INTRODUCTION

PROJECT IDEA

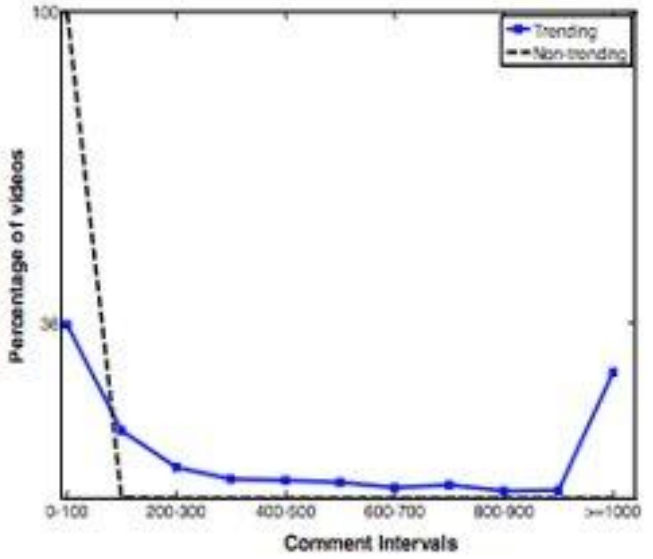
- Recommend content trending on YouTube that has the most “engaging audience”

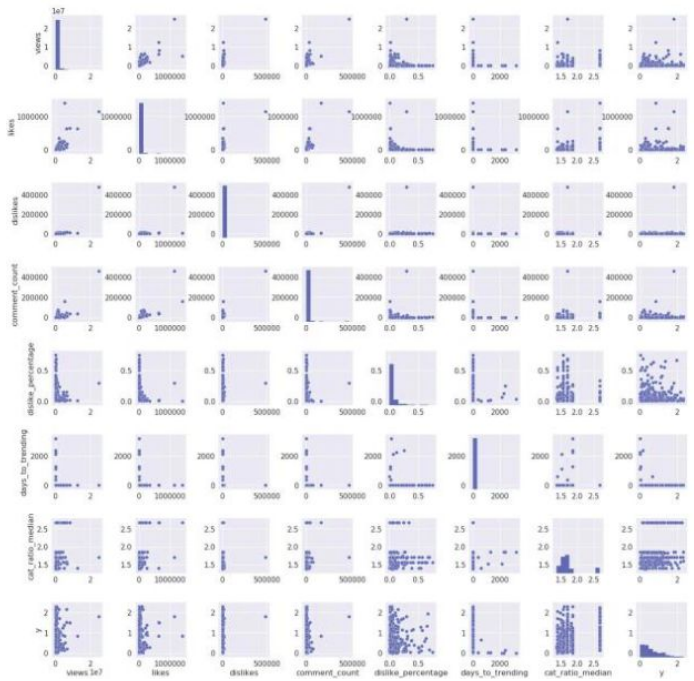
MOTIVATION/PURPOSE

- Possible Flaw in YouTube algorithm
  - YouTube videos with high views does not mean the community is an engaging one
- Profitability
  - High engaging community is profitable for advertiser, content creators and sponsorships
  - Target ads, content and sponsorships to the demographic of the community

Article Recommendation

- **Trending Videos: Measurement and Analysis** (I. Barjasteh, et al 2014)
  - For trending videos, about one half of videos received less than few hundred comments per video; about 30% of trending videos have more than 1,000 comments
- **Youtube Trending Video Metadata Analysis Using Machine Learning** (S. Amudha, et al 2020)
  - Performed data analysis and visualizations to provide insight into latest trends and user engagement in videos with respect to categories and year.

