r-GYC0
1       https://www.youtube.com/watch?v=RB-ARvWD3mk
2       https://www.youtube.com/watch?v=m1bLXk6UVts
3       https://www.youtube.com/watch?v=DWO1pkHgrBM
4       https://www.youtube.com/watch?v=qLvxc83GrM4
…                                               …
```

```
23926  https://www.youtube.com/watch?v=nPoNdKKjZOQ
23927  https://www.youtube.com/watch?v=_sqLYLuAXxM
23928  https://www.youtube.com/watch?v=ZLazJKggCd0
23929  https://www.youtube.com/watch?v=L2uneW6tcyI
23930  https://www.youtube.com/watch?v=UU2AZyf6EGQ

                                                 Channel time_24hr  \
0      [{'name': 'Comedy Central UK', 'url': 'https:/…  15:45:29
1      [{'name': 'Wham Baam Teslacam', 'url': 'https:…  05:10:52
2      [{'name': 'Comedy Central', 'url': 'https://ww…  05:57:16
3      [{'name': 'Comedy Central', 'url': 'https://ww…  05:55:18
4      [{'name': 'Key & Peele', 'url': 'https://www.y…  05:45:08
…                                                    …       …
23926  [{'name': 'WWE', 'url': 'https://www.youtube.c…  10:25:25
23927  [{'name': 'WWE', 'url': 'https://www.youtube.c…  02:39:23
23928  [{'name': 'Saint', 'url': 'https://www.youtube…  05:00:58
23929  [{'name': 'WWE', 'url': 'https://www.youtube.c…  13:50:05
23930  [{'name': 'Annoying Orange', 'url': 'https://w…  10:04:35

             date  13 clusters  Category Predicted date_column year_month  \
0      2024-01-01            0       Entertainment  2024-01-01    2024-01
1      2023-02-03            0  Lifestyle & How-to  2023-02-03    2023-02
2      2023-01-17            0       Entertainment  2023-01-17    2023-01
3      2023-01-17            0       Entertainment  2023-01-17    2023-01
4      2023-01-17            0       Entertainment  2023-01-17    2023-01
…             …            …                   …           …         …
23926  2016-07-26           12     Sports & Gaming  2016-07-26    2016-07
23927  2016-07-24           12       Entertainment  2016-07-24    2016-07
23928  2016-07-23           12       Entertainment  2016-07-23    2016-07
23929  2016-07-22           12     Sports & Gaming  2016-07-22    2016-07
23930  2013-04-27           12       Entertainment  2013-04-27    2013-04

       hour  year  month
0        15  2024      1
1         5  2023      2
2         5  2023      1
3         5  2023      1
4         5  2023      1
…     …   …    …
23926    10  2016      7
23927     2  2016      7
23928     5  2016      7
23929    13  2016      7
23930    10  2013      4

[23931 rows x 12 columns]
```

[244]:
```
yearly_tally = watch_df.groupby('year').size()
plt.figure(figsize=(12, 6))
yearly_tally.plot(kind='bar', color='skyblue', alpha=0.7)
plt.title('Distribution of Videos Watched Over the Years')
plt.xlabel('Year')
plt.ylabel('Number of Videos Watched')
```

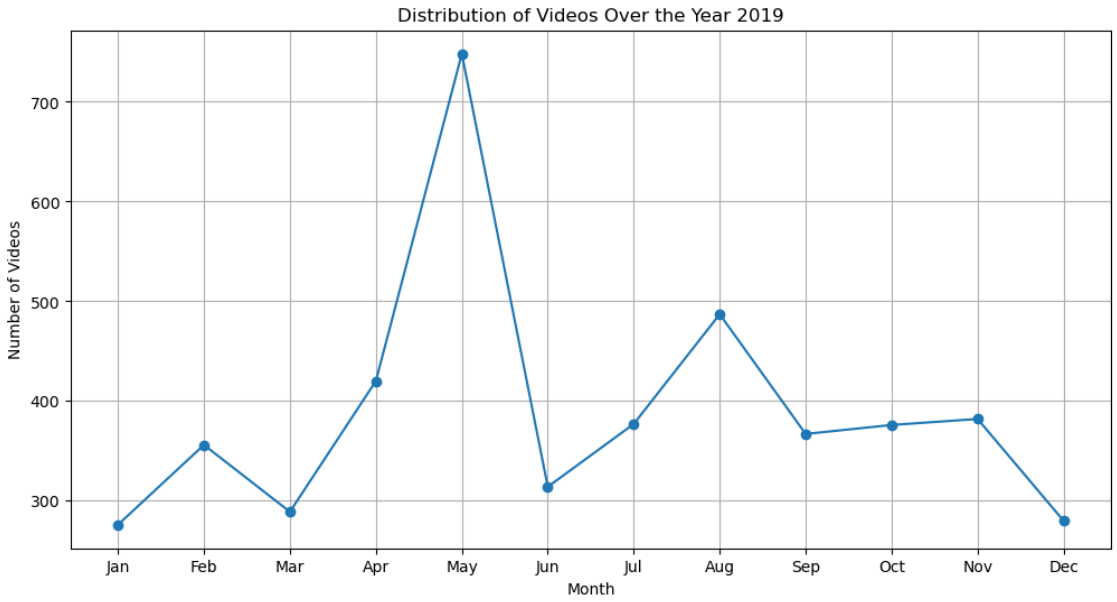
[244]:
```
Text(0, 0.5, 'Number of Videos Watched')
```



Makes sense as 2019-2021 were the peak covid years! Let's examine if the video count in 2019 coincides with the pandemic outbreak.

Month to month in 2019

[251]:
```
watch_df_2019 = watch_df[watch_df['year'] == 2019]
monthly_tally = watch_df_2019.groupby('month').size()

plt.figure(figsize=(12, 6))
monthly_tally.plot(kind='line', marker='o')
plt.title('Distribution of Videos Over the Year 2019')
plt.xlabel('Month')
plt.ylabel('Number of Videos')
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul',␣
 ↪'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.grid(True)
plt.show()
```



No correlation with Covid suggested here!

[ ]: