Final Project Report

TikTok Data Analysis Project – Travel/Visa Niche
Course: Data Analytics for AI Applications
Faculty of Computer Science, TU Chemnitz

Author: Ayodele Moses Omolanke

Matric Number: 877734

Date: 22nd January, 2026

1. Project Overview

The goal of this project is to analyze TikTok video performance using real data from my own TikTok account in the travel and visa niche. TikTok is a platform where content performance depends on many factors such as hook style, content format, video duration, and audience engagement. However, many creators rely on intuition rather than data-driven decisions.

In this project, I aim to use data analytics techniques to understand which factors contribute to higher engagement on TikTok videos related to travel, visas, scholarships, internships, and international opportunities. The final goal is to build a simple machine learning model that can help predict whether a video is likely to perform well based on its characteristics.

The project follows a full data analytics pipeline, starting from dataset design and data collection, moving through exploratory data analysis, and later applying machine learning techniques.

2. Dataset Description

The dataset used in this project is collected manually from TikTok Analytics. It consists of real performance data from my own TikTok videos. At the current stage, the dataset contains 50 TikTok videos collected from November–January 2026, covering a range of low-, medium-, and high-performing content. The dataset was frozen at 50 videos to ensure consistency during analysis.

Each data entry represents one TikTok video and includes both numerical and categorical features. The collected features include:

- Views
- Likes
- Comments
- Shares
- Saves (favorites)
- Profile visits
- Followers gained
- Video duration (seconds)

- Watch time percentage
- Average watch time (seconds)
- Posting date, time, and day of the week
- Hook style (Fomo, Value, Tutorial, Question)
- Content format (text overlay, voiceover, screen recording)
- Niche category (visa, scholarship, internship, conference, study)
- Hashtag information

In addition, a new derived feature called **engagement_rate** was created using the formula:

engagement_rate = (likes + comments + shares + saves) / views

This feature provides a better measure of audience interaction than views alone, since it captures how actively viewers engage with the content.

The dataset is stored in both Excel and CSV formats and is version-controlled using GitHub.

3. Milestones Completed

These milestones were **completed** ahead of the mid-level deadline:

3.1 Dataset Design and Collection

- Designed a structured dataset suitable for data analytics and machine learning.
- Collected real TikTok analytics data manually for 10 videos.
- Ensured data consistency and correctness across all features.

3.2 Feature Engineering

- Created a new engagement metric (**engagement_rate**) to better measure performance.
- Categorized videos by hook style, content format, and niche topic.
- Hook Style Labeling Rubric

Category	Definition
Fomo	Content that creates urgency or fear of missing out (e.g., "Don't miss this opportunity").
Question	Content that begins with a direct question to the viewer.
Value	Content that focuses on providing useful or informative benefits.
Tutorial	Content that explains a process or steps in a structured manner.

Labels were assigned manually using a predefined rubric and rechecked for consistency across the dataset.

3.3 Exploratory Data Analysis (EDA)

Performed initial exploratory analysis using Excel. Created visualizations to analyze engagement patterns across videos and content features.

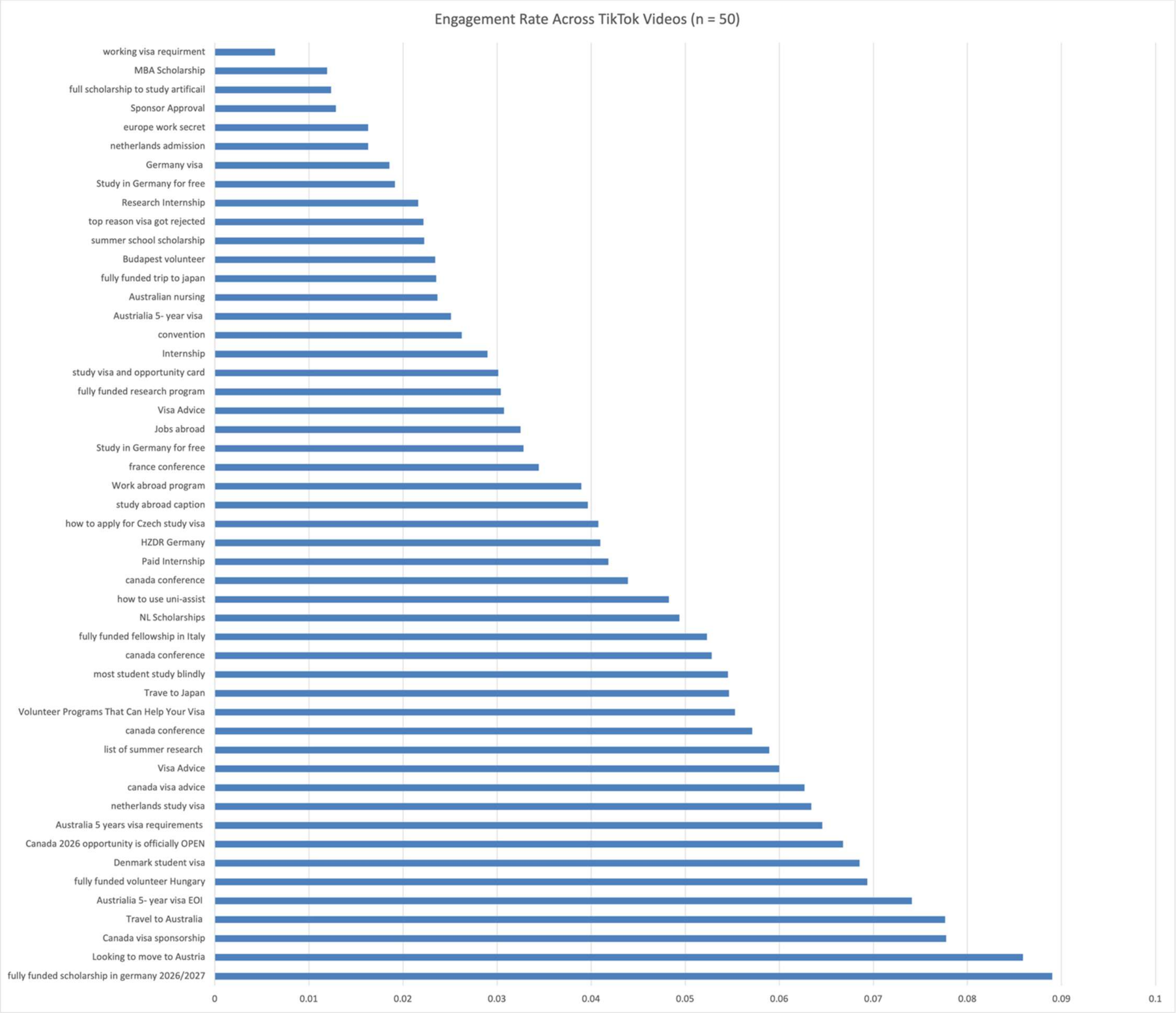

Figure 1: Engagement rate across TikTok videos (n = 50).

For readability, videos are ordered by engagement rate.

Table 1: Average engagement rate by hook style (n = 50)

Hook Style	Average Engagement Rate	Number of Videos (n)
Fomo	0.05371381	6
Question	0.04315071	**19**
Tutorial	0.063211973	**1**
Value	0.036627455	**24**

