DELHI TECHNOLOGICAL UNIVERSITY

DEEP LEARNING

ADOBE CHALLENGE

Task 1: Exploratory Data Analysis

Students:
ANUBHAV BHOUMIK
(2K21/EP/17)
KARTHIK VENKATESHWARAN
(2K21/EP/49)

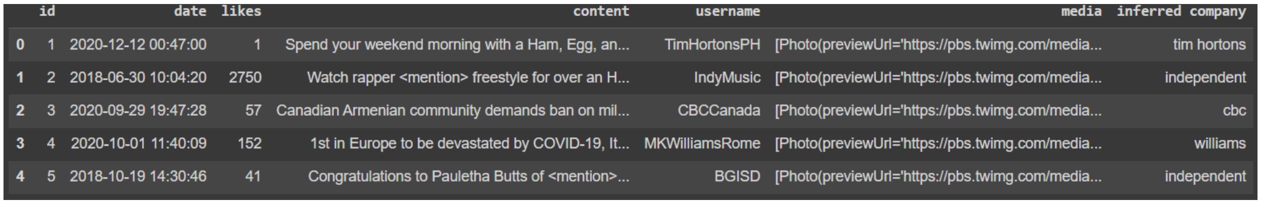
1 Introduction

Exploratory Data Analysis (EDA) is a critical first step in analyzing the data from an experiment or research study. EDA is a way to visualize, summarize, and understand the underlying patterns of the data.

It also helps in identifying any potential problems or anomalies that could affect later stages of analysis, such as outliers, missing values, or the distribution of the data. By employing a combination of graphical techniques and statistical summaries, EDA facilitates a deeper understanding of the data's context, which is essential for developing a sound analytical strategy

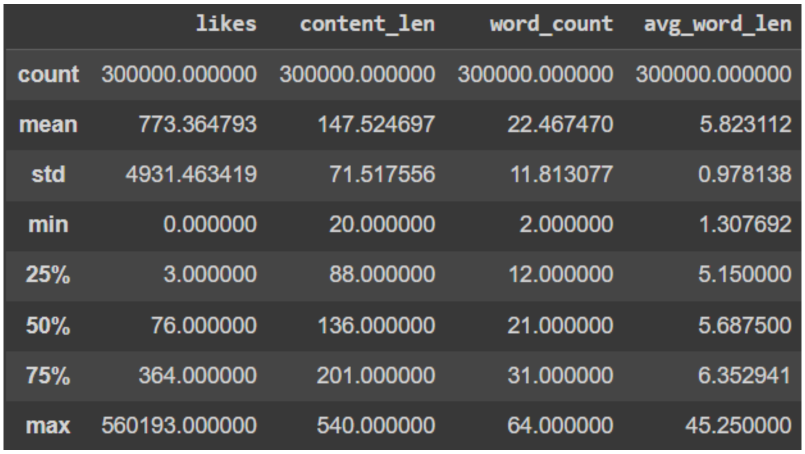
2 Data Description

Brands use Twitter to post marketing content about their products to serve several purposes,including ongoing product campaigns, sales, offers, discounts, brand building, community engagement, etc. User engagement on Twitter is quantified by metrics like user likes, retweets, comments, mentions, follows, clicks on embedded media and links. For this challenge, we have sampled tweets posted in the last five years from Twitter enterprise accounts. Each sample contains tweet ID, company name, username, timestamp, tweet text, media links and user likes.



	id	date	likes	content	username	media	inferred company
0	1	2020-12-12 00:47:00	1	Spend your weekend morning with a Ham, Egg, an...	TimHortonsPH	[Photo(previewUrl='https://pbs.twimg.com/media...	tim hortons
1	2	2018-06-30 10:04:20	2750	Watch rapper <mention> freestyle for over an H...	IndyMusic	[Photo(previewUrl='https://pbs.twimg.com/media...	independent
2	3	2020-09-29 19:47:28	57	Canadian Armenian community demands ban on mil...	CBCCanada	[Photo(previewUrl='https://pbs.twimg.com/media...	cbc
3	4	2020-10-01 11:40:09	152	1st in Europe to be devastated by COVID-19, It...	MKWilliamsRome	[Photo(previewUrl='https://pbs.twimg.com/media...	williams
4	5	2018-10-19 14:30:46	41	Congratulations to Pauletha Butts of <mention>...	BGISD	[Photo(previewUrl='https://pbs.twimg.com/media...	independent

On calculating the word and character counts of the contents in the dataset present in text form we get the following statistical insights.

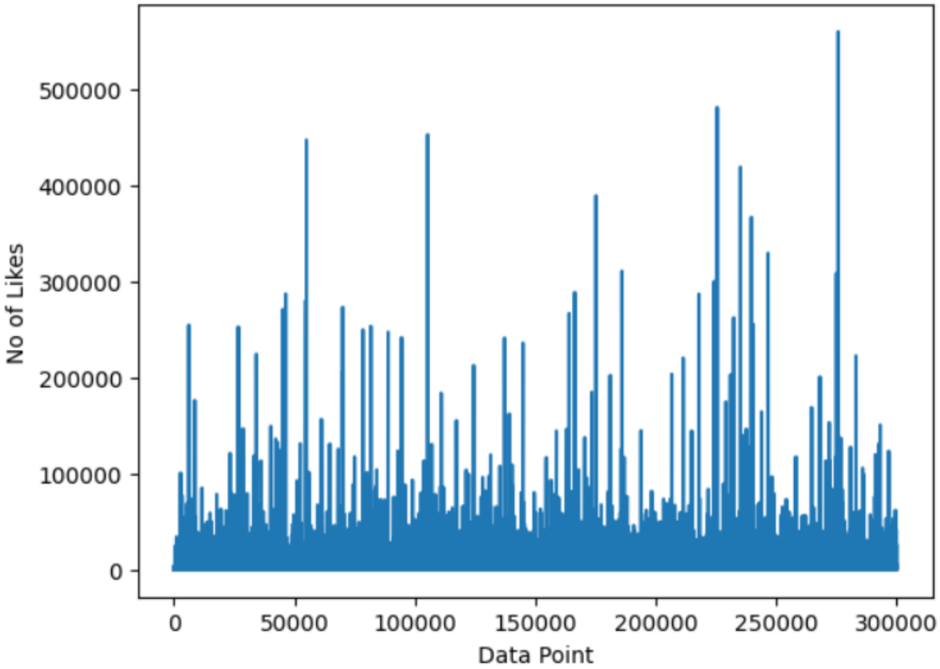


	likes	content_len	word_count	avg_word_len
count	300000.000000	300000.000000	300000.000000	300000.000000
mean	773.364793	147.524697	22.467470	5.823112
std	4931.463419	71.517556	11.813077	0.978138
min	0.000000	20.000000	2.000000	1.307692
25%	3.000000	88.000000	12.000000	5.150000
50%	76.000000	136.000000	21.000000	5.687500
75%	364.000000	201.000000	31.000000	6.352941
max	560193.000000	540.000000	64.000000	45.250000