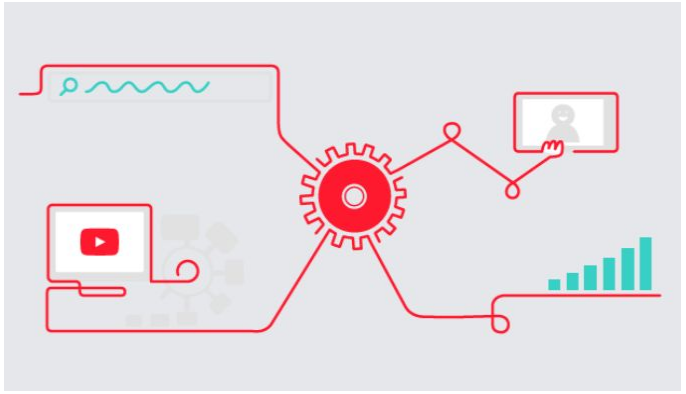


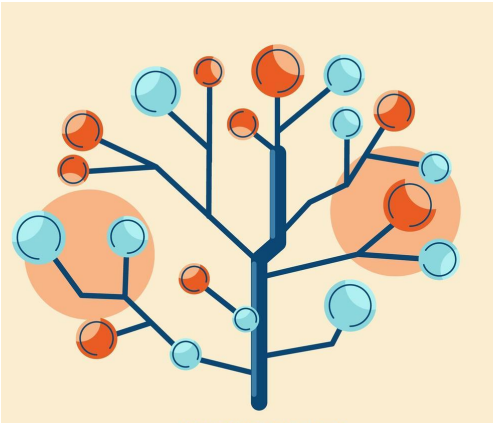


SYSTEM MODEL

YouTube Data API v3

kaggle





1 PHASE 1 - DATA COLLECTION

KAGGLE & YOUTUBE API

2 PHASE 2 - DATA PROCESSING

CLEAN & ANALYZE DATA
CLASSIFY THRESHOLD

3 PHASE 3 - TRAIN/APPLY MODEL

BINARY CLASSIFICATION
RECOMMENDATION
EVALUATION

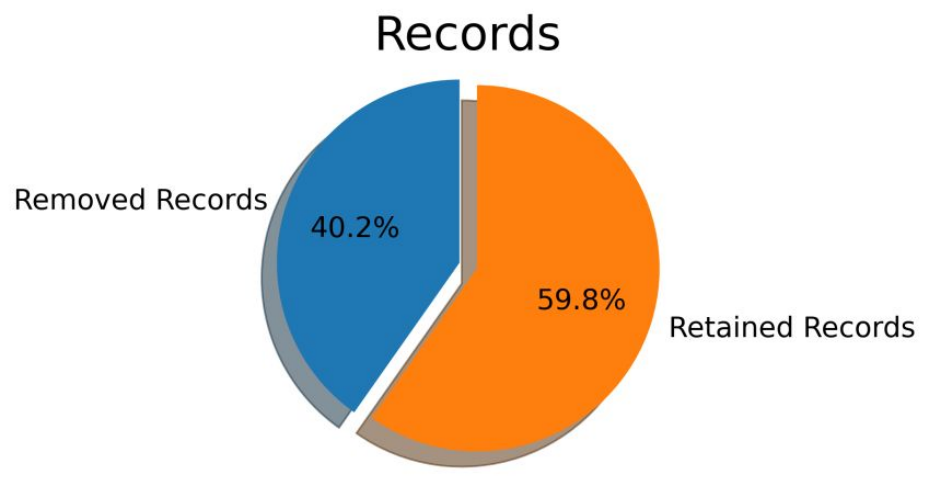
PHASE 1: DATA COLLECTION

- Training Data
  - Kaggle Data Set
  - Records: 375,942 videos
  - 200 Trending Videos per day for 6 months: Q4 2018 - Q1 2019
  - 10 Countries
  - Originated from YouTube Data API

- Test Data
  - YouTube Data API
  - 200 Trending Videos per day for 3 days
  - Dec. 10, 2020 - Dec 12, 2020

