



Course 1: Foundations of Data Science

Background on the TikTok scenario

Welcome to your new role at TikTok! We're thrilled to have you on the data team!

At TikTok, our mission is to inspire creativity and bring joy. Our employees lead with curiosity and move at the speed of culture. Combined with our company's flat structure, you'll be given dynamic opportunities to make a real impact on a rapidly expanding company and grow your career. TikTok users have the ability to report videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. This process generates a large number of user reports that are difficult to address quickly.

TikTok is working on the development of a predictive model that can determine whether a video contains a claim or offers an opinion. With a successful prediction model, TikTok can reduce the backlog of user reports and prioritize them more efficiently.

Project background

TikTok's data team is in the earliest stages of the claims classification project. The following tasks are needed before the team can begin the data analysis process:

- A project proposal identifying the following:
 - Organize project tasks into milestones
 - Classify tasks using the PACE workflow
 - Identify relevant stakeholders

Your assignment

For your first assignment, TikTok is asking for a project proposal that will create milestones for the tasks within the comment classification project. Remember to take into account your audience, team, project goal, and PACE stages of each task in planning your project deliverable.

Team members at TikTok

As a new data analyst, you'll work closely with a talented team of experienced data professionals. You'll also collaborate with TikTok co-workers outside the data team as your project requires.

Data team roles

- Willow Jaffey- Data Science Lead
- Rosie Mae Bradshaw- Data Science Manager
- Orion Rainier- Data Scientist

Cross-functional team members

- Mary Joanna Rodgers- Project Management Officer
- Margery Adebawale- Finance Lead, Americas
- Maika Abadi- Operations Lead



Course 1: Foundations of Data Science

Note: The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. The data shared in this project has been altered for pedagogical purposes.

Your TikTok team includes several managers overseeing operations. It is important to adapt your communication to their roles since their responsibilities are less technical. The members of the data team at TikTok are well versed in data analysis and data science. Messages to these more technical co-workers should be concise and specific.

Meeting notes

Now that you are onboard, you have access to the company network and are set up with a company email account.

You notice an email from your supervisor at the top of your inbox.

From: Rosie Mae Bradshaw

Subject: Your first task: Review meeting notes

If you are able to read this, then your company accounts have been created! It's the perfect time to get started with your first task. Last week, I attended an internal meeting with our leadership team about a new project we are about to begin. You'll receive more information in the next few days, but I would like you to be aware of some needs that were identified by our leadership. Here is an excerpt from the notes I took during the leadership team meeting. I've organized the points by the individual who made the comment:

Mary Joanna Rodgers- Project Management Officer

- The project will need a global document identifying deliverables & milestones.
- We will need to generate some visuals to share with TikTok executives.

Orion Rainier-Data Scientist

- We will need to have an understanding of the data we will be working with on this project.
- It will require someone on the team to inspect the dataset for missing data.
- The data team needs to determine what information will be the most useful in TikTok's data through exploratory data analysis (EDA).
- Eventually, our team will need to run some method of statistical testing.

Willow Jaffey- Data Science Lead

- The goal of the claim classification project is to build and deliver a reliable machine learning model.
- The team will need to determine the type of regression model for this project.
- Once we have a final model, I'll need to know the main talking points going into our presentation to the leadership team.
- The team will need to determine the best method of hypothesis testing for this dataset.

My thoughts and concerns...

- I think it's best to use Python for this project. I'll have someone on my team set that up as soon as we have the project proposal in place.
- I'd suggest the data team will need to check the assumptions made by any regression modeling developed for this project.
- The data team will identify and set up the project's workflow.



Course 1: Foundations of Data Science

There will be more details sent to you very soon.

Welcome to the team,

Rosie Mae Bradshaw

Data Science Manager

TikTok

(P.S. Don't forget about the company-sponsored lunch on Friday. The food is always fantastic. It's a great opportunity to meet more of our wonderful co-workers. Our team meets in the lobby around noon. Please join us!)

TikTok project proposal

Overview

TikTok seeks a way to use machine learning to extract claims or propositions within TikTok videos and comments and classify each data point as a claim or opinion.