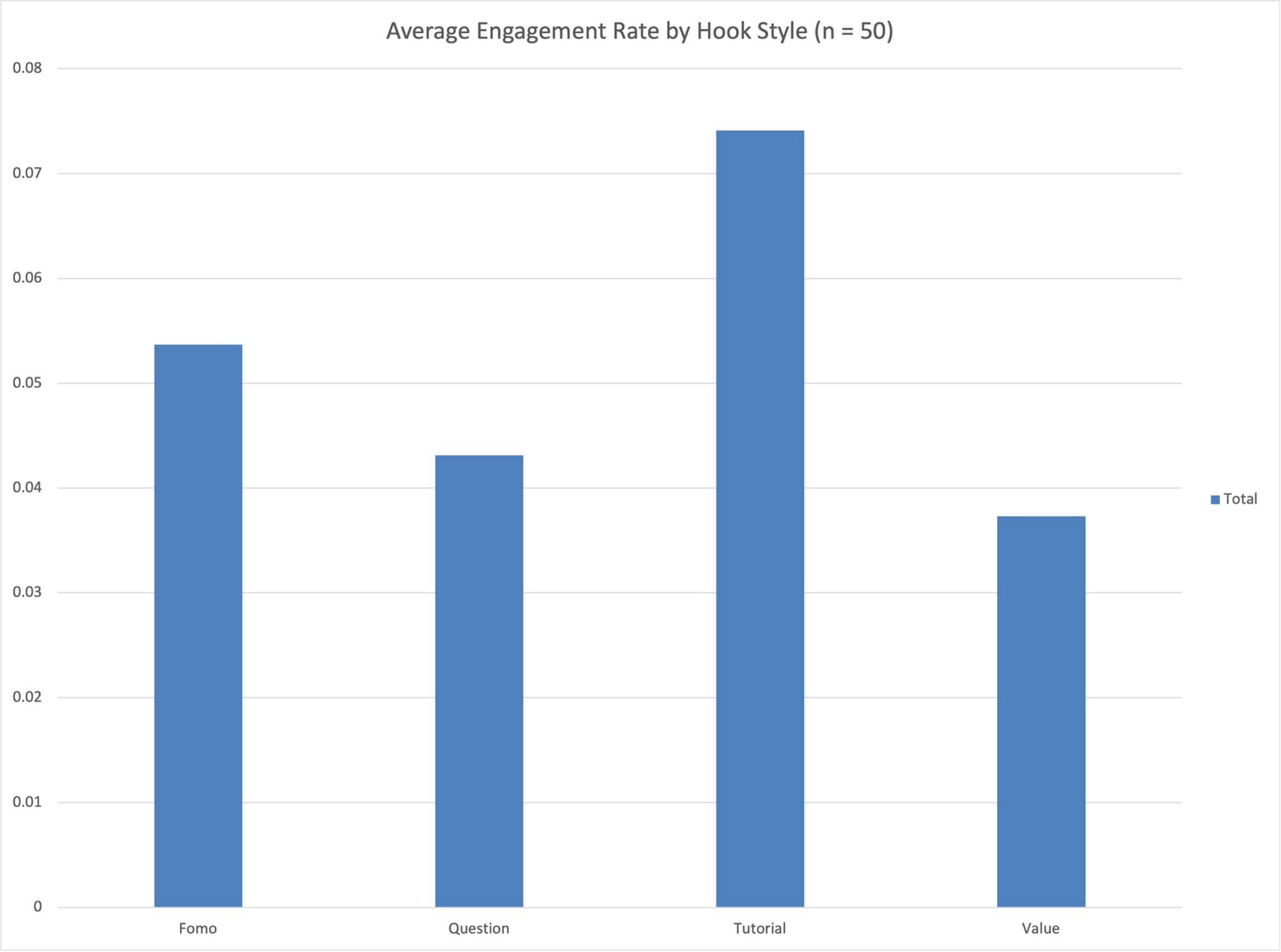

Figure 2: Average engagement rate by hook style (n = 50).

This figure shows the average engagement rate for each hook style category.

Tutorial-style videos show the highest average engagement rate; however, this result is based on a very small sample size and should be interpreted with caution.

Value-based content showed lower average engagement despite having the largest number of videos

This visualization shows clear variation in engagement across videos. Videos using Fomo-style hooks tend to appear more frequently among the highest engagement rates, while value-based hooks show lower average engagement despite having a larger sample size.

Sample sizes vary across hook styles. In particular, the tutorial category has a very small sample size (n = 1), so conclusions for this category should be interpreted cautiously.

Performance Classification

To enable basic performance modeling, videos were classified as high-performing or low-performing based on their engagement rate. A video was labeled as high-performing if its engagement rate exceeded the median engagement rate of the dataset. Otherwise, it was labeled as low-performing.

