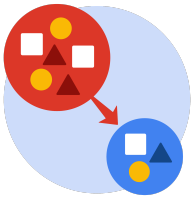
Course Four

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

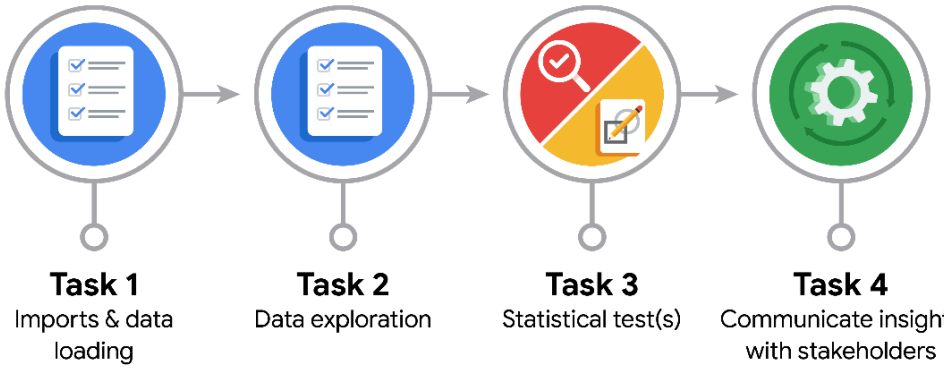
Relevant Interview Questions

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 is to analyze the relationship between `verified_status` and `video_view_count` on TikTok to understand user behavior and inform the development of a machine learning model for claim classification.

- What is your research question for this project?

Is there a statistically significant difference in video view counts between verified and non-verified TikTok accounts?

- What is the importance of random sampling?

Random sampling ensures that the sample is representative of the entire population of TikTok videos. This minimizes bias and allows for valid statistical inferences about the relationship between verification status and view counts. In this particular project, we did not do random sampling, we used the data set given. If we were to collect new data, random sampling would be very important.

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

If you only sampled videos from a specific niche or demographic (e.g., only gaming-related videos or only videos from a particular region), your results might not generalize to the entire TikTok platform. For example, gaming videos might naturally have higher view counts regardless of verification status, leading to a biased conclusion.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

Descriptive statistics provide a summary of the data, allowing us to understand the central tendency (mean, median), dispersion (standard deviation), and distribution of variables. They help identify patterns, outliers, and potential relationships within the data.

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

Computing the mean video view counts for verified and non-verified accounts allowed us to observe a significant difference, which formed the basis for our hypothesis test. It provided a clear picture of the initial relationship between the two variables.

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

The null hypothesis (H0) states that there is no significant difference or relationship between variables. The alternative hypothesis (HA) states that there is a significant difference or relationship.

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

• **Null Hypothesis (H0):** There is no statistically significant difference in the mean video view counts between verified and non-verified TikTok accounts. • **Alternative Hypothesis (HA):** There is a statistically significant difference in the mean video view counts between verified and non-verified TikTok accounts.

- What conclusion can be drawn from the hypothesis test?

The hypothesis test (two-sample t-test) resulted in a very low p-value (2.6088823687177823e-120), which is significantly lower than the standard significance level of 0.05. Therefore, we rejected the null hypothesis and concluded that there is a statistically significant difference in video view counts between verified and non-verified accounts.

PACE: Execute Stage

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

(Note: This was a hypothesis test, not an A/B test, but the insight remains the same.) The key insight is that non-verified accounts tend to have significantly higher video view counts than verified accounts. This suggests that verification status is not a primary driver of video engagement and that other factors, such as content quality or algorithm promotion, may play a more significant role.

- What recommendations do you propose based on your results?

• Build a regression model for `verified_status` to analyze user behavior. • Investigate the factors contributing to the higher view counts of non-verified accounts, such as content characteristics or spam bot activity. • Investigate the algorithms that are used to promote or display videos. • Evaluate the effectiveness of the current verification process.