Milestone	Tasks	Deliverables/Reports	Relevant Stakeholder (Optional)
1	<div>Establish structure for project work</div> <div>Plan ▾</div>	<ul style="list-style-type: none">Global-level project document	Rosie Mae Bradshaw — DS Manager
1a	<div>Write a project proposal</div>		Mary Joanna Rodgers — PM Officer



Course 1: Foundations of Data Science

	Plan ▾		
2	Compile summary information a... Analyze ▾	• Data files ready for EDA	Orion Rainier — DS
2a	Begin exploring the data Analyze ▾		Orion Rainier — DS
3	Data exploration and cleaning Plan ▾ and Analyze ▾	• EDA results	Orion Rainier — DS
3a	Visualization building Analyze ▾ and Construct ▾	• Tableau dashboard/visualizations	Mary Joanna Rodgers — PM Officer
4	Compute descriptive statistics Analyze ▾	• Analysis of testing results between two important variables	Willow Jaffey — DS Lead



Course 1: Foundations of Data Science

4a	Conduct hypothesis testing Analyze and Construct		Orion Rainier — DS
5	Build a regression model Analyze and Construct	• Regression model	Willow Jaffey — DS Lead
5a	Evaluate the model Execute		Rosie Mae Bradshaw — DS Manager
6	Build a machine learning model Construct	• Final model	Willow Jaffey — DS Lead
6a	Communicate final insights with ... Execute	• Report to all stakeholders	Willow Jaffey — DS Lead

Note: The estimated times for the milestones in the example equate to the length of the courses where you will learn the necessary skills. Realistic timelines when working with actual clients and data scientists as a data scientist would most likely have tight deadlines, for example:

Milestone 1: 1-2 days

Milestone 2: 2-3 weeks



Course 1: Foundations of Data Science

Milestone 3: 1 week

Milestone 4: 1 week

Milestone 5: 1-2 weeks

Project goal:

The TikTok data team is developing a machine learning model for classifying claims made in videos submitted to the platform.

Background:

TikTok is the leading destination for short-form mobile video. The platform is built to help imaginations thrive. TikTok's mission is to create a place for inclusive, joyful, and authentic content—where people can safely discover, create, and connect.

Scenario:

As a data analyst on TikTok's data team, you'll help by preparing the data needed for the claims classification project. You'll build a dataframe, organize the claims data for the process of exploratory data analysis, and update the team on your progress and insights.

Course 2 tasks:

- Build a dataframe for the TikTok dataset
- Read in data from TikTok csv file
- Display rows within dataframe
- Examine data type of each column
- Gather descriptive statistics
- Visualize the TikTok data in Python
- Report to TikTok's data team through an executive summary

Note: The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.

This project uses a dataset called `tiktok_dataset.csv`. It contains synthetic data created for this project in partnership with TikTok. Examine each data variable gathered.

19,383 rows – Each row represents a different published TikTok video in which a claim/opinion has been made.

12 columns

Column name	Type	Description
#	int	TikTok assigned number for video with claim/opinion.



Course 1: Foundations of Data Science

claim_status	obj	Whether the published video has been identified as an “opinion” or a “claim.” In this dataset, an “opinion” refers to an individual’s or group’s personal belief or thought. A “claim” refers to information that is either unsourced or from an unverified source.
video_id	int	Random identifying number assigned to video upon publication on TikTok.
video_duration_sec	int	How long the published video is measured in seconds.
video_transcription_text	obj	Transcribed text of the words spoken in the published video.
verified_status	obj	Indicates the status of the TikTok user who published the video in terms of their verification, either “verified” or “not verified.”
author_ban_status	obj	Indicates the status of the TikTok user who published the video in terms of their permissions: “active,” “under scrutiny,” or “banned.”
video_view_count	float	The total number of times the published video has been viewed.
video_like_count	float	The total number of times the published video has been liked by other users.
video_share_count	float	The total number of times the published video has been shared by other users.
video_download_count	float	The total number of times the published video has been downloaded by other users.
video_comment_count	float	The total number of comments on the published video.