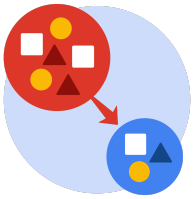
Course Four

From Data to Insight: The Power of Statistics



Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. As a reminder, this document is a resource that you can reference in the future, and a guide to help you consider responses and reflections posed at various points throughout projects.

Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- ☐ Complete the questions in the Course 4 PACE strategy document
- ☐ Answer the questions in the Jupyter notebook project file
- ☐ Compute descriptive statistics
- ☐ Conduct a hypothesis test
- ☐ Create an executive summary for external stakeholders

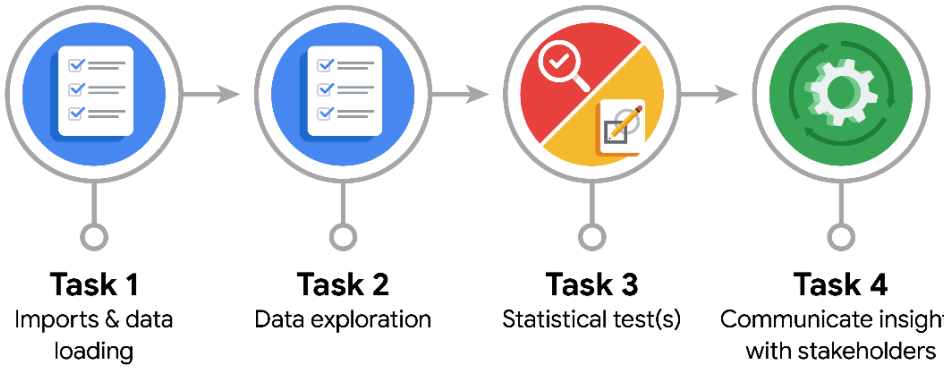
Relevant Interview Questions

Completing this end-of-course project will empower you to respond to the following interview topics:

- How would you explain an A/B test to stakeholders who may not be familiar with analytics?
- If you had access to company performance data, what statistical tests might be useful to help understand performance?
- What considerations would you think about when presenting results to make sure they have an impact or have achieved the desired results?
- What are some effective ways to communicate statistical concepts/methods to a non-technical audience?
- In your own words, explain the factors that go into an experimental design for designs such as A/B tests.

Reference Guide

This project has four tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- What is the main purpose of this project?

The main purpose to develop a machine learning model to categorize submissions on tiktok into 'opinion' or 'claim'. This portion of the project is to perform statistical analysis

- What is your research question for this project?

Can we categorize submissions by 'opinion' and 'claim' with a machine learning algorithm Is there statistical difference in the data between verified and unverified accounts? And is there a significant difference in video views for verified versus unverified accounts?

- What is the importance of random sampling?

The importance of random sampling is that the sample will be more closely representing the population, and therefore the results will be a better answer to the question at hand.

- Give an example of sampling bias that might occur if you didn't use random sampling.

With non random sampling in the scope of this project, we could select more of one type of type of submission than the other creating an over representation bias in the sample.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

Descriptive statistics are useful, since they quickly give us an understanding of the spread and distribution of the data set. We can see the standard deviation, mean, median, mode and shape of the data from these easily generated statistics in pandas.

- How did computing descriptive statistics help you analyze your data?

Descriptive statistics shows an overview of the data, giving an idea for its distribution and potential outliers.

- In hypothesis testing, what is the difference between the null hypothesis and the alternative hypothesis?

A null hypothesis assumes that there is no statistically significance of a change observed in the data, the change is due to random chance. While the alternative hypothesis assumes that there is a significant change or difference in the data, not due to random chance.

- How did you formulate your null hypothesis and alternative hypothesis?

We are determining whether there is a difference in the amount of video views a verified versus an unverified account garners. Therefore, the null hypothesis is that the differences observed between the two groups is due to random chance, while the alternative hypothesis is that the differences are statistically significant and cannot be rejected as being caused by random chance.

- What conclusion can be drawn from the hypothesis test?

The conclusion drawn from the test is that there is a statistically significant difference between the views acquired by verified and unverified accounts. According to the mean of the two groups, unverified gain nearly three times as many views.

PACE: Execute Stage

- What key business or organizational insight(s) emerged from your A/B test?

A/B testing will show how changes impact functional and or user behavior. A testing example for this aspect of the tiktok project could be the implementation of showing one group of users only verified account videos and another group only unverified. This may give some better understanding to the viralness of unverified account videos and whether that holds up when content is filtered.

- What recommendations do you propose based on your results?

I propose that we determine a test to better understand the viralness of unverified account videos. From previous aspects of this project, we know that verified accounts are much less likely to have “claim” videos and “opinion” videos have lower view counts.

A better understanding of why unverified accounts which will likely have more views, be a claim, and have correlation with a greater ban rate is very important to develop the best machine learning model for our classification project.