Google APIs

PHASE 2: DATA PROCESSING



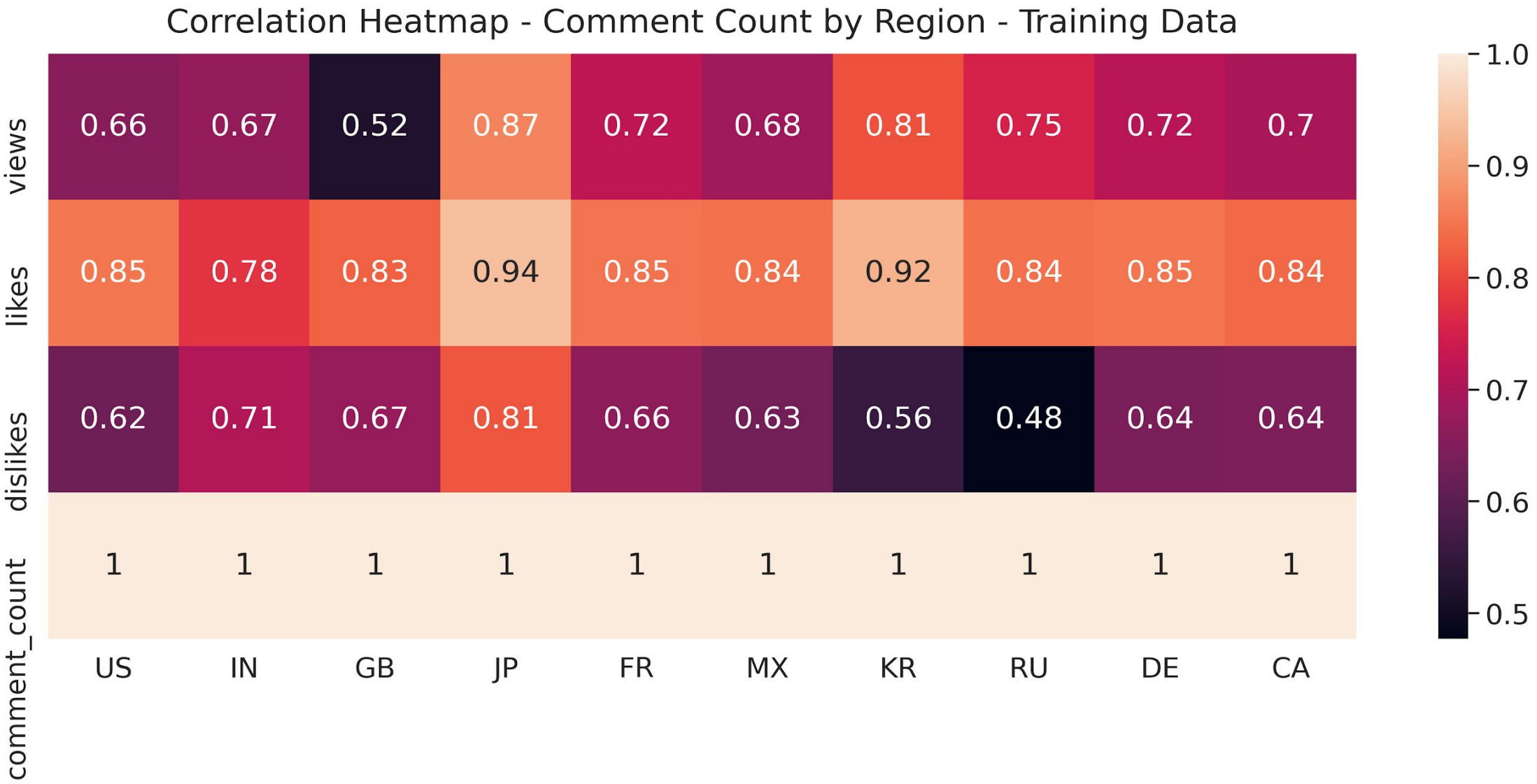
Removal Criterion

- views less than 100,000
- no comments
- no likes
- no dislikes

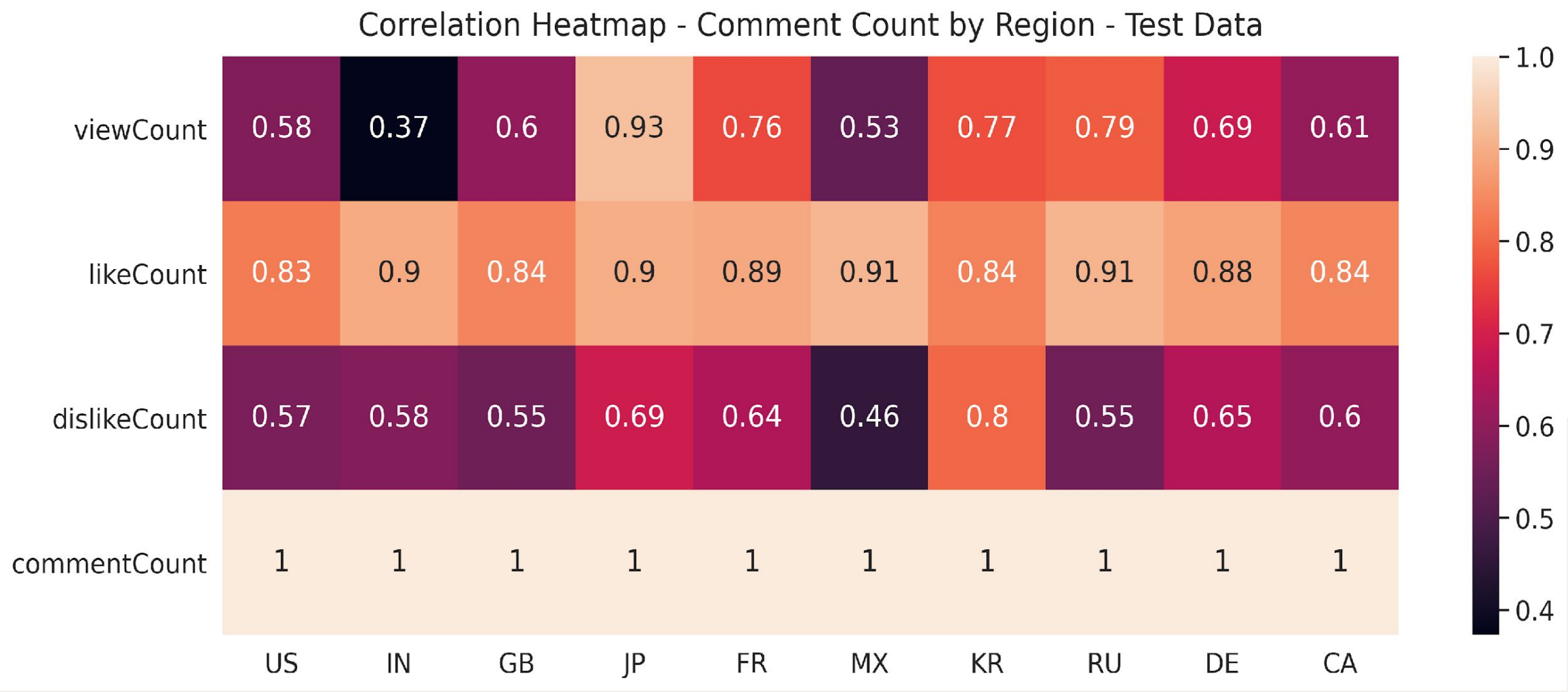
Total Records	Removed Records	Retained Records
375,942	151,201	224,741

Region Region Code	Records	Records Removed	% of Removed Records
United States (US)	40949	6220	15.189626%
India (IN)	37352	8955	23.974619%
United Kingdom (GB)	38916	5886	15.124884%
Japan (JP)	20523	13990	68.167421%
France (FR)	40724	23987	58.901384%
Mexico (MX)	40451	26111	64.549702%
South Korea (KR)	34567	18025	52.145109%
Russia (RU)	40739	25885	63.538623%
Germany (DE)	40840	20329	49.777179%
Canada (CA)	40881	8359	20.447151%

Training Data - Features



Test Data – Features



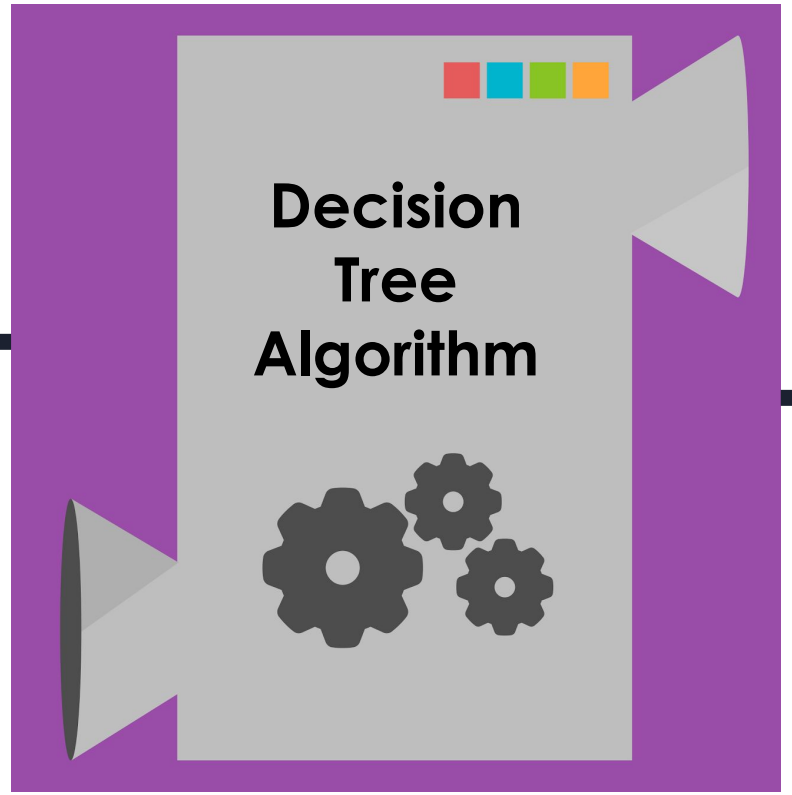
PHASE 3: RECOMMENDATION SYSTEM

INPUT PARAMETERS

# of Views

# of Likes

# of Dislikes

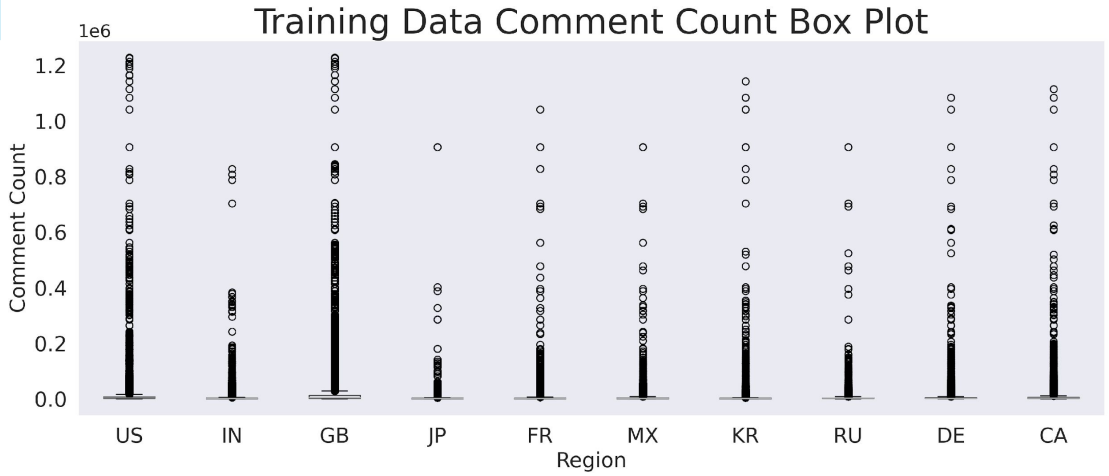


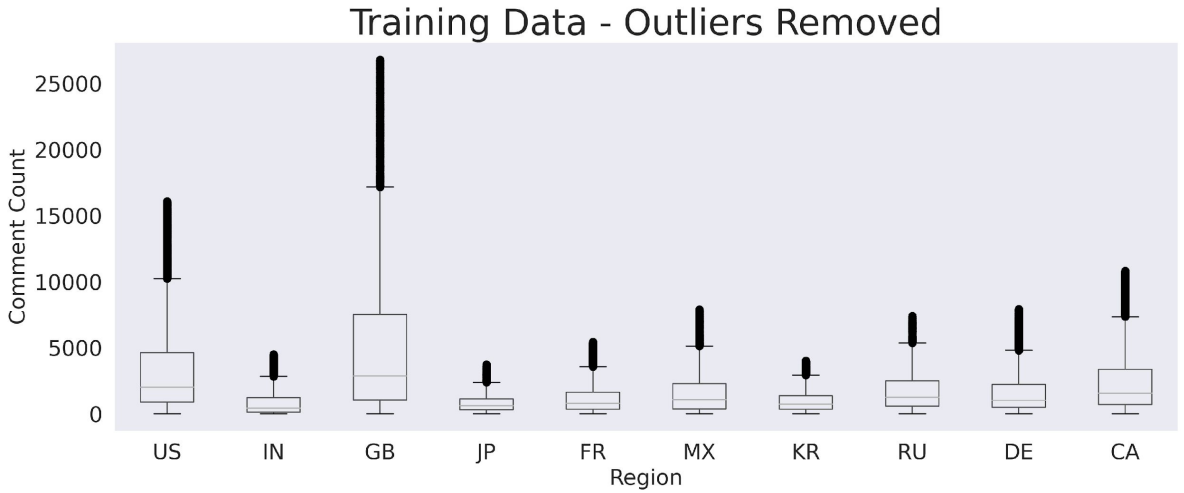
BINARY CLASSIFICATION

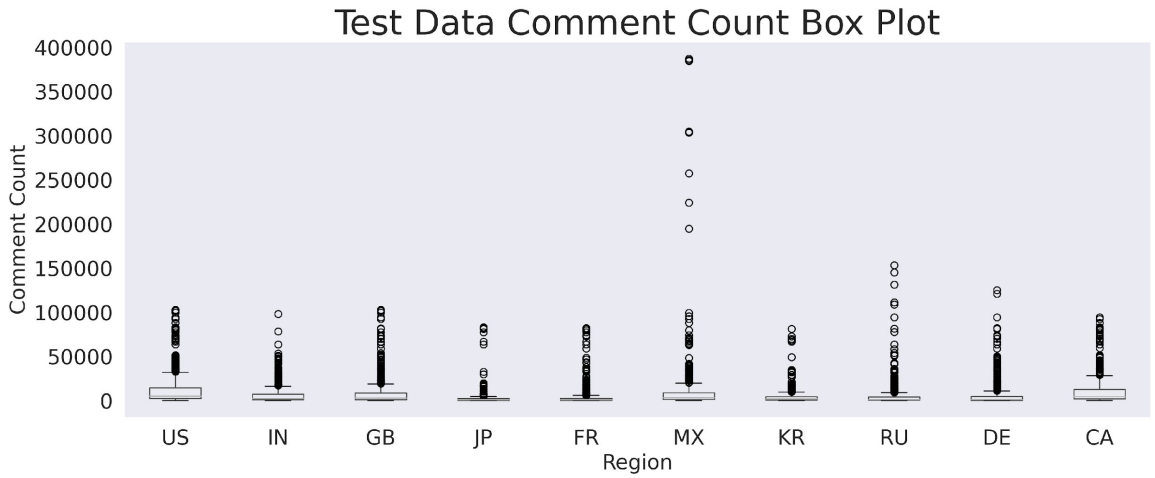
