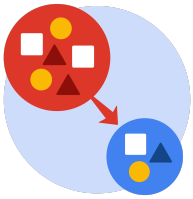
Course Four

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~~

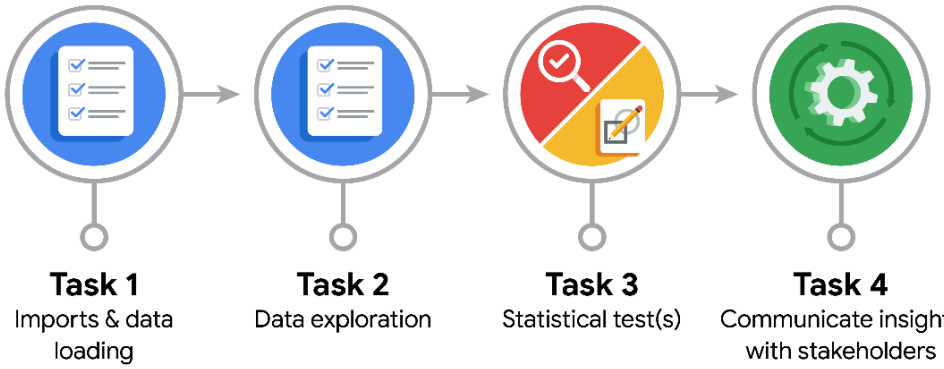
Relevant Interview Questions

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?

To apply descriptive statistics and inferential statistics like probability, probability distributions and hypothesis testing to determine is there are statistically significant difference between verified and non verified users.

- What is your research question for this project?

Is there a statistically significant difference in views of a video from verified and non verified users?

- What is the importance of random sampling?

Random sampling is on the base of all statistic analysis. If done properly, the results from the conducted study will be important for taking more informed decisions. On the contrary, if sampling is

biased, no matter how good is the statistical analysis, it always make assumptions and conclusions that don't reflect the real situations and, depending on the impact, they could be very harmful or completely misdleading for some population.

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

Not representing populations proportions in a proper way, For example, misrepresenting a group within a population; taking only a very specific population group as sample that doesn't reflect the diversity of characteristics or conditions of the whole population. In our TikTok case, to only select users that have made claims, that will misrepresent videos that are classified as opinions, and the result will be biased.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

It helps us to understand how the distribution of data is around to the center and how dispersed is. It also help us to identify outlier. All of these primary explloration will give us a guide of what statistical methods to apply and how to do it.

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

It show us what how data behave around the center, like what's the average, what the most common value and what's the value in the middle. It also show how dispersed is our set, like what's the general distance of data from the center and how consistent are those dispertions.

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

The null hypothesis is a statement that is assumed to be true unless there is enough evidence to reject it. It also tends to express the status quo, supporting that the result is likely due to chance. The alternative hypothesis is the statement that replaces the null hypothesis only if there is enough evidence that former one must be reject it. It expresses a change in the status quo, showing that the result is very unlikely due to chance, that is, it is statistically significant.

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

Null hypothesis, H0 = There is no difference in the number of video views between verified and unverified users.

Alternative hypothesis, Ha = There is difference in the number of video views between verified and unverified users.

- What conclusion can be drawn from the hypothesis test?

The p-value is 2.60e-120 < 0.05 = 5 %. Then we should reject H0. This means the difference in mean video counts between verified and not verified users is statistically significant.

PACE: Execute Stage

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

- The mean of view in videos for verified users is 91,439 and for non verified is 265,664. However, those data do not follow normal distributions, so it's not clear why a t-test should apply to them. In fact both distributions are very skewed to the right, due to the typical behaviour of social networks, where a minority of videos get a great number of views. That's why their standard deviations, 221,139 and 325,682, correspondingly, are even greater than their means.
- If we accept a 2-sample t-test is appropriate, which doesn't seem the case, then at a 5 % of significance level, we can conclude that the average video views for verified users is different from not verified ones. This also means user's verified status is an important variable to keep in mind for machine learning models in our claim classification project.

- What recommendations do you propose based on your results?

- I would recommend performing more hypothesis tests that could be applied to non normal distributions and that could show if verified status affects other variables in our data set like videos classified as claim or opinion.