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 hypothesis testing.

The purpose of this project is to demostrate 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?

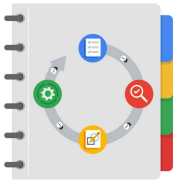
Part 3: Communicate insights with stakeholders

- 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.

Data exploration and hypothesis testing



PACE stages

Throughout these project notebooks, you'll see references to the problem-solving framework PACE. The following notebook components are labeled with the respective PACE stage: Plan, Analyze, Construct, and Execute.

PACE: Plan

Consider the questions in your PACE Strategy Document and those below to craft your response.

Question: What is your research question for this data project? Later on, you will need to formulate the null and alternative hypotheses as the first step of your hypothesis test. Consider your research question now, at the start of this task.

- The overall question being addressed in this project is: How can we predict whether a video contains a claim or opinion? For this portion of the project, the question is: Is there a statistically significant difference between the engagement of videos authored by verified users versus unverified users?

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.

▶

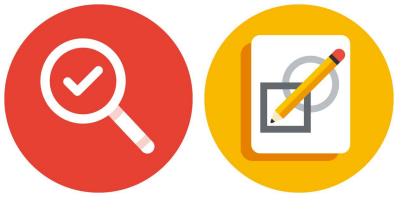
Hint:

In [2]:
```
# Import packages for data manipulation
import numpy as np
import pandas as pd

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

# 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 [3]:
```
# Load dataset into dataframe
data = pd.read_csv("tiktok_dataset.csv")
```



PACE: Analyze and Construct

Question: Consider the questions in your PACE Strategy Document and those below to craft your response: Data professionals use descriptive statistics for Exploratory Data Analysis. How can computing descriptive statistics help you learn more about your data in this stage of your analysis?

- Descriptive statistics provide a comprehensive summary of the dataset, offering insights into various aspects of the data. Key metrics include measures of central tendency such as mean, median, and mode, which indicate the average values and the most common occurrences within the data. Additionally, descriptive statistics reveal the distribution of variables, helping to identify whether the data follows a normal distribution, is skewed, or has other patterns.

- Furthermore, these statistics help in detecting the presence of outliers, which are data points that deviate significantly from the rest of the dataset. Identifying outliers is crucial as they can impact the results of statistical analyses. Descriptive statistics also highlight the existence of missing values, which need to be addressed to ensure the accuracy and reliability of the analysis.

- Overall, by summarizing the main features of the data, descriptive statistics enable a better understanding of the dataset's structure and inform subsequent steps in the data analysis process.

Task 2. Data exploration

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

▶

Hint:

Inspect the first five rows of the dataframe.

In [4]:
```
# Display first few rows
data.head()
```

Out[4]:

	#	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...	not verified
1	2	claim	4014381136	32	someone shared with me that there are more mic...	not verified
2	3	claim	9859838091	31	someone shared with me that american industria...	not verified
3	4	claim	1866847991	25	someone shared with me that the metro of st. p...	not verified
4	5	claim	7105231098	19	someone shared with me that the number of busi...	not verified

In [5]:
```
# Generate a table of descriptive statistics about the data
data.describe()
```

Out[5]:

	#	video_id	video_duration_sec	video_view_count	video_like_count
count	19382.000000	1.938200e+04	19382.000000	19084.000000	19084.000000
mean	9691.500000	5.627454e+09	32.421732	254708.558688	84304.636030
std	5595.245794	2.536440e+09	16.229967	322893.280814	133420.546814
min	1.000000	1.234959e+09	5.000000	20.000000	0.000000
25%	4846.250000	3.430417e+09	18.000000	4942.500000	810.750000
50%	9691.500000	5.618664e+09	32.000000	9954.500000	3403.500000
75%	14536.750000	7.843960e+09	47.000000	504327.000000	125020.000000
max	19382.000000	9.999873e+09	60.000000	999817.000000	657830.000000

Check for and handle missing values.

In [6]:
```
# Check for missing values
data.isnull().sum()
```

Out[6]:
```
#                           0
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 [7]:
```
# Drop rows with missing values
data = data.dropna()
```

