

Executive Summary

Data Analysis

Overview

This summary describes the findings from the data analysis stage of the claims classification project. The columns of the dataset are examined and descriptive statistics are gathered to understand the variables in preparation for exploratory analysis.

Objective

- Examine the dataset of the claims classification project.
- Gather descriptive statistics for the dataset.
- Understand which variables may contain predictive information.

Results

- The variables 'claim_status' and 'author_ban_status' contain useful information.
- The number of claims and opinions in the dataset are similar, i.e., the two categories are balanced.

```
data['claim_status'].value_counts()
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

- Viewer engagement with the videos also provided useful information. The median view, likes and share counts for banned authors were much higher than active authors, indicating a correlation between author ban and content engagement.

	video_view_count		video_like_count		video_share_count	
	mean	median	mean	median	mean	median
author_ban_status						
active	215927.039524	8616.0	71036.533836	2222.0	14111.466164	437.0
banned	445845.439144	448201.0	153017.236697	105573.0	29998.942508	14468.0
under review	392204.836399	365245.5	128718.050339	71204.5	25774.696999	9444.0

Next Steps

- With the most useful variables having been identified, the data is ready for further exploratory analysis.