Viewer Engagement with YouTube News Videos

Data Wranglers

Umang Prajapati, Deep Patel, Kunal Patel, Reeve Lood, Rimsha Rizvi

Project Introduction

- **Data**: YouTube engagement survey by Wu et al.

- **Source file**: top_news.json from the quality videos dataset
  - Info about videos from top 100 most viewed news channels

- **Question**: *How does the video topic affect viewer engagement?*

Data Cleaning

topicDetails	statistics	contentDetails	snippet	id	insights
{'topicCategories': ['https://en.wikipedia.org...	{'commentCount': '0', 'viewCount': '701', 'fav...	{'duration': 'PT11M5S', 'definition': 'sd', 'd...	{'description': '⬇ Descargá la App de C5N para...	JEibcX5VF1M	{'startDate': '2016-08-21', 'dailyShare': '6,2...
{'topicCategories': ['https://en.wikipedia.org...	{'commentCount': '0', 'viewCount': '889', 'fav...	{'duration': 'PT4M9S', 'definition': 'sd', 'di...	{'description': '⬇ Descargá la App de C5N para...	CJwOo9Yp3Jk	{'startDate': '2016-08-10', 'dailyShare': '0,0...
{'topicCategories': ['https://en.wikipedia.org...	{'commentCount': '4', 'viewCount': '12720', 'f...	{'duration': 'PT13M11S', 'definition': 'hd', '...	{'description': '', 'title': 'C5N - Road Music...	CbdrzQPgGpl	{'startDate': '2016-07-31', 'dailyShare': '5,2...
{'topicCategories': ['https://en.wikipedia.org...	{'commentCount': '3', 'viewCount': '4367', 'fa...	{'duration': 'PT32M18S', 'definition': 'sd', '...	{'description': '⬇ Descargá la App de C5N para...	J9AqmuhrHAw	{'startDate': '2016-07-17', 'dailyShare': '0,1...
{'topicCategories': ['https://en.wikipedia.org...	{'commentCount': '0', 'viewCount': '3797', 'fa...	{'duration': 'PT8M7S', 'definition': 'hd', 'di...	{'description': '', 'title': 'C5N - Atentado e...	o2ju4DluRLs	{'startDate': '2016-07-02', 'dailyShare': '5,4...

Before

description	title	channelId	channelTitle	publishedAt	categoryId	detectLang
⬇ Descargá la App de C5N para Android https://...	C5N - Alerta Verde: Programa 21/08/2016 (Parte 1)	UCFgk2Q2mVO1BklRQhSv6p0w	c5n	2016-08-21T18:51:46.000Z	25	es
⬇ Descargá la App de C5N para Android https://...	C5N - No te Duermas: El Polaco, ¡de la cumbia ...	UCFgk2Q2mVO1BklRQhSv6p0w	c5n	2016-08-10T14:30:25.000Z	25	es
	C5N - Road Music: Barbie Simons en España con ...	UCFgk2Q2mVO1BklRQhSv6p0w	c5n	2016-07-31T01:17:02.000Z	17	NA
⬇ Descargá la App de C5N para Android https://...	C5N - Right Now: Programa 16/07/2016	UCFgk2Q2mVO1BklRQhSv6p0w	c5n	2016-07-17T01:36:27.000Z	25	es
	C5N - Atentado en Bangladesh: habla Diego, el ...	UCFgk2Q2mVO1BklRQhSv6p0w	c5n	2016-07-02T00:40:42.000Z	17	NA