3 Plots for various data

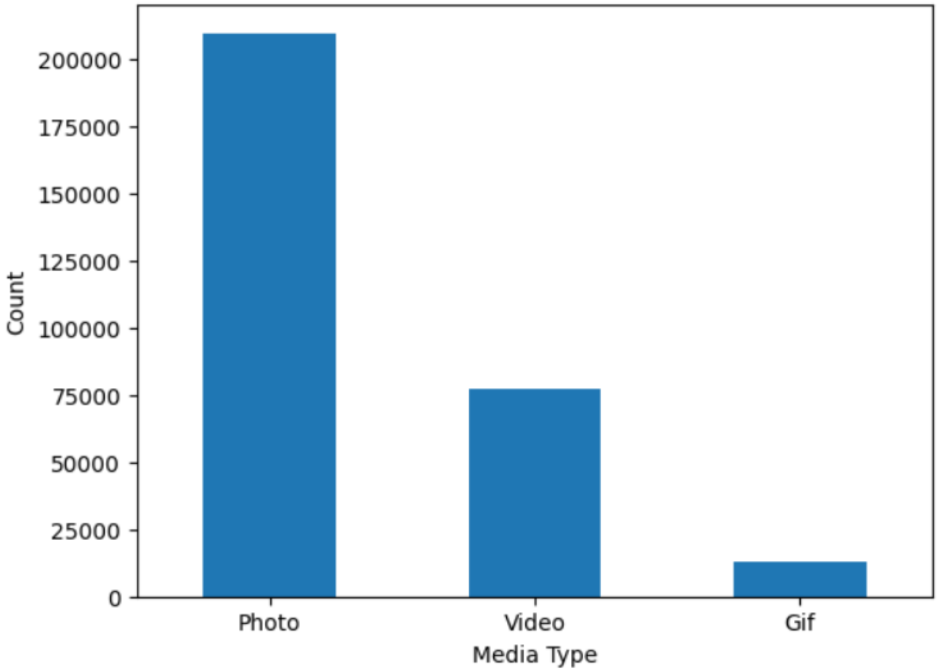
3.1 Like count

Count of likes for each datapoint present in the dataset.



3.2 Media Type count

Count of the type of media present in the dataset.

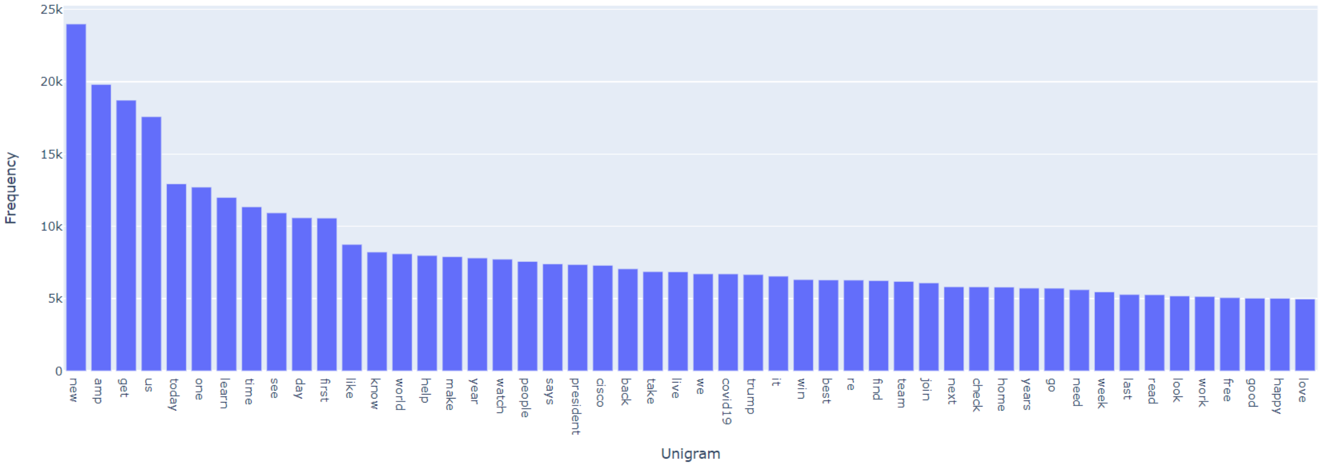


3.3 Most Frequent Words

We cleaned the dataset of text data by:

1. Lowercasing the text data
2. Removed punctuations
3. Removed stop words
4. Removed outliers (words such as 'hyperlink' and 'mention')

After cleaning, we get the following words as the most frequently occuring in our dataset.

