# Introduction

In this project, I developed an interactive Visual Analytics dashboard to study and explore engagement data across multiple social media platforms. My objective was to create a system that combines machine learning techniques, such as t-SNE for dimensionality reduction and KMeans for clustering, with interactive and synchronized visualizations. By building this dashboard, I wanted to better understand how engagement rates differ across clusters, demographic groups, platforms, and sentiment categories, while making it easy to identify patterns, outliers, and valuable audience insights.

# Data Preprocessing

Before constructing the dashboard, I performed a full cleaning and transformation process on the dataset, which originally contained 100,000 rows and 18 columns. Some fields had very high percentages of missing values, such as Influencer ID with about 90% null values and Campaign ID with around 80%. Since those were administrative fields with no analytical relevance, I removed them entirely. The Sentiment column had nearly 50% missing entries, and while I first considered filling them with an 'Unknown' label, I ultimately chose to drop those rows, as the analysis is based on a sample of 5,000 rows and removing them still left me with enough data for the t-SNE and clustering. The remaining important columns like Engagement Rate, Platform, Age, Gender, and Audience Location had no significant missing data, so I preserved them. Once the null values were handled, I standardized the dataset by grouping Age into five brackets (18–24, 25–34, 35–44, 45–54, 55+), encoding Gender as 0 for male and 1 for female, merging all U.S. states into a single 'United States' category, and normalizing Sentiment into lowercase categories (positive, neutral, and negative). I also converted Post Timestamp into a datetime format so I can add time-based analyses later, even though this feature is not yet displayed.

# Dimensionality Reduction and Clustering

After the dataset was cleaned and standardized, I selected a random sample of 5,000 rows to make the processing more efficient. Using this sample, I applied t-SNE to generate two coordinates, TSNE1 and TSNE2, which allowed me to represent the data in a two-dimensional space while preserving the structure of relationships between points. Then I used KMeans to cluster the data into five distinct groups, and these clusters were later represented with different colors in the scatterplot to make them easy to distinguish and explore visually.

# Interactive Dashboard

The final dashboard, which I implemented in tsne\_clusters.html, brings together four fully connected visualizations: a t-SNE scatterplot, a bar chart showing how posts are distributed by platform, another bar chart displaying gender distribution, and a histogram showing engagement rate distribution. The scatterplot is the central interactive element. I can draw a selection region on it to focus on a specific cluster or group of points, and as soon as I do, the bar charts and histogram automatically update to reflect only the selected data. When I clear the selection, the charts return to showing the complete filtered dataset, and if I double-click on the scatterplot, the view resets to its default zoom and position.

# Filters, Interaction, and Conclusion

To make the analysis flexible, I included several filtering options. I can choose specific age groups and regions from dropdowns, switch between positive, negative, neutral, or all sentiment using dedicated buttons, and even use the bar charts to activate additional filters by clicking on individual platforms or genders. To simplify navigation, I added a Reset button placed to the left of the age filter, which clears all active filters and any scatterplot selection, bringing the dashboard back to its default state so I can start a new exploration whenever I want. With these tools, I can explore engagement both broadly and at the cluster level, finding groups where engagement is particularly high, understanding which demographics or platforms are most relevant, and identifying smaller, specialized segments for targeted campaigns. This combination of dimensionality reduction, clustering, and interactive visualizations makes the dashboard a powerful and user-friendly way to analyze large and complex datasets, and I plan to expand it in the future with automatic cluster summaries, exportable reports, time series visualizations, and more advanced navigation controls like zooming and panning on the scatterplot.