Analyzing YouTube's Trending Page

Presented by:
Ethan Villalovoz
Logan Sutton
Berkeley Conkling
Roy Zabetski
Kyle Hawkins
Chance Bradford
Matthew Bruggeman
Wenjie Wang
Silvestre Pamatz-Rangel

Group 1

The Data

- “Trending Youtube Videos Statistics and Comments”
  - Includes data on daily trending YouTube videos
    - Comments (over 690,000)
    - Views
    - Likes
    - Video Category
    - Tags
  - Example data format:

video_id	title	channel_title	category_id	tags	views	likes	dislikes	comment_total
cLdxuaxaQwc	My Response	PewDiePie	22	[none]	5845909	576597	39774	170708

Our agenda

- Used three algorithms to analyze trending page
  1. Analyzing Relationships
  2. Association Rule Mining (Apriori)
  3. Regression Analysis
     a. Linear Regression
     b. Logistic Regression
- Summarize findings and difficulties

Analyzing Relationships

Analyzing Relationships: Our Goal

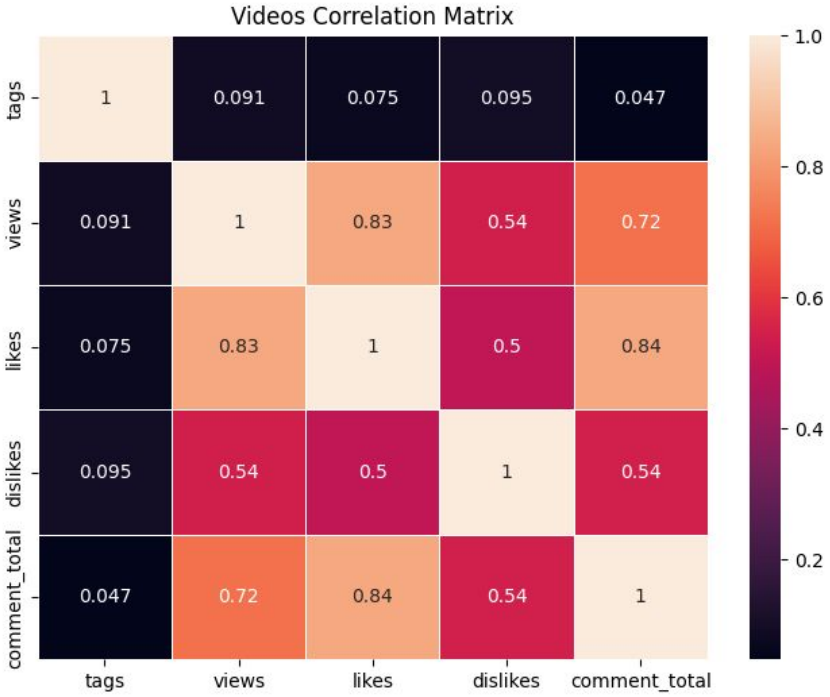
- Find relationships between numerical data categories in terms of strength and direction of the linear relationship between two variables
- Gain valuable insight into dataset relationships to aid in further data mining tasks
- Discover relationships of the video category classifier
- Aid in classification

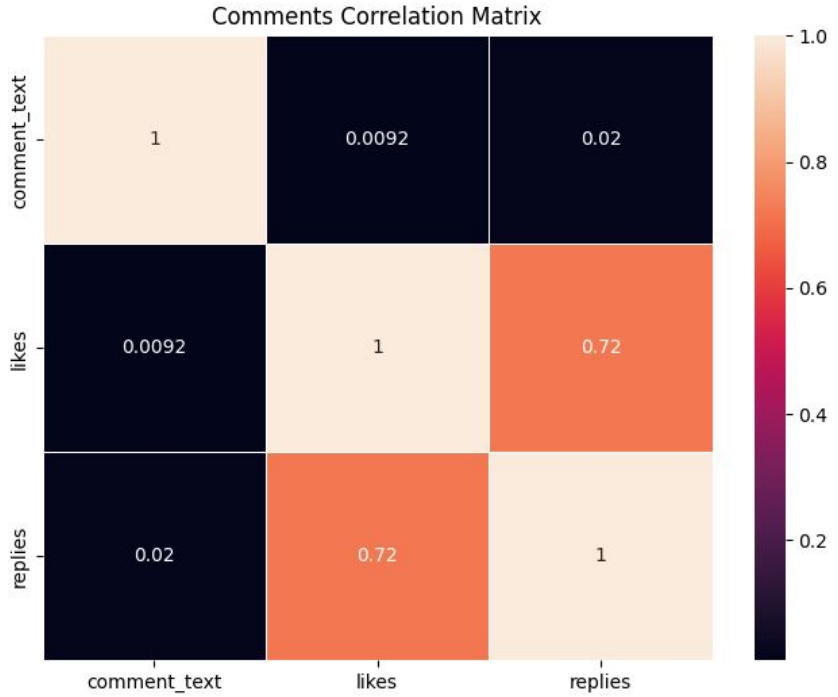
How it Works

- Correlation Analysis:
  - Used to find relationship between two variables/datasets
  - Strong relationships are close to 1
  - Utilized Pearson Correlation Coefficient