Performance Label	Number of Videos
High	25
Low	25

Using the median as a threshold provides an objective and data-driven performance definition.

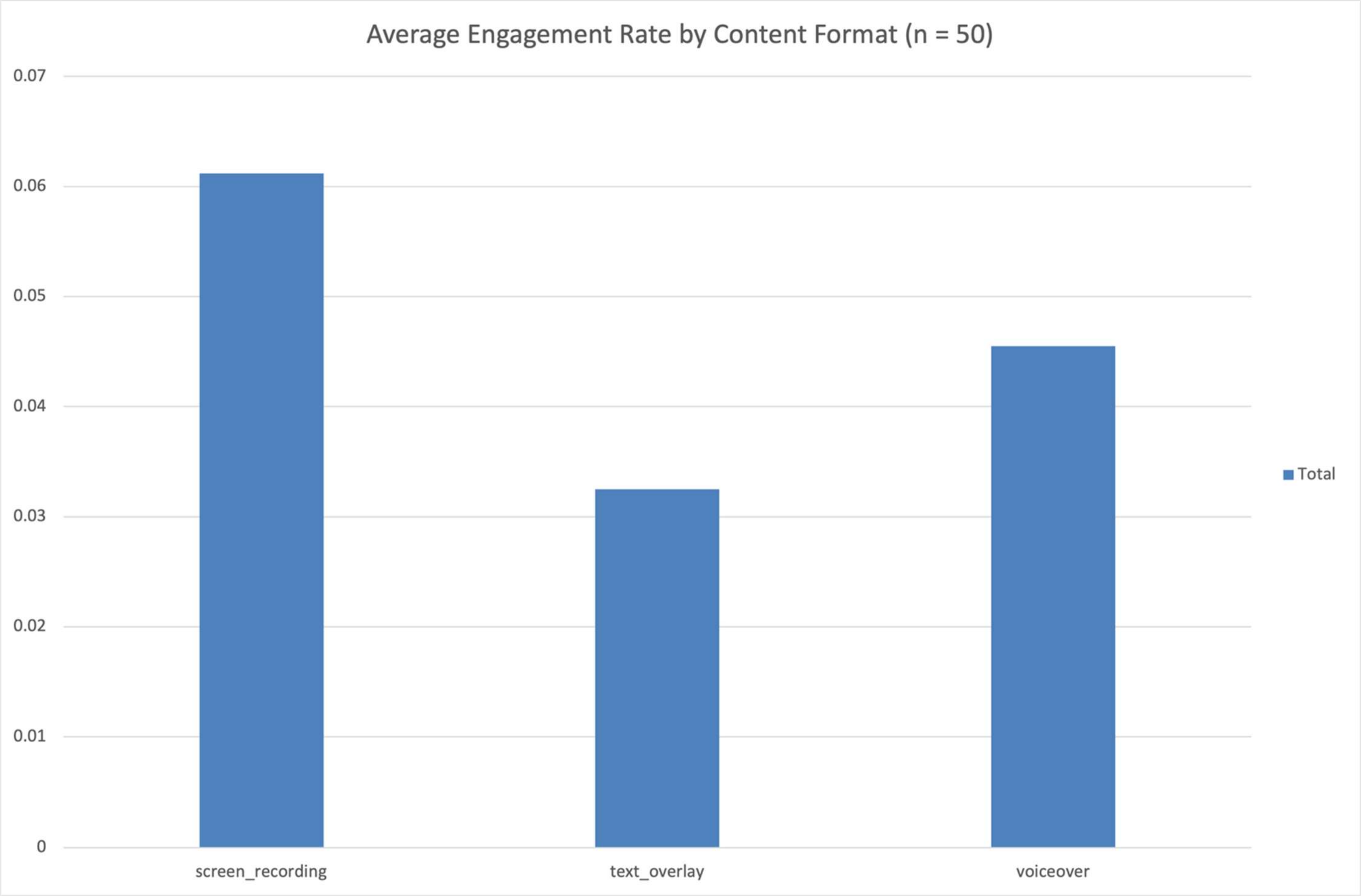

Figure 3: Average engagement rate by content format (n = 50)

Figure 3 shows the average engagement rate across different content formats.
Screen recording videos achieved the highest average engagement rate, followed by voiceover content.
Text-overlay-only videos showed comparatively lower engagement.

