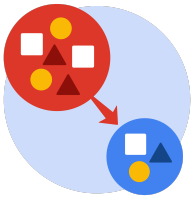
Course Four

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

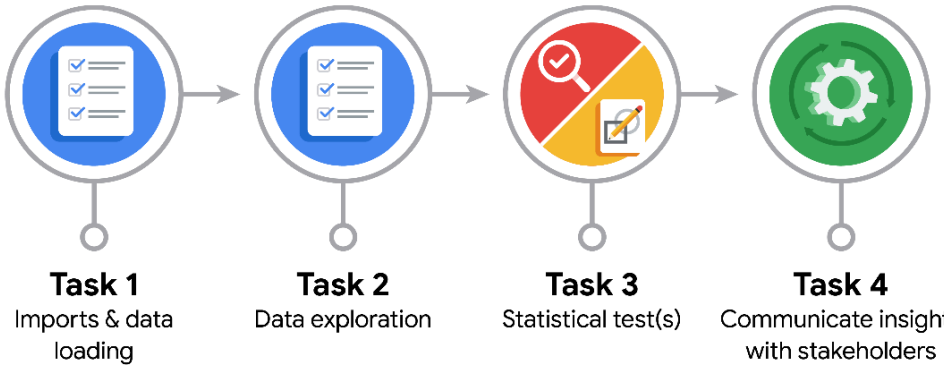
Relevant Interview Questions

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 purpose of this project is to demonstrate knowledge of how to prepare, create, and analyze hypothesis tests where our hypotheses for this data project are the null hypothesis (there is no difference in the mean video_view_count between 'not verified' and 'verified') and the alternate hypothesis (there is a difference in the mean video_view_count between 'not verified' and 'verified').

- What is your research question for this project?

Our research question for this data project is whether there is a statistically significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts.

- What is the importance of random sampling?

Random sampling is a fundamental technique in data analysis, especially in the Plan stage of the PACE workflow. It plays a crucial role in ensuring data accuracy, efficiency, and representativeness (minimizing bias that can occur, ensuring that the sample selected accurately represents the entire population, and, as a result, drawing meaningful conclusions from the data). Plus, findings from a randomly selected sample can be generalized to the entire population and provide valuable insights that can inform broader strategies and future decisions.

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

Sampling bias occurs when a sample is not representative of the population it aims to represent. This can lead to inaccurate conclusions and biased results. Here is an example of sampling bias that might occur if random sampling is not used. Our goal is to assess student satisfaction with the university's dining services. The population is all students at this university. If we use Non-Random Sampling Method such as Convenience Sampling where we survey only students who frequent the dining halls or are easily accessible (e.g., students living on campus), the two potential biases that may occur are Overrepresentation (Students who frequent the dining halls may have a more positive view of the services compared to those who rarely or never use them) or Underrepresentation (Students who have negative experiences with the dining services might be less likely to participate in the survey, leading to a biased sample). Thus, using a non-random sampling method like convenience sampling could lead to a biased result, as the sample would not accurately represent the entire student population.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

Computing descriptive statistics helps one learn more about our data in this stage of your analysis by measuring the central tendency, dispersion, and position of our data.

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

By effectively using descriptive statistics, we gain valuable insights into our data, make informed decisions about our analysis, and lay the groundwork for successful model building in the Construct stage. For example, with respect to numerical data we calculate and find the mean, median, mode, standard deviation, range, and percentiles and with respect to categorical data, we determine the frequency counts and proportions. Plus, descriptive statistics help us identify unusual data points that might skew results, guide the choice of appropriate models based on data distribution and relationships, and suggest potential relationships or patterns to explore.

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

In hypothesis testing, we make statements about a population parameter. These statements are typically divided into two hypotheses: the null hypothesis and the alternative hypothesis. Often, the null hypothesis is a statement of "no change" or "no relationship." It is the hypothesis that we aim to disprove. And, the alternative hypothesis is a statement of an effect or a difference. It is the hypothesis we are trying to prove. The alternative hypothesis can be one-sided (e.g., "greater than," "less than") or two-sided (e.g., "not equal to"). Finally, the null hypothesis is always assumed to be true until proven otherwise. We accept the alternative hypothesis once we reject the null hypothesis.

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

Since our research question for this data project is whether there is a statistically significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts, we formulated a null hypothesis stating of "no change" or "no difference" and a two-sided (e.g., "not equal to") alternative hypothesis stating of a difference.

- What conclusion can be drawn from the hypothesis test?

With a p-value (2.6088823687177823e-120) of less than 0.05 (as the significance level is 5%), we reject the null hypothesis in favor of the alternative hypothesis. Therefore, at the 5% significance level there is a difference in the mean video_view_count between 'not verified' and 'verified '.

PACE: Execute Stage

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

The key business or organizational insight(s) emerged from our not A/B test but our two-sample t-test is that based on the tests, we reject the null hypothesis and conclude that the mean video_view_count are different for each group in `verified_status`. Thus, it is likely that there is a statistically significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts.

- What recommendations do you propose based on your results?

The TikTok data team recommends building a regression model on verified_status that helps us analyze user behavior. Plus, since our goal is to make predictions on claim status (a categorical variable), we recommend building a logistic regression model. The TikTok data team recommends investigating the root cause of the behavioral difference between verified accounts and unverified accounts. For example, are unverified accounts associated with bots helping inflate view counts?