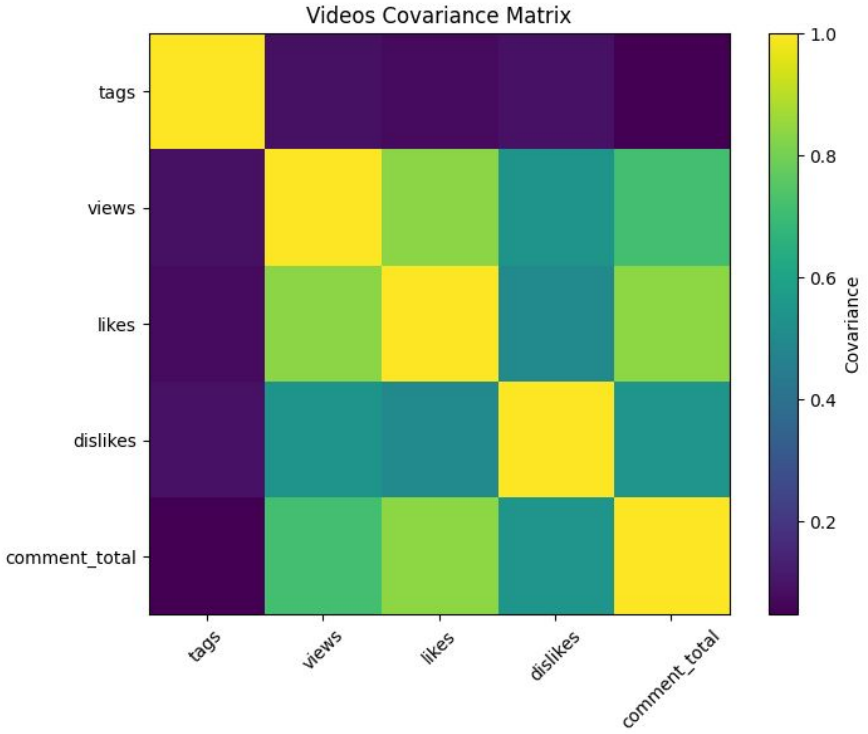
- Covariance Analysis
  - Measures the direction of the relationship between two variables
    - Positive covariance means both variables tend to be high or low at the same time.
    - Negative covariance means that when one variable is high, the other tends to be low
  - Normalized dataset values for calculation

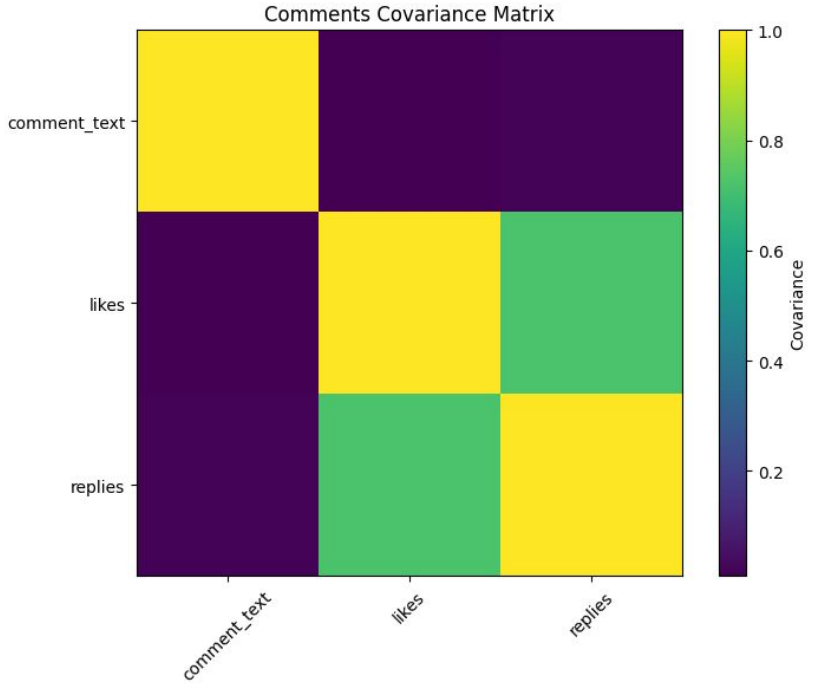
Correlation matrices:





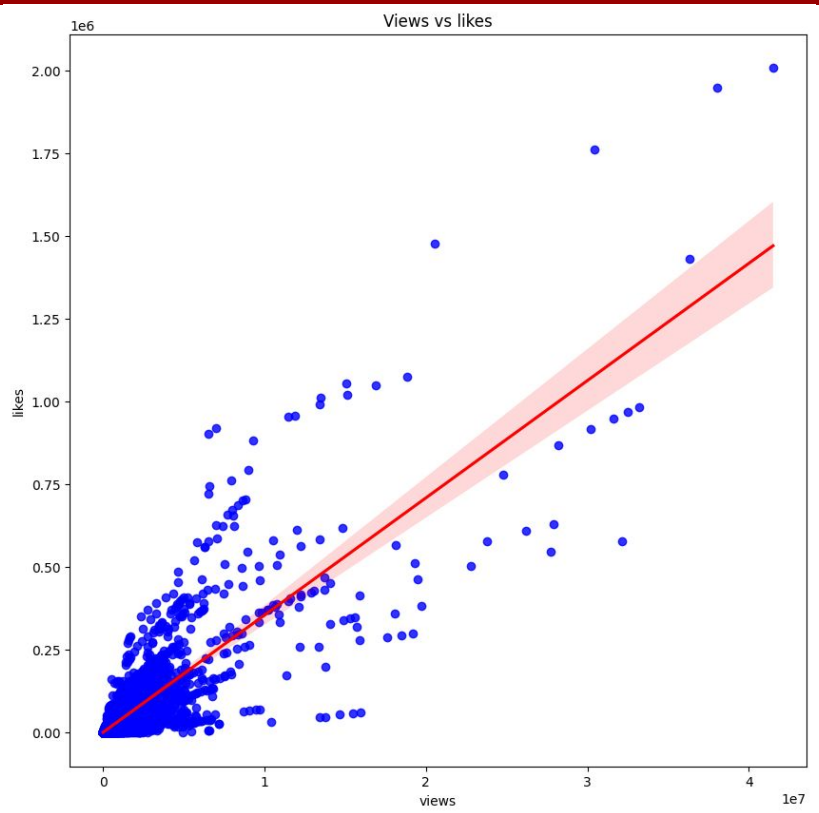
Covariance matrices:



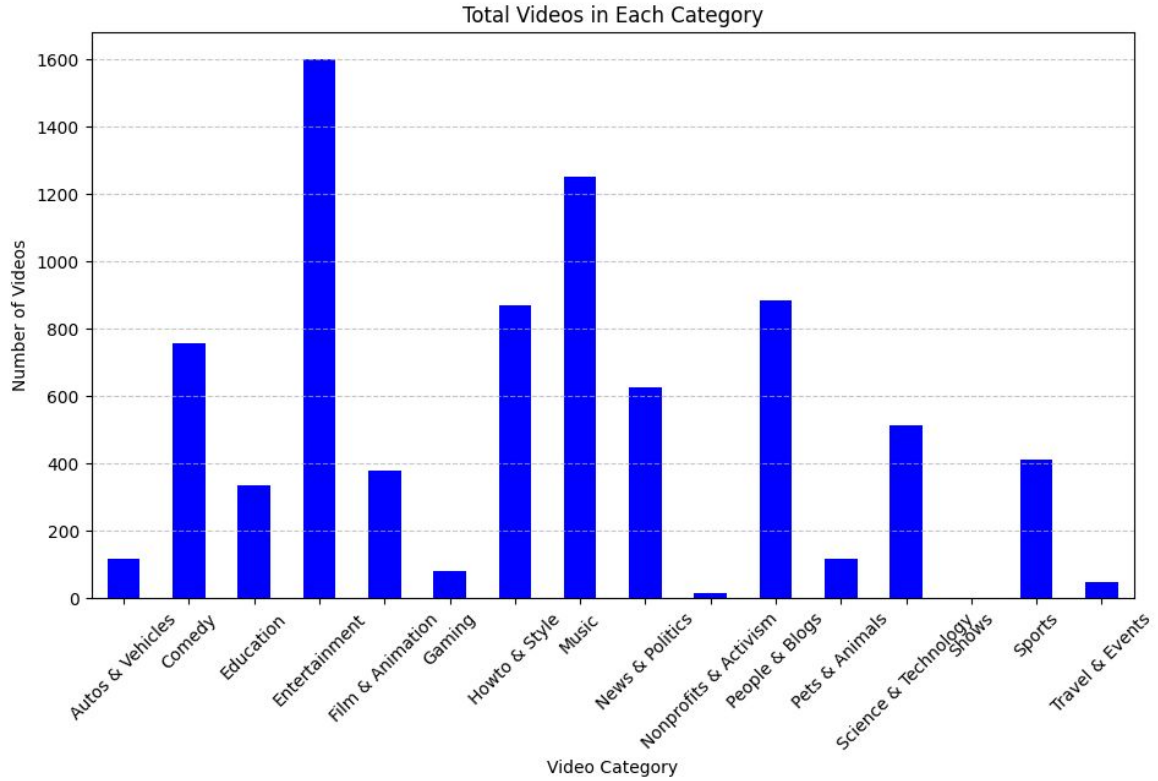


Correlation Results

- Appears to be strong relationship between view count and likes (user engagement)
- View to like ratio is linearly related
- Slope = 0.0354

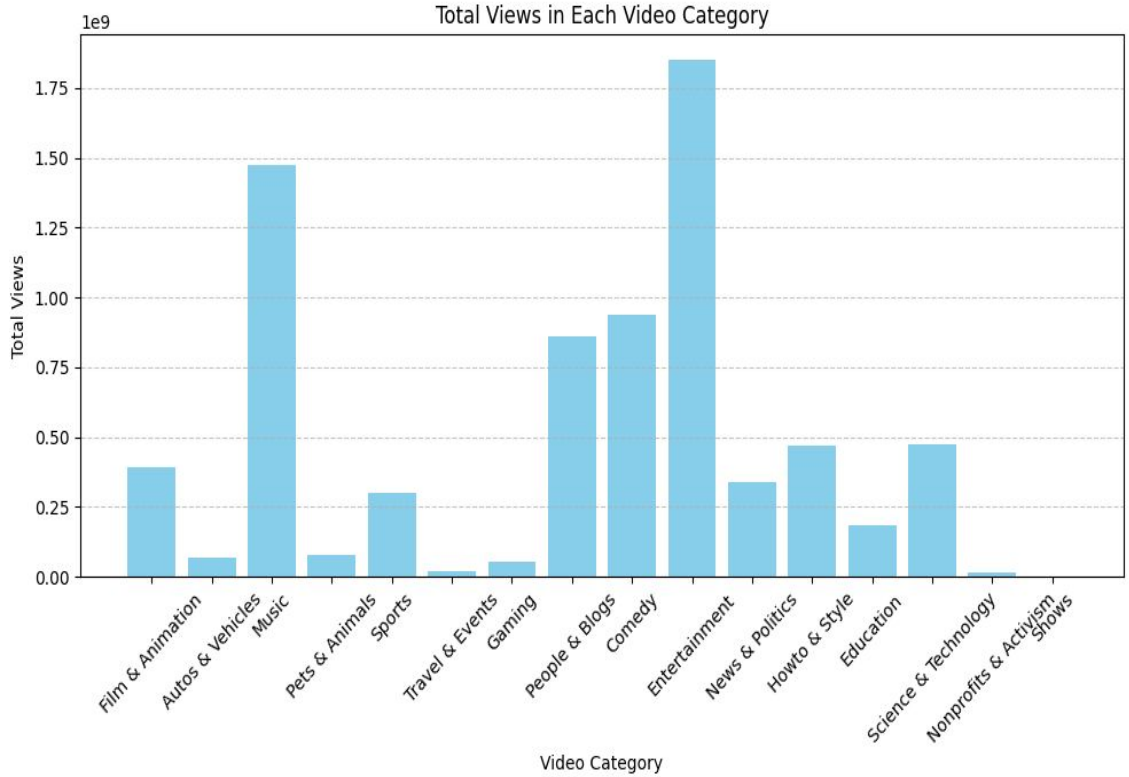


Results



- Entertainment, music, howto, and nonprofits are some of the most commonly trending categories
- Gaming, news, travel, and shows are rarely on the trending page

Results Cont.



- Entertainment, music and comedy receive the most views per category
- Results are inline with total videos per category

Analyzing Relationships: Summary/Reflection

- **Correlation matrix analysis**
  - User engagement (likes, dislikes, comments) is highly correlated with video views
    - i.e. more engagement means more views
  - Likes and replies are strongly correlated for comment data
    - i.e. more likes a comment receives mean more replies
- **Covariance matrix analysis**
  - Similar findings to correlation matrix
    - As views go up so does viewer engagement (likes, dislikes, comments)
  - Higher number of video tags do not correlate to increased video views
  - Number of words in comments does not correlate to comment likes or replies

Association Rule Mining

Association Rule (Apriori Algorithm)

- Our goal:
  - Find which categories of YouTube videos tend to attract high view counts more frequently
- How we will do this:
  - Discover frequent items that are associated with high, medium, and low views
  - Example: ['Comedy', 'High Views']

Additional Findings

- Is there association between video category and view amount?
  - Partition view amounts into thresholds
- What are frequent words found in a video's title?
  - Exclude trivial words such as "the", "and", "a", etc.
- Is there a linear relationship between video title length, number of tags, and view count?

Additional Findings

- Top correlations between video category and view count
  - Expect funny videos to have higher view count
- Most common non-trivial words found in video's titles
  - Expecting words like: "music", "trailer", etc.
- Video title length might have a small relationship with view count
- Higher number of tags could be attributed to higher view count

Association Rule

- How it works
  - Equal depth partitioning: three bins of low, medium, and high views
  - Two key functions: apriori() and getConfidenceItemsTop5()