After

Exploratory Data Analysis

- Column values of original dataframe are type **dict**

- Each record represents info about **one** video

```
In [5]:  snippets['channelTitle'].value_counts().head()

Out[5]:  Fox News            483
         The Young Turks     481
         ABP NEWS            460
         ABS-CBN News        459
         Zee News            458
         Name: channelTitle, dtype: int64

In [6]:  snippets['detectLang'].value_counts().head()

Out[6]:  en      15966
         es       2083
         ar       1432
         fr       1141
         ru        938
         Name: detectLang, dtype: int64

In [8]:  statistics.median()  # one number from the five-number summary

Out[8]:  commentCount       20.0
         viewCount       17459.0
         favoriteCount       0.0
         dislikeCount        9.0
         likeCount          77.0
         dtype: float64
```
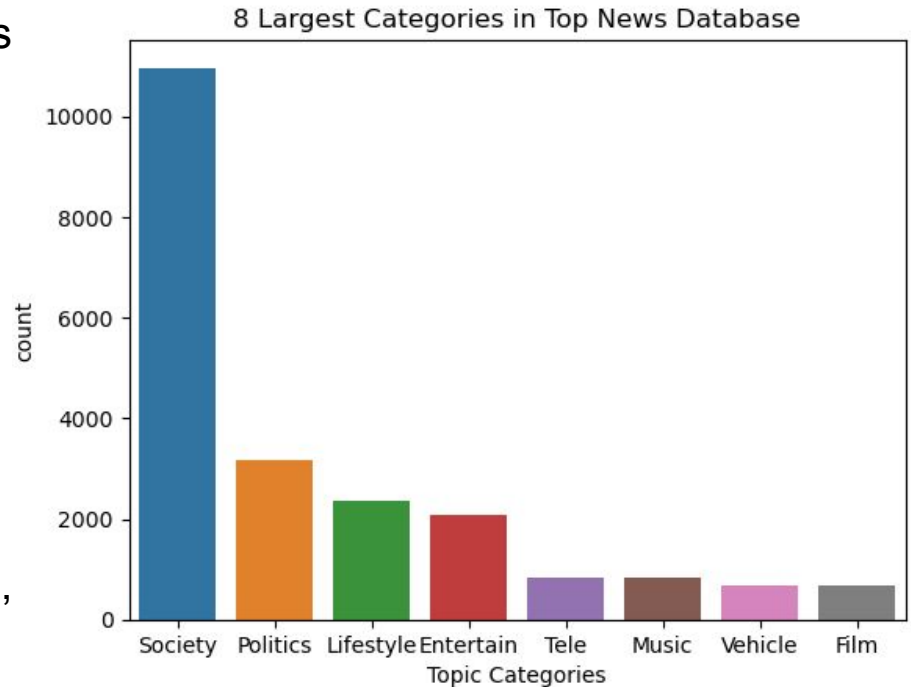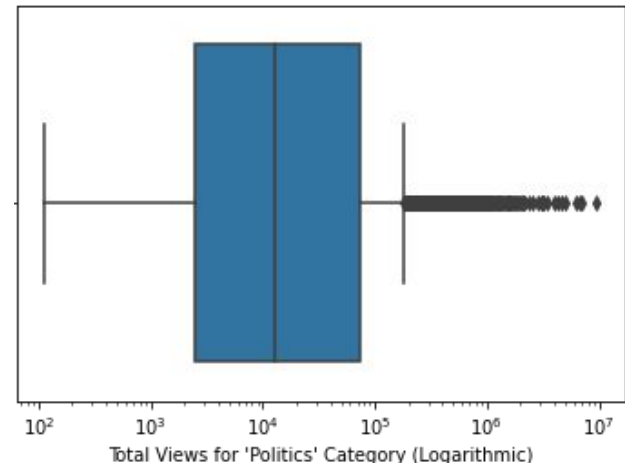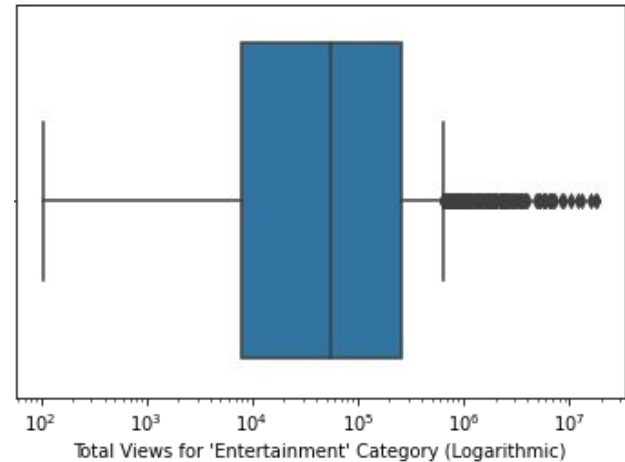
Visualization - Bar Chart

- Visualize the **largest** news categories

- Each video has a set of categories chosen during upload

- Most popular is the "**Society**" tag

- Popular genres also include "Politics", "Lifestyle", "Music", and "Film"

8 Largest Categories in Top News Database

Visualization - Box Plot

- Visualization of the **total views** of the "**Entertainment**" and "**Politics**" categories
- Logarithmic scale
- Many outliers as there are ~2000 entries for the "Entertainment" category and ~3000 entries for the "Politics" category
- On average, more viewers for news videos labeled as "Entertainment" than those labeled "Politics"

Total Views for 'Entertainment' Category (Logarithmic)

Total Views for 'Politics' Category (Logarithmic)

ML Analysis - Decision Tree

The Big 5

Society	10385
Politics	3075
Lifestyle_(sociology)	2276
Entertainment	1984
Music	799

	commentCount	viewCount	favoriteCount	dislikeCount	likeCount	target		target
0	0	701	0	0	9	Pet		0
1	0	889	0	0	10	Entertainment		1
2	4	12720	0	4	213	Music		1
3	3	4367	0	1	4	Entertainment		1
4	0	3797	0	2	10	Politics		1
...	...	...	...	...	...	...	→	...
23240	68	5946	0	9	166	Film		0
23241	575	121063	0	94	1125	Food		0
23242	249	18904	0	55	291	Society		1
23243	103	8432	0	17	181	Food		0
23244	53	7949	0	13	205	Lifestyle_(sociology)		1

23245 rows × 6 columns

Preparing the data

ML Analysis - Decision Tree (cont.)

