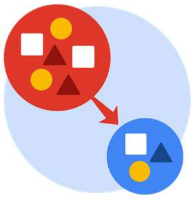
Course Four

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

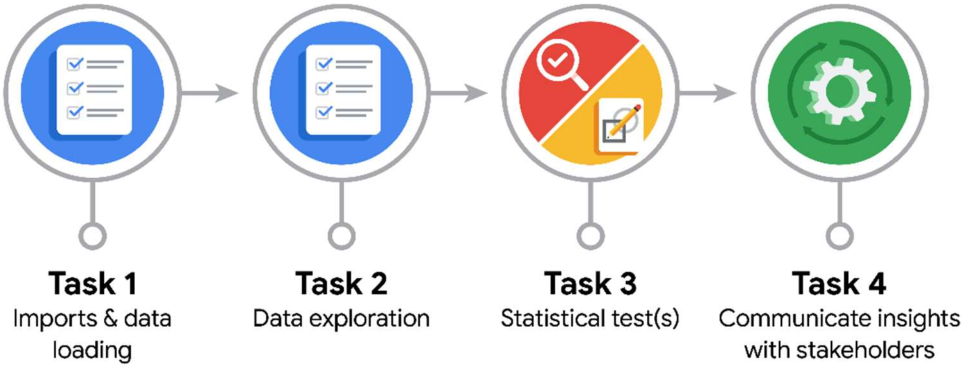
Relevant Interview Questions

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?

This project aims to develop a model that predicts whether video content expresses a claim or an opinion. This will help moderators identify potentially misleading or harmful videos for further investigation.

- What is your research question for this project?

This phase of the project addresses the research question of whether a statistically significant difference exists in video engagement between unverified and verified authors.

- What is the importance of random sampling?

Random sampling is crucial for minimizing bias and ensuring that the sample is representative of the population, giving each unit an equal chance of selection.

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

Non-random sampling methods can introduce bias by disproportionately selecting individuals from specific subgroups. This means that the characteristics observed in the sample may not accurately reflect the distribution of those characteristics in the overall population, limiting the generalizability of any findings.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

Descriptive statistics offer a valuable first look at data, summarizing key features like the mean, median, mode, distribution, outliers, and missing values, and helping to identify initial trends and patterns.

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

Calculation of descriptive statistics identified outliers with values considerably different from the mean and median. Examination of variable distributions revealed a pronounced right skew for several variables. Missing values were also identified and handled accordingly. Analysis of subgroups within variables such as author_ban_status and verified_status provided initial insights into imbalances present within the dataset.

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

The null hypothesis represents the default position, which we attempt to disprove. The alternative hypothesis represents the claim we are trying to establish, contradicting the null hypothesis.

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

The hypothesis test aimed to determine whether a statistically significant difference exists between mean video view counts of verified and unverified users. The null hypothesis (H_0) states there is no statistically significant difference, while the alternative hypothesis (H_1) states there is.

- What conclusion can be drawn from the hypothesis test?

The hypothesis test revealed a statistically significant difference in video view counts between verified and unverified users, leading to the rejection of the null hypothesis.

PACE: Execute Stage

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

The substantial and statistically significant difference in video view counts between unverified and verified users warrants further investigation as a potential predictor of claim status. This difference suggests that unverified users' videos may generate greater engagement, potentially indicating the presence of a claim rather than an opinion.

- What recommendations do you propose based on your results?

Further analysis is recommended for imbalanced variables such as verified_status and author_ban_status, with particular attention to potential collinearity. The next step will involve developing a regression model to investigate these relationships further.