TikTok Data exploration and hypothesis testing

December 13, 2024

1 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.

2 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.

3 Data exploration and hypothesis testing

4 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.

4.1 PACE: Plan

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

1. 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.

Some research questions:

- Are videos posted by verified accounts watched more on average compared to those from unverified accounts?
- Can being a verified account impact the number of views on its videos?

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

4.1.1 Task 1. Imports and Data Loading

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

Hint:

Be sure to import `pandas`, `numpy`, `matplotlib.pyplot`, `seaborn`, and `scipy`.

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

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

# Import packages for statistical analysis/hypothesis testing
```

```
### YOUR CODE HERE ###
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.

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

4.2 PACE: Analyze and Construct

Consider the questions in your PACE Strategy Document and those below to craft your response:
1. 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?

Answer:

Using descriptive statistics lets me easily explore and summarize lots of data. In this example, they help me calculate the average video views for each verification status group.

4.2.1 Task 2. Data exploration

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

Hint:

Refer back to *Self Review Descriptive Statistics* for this step-by-step proccess.

Inspect the first five rows of the dataframe.

[3]:
```
# Display first few rows
### YOUR CODE HERE ###
data.head(5)
```

[3]:

	#	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		

```
4  someone shared with me that the number of busi…    not verified

  author_ban_status  video_view_count  video_like_count  video_share_count  \
0      under review          343296.0           19425.0              241.0
1            active          140877.0           77355.0            19034.0
2            active          902185.0           97690.0             2858.0
3            active          437506.0          239954.0            34812.0
4            active           56167.0           34987.0             4110.0

   video_download_count  video_comment_count
0                   1.0                 0.0
1                1161.0                684.0
2                 833.0                329.0
3                1234.0                584.0
4                 547.0                152.0
```

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

[5]:
```
                #      video_id  video_duration_sec  video_view_count  \
count  19382.000000  1.938200e+04        19382.000000      19084.000000
mean    9691.500000  5.627454e+09           32.421732     254708.558688
std     5595.245794  2.536440e+09           16.229967     322893.280814
min        1.000000  1.234959e+09            5.000000         20.000000
25%     4846.250000  3.430417e+09           18.000000       4942.500000
50%     9691.500000  5.618664e+09           32.000000       9954.500000
75%    14536.750000  7.843960e+09           47.000000     504327.000000
max    19382.000000  9.999873e+09           60.000000     999817.000000

       video_like_count  video_share_count  video_download_count  \
count      19084.000000       19084.000000          19084.000000
mean       84304.636030       16735.248323           1049.429627
std       133420.546814       32036.174350           2004.299894
min            0.000000           0.000000              0.000000
25%          810.750000         115.000000              7.000000
50%         3403.500000         717.000000             46.000000
75%       125020.000000       18222.000000           1156.250000
max       657830.000000      256130.000000          14994.000000

       video_comment_count
count         19084.000000
mean            349.312146
std             799.638865
min               0.000000
25%               1.000000
50%               9.000000
```

```
75%             292.000000
max            9599.000000
```

Check for and handle missing values.

[6]:
```
# Check for missing values
### YOUR CODE HERE ###
data.isna().sum()
```

[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
```

[7]:
```
# Drop rows with missing values

### YOUR CODE HERE ###
data = data.dropna(axis = 0)
```

[8]:
```
# Display first few rows after handling missing values

### YOUR CODE HERE ###
data.head(5)
```

[8]:

	#	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

```
  author_ban_status  video_view_count  video_like_count  video_share_count  \
0      under review          343296.0           19425.0              241.0
1            active          140877.0           77355.0            19034.0
2            active          902185.0           97690.0             2858.0
3            active          437506.0          239954.0            34812.0
4            active           56167.0           34987.0             4110.0

   video_download_count  video_comment_count
0                   1.0                 0.0
1                1161.0                684.0
2                 833.0                329.0
3                1234.0                584.0
4                 547.0                152.0
```

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.

[9]:
```
# Compute the mean `video_view_count` for each group in `verified_status`
### YOUR CODE HERE ###
data.groupby("verified_status")["video_view_count"].mean()
```

[9]:
```
verified_status
not verified    265663.785339
verified         91439.164167
Name: video_view_count, dtype: float64
```

4.2.2 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?

For Null hypothesis:

- Verified and unverified TikTok accounts get the same number of views on their videos. Any difference seen is just by chance.

For Alternative hypothesis:

- Verified TikTok accounts have a different number of views on their videos compared to unverified accounts. The difference is real and not due to chance.

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

==> ENTER YOUR NULL AND ALTERNATIVE HYPOTHESES HERE (Double Click)

Null Hypothesis (H_0):

Verified and unverified TikTok accounts get the same number of views. Any difference is because of random variation.

Alternative Hypothesis (H_A):

Verified and unverified TikTok accounts have different view counts. The difference is due to actual differences in the population.

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

[12]:
```
# Conduct a two-sample t-test to compare means
### YOUR CODE HERE ###
not_verified = data[data["verified_status"] == "not␣
↪verified"]["video_view_count"]
verified = data[data["verified_status"] == "verified"]["video_view_count"]

statistic, p_value = stats.ttest_ind(a = not_verified, b = verified, equal_var␣
↪= False)
print(statistic)
print(p_value)
```

```
25.499441780633777
2.6088823687177823e-120
```

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

Answer:

With a p-value much smaller than the 5% significance level, I reject the null hypothesis. This shows there's a significant difference in average video views between verified and unverified TikTok accounts.

4.3 PACE: Execute

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

4.4 Step 4: Communicate insights with stakeholders

Ask yourself the following questions:

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

The analysis shows a notable difference in views between verified and unverified accounts' videos, suggesting they might act differently.

Delving into the root causes of this behavioral difference would be worthwhile. For example: * Finding out why could be interesting. * Do unverified accounts post more clickbait videos? * Or are they linked to spam bots that boost views?

The next step is creating a regression model on verified_status to predict claim status and analyze verified users. A logistic regression model is needed because the data is skewed.

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.