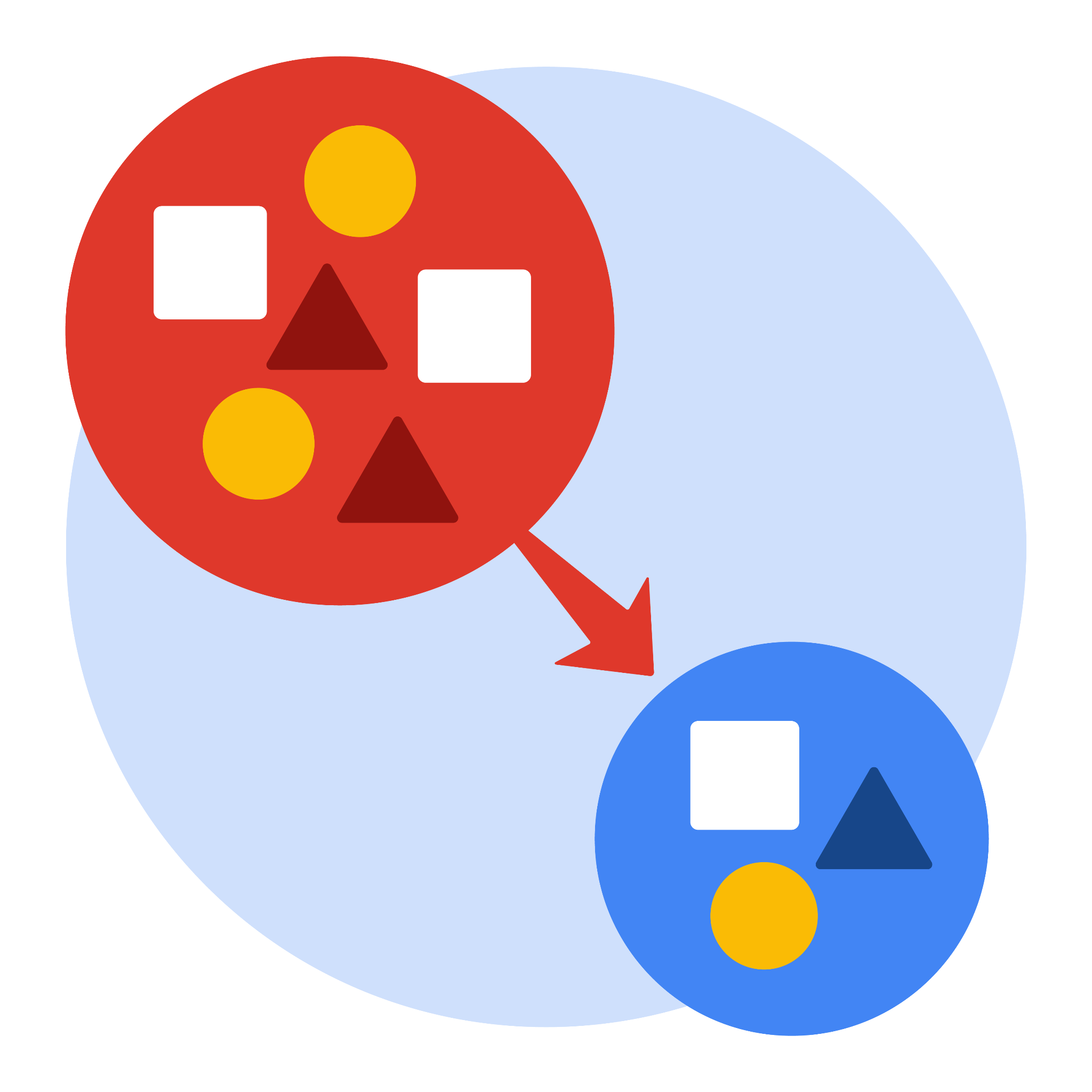
**Course Four**

# 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

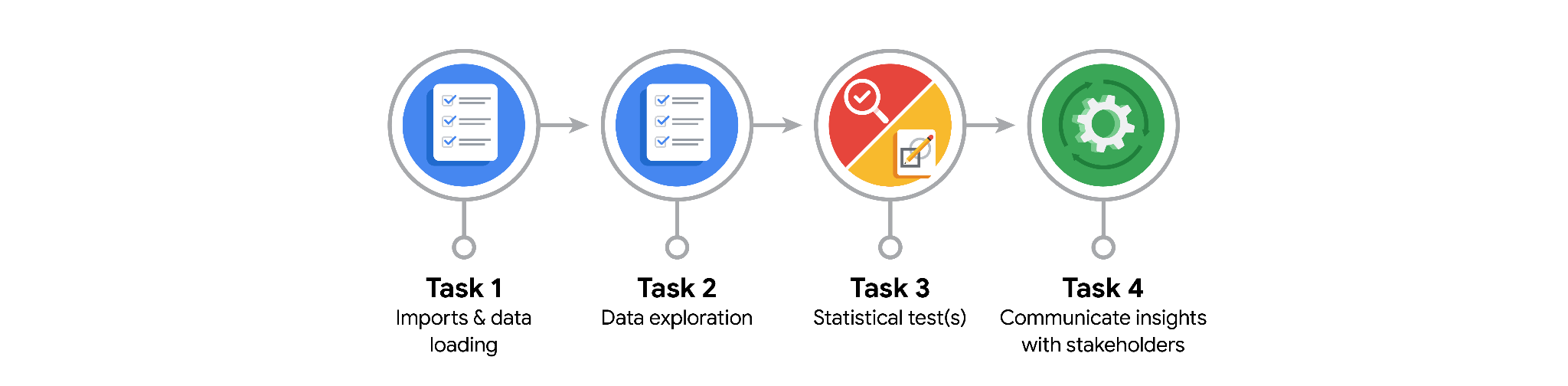
# Relevant Interview Questions

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 conduct a statistical analysis, specifically an A/B test (two-sample t-test), on the NYC TLC taxi dataset to investigate the relationship between payment type and fare amount. It aims to rigorously test the client's hypothesis that customers who use credit cards pay significantly higher fares than those who use cash, ultimately providing data-driven insights and potential recommendations communicated through an executive summary.

