Activity_Course 4 TikTok project lab

December 18, 2023

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.

The research question is, whether there is a statistical difference in the data between verified and unverified accounts.

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`.

[31]:
```
# Import packages for data manipulation
import pandas as pd
import numpy as np


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


# Import packages for statistical analysis/hypothesis testing
import statsmodels.api as sm
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.

[32]:
```
# 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?

Computing descriptive statistics can be help us understand the data further by allowing us to draw conclusions about the data. For instance in this portion of the project we are determaning whether there is a statistical difference in the data between verified and unverified accounts. Doing so will give us information on whether or not that statistical difference is a result of chance or possibly something more significant which would require further investigation.

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.

[33]:
```
# Display first few rows

data.head(10)
```

[33]:

	#	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		
3	4	claim	1866847991	25		
4	5	claim	7105231098	19		
5	6	claim	8972200955	35		
6	7	claim	4958886992	16		
7	8	claim	2270982263	41		
8	9	claim	5235769692	50		
9	10	claim	4660861094	45		

```
2  someone shared with me that american industria…    not verified
3  someone shared with me that the metro of st. p…    not verified
4  someone shared with me that the number of busi…    not verified
5  someone shared with me that gross domestic pro…    not verified
6  someone shared with me that elvis presley has …    not verified
7  someone shared with me that the best selling s…    not verified
8  someone shared with me that about half of the …    not verified
9  someone shared with me that it would take a 50…        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
5      under review          336647.0          175546.0            62303.0
6            active          750345.0          486192.0           193911.0
7            active          547532.0            1072.0               50.0
8            active           24819.0           10160.0             1050.0
9            active          931587.0          171051.0            67739.0

   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                4293.0               1857.0
6                8616.0               5446.0
7                  22.0                 11.0
8                  53.0                 27.0
9                4104.0               2540.0
```

[34]:
```
# Generate a table of descriptive statistics about the data
data.describe(include='all')
```

[34]:
```
                  # claim_status      video_id  video_duration_sec  \
count   19382.000000        19084  1.938200e+04        19382.000000
unique           NaN            2           NaN                 NaN
top              NaN        claim           NaN                 NaN
freq             NaN         9608           NaN                 NaN
mean     9691.500000          NaN  5.627454e+09           32.421732
std      5595.245794          NaN  2.536440e+09           16.229967
min         1.000000          NaN  1.234959e+09            5.000000
25%      4846.250000          NaN  3.430417e+09           18.000000
50%      9691.500000          NaN  5.618664e+09           32.000000
75%     14536.750000          NaN  7.843960e+09           47.000000
```

```
max     19382.000000          NaN  9.999873e+09           60.000000

                                video_transcription_text verified_status  \
count                                              19084           19382
unique                                             19012               2
top     a friend read  in the media a claim that badmi…    not verified
freq                                                   2           18142
mean                                                 NaN             NaN
std                                                  NaN             NaN
min                                                  NaN             NaN
25%                                                  NaN             NaN
50%                                                  NaN             NaN
75%                                                  NaN             NaN
max                                                  NaN             NaN

       author_ban_status  video_view_count  video_like_count  \
count              19382      19084.000000      19084.000000
unique                 3               NaN               NaN
top               active               NaN               NaN
freq               15663               NaN               NaN
mean                 NaN     254708.558688      84304.636030
std                  NaN     322893.280814     133420.546814
min                  NaN         20.000000          0.000000
25%                  NaN       4942.500000        810.750000
50%                  NaN       9954.500000       3403.500000
75%                  NaN     504327.000000     125020.000000
max                  NaN     999817.000000     657830.000000

        video_share_count  video_download_count  video_comment_count
count        19084.000000          19084.000000         19084.000000
unique                NaN                   NaN                  NaN
top                   NaN                   NaN                  NaN
freq                  NaN                   NaN                  NaN
mean         16735.248323           1049.429627           349.312146
std          32036.174350           2004.299894           799.638865
min              0.000000              0.000000             0.000000
25%            115.000000              7.000000             1.000000
50%            717.000000             46.000000             9.000000
75%          18222.000000           1156.250000           292.000000
max         256130.000000          14994.000000          9599.000000
```

Check for and handle missing values.

[35]:
```
# Check for missing values
data.isna().sum()
```

[35]:
```
#                           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
```

[36]:
```
# Drop rows with missing values
data = data.dropna()
```

[38]:
```
# Display first few rows after handling missing values

print(data.head(3))
print()
print('**********************************************************************')
print()
print(data.isna().sum())
```

```
   # claim_status    video_id  video_duration_sec  \
0  1        claim  7017666017                  59
1  2        claim  4014381136                  32
2  3        claim  9859838091                  31

                          video_transcription_text verified_status  \
0  someone shared with me that drone deliveries a…    not verified
1  someone shared with me that there are more mic…    not verified
2  someone shared with me that american industria…    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

   video_download_count  video_comment_count
0                   1.0                 0.0
1                1161.0               684.0
2                 833.0               329.0

**********************************************************************
```

```
#                           0
claim_status                0
video_id                    0
video_duration_sec          0
video_transcription_text    0
verified_status             0
author_ban_status           0
video_view_count            0
video_like_count            0
video_share_count           0
video_download_count        0
video_comment_count         0
dtype: int64
```

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.

[29]:
```
# Compute the mean `video_view_count` for each group in `verified_status`
not_verififed = data[data['verified_status'] == 'not verified']
verififed = data[data['verified_status'] == 'verified']

not_verififed_mean = not_verififed['video_view_count'].mean()
verififed_mean = verififed['video_view_count'].mean()

print('Not verified mean view count: ' + str(not_verififed_mean))
print('Verified mean view count: ' + str(verififed_mean))

not_verififed['video_view_count'].describe()
```

```
Not verified mean view count: 265663.78533885034
Verified mean view count: 91439.16416666667
```

[29]:
```
count     17884.000000
mean     265663.785339
std      325681.881915
min          20.000000
25%        5160.000000
50%       46723.000000
75%      523099.500000
max      999817.000000
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?

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).

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

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).
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.

[25]:
```
# Conduct a two-sample t-test to compare means
stats.ttest_ind(a=verififed['video_view_count']␣
↪,b=not_verififed['video_view_count'] , equal_var=False)
```

[25]:
```
Ttest_indResult(statistic=-25.499441780633777, pvalue=2.6088823687177823e-120)
```

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

Since the p-value is extremely small (2.608e-120) (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.

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?

Since the p-value is extremely small (2.608e-120) (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 and that there should be further investigation into verified vs unverified accounts.

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.

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.