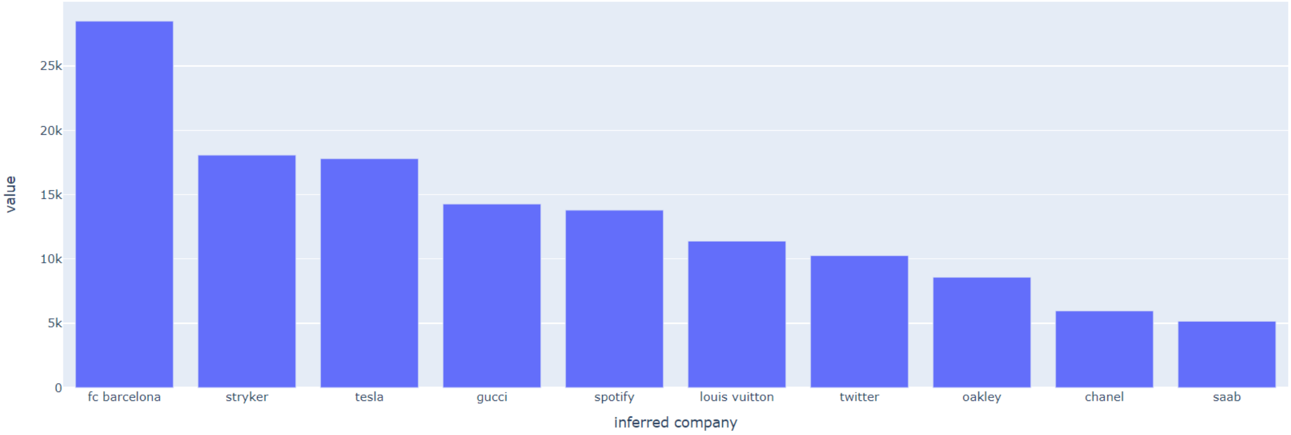


3.4 Bivariate Analysis

Bivariate analysis is a statistical method that involves the analysis of two variables at a time, for the purpose of determining the empirical relationship between them. It's used to test hypotheses about relationships between these variables or to estimate the strength and direction of associations.

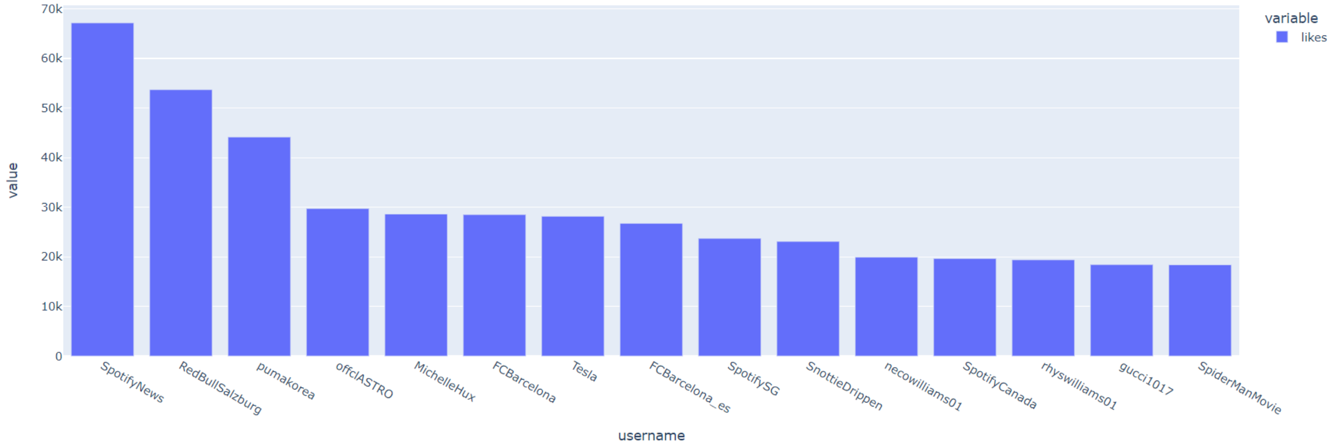
3.4.1 Company vs Likes

This graph tells us the average likes received on a tweet by a company.



