# Adobe – DTU Behavior & Content Simulation Challenge: Phase -1 Exploratory Data Analysis


Emad Abd Al Fatah
*2K21/CO/165*
*Dept of computer science and engineering*
Delhi Technological university
emadfatah123@gmail.com

Yaseen Mohammed Ahmed
*2K21/CO/538*
*Dept of computer science and engineering*
Delhi Technological university
yaseenalsalami3@gmail.com

Raad Ghazi
*2K21/CO/360*
*Dept of computer science and engineering*
Delhi Technological university
alzaeemraad9@gmail.com


## I. Introduction

This report presents the detailed analysis and preprocessing steps undertaken for the dataset which is provided as part of the Adobe – DTU Behavior & Content Simulation Challenge. The challenge aims to simulate user behavior and content interactions to derive insights and improve user experience in digital platforms. The dataset provided comprises information such as id, date, likes, content, username, media, and inferred company, reflecting user engagement and content attributes.

The primary goal of this analysis is to prepare the dataset for further modeling and analysis by addressing data quality issues, understanding data characteristics, and identifying patterns or anomalies. By performing exploratory data analysis (EDA) and data cleaning, we aim to gain insights into user behavior, content preferences and company interactions within the digital platform.

The insights derived from this analysis will be involved in developing strategies to enhance user engagement, personalize content recommendations, and optimize content delivery. The findings will also contribute to improving the overall user experience and platform performance, aligning with the objectives of the Behavior & Content Simulation Challenge.

## II. Data Cleaning and Preprocessing

The data cleaning and preprocessing phase involved several steps to ensure the dataset's quality, consistency, and readiness for further analysis. Each step was thoroughly executed to address data anomalies, handle missing values, standardize the data format and remove unwanted text and emoji.

### *A. Exploratory Data analysis (EDA)*

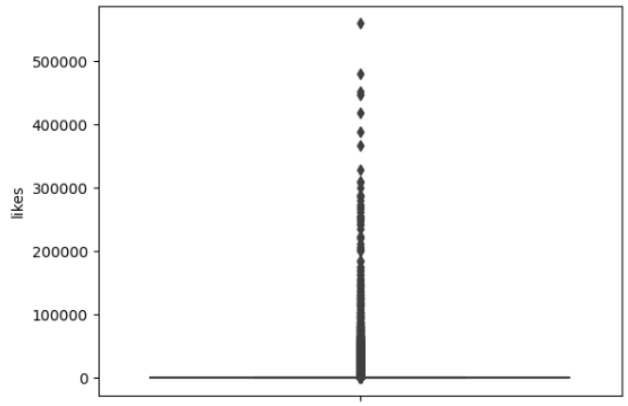
The EDA process began with an assessment of the dataset's general characteristics, focusing on central tendencies, variability, and distribution. we analyzed the 'likes' and the 'id' column to understand its distribution and identify outliers using

measures like variance, standard deviation, and IQR. Additionally, multivariate analysis included visualizations such as count plots for categorical variables like 'inferred company,' providing insights into company distribution and engagement levels.

**Univariate Analysis**: The 'likes' column was analyzed to understand its spread, modality, and presence of outliers. Measures of spread, including variance, standard deviation, and IQR, were calculated to assess the variability of likes across the dataset.

Ex: Likes outliers

```
Likes Outliers:
1         2750
19        1138
21         907
23        1309
26         961
          ...
299957    1332
299974    1272
299977    2070
299978    1433
299989    6028
Name: likes, Length: 40577, dtype: int64
Mode of Inferred Company: independent
Most Common Company: independent
```




| | id | likes |
|---|---|---|
| count | 300000.000000 | 300000.000000 |
| mean | 150000.500000 | 773.364793 |
| std | 86602.684716 | 4931.463419 |
| min | 1.000000 | 0.000000 |
| 25% | 75000.750000 | 3.000000 |
| 50% | 150000.500000 | 76.000000 |
| 75% | 225000.250000 | 364.000000 |
| max | 300000.000000 | 560193.000000 |

**Multivariate Analysis**: Visualizations such as count plots were generated to explore categorical variables like 'inferred

company.' These plots provided insights into the distribution of companies in the dataset and their respective engagement levels based on likes.