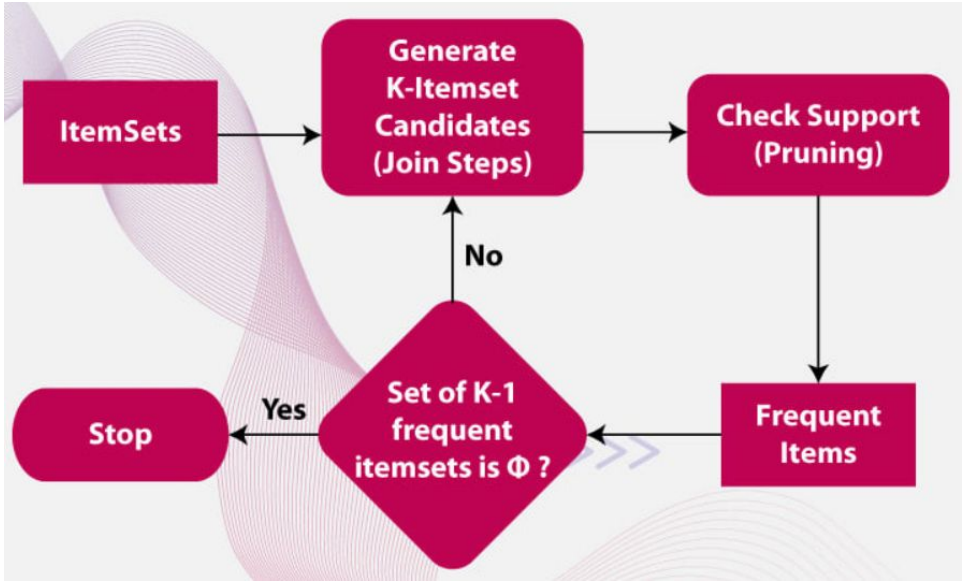


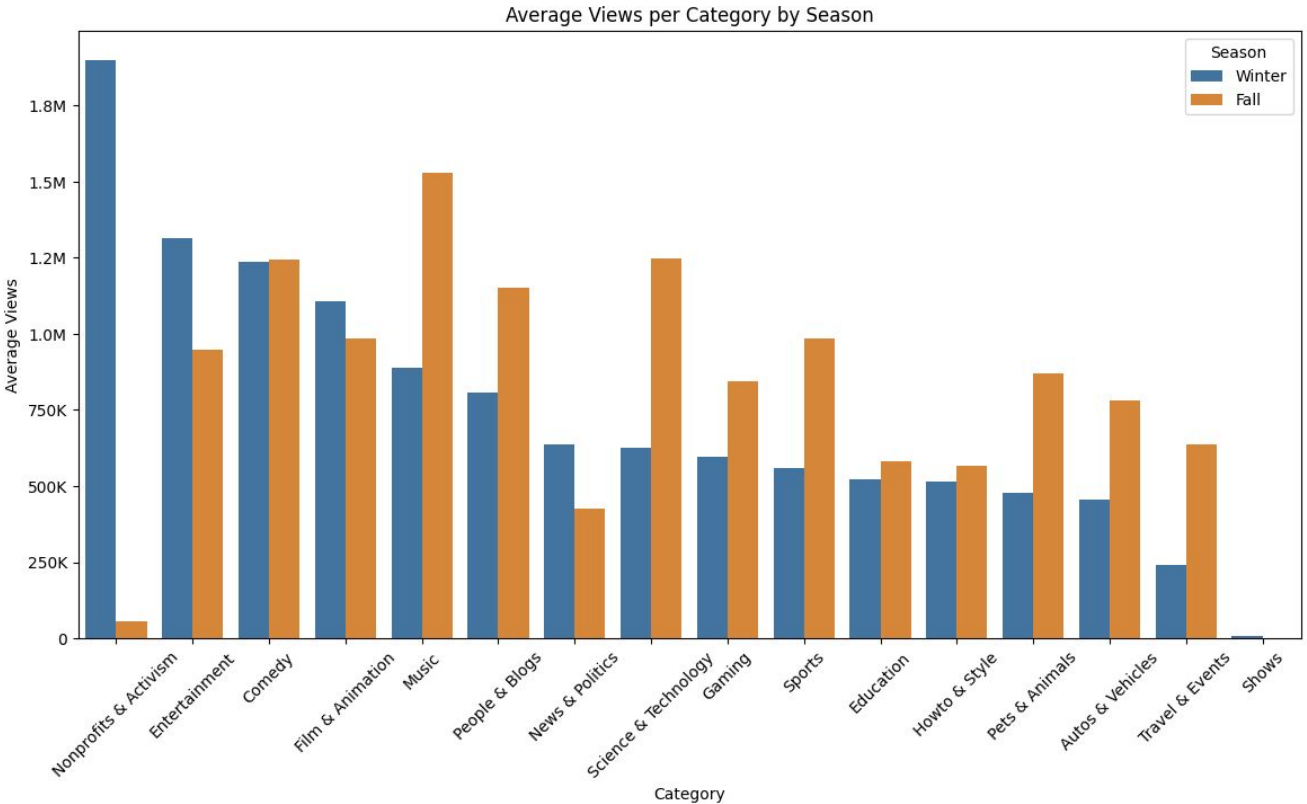
Image Source: https://www.analytixlabs.co.in/blog/apriori-algorithm-in-data-mining/

Association Rule Results

- Parameters:
  - Get top 5 confidence itemsets
  - min_sup = 100 (the pair occurs at least 100 times)
  - K_max = 2

Frequent Items	Confidence	Support of the Item (X_Y)	Support of First Subitem (X)
['Comedy', 'High View']	0.5820	440	756
['Sports', 'Low View']	0.4951	203	410
['Education', 'Medium View']	0.4641	155	334
['News & Politics', 'Low View']	0.4504	282	626
['Film & Animation', 'High View']	0.4206	159	378

Average Views by Season



Video tag Association Rules:

- Goal: Find frequent patterns of tags across entire data along with association rules
- Observations:
  - Deemed most rules as trivial due to overlapping tags
    - ex: Funny, Jokes, Comedy, Humor
    - Affected convection, lift, antecedent & consequent supports
- Future improvements:
  - Employ a clustering algorithm like k-means to group similar tags to combat the large diversity in the tags

Results: Common video tag Sets

Association Rules (Confidence >= 70%)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	z
0	(celebrities)	(funny video)	0.024958	0.022843	0.019882	0.796610	34.873823	0.019311	4.804357	
1	(funny video)	(celebrities)	0.022843	0.024958	0.019882	0.870370	34.873823	0.019311	7.521755	
2	(celebrities)	(humor)	0.024958	0.038917	0.020305	0.813559	20.904937	0.019333	5.154899	
3	(jokes)	(celebrities)	0.019882	0.024958	0.019459	0.978723	39.215290	0.018962	45.826988	
4	(celebrities)	(jokes)	0.024958	0.019882	0.019459	0.779661	39.215290	0.018962	4.448230	
5	(hollywood)	(celebrity)	0.019459	0.027073	0.015651	0.804348	29.710598	0.015125	4.972739	
6	(clip)	(comedic)	0.018190	0.016920	0.016497	0.906977	53.602326	0.016190	10.568105	
7	(comedic)	(clip)	0.016920	0.018190	0.016497	0.975000	53.602326	0.016190	39.272420	
8	(comedian)	(funny)	0.023266	0.091794	0.021574	0.927273	10.101718	0.019438	12.487838	
9	(comedy)	(funny)	0.068951	0.091794	0.049069	0.711656	7.752792	0.042740	3.149737	
10	(sketch)	(comedy)	0.018190	0.068951	0.017343	0.953488	13.828506	0.016089	20.017555	
11	(humor)	(funny)	0.038917	0.091794	0.027919	0.717391	7.815267	0.024346	3.213654	
12	(late night)	(funny)	0.019036	0.091794	0.019036	1.000000	10.894009	0.017288	inf	
13	(sketch)	(funny)	0.018190	0.091794	0.015228	0.837209	9.120566	0.013559	5.578982	
14	(talk show)	(funny)	0.019459	0.091794	0.015651	0.804348	8.762573	0.013865	4.641944	
15	(funny video)	(humor)	0.022843	0.038917	0.020305	0.888889	22.840580	0.019416	8.649746	
16	(jokes)	(funny video)	0.019882	0.022843	0.019459	0.978723	42.846336	0.019004	45.926396	
17	(funny video)	(jokes)	0.022843	0.019882	0.019459	0.851852	42.846336	0.019004	6.615799	
18	(jokes)	(humor)	0.019882	0.038917	0.019459	0.978723	25.148936	0.018685	45.170897	
19	(celebrities, humor)	(funny video)	0.020305	0.022843	0.019882	0.979167	42.865741	0.019418	46.903553	
20	(celebrities, funny video)	(humor)	0.019882	0.038917	0.019882	1.000000	25.695652	0.019108	inf	
21	(humor, funny video)	(celebrities)	0.020305	0.024958	0.019882	0.979167	39.233051	0.019375	46.802030	
22	(celebrities)	(humor, funny video)	0.024958	0.020305	0.019882	0.796610	39.233051	0.019375	4.816836	
23	(funny video)	(celebrities, humor)	0.022843	0.020305	0.019882	0.870370	42.865741	0.019418	7.557650	
24	(jokes, celebrities)	(funny video)	0.019459	0.022843	0.019459	1.000000	43.777778	0.019014	inf	
25	(jokes, funny video)	(celebrities)	0.019459	0.024958	0.019459	1.000000	40.067797	0.018973	inf	

Highest lifts:

- Comedic -> clip
- Jokes -> funny Video
- Jokes -> (celebrities, funny video)
- (Celebrities, funny) -> funny

Perfect Conviction (inf):

- Late night -> funny
- (jokes, funny video) -> jokes
- (jokes, celebrities) -> funny video
- (Celebrities, funny video) -> humor

Frequent tags

- Goal: See if and how the seasons impact the most frequent tags that lead to a trending video
- Process:
  - Created seasonal bins (fall, winter, spring, summer)
  - Run Apriori on each subset
    - Transaction: video
    - Item: tag
- Observations:
  - While patterns did shift, there wasn't a significant change in theme around the tags. This can can due to the data only holding Sept (before halloween) and Jan (after christmas)

Results: Most frequent tags by Season

Fall:

support	itemsets
0.086149	(funny)
0.070101	(comedy)
0.061655	([none])
0.048986	(comedy, funny)
0.038851	(2017)
0.037162	(humor)
0.035473	(makeup)
0.032095	(video)
0.031250	(vlog)
0.030405	(tutorial)
0.030405	(review)
0.030405	(news)
0.030405	(how to)
0.027027	(celebrity)
0.027027	(humor, funny)
0.027027	(beauty)
0.026182	(music)
0.026182	(celebrities)
0.024493	(interview)
0.024493	(Pop)
0.023649	(trailer)
0.022804	(NBC)

Winter:

support	itemsets
0.097285	(funny)
0.068627	(comedy)
0.056561	([none])
0.049020	(comedy, funny)
0.040724	(humor)
0.038462	(2017)
0.038462	(how to)
0.036199	(music)
0.034691	(interview)
0.032428	(halloween)
0.031674	(vlog)
0.030166	(makeup)
0.030166	(celebrity)
0.029412	(humor, funny)
0.028658	(tutorial)
0.027903	(video)
0.026395	(science)
0.026395	(food)
0.024887	(comedian)
0.024133	(beauty)
0.024133	(celebrities)
0.023379	(funny video)

Frequent Title keywords:

- Goal: Extract keyword patterns from trending video titles
- Process
  - Remove punctuation and “Stopping words” (“The”, “of”, “a”,etc..)
    - Words with no significant meaning
  - Run Apriori on data set
    - Transaction: Video
    - Item: words
- Observations:
  - Most frequent: trailer, video, official

Results: frequent title keywords

support	itemsets
0.073604	(official)
0.054569	(video)
0.045685	(2017)
0.041455	(trailer)
0.037648	(video, official)
0.032149	(vs)
0.026650	(first)
0.026650	(ft)
0.024112	(makeup)
0.023689	(trailer, official)
0.023266	(audio)
0.020305	(2)
0.019882	(live)
0.019459	(music)
0.019459	(halloween)
0.018190	(new)
0.015228	(music, video)
0.015228	(10)
0.015228	(full)
0.015228	(hd)
0.014382	(week)
0.014382	(1)
0.014382	(game)

Summary: Association Rule

- Comedy and Film & Animation videos tend to have higher view counts
- Education videos tend to have medium view counts
- Sports and News & Politics videos tend to have low view counts
- The large diversity of tags heavily hindered our association rules
- Seasons had no apparent change in theme around tags due to a lack of diversity (only held 2 months)

Regression Analysis
(Linear Regression/Logistic Regression)

Image Credit:https://www.g2.com/articles/regression-analysis

Regression Analysis Pre-Processing

- Data split (split randomly):
  - 80% dedicated to training
  - 20% dedicated to testing
- Views are min-max normalized to be more readable

Regression Analysis (Logistic Regression)

- Results (Pre-Normalization)
  - Logistic model has generated very great accuracy!
- Conclusion
  - Classification report doesn't tell us the answers we want to find
  - How does each feature influence our model?

```
[6] # Load the dataset
    data = pd.read_csv("./USvideos.csv")

    # Define target variable (y)
    median_views = data['views'].median()
    data['viral'] = (data['views'] >= median_views).astype(int)  # Binary classification: 1 for viral, 0 for non-viral

    # Prepare the features
    X = data[['views', 'likes', 'dislikes', 'comment_total', 'category_id']]  # Numerical and categorical features
    # Preprocess text features like 'tags' if needed and add to X

    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, data['viral'], test_size=0.2, random_state=42)

    # Initialize and train the logistic regression model
    model = LogisticRegression()
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)

    # Evaluate the model
    accuracy = accuracy_score(y_test, y_pred)
    print("Accuracy:", accuracy)

    # Print classification report
    print("Classification Report:")
    print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.9275
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.90      0.93       803
           1       0.90      0.95      0.93       797

    accuracy                           0.93      1600
   macro avg       0.93      0.93      0.93      1600
weighted avg       0.93      0.93      0.93      1600
```

Regression Analysis (Logistic Regression)

```python
# Get the coefficients of the logistic regression model
coefficients = model.coef_[0]

# Map coefficients to feature names
feature_names = X.columns

# Create a DataFrame to display coefficients and feature names
coefficients_df = pd.DataFrame({'Feature': feature_names, 'Coefficient': coefficients})
coefficients_df = coefficients_df.sort_values(by='Coefficient', ascending=False)

# Print the DataFrame
print("Feature Coefficients:")
print(coefficients_df)
```

```
Feature Coefficients:
         Feature  Coefficient
2       dislikes     0.000146
3  comment_total     0.000109
0          views     0.000017
1          likes    -0.000032
4    category_id    -0.243570
```

Regression Analysis (Logistic Regression)

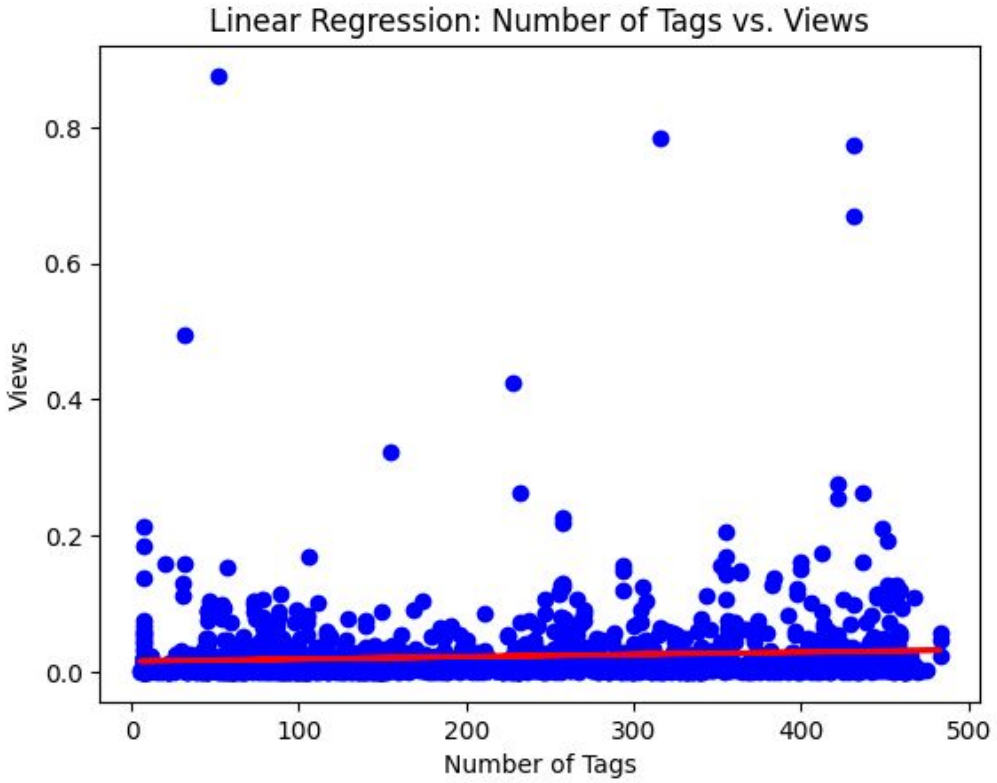
- Results (Post-Normalization)
  - Accuracy got even better!
- Conclusion
  - The feature that had the greatest influence on what is a viral video or not is views
    - May not tell us a lot so likes is a better feature
  - Category_id has the least impact on what is a viral video or not

```
# Prepare the features
X = data[['views', 'likes', 'dislikes', 'comment_total', 'category_id']]  # Numerical and categorical features

# Scale the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

```
Accuracy: 0.976875
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       803
           1       0.99      0.96      0.98       797

    accuracy                           0.98      1600
   macro avg       0.98      0.98      0.98      1600
weighted avg       0.98      0.98      0.98      1600