- Decision tree classifier trained, criterion = **entropy**, max_depth = **4**
- 70% training data, 10% validation data, and 20% test data

```python
# train accuracy
print(1 - np.sum(np.abs(y_train - clf.predict(x_train))) / x_train.shape[0])
```
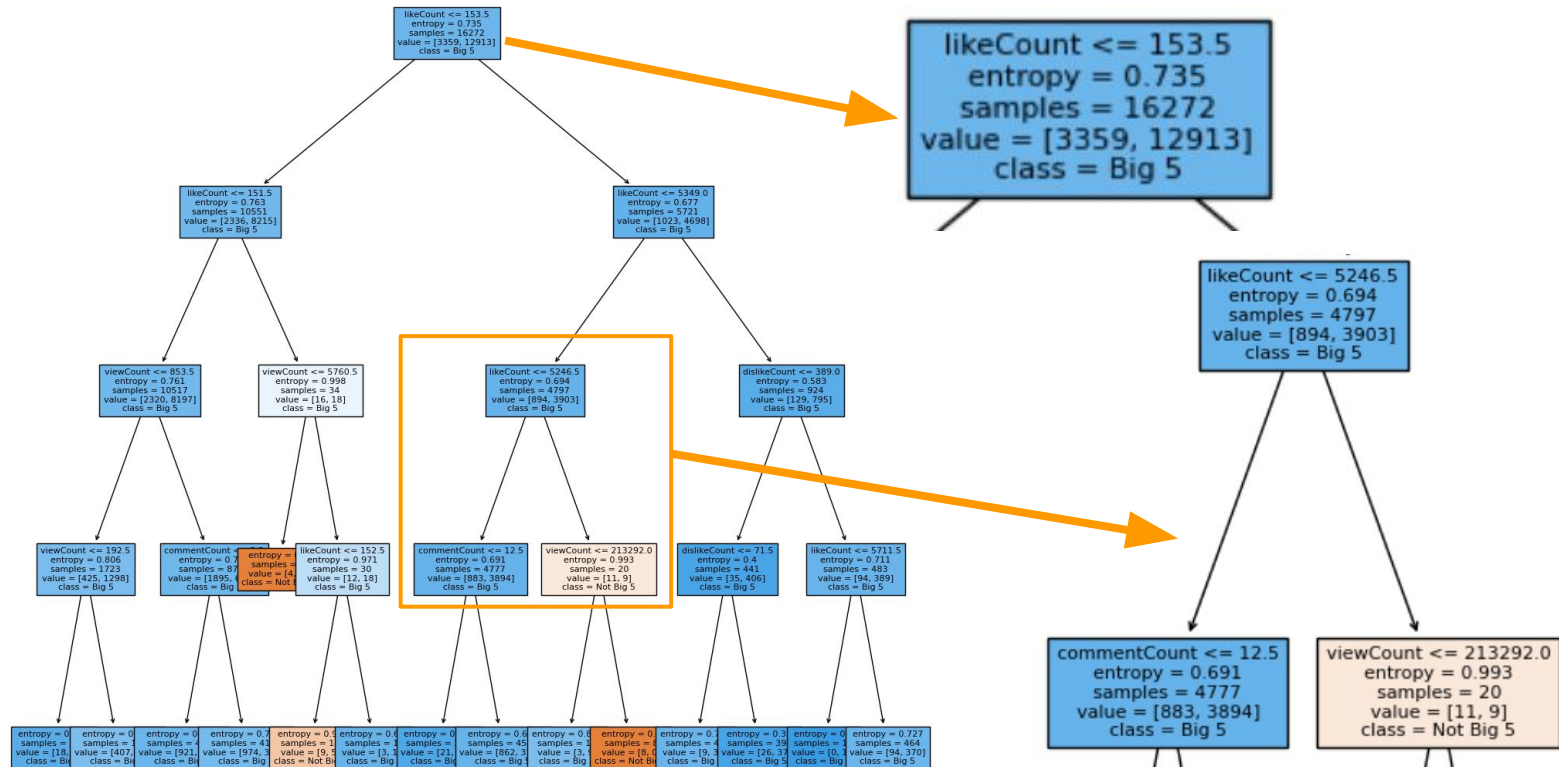
0.794555063913471

```python
# validation accuracy
print(1 - np.sum(np.abs(y_val - clf.predict(x_val))) / x_val.shape[0])
```

0.7341935483870967

```python
# test accuracy
print(1 - np.sum(np.abs(y_test - clf.predict(x_test))) / x_test.shape[0])
```

0.8356282271944923

ML Analysis - Decision Tree (cont.)

Takeaways & Retrospective

- **Video genre** affects viewer engagement: Important to understand viewer interests when creating news content on YouTube
- **Embracing unconventional**: "Entertainment" genre surpassing "Politics" indicates need to consider unconventional factors when analyzing viewer engagement
- "**Society**" takes the lead: Indicating viewers' strong interest in social and current issues
- **Patterns with logarithmic scale**: Easier to notice trends and outliers in comparison to linear scale

- **Impact of outliers**: Can represent viral videos that may have skewed our results
- Emphasising the importance of **data quality**: Meticulous data cleaning ensures accuracy and reliability for insightful decisions
- Valuable lessons learned when analyzing data: It is *not* a straightforward process

References

- **Database Google Drive**:

  https://drive.google.com/drive/folders/1wZwDIR18IHPPTiH1C0dyBbGPR-3MktI7

- **Database GitHub**:

  https://github.com/avalanchesiqi/youtube-engagement/blob/master/data/README.md

Questions?