### *B. Data cleaning*

The data cleaning process aimed to rectify data inconsistencies, handle missing values, and standardize text data while removing unwanted text and emojis.

a. Normalizing Text Data:

The 'content' column, containing text data, was normalized to lowercase to ensure uniformity. Punctuations and emojis were removed from the text using Python's string manipulation functions, ensuring clean and standardized text format. A custom function remove_emoticons was applied to remove emojis from text data, enhancing the readability and analysis of textual content.

b. Removing Unwanted Text (Noise):

Unwanted text such as URLs, mentions, hashtags, and special characters were removed from the 'content' column to eliminate noise and focus on relevant content. Another custom function remove_noise was used to clean text data by removing unwanted text elements, enhancing data quality and analysis accuracy.

c. Handling Missing Values:

Rows with missing values were dropped from the dataset to maintain data integrity and avoid bias in subsequent analyses. Numerical columns with missing values were imputed with mean values to preserve data completeness while minimizing impact on statistical measures.

## III. RESULT AND ANALYSIS

1. Descriptive Statistics

Descriptive statistics were calculated to understand the data's central tendency, spread, and distribution Measures of spread (variance, standard deviation, IQR) were calculated for numerical columns like 'likes' to assess their variability. Skewness and kurtosis were analyzed to understand the data's shape and distribution.

Most Mentioned Companies: "Independent" has the highest skewness (7598.56), suggesting it's the company Kurt mentions the most. Companies like "AAA" (6073.06), "CNN" (2806.95), and "Cisco" (1829.51) also seem to be frequently mentioned.

Moderately Mentioned Companies: Companies like "Amazon" (1460.62), "Cameron" (1315.56), "Microsoft" (1179.93), and "Pfizer" (1095.97) have skewness values in the middle range, indicating a moderate level of mentions.

Least Mentioned Companies: "WWF" (888.21) and "Toyota" (780.18) have the lowest skewness values, suggesting they are mentioned by Kurt the least frequently compared to the other companies

```
Company: ('independent', 'kurt')
Skewness: 7598.56
----------------------------------
Company: ('aaa', 'kurt')
Skewness: 6073.06
----------------------------------
Company: ('cnn', 'kurt')
Skewness: 2806.95
----------------------------------
Company: ('cisco', 'kurt')
Skewness: 1829.51
----------------------------------
Company: ('amazon', 'kurt')
Skewness: 1460.62
----------------------------------
Company: ('cameron', 'kurt')
Skewness: 1315.56
----------------------------------
Company: ('microsoft', 'kurt')
Skewness: 1179.93
----------------------------------
Company: ('pfizer', 'kurt')
Skewness: 1095.97
----------------------------------
Company: ('wwf', 'kurt')
Skewness: 888.21
----------------------------------
Company: ('toyota', 'kurt')
Skewness: 780.18
----------------------------------
```

2. Inferred Company Analysis

Analysis of the 'inferred company' column revealed insights into the distribution of companies and their associated likes. Count plots and bar plots were used to visualize the distribution of inferred companies and their corresponding likes. Top companies based on likes were identified and analyzed for their average likes and distribution over time.

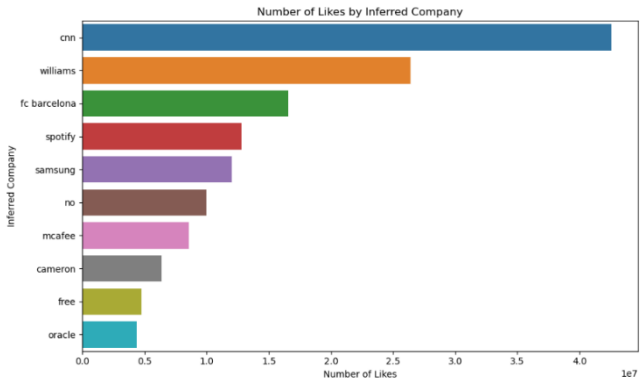
- Count Plots and Bar Plots:

Count plots and bar plots were utilized to visualize the distribution of inferred companies and their corresponding likes. These visualizations offered a clear depiction of company engagement levels and popularity based on the number of likes.

Top Liked Companies: Based on the y-axis order, "cnn" appears at the top, indicating it received the most likes among the listed companies.

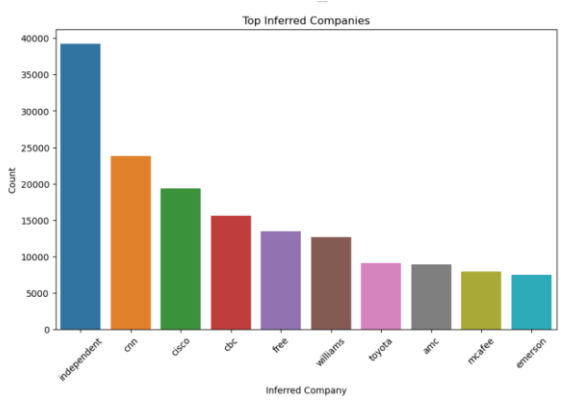
Moderately Liked Companies: Companies like "Spotify", "Samsung", "FC Barcelona", "Spotify", and "Williams" follow "CNN on the y-axis, suggesting they garnered a moderate number of likes compared to "CNN" but likely more than the companies below them.

Least Liked Companies: Companies positioned towards the bottom of the y-axis, including "free" ,"Cameron", "Oracle", likely received the fewest likes in this sample.

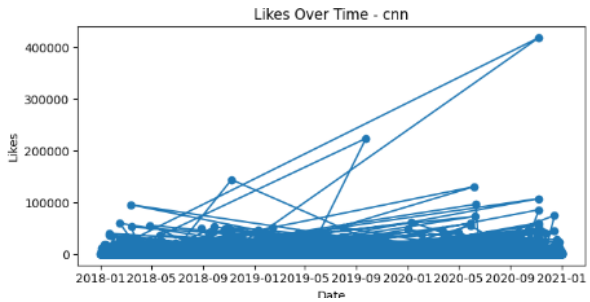


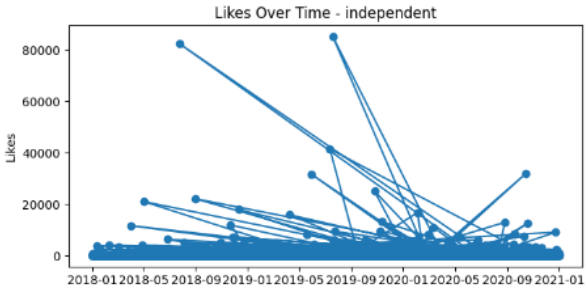

- Top Companies Analysis:

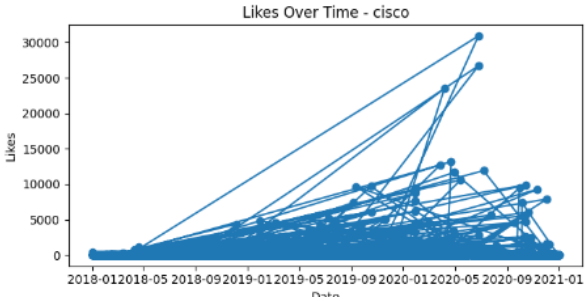
Top companies based on likes were identified and analyzed for their average likes and distribution over time. This analysis helped identify key players in terms of user engagement and popularity within the dataset.




- Line Graphs for Trends Over Time:



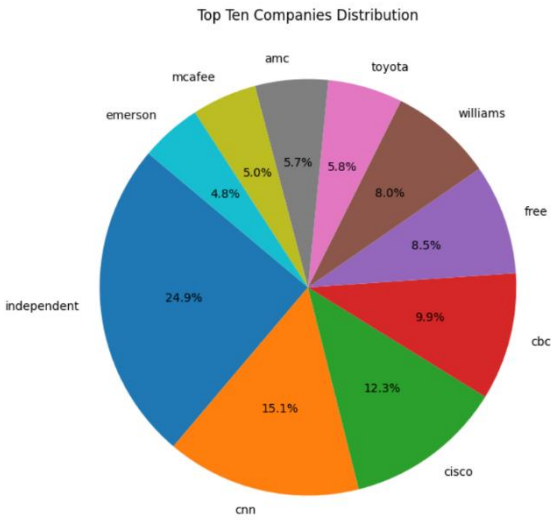




Line graphs were employed to observe trends over time for the top companies. This analysis allowed us to track changes in likes and user engagement patterns across different time periods, providing insights into temporal trends and fluctuations.

- Pie Chart for Data Proportions:

A pie chart was used to represent the proportions data proportions of each company that is present in the dataset. We represented the data proportions of the first 10 companies in the dataset where independent is the leading in the size of data acquired.




### A. *Abbreviations and Acronyms*

EDA - Exploratory Data Analysis
MSE - Mean Squared Error

MAE - Mean Absolute Error
IQR - Interquartile Range

## REFERENCES


[1] Kwon, N., Lee, J., & Park, Y. (2019). Social media engagement: A literature review and framework for future research. *Journal of Information Science*, 45(3), 392-405.

[2] What is Exploratory Data Analysis by Prasad Patil published in Towards Data Science

[3] EDA and models . Python · IEEE-CIS Fraud Detection by ANDREW LUKYANENKO- kaggle

[4] Social media | EDA | cleaning | data preparation by ELENI G -Kaggle

[5] P. Pushkar and U. Mittal, "User Behavior Analysis based on their Social Media Interaction," *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, 2022, pp. 107-110, doi: 10.1109/ICIEM54221.2022.9853113.