Feature Coefficients:
         Feature  Coefficient
0          views    23.848196
1          likes     3.788244
2       dislikes     3.014816
3  comment_total     1.309091
4    category_id     0.096754
```

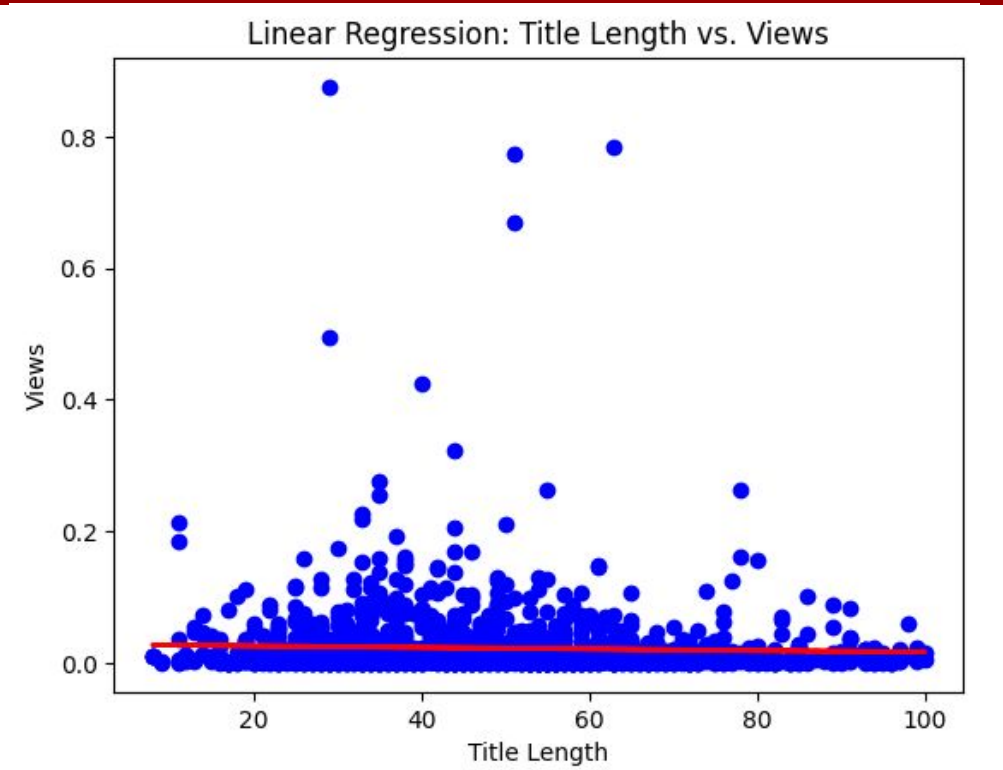
Regression Analysis (Linear Regression)

- Results
  - Correlation Coefficient: 0.099
  - Correlation not strong enough to say there is a linear relationship
- Conclusion
  - # of tags does not affect # of views
  - Future improvements: eliminating outliers and using a better training/test splitting method like tenfold



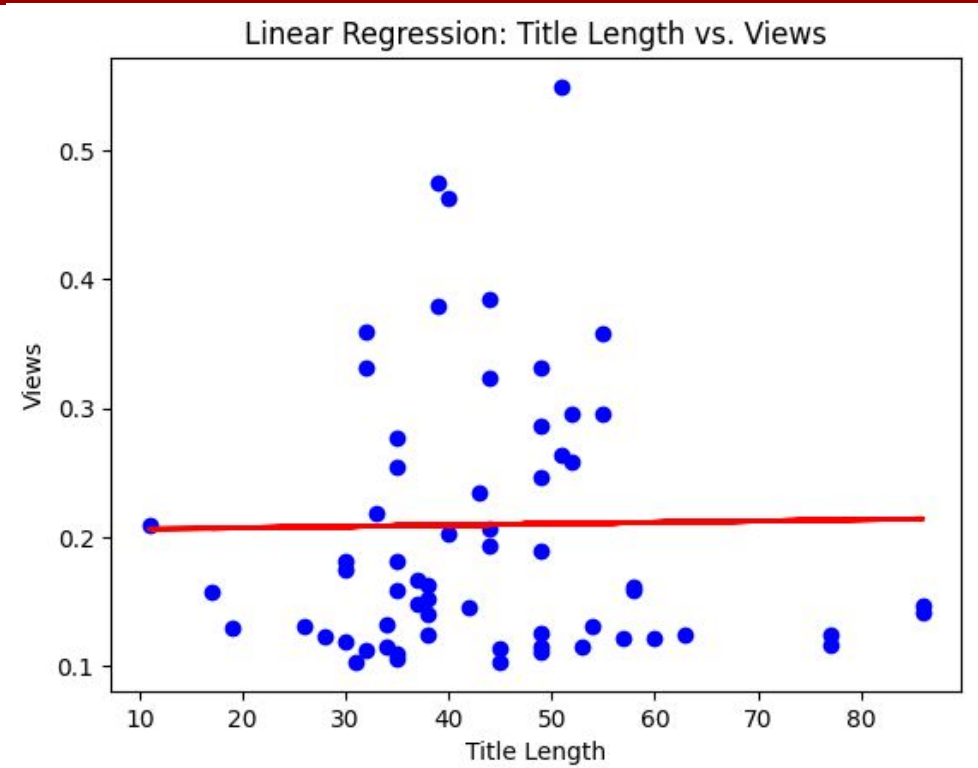
Regression Analysis (Linear Regression)

- Results
  - Correlation Coefficient: -0.0479
  - Again, correlation not strong enough to say there is a linear relationship
- Observation
  - Low view videos may be influencing outcome
  - Remove low view videos and try again



Regression Analysis (Linear Regression)

- Results
  - Not promising, still no valuable relationship
- Conclusion
  - Title length does not affect # of views
  - Future improvements: eliminating outliers and using a better training/test splitting method like tenfold



Conclusion

Data challenges

- Many categories were not numerical and were either removed or converted to numerical form prior to calculations
- Original dataset was missing helpful categories like “shares”, “subscriptions” etc… that could improve findings and predictions.
- Null values had to be dropped from comments data to not skew results

Project Summary

- Key findings:
  - Views and likes have a positive linear relationship
  - Frequent words in video titles
    - Trailer
    - Music
    - Official
  - Overlapping in tags made most association rules trivial
  - Comedy videos typically have high view count
  - Tags and title length do not affect view count

Project Summary

- Key takeaways:
  - The dataset is insufficient for predicting whether or not a YouTube video will be trending
  - Additional data could be helpful
    - Numerical data
      - Share count
      - Subscriber count
    - Non-numerical data
      - Direct thumbnail image file
      - Content of video

Questions?

Thank You Dr. Wei

Sources:

1. “Trending YouTube Video Statistics and Comments,” Kaggle, Oct. 25, 2017. https://www.kaggle.com/datasets/datasnaek/youtube

2. Wei, Dr. Honghao “Mining Frequent Patterns, Association and Correlations: Basic Concepts and Methods.” Class lecture, CPT_S 315 Introduction to Data Mining, Washington State University, Pullman, Washington, February 5, 2024.

3. GfG, “Linear Regression in Machine learning,” GeeksforGeeks, Mar. 14, 2024. https://www.geeksforgeeks.org/ml-linear-regression/

4. GfG, “Logistic regression in machine learning,” GeeksforGeeks, Jan. 30, 2024. https://www.geeksforgeeks.org/understanding-logistic-regression/