1 - Engaging Community

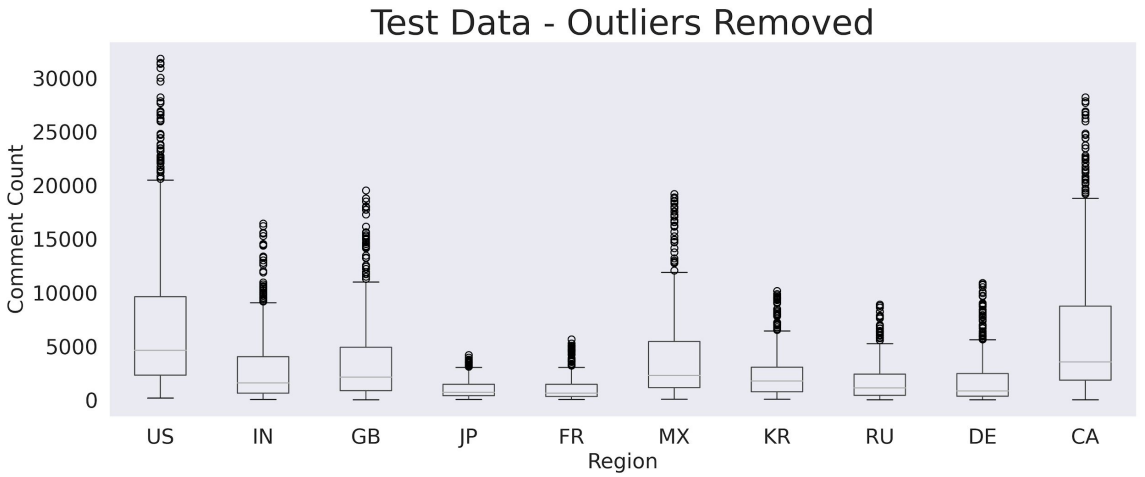
0 - Non-Engaging Community

Threshold: Comment Count Means



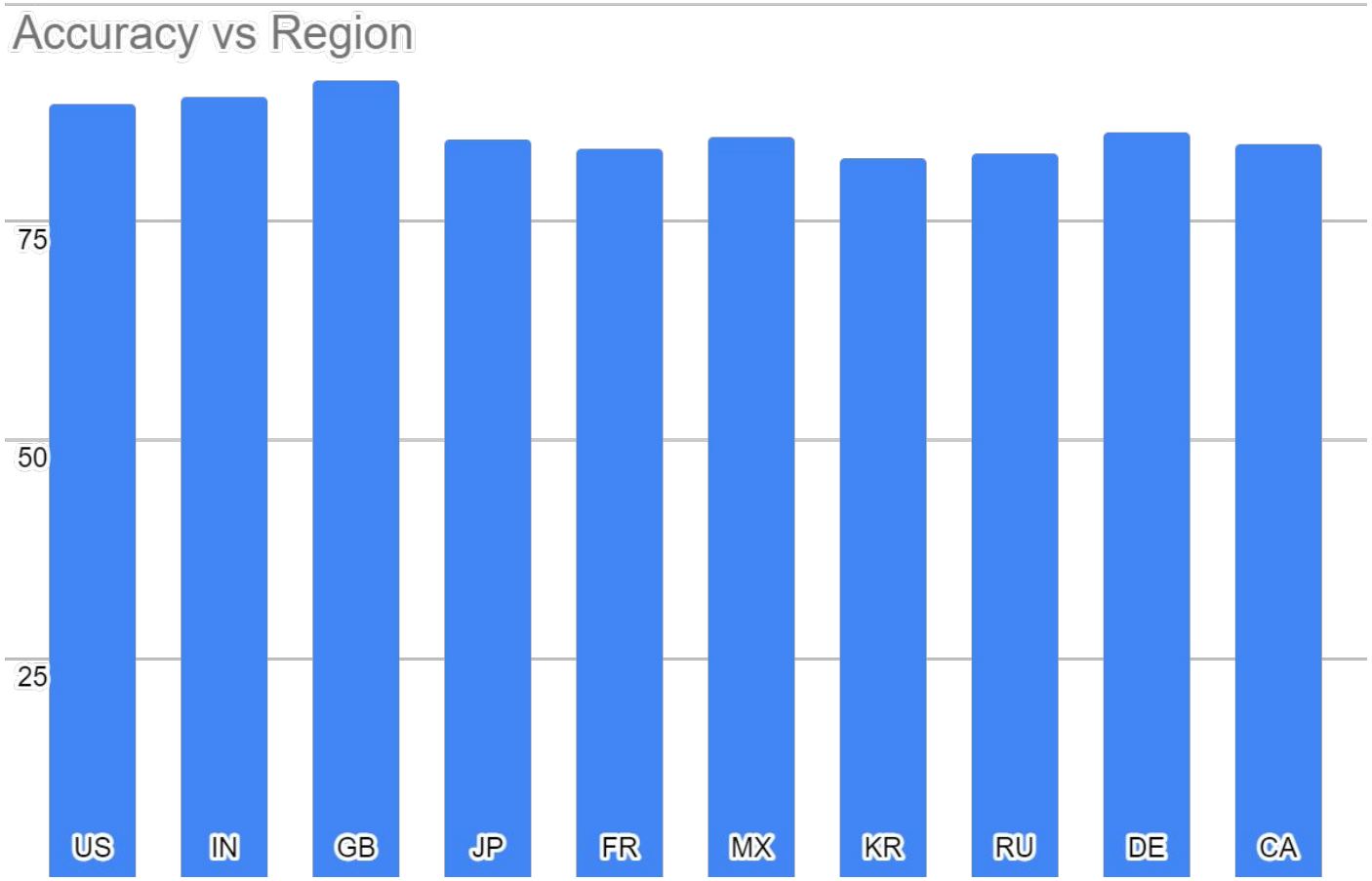






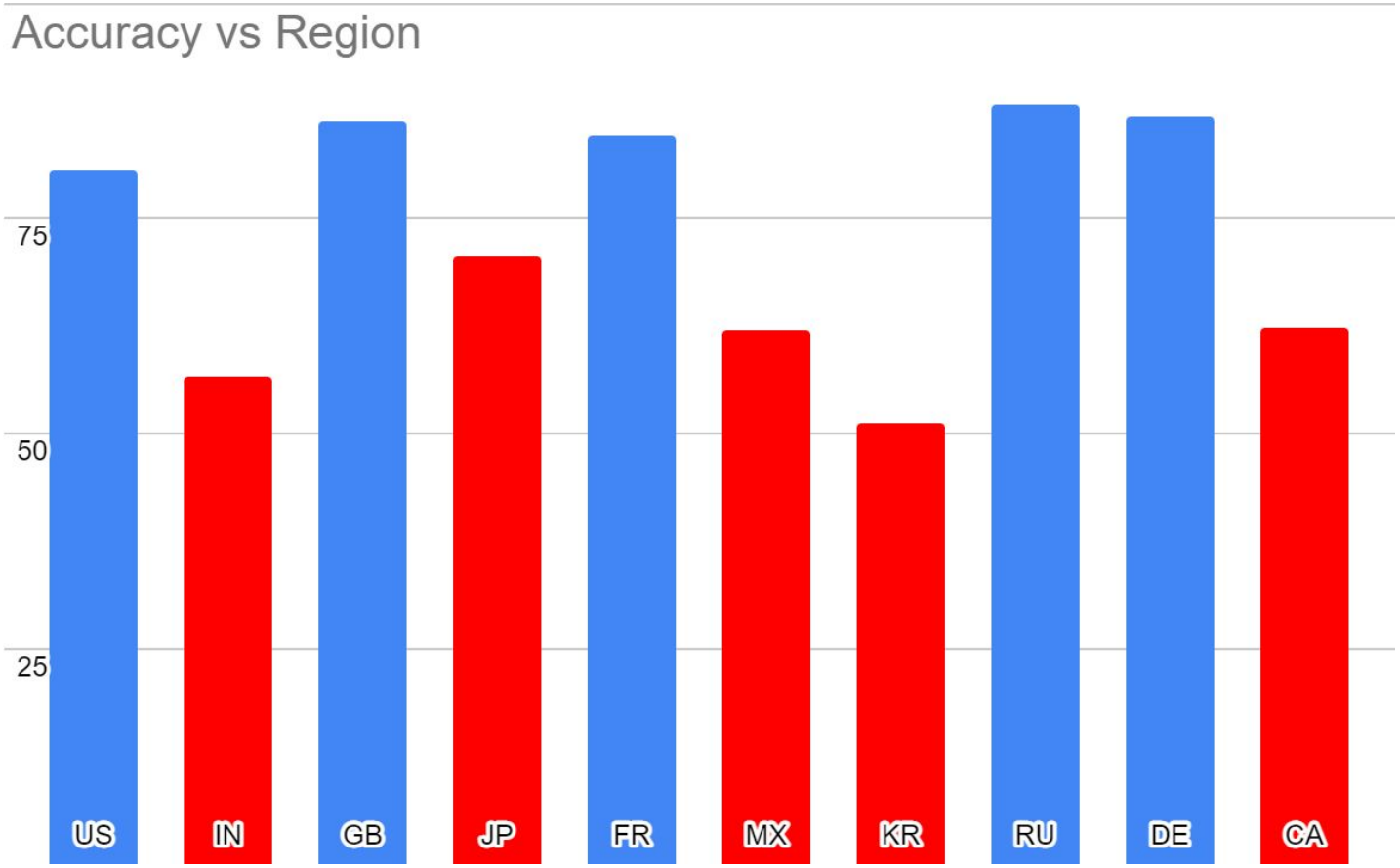
Training Data 80/20 Split

Region: US	Region: MX
Accuracy = 88.3484 Error = 11.6516	Accuracy = 84.6207 Error = 15.3793
[[3767. 544.]	[[1425. 262.]
[268. 2390.]]	[180. 1007.]]
---	---
Region: IN	Region: KR
Accuracy = 89.1882 Error = 10.8118	Accuracy = 82.2077 Error = 17.7923
[[3059. 359.]	[[1586. 304.]
[251. 1973.]]	[294. 1177.]]
---	---
Region: GB	Region: RU
Accuracy = 91.1202 Error = 8.87984	Accuracy = 82.6493 Error = 17.3507
[[3573. 277.]	[[1466. 284.]
[312. 2471.]]	[236. 1011.]]
---	---
Region: JP	Region: DE
Accuracy = 84.5004 Error = 15.4996	Accuracy = 85.2342 Error = 14.7658
[[615. 111.]	[[2097. 371.]
[86. 459.]]	[231. 1378.]]
---	---
Region: FR	Region: CA
Accuracy = 83.3633 Error = 16.6367	Accuracy = 83.7934 Error = 16.2066
[[1695. 331.]	[[3319. 615.]
[225. 1091.]]	[430. 2084.]]



