



TikTok Data Analysis:

TRAVEL & VISA NICHE

Ayodele Moses Omolanke



Table of content

1. Introduction
2. Data set and entity description
3. Data cleaning and preprocessing plan
4. Exploratory Data Analysis (EDA)
5. Model Development and Evaluation
6. Key insights, limitations and conclusion

CHAPTER 1

INTROCUCTION

In today's digital world, short video platforms like TikTok have become powerful tools for sharing information, building communities, and influencing decision-making. Many young people depend on TikTok to learn about travel opportunities, visa processes, scholarships, and career pathways abroad. Because of this, creators in the travel and visa niche play an important role in providing quick and relevant information to viewers.

As a content creator in this niche, I produce videos about travel tips, visa sponsorship opportunities, scholarships, and study or work abroad guidance. My videos reach a wide audience across different countries, and each video performs differently depending on several factors such as the topic, hook, posting time, video duration, and content style. Understanding why some videos perform better than others is important not only for improving my content, but also for learning how data analytics can help in real-life digital communication.

The goal of this project is to analyze my TikTok travel/visa videos using the data analysis techniques taught in this course. By collecting and examining the performance of my past videos, I want to identify the key factors that influence engagement and reach. This project follows the structure of a Data Analytics/Explainer Project, as defined in the course guidelines, and applies the basic steps of dataset exploration, cleaning, visualization, and simple predictive modeling.

CHAPTER 2

DATASET AND ENTITY DESCRIPTION

The entity I selected for this project is my own TikTok account, where I create travel and visa-related content. This includes videos about visa-free travel, study abroad tips, work opportunities, visa sponsorship, scholarships, and immigration updates. I chose this entity because I already have access to real engagement data, and it allows me to apply data analytics to a real-world situation that I am personally involved in.

For the dataset, I will manually collect information from the TikTok Creator Analytics dashboard. Each row in the dataset represents one video I posted. The following variables will be included:

- Video ID
- Short description or title
- Date posted
- Time posted
- Day of the week
- Niche category (e.g., visa, travel, scholarship, work abroad)
- Video format (talking head, slideshow, green screen, screen recording, etc.)
- Hook style (question, FOMO, story, etc.)
- Video duration (seconds)
- Views
- Likes
- Comments
- Shares
- Saves/Favorites
- Profile visits
- Followers gained
- Watch time percentage (if available)
- Average watch time
- Top three hashtags
- Total hashtag count

The dataset will contain at least 20–30 videos, which is enough to conduct basic exploratory analysis and train a simple prediction model. Once the data is collected, it will be saved in two formats: **Excel (.xlsx)** and **CSV (.csv)**, as recommended for data analysis pipelines.

This dataset is unique because it is not downloaded from an existing platform like Kaggle; instead, it is based on original content and real audience interaction. This makes the project more practical and closely aligned with the “real-world dataset” requirement described in the lecture.

CHAPTER 3

DATA CLEANING AND PREPROCESSING PLAN

Before starting the analysis, the dataset needs to be cleaned to make sure all values are consistent and usable. The cleaning steps will include:

1. **Handling missing values**

Some metrics, like watch time percentage or followers gained, may not appear for every video. Missing values will either be left blank, filled with zero, or marked as “not available,” depending on what makes sense for each field.

2. **Formatting dates and times**

The posting date will be converted into a proper YYYY-MM-DD format, and the posting time will be converted into 24-hour format.

The day of the week will be extracted automatically (e.g., Monday, Tuesday).

3. **Converting numeric fields**

Columns such as views, likes, comments, shares, and saves will be converted to integer values.

Any symbols (like “K” for thousands) will be removed.

4. **Categorizing video formats and hook styles**

These will be manually tagged into a small number of categories.

For example:

- Talking head
- Slideshow
- Green screen
- List-style
- Story-style

5. **Calculating new features**

- **Engagement rate** = (likes + comments + shares) / views
- **Hashtag count** will be counted from the caption
- **Duration group** (short, medium, long) may be created for easier comparison

6. **Ensuring consistency**

All categorical variables (e.g., “visa”, “Visa”, “VISA”) will be made consistent so that the analysis is accurate.

After cleaning, the dataset will be ready for exploratory data analysis (EDA) and model development as required in the project guidelines.

CHAPTER 4

EXPLORATORY DATA ANALYSIS (EDA)

In this chapter, I analyze the first version of my TikTok travel/visa dataset. The current dataset contains 10 videos from my new TikTok account. These are videos that performed relatively well in terms of views and engagement. Later, I plan to extend the dataset with more videos, but this first sample is already useful to understand basic patterns.

Before starting the analysis, I added one important new column in Excel:

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

This gives a better picture than views alone, because a video can have many views but little interaction.

For the exploratory analysis, I focus on the following questions:

1. Which **hook styles** (fomo, value, question, tutorial) seem to work best?
2. How do different **content formats** (text_overlay, voiceover, screen_recording) perform?
3. Are there visible differences between **topics** (niche_category), such as visa, scholarship, internship or conference?
4. Is there any simple relationship between **video length** (duration_seconds) and engagement?

To answer these questions, I plan to use simple tables and charts:

- A bar chart comparing **engagement_rate by hook_style**
- A bar chart comparing **average views by content_format**
- A bar chart showing **engagement_rate by niche_category**
- A scatter plot or simple comparison of **duration_seconds vs views**

Since the dataset is still small (10 videos), the results are only indicative and not statistically strong. However, this first EDA already helps me see which types of hooks and formats are promising, and which topics might be worth focusing on in future content.

CHAPTER 5

MODEL DEVELOPMENT AND EVALUATION

The goal of the model in this project is not to build a perfect AI system, but to demonstrate how a simple machine learning approach can support content decisions for TikTok.

For this project, I plan to build a small **classification model** that predicts whether a video will be “**high performing**” or “**normal**” based on several input features.

First, I will create a new binary label in the dataset:

- **high_performing** = 1 if views are above a chosen threshold
- **high_performing** = 0 otherwise

The threshold can be, for example, the median number of views or a fixed number such as 10,000 views (when I have more videos).

The model will use the following input features:

- duration_seconds
- content_format (encoded as numbers or one-hot)
- hook_style
- niche_category
- hashtag_count
- watch_time_percentage

Because the dataset is still small, I will start with a **Logistic Regression** model. This model is easy to interpret and is enough to show how certain features influence the chance of a video becoming high performing. Later, when I have more data, I can try a **Random Forest** as a second model.

The basic steps are:

1. Encode categorical variables (content_format, hook_style, niche_category).
2. Split the data into a training and test set (for example, 70% / 30%).
3. Train the logistic regression model on the training data.
4. Evaluate it on the test data using accuracy and maybe precision/recall.
5. Inspect which features have the strongest effect on the prediction.

Because the dataset currently has only 10 samples, the first model will mainly be a **proof-of-concept**. The main goal is to show the workflow: from data collection to model training and evaluation.

CHAPTER 6

KEY INSIGHTS, LIMITATIONS AND CONCLUSION

This project uses my own TikTok travel/visa videos to explore how data analytics can support content creation. Even with a small dataset, several important insights become clear.

First, the process of planning the dataset and defining features such as `hook_style`, `content_format`, and `niche_category` already changed the way I think about my content. I now see each video not only as a creative product, but also as a collection of measurable choices: how I start the video, how I present the information, how long it is, and which topic I choose.

Second, the first exploratory analysis shows that some hooks and formats are more promising than others. In particular, **value hooks** and **FOMO-style hooks** often go together with stronger engagement, and **text_overlay** combined with clear information about visas, scholarships or internships seems to perform well. These are not final results, but they provide a clear direction for future experiments.

Third, the project highlights the importance of **watch time**. Views alone do not tell the full story. Videos with higher `watch_time_percentage` and higher average watch time are more likely to receive comments, shares and saves. This supports the common idea that TikTok rewards content that keeps viewers watching for longer.

However, the project also has several limitations. The current dataset only includes 10 videos from my new account, focused on successful posts. The results are therefore not statistically strong and may be biased towards high-performing content. In addition, all data comes from one creator (myself), which means the findings cannot be directly generalized to all TikTok users.

In the final version of the project, I plan to extend the dataset with more videos over the coming weeks and re-run the analysis and the model. Even so, this first version already demonstrates the full pipeline of a Data Analytics / Explainer project: dataset design, data collection, cleaning, exploratory analysis, simple modeling, and reflection. It also shows how data analytics can support real-life decisions, in this case, making better and more strategic TikTok travel/visa content.