In [12]:
```
# Display first few rows after handling missing values
data.head()
```

Out[12]:

	#	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...	not verified
1	2	claim	4014381136	32	someone shared with me that there are more mic...	not verified
2	3	claim	9859838091	31	someone shared with me that american industria...	not verified
3	4	claim	1866847991	25	someone shared with me that the metro of st. p...	not verified
4	5	claim	7105231098	19	someone shared with me that the number of busi...	not verified

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

In [18]:
```
# Compute the mean `video_view_count` for each group in `verified_status`
print(data.groupby('verified_status')['video_view_count'].mean())
print()
print(data.groupby('author_ban_status')['video_view_count'].mean())
```

```
verified_status
not verified    265663.785339
verified         91439.164167
Name: video_view_count, dtype: float64

author_ban_status
active          215927.039524
banned          445845.439144
under review    392204.836399
Name: video_view_count, dtype: float64
```

Task 3. Hypothesis testing

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 signficance level
3. Find the p-value
4. Reject or fail to reject the null hypothesis

Null Hypothesis (H0): There is no statistically significant difference in the number of views between videos posted by verified accounts and videos posted by unverified accounts.

Alternative Hypothesis (H1): There is a statistically significant difference in the number of views between videos posted by verified accounts and videos posted by unverified accounts.

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

In [19]:
```
# Standardize the case of the 'author_ban_status' column values
data['author_ban_status'] = data['author_ban_status'].str.lower()
```

In [20]:
```
# Standardize the case of the 'verified_status' column values
data['verified_status'] = data['verified_status'].str.lower()
```

In [26]:
```
# Conduct a two-sample t-test to compare the mean `video_view_count` between the tw
active = data[data['author_ban_status'] == 'active']['video_view_count']
banned = data[data['author_ban_status'] == 'banned']['video_view_count']

t_statistic, p_value = stats.ttest_ind(active, banned, equal_var=False)

print("Author Ban Status and Video View Count")
print("T-Statistic:", t_statistic)
print("P-Value:", p_value)
```

```
Author Ban Status and Video View Count
T-Statistic: -28.105495839234667
P-Value: 1.2902882827873965e-146
```

In [27]:
```
# Conduct a two-sample t-test to compare the mean `video_view_count` between the tw
verified = data[data['verified_status'] == 'verified']['video_view_count']
not_verified = data[data['verified_status'] == 'not verified']['video_view_count']
```

```
t_statistic, p_value = stats.ttest_ind(verified, not_verified, equal_var=False)

print("Verified Status and Video View Count")
print("T-Statistic:", t_statistic)
print("P-Value:", p_value)
```

```
Verified Status and Video View Count
T-Statistic: -25.499441780633777
P-Value: 2.6088823687177823e-120
```

Question: Based on the p-value you got above, do you reject or fail to reject the null hypothesis?

- The p-value is very small, significantly less than the threshold of 0.05. Therefore, we reject the null hypothesis in favor of the alternative. There is a statistically significant difference between the mean video view counts of verified accounts and unverified accounts. Additionally, there is also a statistically significant difference between the mean video view counts of banned users and active users.

PACE: Execute

Consider the questions in your PACE Strategy Document to reflect on the Execute stage.

Step 4: Communicate insights with stakeholders

Ask yourself the following questions:

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

The conclusion of the hypothesis test indicates there is a statistically significant difference between the mean video view counts of verified and unverified users. Based on descriptive statistics generated earlier, there are more than twice as many video views on average from unverified users than from verified users. Additionally, there are also approximately twice as many video views from banned authors than from active authors. This suggests a potential correlation between 'verified_status' and 'author_ban_status'. If true, this could indicate that a significant number of views may contain harmful content, which is being viewed and shared more frequently compared to regular videos

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged.