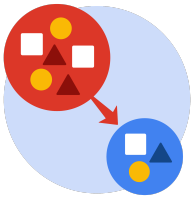
Course Four

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

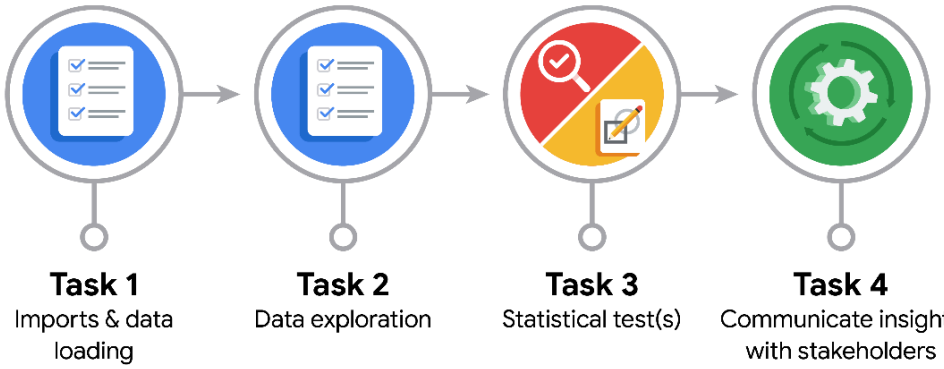
Relevant Interview Questions

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 of this project is to determine whether there is a statistically significant difference between the number of views for TikTok videos posted by verified accounts versus those unverified.

- What is your research question for this project?

Is there a statistically significant difference in the number of views for TikTok videos posted by verified accounts compared to those posted unverified accounts?

- What is the importance of random sampling?

Random Sampling is crucial in research and data analysis for a number of reasons:

1. Reduces Bias: Random sampling minimizes selection bias, ensuring each individual in the population has an equal chance of being included in the sample and this makes the sample more representative.
2. Generalizability
3. Improves Validity: Validate internal data and externally the findings.
4. Facilitates Statistical Analysis: Assumptions for some statistics is random sampling.
5. Reduces Sampling Error: Random sampling helps reduce sampling error, the difference between the sample statistic and the actual population parameter.

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

Suppose we collect data only from TikTok videos posted by accounts with a large number of followers, assuming that these accounts are more likely to be representative of general account usage.

In this case, your sample might overrepresent accounts that are already popular and more likely to be verified. Since verified accounts often have more followers and higher engagement, this could lead to an inflated view count for verified accounts in your sample.

This introduces selection bias because your sample disproportionately includes videos from accounts that are more successful, ignoring videos from smaller or less popular accounts (which are often unverified). As a result, your analysis might incorrectly suggest that verified accounts have significantly higher view counts than unverified accounts, simply because your sample did not include a fair representation of unverified accounts.

This bias would distort the true relationship between account verification status and video views, making your findings less reliable and less generalizable to the entire population of TikTok users.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

Descriptive statistics are useful for several reasons:

1. Data Summary

2. Data Interpretation

3. Foundation for further analysis

4. Comparison across groups

5. Identification of Data Quality Issues such as outliers, missing values, skewed distributions, which may require further study or cleaning before proceeding with analysis.

6. Communication of Results with non-technical and stakeholders.

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

1. Understanding Data Distribution

2. Identification of Outliers

3. Comparing Groups

4. Guiding Hypothesis Testing

5. Summarizing Findings

6. Assisting in Data Cleaning

In the context of the TikTok project, computing descriptive statistics would help in understanding the basic characteristics of the view counts for both verified and unverified accounts, laying the groundwork for more detailed analysis and hypothesis testing.

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

1. Null Hypothesis (H_0): The null hypothesis is a statement that there is no effect, no difference, or no relationship between variables. It represents the default or status quo assumption.

2. Alternative Hypothesis (H_1 or Ha): The alternative hypothesis is a statement that contradicts the null hypothesis. It suggests that there is an effect, a difference, or a relationship between variables.

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

1. H0: There is no statistically significant difference in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts

2. H1: There is a statistically significant difference in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts.

- What conclusion can be drawn from the hypothesis test?

The conclusion drawn from a hypothesis test depends on whether the test provides sufficient evidence to reject the null hypothesis.

1. Reject the null hypothesis (H0): If the test results in a p-value that is below a predetermined significance level, you reject the null hypothesis.

Conclusion: There is statistically significant evidence to suggest a difference in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts. This means that verification status likely impacts view counts.

2. Fail to Reject the Null Hypothesis: If the p-value is above the significance level, you fail to reject the null hypothesis.

Conclusion: There is not enough statistical evidence to support the claim that there is a difference in the number of views between verified and unverified accounts. This suggests that verification status may not significantly affect the view counts, based on the data analyzed.

PACE: Execute Stage

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

1. Promote Verification Benefits: Highlight the advantages of being a verified user through platform communications, encouraging more creators to apply for verification. Additionally, a revision of verification guidelines to facilitate more users.

2. Refine Algorithms: Consider adjusting recommendation algorithms to account for the higher engagement of verified accounts.

3. Encouraging Verification: By demonstrating the benefits of verification, TikTok can encourage more users, especially influencers and content creators, to seek verification, thereby increasing the number of verified accounts on the platform.

- What recommendations do you propose based on your results?

1. Enhance Verification Algorithm.

2. Adjust Content Recommendation Algorithm.

3. Develop Targeted Marketing Strategies.

4. Educate users to have verified accounts through simplified guidelines.