

TikTok Project

Course 4 - The Power of Statistics

You are a data professional at TikTok. The current project is reaching its midpoint; a project proposal, Python coding work, and exploratory data analysis have all been completed.

The team has reviewed the results of the exploratory data analysis and the previous executive summary the team prepared. You received an email from Orion Rainier, Data Scientist at TikTok, with your next assignment: determine and conduct the necessary hypothesis tests and statistical analysis for the TikTok classification project.

A notebook was structured and prepared to help you in this project. Please complete the following questions.

Course 4 End-of-course project: Data exploration and hypothesis testing

In this activity, you will explore the data provided and conduct a hypothesis testing.

The purpose of this project is to demonstrate knowledge of how to prepare, create, and analyze hypothesis tests.

The goal is to apply descriptive and inferential statistics, probability distributions, and hypothesis testing in Python.

This activity has three parts:

Part 1: Imports and data loading

- What data packages will be necessary for hypothesis testing?

Part 2: Conduct hypothesis testing

- How will descriptive statistics help you analyze your data?
- How will you formulate your null hypothesis and alternative hypothesis?

Part 3: Communicate insights with stakeholders

- What key business insight(s) emerge from your hypothesis test?

- What business recommendations do you propose based on your results?

Follow the instructions and answer the questions below to complete the activity. Then, complete an executive summary using the questions listed on the PACE Strategy Document.

Be sure to complete this activity before moving on. The next course item will provide you with a completed exemplar to compare to your own work.

Complete the following steps to perform statistical analysis of your data:

Task 1. Imports and Data Loading

Import packages and libraries needed to compute descriptive statistics and conduct a hypothesis test.

```
In [1]: # Import packages for data manipulation
import pandas as pd
import numpy as np

# Import packages for data visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Import packages for statistical analysis/hypothesis testing
from scipy import stats
```

Load the dataset.

Note: As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

```
In [2]: # Load dataset into dataframe
data = pd.read_csv("tiktok_dataset.csv")
```

response:

In general, descriptive statistics are useful because they let you quickly explore and understand large amounts of data. In this case, computing descriptive statistics helps you quickly compute the mean values of video_view_count for each group of verified_status in the sample data.

Task 2. Data exploration

Use descriptive statistics to conduct Exploratory Data Analysis (EDA).

Inspect the first five rows of the dataframe.

```
In [3]: # Display first few rows
data.head()
```

```
Out[3]:
```

	#	claim_status	video_id	video_duration_sec	video_transcription_text	verified
0	1	claim	7017666017	59	someone shared with me that drone deliveries a...	no
1	2	claim	4014381136	32	someone shared with me that there are more mic...	no
2	3	claim	9859838091	31	someone shared with me that american industria...	no
3	4	claim	1866847991	25	someone shared with me that the metro of st. p...	no
4	5	claim	7105231098	19	someone shared with me that the number of busi...	no

```
In [4]: # Generate a table of descriptive statistics about the data
data.describe()
```

```
Out[4]:
```

	#	video_id	video_duration_sec	video_view_count	video_like
count	19382.000000	1.938200e+04	19382.000000	19084.000000	19084.0
mean	9691.500000	5.627454e+09	32.421732	254708.558688	84304.6
std	5595.245794	2.536440e+09	16.229967	322893.280814	133420.
min	1.000000	1.234959e+09	5.000000	20.000000	0.0
25%	4846.250000	3.430417e+09	18.000000	4942.500000	810.
50%	9691.500000	5.618664e+09	32.000000	9954.500000	3403.5
75%	14536.750000	7.843960e+09	47.000000	504327.000000	125020.0
max	19382.000000	9.999873e+09	60.000000	999817.000000	657830.0

Check for and handle missing values.

```
In [5]: # Check for missing values
data.isna().sum()
```

```
Out[5]: #
claim_status      298
video_id          0
video_duration_sec 0
video_transcription_text 298
verified_status    0
author_ban_status  0
video_view_count   298
video_like_count   298
video_share_count  298
video_download_count 298
video_comment_count 298
dtype: int64
```

```
In [6]: # Drop rows with missing values
data = data.dropna(axis=0)
```

```
In [7]: # Display first few rows after handling missing values
data.head()
```

```
Out[7]:
```

	#	claim_status	video_id	video_duration_sec	video_transcription_text	verified_status
0	1	claim	7017666017	59	someone shared with me that drone deliveries a...	no
1	2	claim	4014381136	32	someone shared with me that there are more mic...	no
2	3	claim	9859838091	31	someone shared with me that american industria...	no
3	4	claim	1866847991	25	someone shared with me that the metro of st. p...	no
4	5	claim	7105231098	19	someone shared with me that the number of busi...	no

You are interested in the relationship between `verified_status` and `video_view_count`. One approach is to examine the mean values of `video_view_count` for each group of `verified_status` in the sample data.

```
In [8]: # Compute the mean `video_view_count` for each group in `verified_status`
#### YOUR CODE HERE ####
data.groupby("verified_status")["video_view_count"].mean()
```

```
Out[8]: verified_status
not verified    265663.785339
verified       91439.164167
Name: video_view_count, dtype: float64
```

Task 3. Hypothesis testing

Before you conduct your hypothesis test, consider the following questions where applicable to complete your code response:

1. Recall the difference between the null hypothesis and the alternative hypotheses. What are your hypotheses for this data project?

Exemplar response:

- **Null hypothesis:** There is no difference in number of views between TikTok videos posted by verified accounts and TikTok videos posted by unverified accounts (any observed difference in the sample data is due to chance or sampling variability).
- **Alternative hypothesis:** There is a difference in number of views between TikTok videos posted by verified accounts and TikTok videos posted by unverified accounts (any observed difference in the sample data is due to an actual difference in the corresponding population means).

Your goal in this step is to conduct a two-sample t-test. Recall the steps for conducting a hypothesis test:

1. State the null hypothesis and the alternative hypothesis
2. Choose a significance level
3. Find the p-value
4. Reject or fail to reject the null hypothesis

H_0 : There is no difference in number of views between TikTok videos posted by verified accounts and TikTok videos posted by unverified accounts (any observed difference in the sample data is due to chance or sampling variability).

H_A : There is a difference in number of views between TikTok videos posted by verified accounts and TikTok videos posted by unverified accounts (any observed difference in the sample data is due to an actual difference in the corresponding population means).

You choose 5% as the significance level and proceed with a two-sample t-test.

```
In [9]: # Conduct a two-sample t-test to compare means
      ### YOUR CODE HERE ###

      # Save each sample in a variable
      not_verified = data[data["verified_status"] == "not verified"]["video_view_c
      verified = data[data["verified_status"] == "verified"]["video_view_count"]

      # Implement a t-test using the two samples
      stats.ttest_ind(a=not_verified, b=verified, equal_var=False)
```

```
Out [9]: Ttest_indResult(statistic=25.499441780633777, pvalue=2.6088823687177823e-12
      0)
```

Response:

Since the p-value is extremely small (much smaller than the significance level of 5%), you reject the null hypothesis. You conclude that there **is** a statistically significant difference in the mean video view count between verified and unverified accounts on TikTok.

Task 4. Communicate insights with stakeholders

Ask yourself the following question:

- What business insight(s) can you draw from the result of your hypothesis test?

Response:

The analysis shows that there is a statistically significant difference in the average view counts between videos from verified accounts and videos from unverified accounts. This suggests there might be fundamental behavioral differences between these two groups of accounts.

It would be interesting to investigate the root cause of this behavioral difference. For example, do unverified accounts tend to post more clickbait-y videos? Or are unverified accounts associated with spam bots that help inflate view counts?

The next step will be to build a regression model on `verified_status`. A regression model is the natural next step because the end goal is to make predictions on claim status. A regression model for `verified_status` can help analyze user behavior in this group of verified users. Technical note to prepare regression model: because the data is skewed, and there is a significant difference in account types, it will be key to build a logistic regression model.