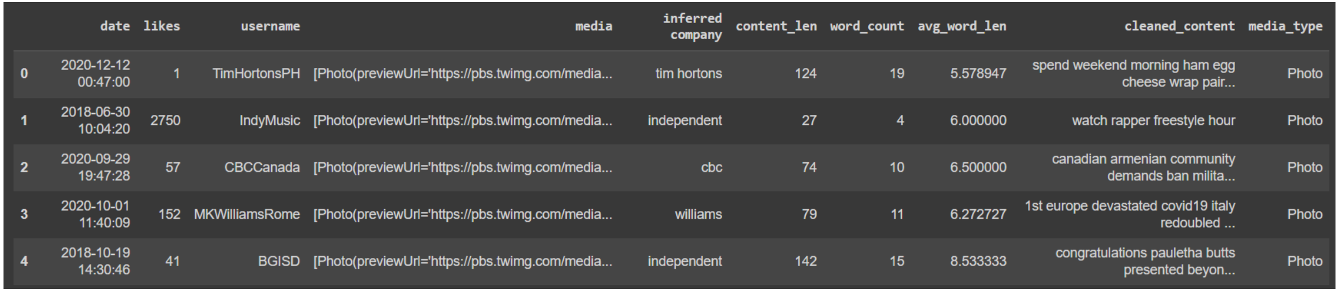
3.4.2 Username vs Likes

This graph tells us the average likes received by usernames on the twitter .



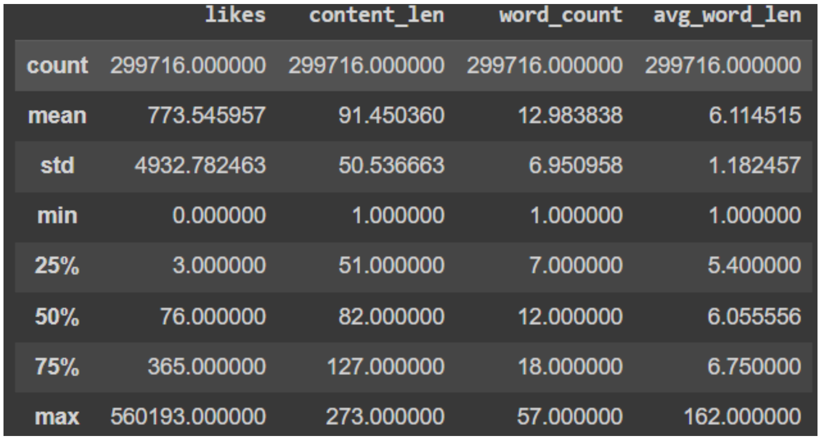
4 Result

After doing EDA and data cleaning of the dataset, We get the following dataset:



	date	likes	username	media	inferred company	content_len	word_count	avg_word_len	cleaned_content	media_type
0	2020-12-12 00:47:00	1	TimHortonsPH	[Photo(previewUrl='https://pbs.twimg.com/media...	tim hortons	124	19	5.578947	spend weekend morning ham egg cheese wrap pair...	Photo
1	2018-06-30 10:04:20	2750	IndyMusic	[Photo(previewUrl='https://pbs.twimg.com/media...	independent	27	4	6.000000	watch rapper freestyle hour	Photo
2	2020-09-29 19:47:28	57	CBCCanada	[Photo(previewUrl='https://pbs.twimg.com/media...	cbc	74	10	6.500000	canadian armenian community demands ban milita...	Photo
3	2020-10-01 11:40:09	152	MKWilliamsRome	[Photo(previewUrl='https://pbs.twimg.com/media...	williams	79	11	6.272727	1st europe devastated covid19 italy redoubled ...	Photo
4	2018-10-19 14:30:46	41	BGISD	[Photo(previewUrl='https://pbs.twimg.com/media...	independent	142	15	8.533333	congratulations pauletha butts presented beyon...	Photo

And below is the various statistical measures which changed after EDA.



	likes	content_len	word_count	avg_word_len
count	299716.000000	299716.000000	299716.000000	299716.000000
mean	773.545957	91.450360	12.983838	6.114515
std	4932.782463	50.536663	6.950958	1.182457
min	0.000000	1.000000	1.000000	1.000000
25%	3.000000	51.000000	7.000000	5.400000
50%	76.000000	82.000000	12.000000	6.055556
75%	365.000000	127.000000	18.000000	6.750000
max	560193.000000	273.000000	57.000000	162.000000

Our exploratory data analysis of the tweets dataset revealed significant insights into user sentiments and trends, underscoring the immense potential of social media data in understanding public discourse. Future efforts should focus on expanding the dataset and refining analytical techniques to further explore the dynamic landscape of social media interactions.

References

1.https://pythonspot.com/nltk-stop-words/

2.https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

3.https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.html

4.https://plotly.com/python/plotly-express/