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

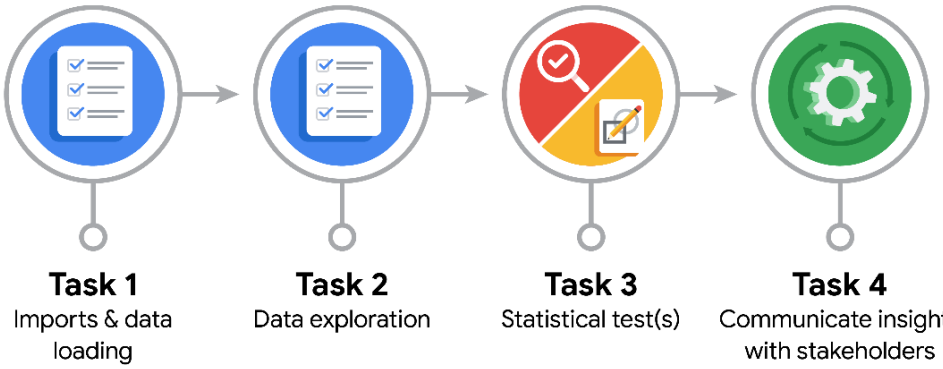
Completing this end-of-course project will empower you to respond to the following interview topics:

- How would you explain an A/B test to stakeholders who may not be familiar with analytics?
    - Two groups are asked to do the same task, but using different tools. At the end, we see the differences in each group and determine if the different tools had any effect on the outcome
    - For example, we might as two groups to visit an ecommerce site, but we give one group a webpage where the ‘add to cart’ button is very big and the other group a webpage where the ‘add to cart button is very small’. Then we see if there is a noticeable difference in the number of people who purchased items from both groups, and if we can attribute such differences to the differently designed web pages.
- If you had access to company performance data, what statistical tests might be useful to help understand performance?
    - Descriptive statistics would be enough to understand the performance of a company based on the historical data. We could find the mean, min, max, and standard deviations for the data

- What considerations would you think about when presenting results to make sure they have an impact or have achieved the desired results?
    - To determine if a result is statistically significant we calculate the p-value of the test. If the p-value is smaller than the confidence level, we can reject the null hypothesis and say that the variable does indeed have an impact on our study
- What are some effective ways to communicate statistical concepts/methods to a non-technical audience?
    - Visualizations of probability distributions and the areas for confidence levels would probably help non-technical audience members understand statistical methods
- In your own words, explain the factors that go into an experimental design for designs such as A/B tests.
    - Choose a sample population randomly to ensure that it is representative of the larger population
    - The null and alternative hypotheses need to be clearly defined
    - The significance level has to be chosen
    - Finally we calculate the p-value and then determine if we should accept or reject the null hypothesis

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 if the method of payment for taxi cab fares has an impact on the amount paid by the passengers

- What is your research question for this project?

Does the method of payment affect the total fare paid?

- What is the importance of random sampling?

Random sampling ensures that the sample is not biased but reflects the larger population. Having an unbiased sample is crucial as it allows data analysts to make reliable inferences about the data.

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

Sampling bias is when some members of the population is overrepresented, and thus leads to bias in the data. For example, convenience sampling can lead to undercoverage bias. Undercoverage bias occurs when some members of a population are inadequately represented in the sample. For example, a convenience sample of shoppers at a mall will underrepresent people who don't like to shop at malls, or prefer to shop at a different mall, or don't visit the mall because they lack transportation.

PACE: Analyze & Construct Stages

- In general, why are descriptive statistics useful?

Descriptive statistics allow us to summarize, explore and understand large amounts of data quickly

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

By calculating the mean of the total fare by payment types, we can see that there is some variation. This can help us identify the variables we might choose to use in our hypothesis testing. In this example, based on the variations of total fare by payment type, we chose to investigate further if there is any significant difference between pay with cash or paying with credit card

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

The null hypothesis represents the situation we want to disprove. It represents the status quo, and assumes the observations have happened by chance.

The alternative hypothesis is everything that the null hypothesis is not. It assumes that the observation did not happen by chance but rather as a result of the changes in the variables.

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

The null hypothesis assumes that the observations happened by chance, therefore the null hypothesis is that the difference in mean total fare between cash and credit card has occurred by chance, and there is no significant difference in paying with cash or paying with card.

The alternative hypothesis, as the opposite of the null hypothesis, is that there *is* a significant difference between paying with cash and paying with card.

- What conclusion can be drawn from the hypothesis test?

We concluded that there is a statistically significant difference between paying with cash and card. Passengers who pay with card will pay a higher total fare on average.

PACE: Execute Stage

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

The test suggests that payments with card result in higher total fare which means a higher revenue for taxi drivers.

- What recommendations do you propose based on your results?

It is recommended that the New York City TLC encourages passengers to pay with credit card instead of cash. For example they may have the taxi drivers remind passengers that credit card is preferred as a payment method.