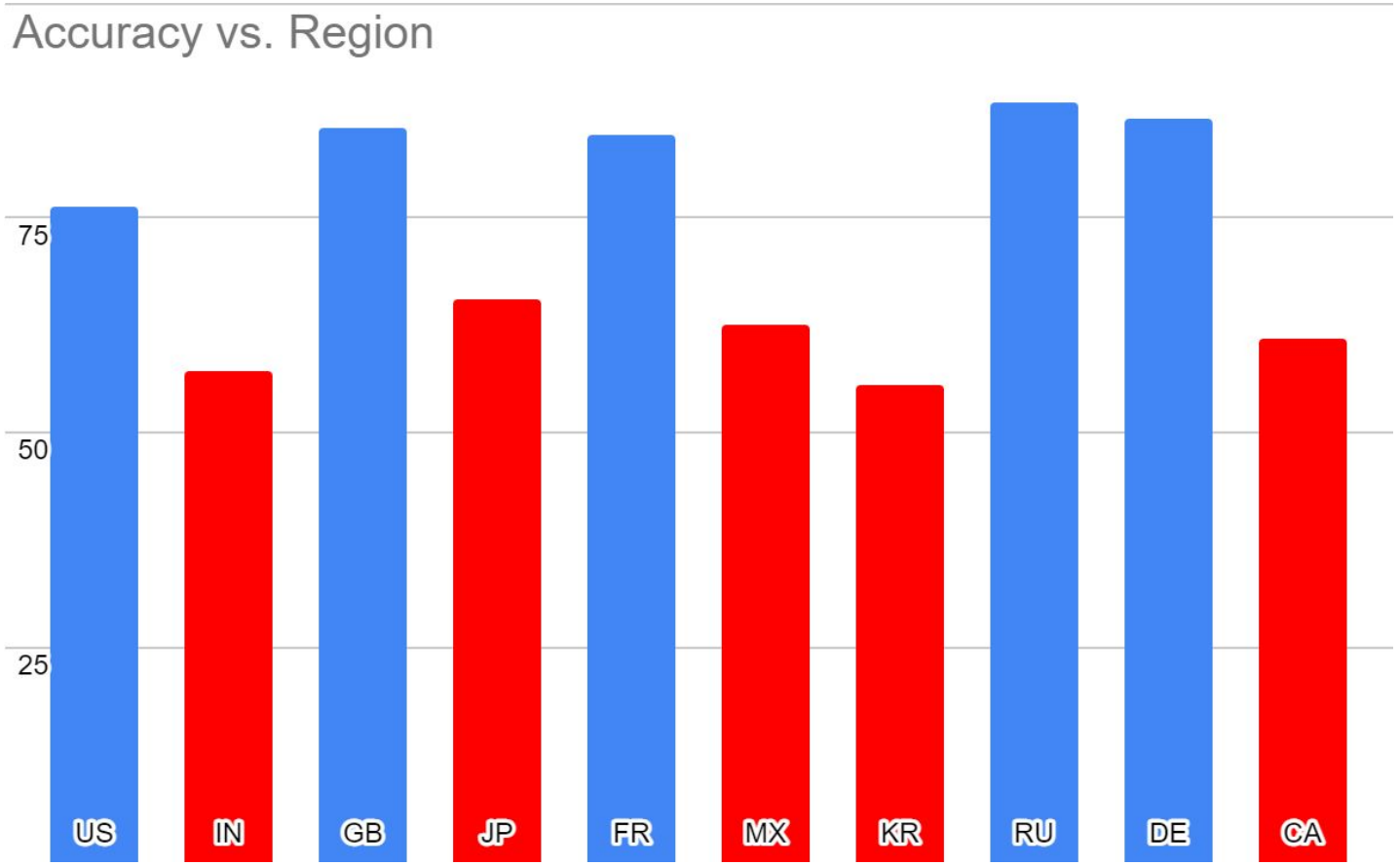
80% Split with Real Test Data

Region: US	Region: MX
Accuracy = 80.6838 Error = 19.3162	Accuracy = 62.0748 Error = 37.9252
[[245. 5.]	[[124. 0.]
[108. 227.]]	[223. 241.]]
---	---
Region: IN	Region: KR
Accuracy = 56.6794 Error = 43.3206	Accuracy = 51.0888 Error = 48.9112
[[92. 3.]	[[65. 7.]
[224. 205.]]	[285. 240.]]
---	---
Region: GB	Region: RU
Accuracy = 86.2306 Error = 13.7694	Accuracy = 88.1905 Error = 11.8095
[[323. 61.]	[[262. 24.]
[19. 178.]]	[38. 201.]]
---	---
Region: JP	Region: DE
Accuracy = 70.5151 Error = 29.4849	Accuracy = 86.927 Error = 13.073
[[172. 19.]	[[303. 26.]
[147. 225.]]	[51. 209.]]
---	---
Region: FR	Region: CA
Accuracy = 84.5501 Error = 15.4499	Accuracy = 62.3288 Error = 37.6712
[[289. 43.]	[[143. 3.]
[48. 209.]]	[217. 221.]]



100% Training Data with Real Test Data

Region: US	Region: MX
Accuracy = 76.4103 Error = 23.5897	Accuracy = 62.585 Error = 37.415
[[219. 4.]	[[130. 3.]
[134. 228.]]	[217. 238.]]
---	---
Region: IN	Region: KR
Accuracy = 57.0611 Error = 42.9389	Accuracy = 55.6114 Error = 44.3886
[[94. 3.]	[[92. 7.]
[222. 205.]]	[258. 240.]]
---	---
Region: GB	Region: RU
Accuracy = 85.3701 Error = 14.6299	Accuracy = 88.381 Error = 11.619
[[323. 66.]	[[267. 28.]
[19. 173.]]	[33. 197.]]
---	---
Region: JP	Region: DE
Accuracy = 65.3641 Error = 34.6359	Accuracy = 86.5874 Error = 13.4126
[[146. 22.]	[[304. 29.]
[173. 222.]]	[50. 206.]]
---	---
Region: FR	Region: CA
Accuracy = 84.5501 Error = 15.4499	Accuracy = 60.9589 Error = 39.0411
[[289. 43.]	[[132. 0.]
[48. 209.]]	[228. 224.]]



FUTURE WORK

- Improved Accuracy
- Models adjusted to account for regional differences
- Account for more up-to-date data
- Analyze text information of comments and description of video and their impact on video's viewership
- Explore other methods of visualizing data

CONCLUSION

- Communities are affected by their region
- More data is NOT always best for our model, requires further tuning
- Viewership does not guarantee an engaging community
- Profitability is not solely dependant on viewership, established-engaging communities are also important
- YouTube Channels with high community engagement and relatively lower views can also trend on YouTube.