3.4 Project Organization and Documentation

- Created a well-structured GitHub repository including folders for data, documentation, notebooks, source code, and presentation.
- Uploaded dataset files, report drafts, and presentation slides.
- Prepared and delivered a successful progress presentation, which received positive feedback from the lecturer.

These completed milestones demonstrate that the project is progressing according to plan and that the foundational data analysis work is already in place.

4. Modelling Approach

4.1 Feature and Target Definition

Feature Variables (X)

The model uses a small set of simple and interpretable features derived from the dataset.

Categorical features:
• *Hook style* (FOMO, Question, Value, Tutorial)
• *Content format* (voiceover, text_overlay, screen_recording)

These features were selected due to their relevance to content strategy and their interpretability in explaining model predictions.

The prediction task is formulated as a binary classification problem.
The target variable is **performance_label**, indicating whether a video is high- or low-performing based on whether its engagement rate exceeds the dataset median.

The feature set includes content-related variables available at posting time: **hook style**, **content format**, and **video duration**. Engagement outcome variables such as views, likes, comments, and shares were excluded to prevent data leakage.

4.2 Model Training

A logistic regression model was trained to classify videos as high- or low-performing. Categorical features were transformed using one-hot encoding, while numerical features were used directly.
The dataset was split into training (70%) and testing (30%) subsets to evaluate generalization performance.
Model performance was evaluated using classification accuracy.

4.3 Model results and Interpretation

The model achieved an accuracy in the range of approximately 60–65%.

The logistic regression model achieved moderate classification accuracy when distinguishing between high- and low-performing videos. Accuracy represents the proportion of correctly classified videos in the test set. The results indicate that content-related features such as hook style, content format, and video duration provide some predictive signal, but do not fully explain engagement outcomes.

The results suggest that while content design influences engagement, additional factors such as audience behavior, platform dynamics, and timing may also play a significant role.

Due to the relatively small dataset size (n = 50), the model results should be interpreted cautiously. The balanced class distribution is a consequence of the median-based labeling strategy rather than a reflection of real-world class frequencies. More complex models and stronger generalization would require a larger and more diverse dataset.

4.4 Handling Class Imbalance and Consistency Checks

During the analysis, the dataset showed a mild class imbalance between high-performing and low-performing videos, with a larger proportion of videos labelled as *High* engagement.

To address this, the performance threshold was defined objectively using the median engagement rate, ensuring consistent and data-driven labelling across all videos. This prevents subjective bias and keeps the class definition stable.

In addition, model results were interpreted cautiously, focusing on relative trends (e.g. which hook styles or content formats perform better) rather than absolute prediction accuracy.

Given the exploratory nature of the project and the limited dataset size (n = 50), no aggressive re-sampling techniques were applied. Instead, consistency was ensured through:

- A fixed labeling rule (median-based threshold)
- Manual verification of categorical labels
- Cross-validation during model training to reduce overfitting

This approach balances interpretability and robustness while remaining appropriate for a small real-world dataset.

5. Conclusion

This project analyzed TikTok engagement patterns in the travel and visa content niche using a dataset of 50 videos collected from real account analytics. Engagement rate was used as the primary performance metric, and videos were objectively classified into high- and low-performing categories using a median-based threshold.

Exploratory data analysis showed that engagement varies substantially across videos and is influenced by both hook style and content format. FOMO-style hooks and tutorial content showed higher average engagement, although some categories were limited by small sample sizes. Screen-recorded videos also demonstrated stronger engagement compared to text-overlay formats.

A simple logistic regression model was trained to predict video performance using content-related features available at posting time. The model achieved moderate predictive performance, indicating that while content design contributes to engagement outcomes, it does not fully explain them.

Overall, the results highlight the importance of data-driven content analysis while also emphasizing the limitations of small datasets. Future work can extend this analysis with larger datasets, additional features, and more advanced validation techniques.