* What is your research question for this project?

Is there a statistically significant difference in the average fare amount between NYC taxi rides paid for with credit cards compared to those paid for with cash?

* What is the importance of random sampling?

Random sampling is crucial because it minimizes selection bias, ensuring the sample accurately represents the population and allowing findings to be reliably generalized. This representativeness forms the foundation for valid statistical inference, enabling meaningful hypothesis testing and estimation. In experimental designs like A/B tests, the related concept of random assignment is essential for creating comparable groups, which allows researchers to more confidently attribute observed differences to the factor being tested rather than pre-existing group variations.

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

An example of sampling bias that might occur without random sampling in this project is convenience sampling, such as analyzing only taxi trip data collected between 10 AM and 4 PM on weekdays. This approach would systematically exclude rides from early mornings, evenings, late nights, and weekends, periods likely having different average fare amounts and payment type distributions (e.g., due to surcharges, trip purposes, or cash availability). Consequently, the sample would overrepresent mid-day weekday trips, providing a biased view that isn't generalizable to the entire population of NYC taxi rides.



 **PACE: Analyze & Construct Stages**

* In general, why are descriptive statistics useful?

Descriptive statistics are useful because they summarize and simplify large datasets into easily understandable metrics, allowing analysts to quickly grasp core characteristics like central tendency (e.g., mean, median) and dispersion (e.g., standard deviation, range). This provides an essential initial understanding of the data's patterns, helps identify potential outliers or issues, and forms a necessary foundation for more complex inferential statistical analysis.

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

Computing descriptive statistics, specifically calculating the mean fare\_amount for credit card and cash payment types using groupby(), directly helped analyze the data by providing a clear quantitative comparison between the two groups central to the research question. This revealed that the average fare amount was higher for credit card users (approx. $13.43) than for cash users (approx. $12.21) within this sample. This initial finding offered preliminary insight into the relationship being investigated and established the observed difference that the subsequent hypothesis test would formally evaluate for statistical significance.

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

In hypothesis testing, the null hypothesis (H₀) is a statement representing the default assumption of no effect, no difference, or no relationship between variables; it's essentially the status quo that the statistical test aims to potentially disprove. Conversely, the alternative hypothesis (Hₐ) is a statement that contradicts the null hypothesis, proposing that there *is* an effect, a difference, or a relationship; it represents what the researcher typically seeks to find evidence for. The entire testing process evaluates sample data to determine if there's enough statistical evidence to reject the null hypothesis in favor of the alternative hypothesis.

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

The null hypothesis (H₀) was formulated to represent the default assumption of no difference, stating that the average fare amount for customers paying with credit cards is equal to the average fare amount for those paying with cash. The alternative hypothesis (Hₐ) was formulated to directly contradict the null and address the core research question, stating that there *is* a difference in the average fare amounts between the two payment types; this specific two-tailed formulation (checking for any difference, not just higher) was used for the t-test conducted in the lab analysis, directly challenging the H₀ based on the client's interest in the relationship between payment type and fare amount.

* What conclusion can be drawn from the hypothesis test?

Based on the hypothesis test conducted in the lab, which resulted in an extremely small p-value (approximately 6.79 x 10⁻¹², as shown in image stats5.png), we reject the null hypothesis. The conclusion drawn is that there is a statistically significant difference in the average fare amount between customers who use credit cards and those who use cash. Furthermore, considering the sample means (approx. $13.43 for credit card vs. $12.21 for cash, from image stats4.jpg), the data indicates that customers paying by credit card have a significantly higher average fare amount than those paying by cash within this dataset

**PACE: Execute Stage**

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

The key business insight derived directly from the A/B test result is that credit card payments are associated with statistically significantly higher average fare amounts compared to cash payments within this dataset (approx. $13.43 vs $12.21). This suggests a potential opportunity for the business: strategies aimed at encouraging customers to use credit cards *could* potentially lead to increased revenue per ride for taxi drivers. However, this insight must be heavily qualified by the significant limitations of using this test on observational data, primarily the strong possibility that higher fare amounts actually determine the choice to use a credit card, rather than the payment type influencing the fare amount.

* What recommendations do you propose based on your results?

Based on the results and the significant limitations identified, the primary recommendation is caution. While the analysis showed a statistically significant association between credit card use and higher average fares, this relationship is likely driven by factors not controlled for (such as longer trips necessitating card payments) rather than payment type influencing the fare itself. Therefore, it is strongly recommended *not* to implement potentially costly strategies aimed at encouraging credit card use solely based on this finding with the expectation of increasing revenue. Instead, further investigation is proposed: conduct more sophisticated analyses (e.g., multivariate regression) to explore the influence of confounding variables like trip distance, duration, time of day, and location on both fare amount and payment choice to understand the underlying dynamics more accurately before considering any interventions.