**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 develop a regression model that estimates taxi fares before the ride based on the data gathered by the New York City Taxi & Limousine Commission (TLC). The project aims to provide the TLC with a reliable and accurate tool for estimating fares, which can be used for various purposes such as optimizing pricing strategies, improving cost transparency for passengers, and supporting decision-making processes within the TLC. Ultimately, the project's purpose is to leverage data analysis and modeling techniques to enhance the efficiency and effectiveness of fare estimation in the taxi industry.

* What is your research question for this project?

The research question for this project is: "Can we accurately estimate taxi fares before the ride based on the available data?"

* What is the importance of random sampling?

Random sampling is of utmost importance in research and data analysis as it ensures the representativeness and generalizability of the findings. By selecting samples randomly from a population, every individual or unit in the population has an equal chance of being included in the sample. This minimizes bias and allows the sample to be a fair and unbiased representation of the larger population. Random sampling helps reduce the likelihood of systematic errors and ensures that the sample is more likely to reflect the true characteristics of the population. This is crucial for making accurate inferences and generalizations from the sample to the population, enhancing the external validity of the study. Additionally, random sampling enables the use of statistical techniques that rely on the assumption of random selection, allowing for valid and reliable analyses of the collected data.

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

In the context of the current project scenario, if random sampling is not used, a potential sampling bias could arise if the data is collected only from certain time periods or specific regions within New York City. For example, if the data is collected only during peak tourist seasons or from heavily populated areas with high taxi usage, it may not accurately represent the entire population of taxi trips in New York City throughout the year. This could lead to biased estimates and incorrect conclusions about taxi fares. Without random sampling, the sample may disproportionately represent certain demographics or geographic areas, leading to an overrepresentation or underrepresentation of specific groups in the analysis. Random sampling helps to mitigate this bias by ensuring that every taxi trip has an equal chance of being included in the sample, regardless of time, location, or other factors, providing a more accurate representation of the population of taxi trips in New York City.



 **PACE: Analyze & Construct Stages**

* In general, why are descriptive statistics useful?

Descriptive statistics are useful because they provide a concise summary of a dataset, allowing us to understand and interpret the data in a meaningful way. Here are some reasons why descriptive statistics are important:

1. Summarizing data: Descriptive statistics provide key summary measures such as measures of central tendency (mean, median, mode) and measures of dispersion (standard deviation, range). These measures help us understand the typical values and variability in the data.

2. Data exploration: Descriptive statistics allow us to explore the characteristics and patterns present in the data. We can identify outliers, patterns, trends, and distributions, which can inform further analysis and decision-making.

3. Data comparison: Descriptive statistics enable us to compare different groups or variables within the dataset. We can examine differences in means, variances, or proportions, helping us identify significant variations and make informed comparisons.

4. Communicating findings: Descriptive statistics provide a clear and concise way to communicate the main features of the dataset to others. They help in presenting key findings and insights in a simple and understandable manner, making complex data more accessible to a wider audience.

5. Decision-making: Descriptive statistics provide a basis for decision-making by providing objective information about the dataset. By understanding the central tendencies, variations, and distributions in the data, we can make informed decisions and take appropriate actions.

Moreover, descriptive statistics play a crucial role in summarizing and interpreting data, aiding in data exploration, comparison, communication, and decision-making processes. They provide valuable insights and help us draw meaningful conclusions from the data at hand.

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

Computing descriptive statistics helped in analyzing the data for the project by providing key insights and summaries that allowed for a better understanding of the dataset. Here are some ways in which descriptive statistics helped in the analysis of the data for this project:

1. Understanding the distribution of total fare amounts: By computing descriptive statistics such as the mean, standard deviation, and quartiles, we gained an understanding of the central tendency, spread, and shape of the distribution of total fare amounts. This information helped in assessing the typical fare amounts and the variability in the data.

2. Identifying outliers: Descriptive statistics like the minimum and maximum values helped identify any extreme values or outliers in the total fare amounts. Outliers could indicate unusual or erroneous data points that might require further investigation or data cleaning.

3. Comparing payment types: Descriptive statistics allowed for a comparison of the average total fare amounts across different payment types. By examining the means and other summary statistics for each payment type, we could identify any significant differences in fare amounts based on the payment method used.

4. Summarizing the data for communication: Descriptive statistics provided a concise summary of the total fare amounts, allowing for effective communication of the main features of the dataset to stakeholders. The summary statistics could be presented in reports or visualizations to convey the key findings and provide a comprehensive overview of the data.

In summary, computing descriptive statistics helped in analyzing the data by providing insights into the distribution, outliers, comparisons, and summarization of the total fare amounts. These statistics formed the foundation for understanding and interpreting the data, enabling more informed decision-making and effective communication of the findings.

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

In hypothesis testing, the null hypothesis (H0) and the alternative hypothesis (HA) are two contrasting statements that represent different claims about a population parameter.

The null hypothesis is the default position or assumption, which states that there is no significant difference or relationship between variables or that any observed difference is due to chance. It represents the status quo or the absence of an effect. The null hypothesis is often denoted as "H0" and is the hypothesis that is tested against the alternative hypothesis.

The alternative hypothesis, on the other hand, contradicts the null hypothesis and represents the claim or hypothesis that researchers or analysts want to support. It suggests that there is a significant difference, effect, or relationship between variables, and that any observed difference is not due to chance alone. The alternative hypothesis is denoted as "HA" or "H1."

In hypothesis testing, the goal is to gather evidence from the sample data to either reject the null hypothesis in favor of the alternative hypothesis or fail to reject the null hypothesis due to insufficient evidence. The decision to accept or reject the null hypothesis is based on the analysis of the sample data and the calculation of statistical test statistics, such as p-values or confidence intervals.

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

In this project, the formulation of the null hypothesis (H0) and alternative hypothesis (HA) is based on the specific research question and the objective of the analysis.

For example, in the hypothesis testing scenario related to the average total fare amount between customers who use credit cards and customers who use cash,

H0: The null hypothesis (H0) states that there is no difference in the average total fare amount between the two payment types.

HA: The alternative hypothesis (HA) suggests that there is a difference in the average total fare amount between the two payment types.

* What conclusion can be drawn from the hypothesis test?

From the hypothesis test conducted in this project, we can draw the conclusion that there is a statistically significant difference in the average total fare amount between customers who use credit cards and customers who use cash. The extremely low p-value obtained (P-value: 9.601326356411809e-73) indicates strong evidence against the null hypothesis, suggesting that the null hypothesis can be rejected.

In practical terms, this means that there is a significant difference in the fares paid by customers who use credit cards compared to customers who use cash. However, the direction of the difference (whether credit card users pay more or less than cash users) is not determined by the hypothesis test alone. To understand the direction of the difference, further analysis and interpretation of the descriptive statistics are required.

In summary, the hypothesis test provides evidence that credit card usage and cash usage are associated with different fare amounts, but additional investigation is needed to determine the nature and magnitude of this difference.

**PACE: Execute Stage**

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

This business insight suggests that payment type plays a role in determining the fare amount. The average total fare amount for credit card users (13.43) is higher compared to cash users (12.21). This information can be valuable for the New York City TLC to consider when making decisions related to pricing, promotions, or incentives for different payment methods. They may want to incentivize or promote the use of certain payment types to maximize revenue or improve customer experience.

Additionally, this insight can inform the development of predictive models or pricing strategies based on payment type. It allows the TLC to better understand customer behavior and preferences, enabling them to optimize their operations and provide tailored services to different customer segments.

Overall, the results provide actionable insights for the New York City TLC to make informed decisions related to fare pricing, payment options, and customer engagement strategies.

* What recommendations do you propose based on your results?

Based on the results of the hypothesis test that showed a statistically significant difference in the average total fare amount between credit card users and cash users, I would recommend the following:

1. Promote digital payment options: Since credit card users tend to have higher average fares compared to cash users, it would be beneficial for the New York City TLC to encourage and promote digital payment options such as credit cards or mobile payment systems. This can be done through awareness campaigns, incentives, or partnerships with payment service providers. By promoting digital payments, the TLC can potentially increase revenue and streamline the fare collection process.

2. Analyze fare structure: The results of the hypothesis test indicate that there is a difference in fare amounts between payment types. To optimize revenue and ensure fairness, it would be valuable for the TLC to conduct a detailed analysis of the fare structure for credit card payments and cash payments. This analysis can help identify any discrepancies or potential adjustments needed in the fare calculation method to ensure that both payment types are charged appropriately.

3. Provide incentives for cash payments: While credit card payments may result in higher fares, it is important to consider the convenience and accessibility of cash payments for certain demographics or situations. To cater to customers who prefer or rely on cash payments, the TLC could explore providing incentives such as discounts or loyalty programs specifically for cash users. This can help maintain a balance between different payment options and cater to a diverse customer base.

4. Monitor and evaluate payment trends: It is crucial for the TLC to continuously monitor and evaluate payment trends to stay informed about changing customer preferences and behaviors. This can be achieved by analyzing data on payment types over time and conducting regular statistical analyses. By staying updated on payment trends, the TLC can adapt its policies and strategies to align with customer needs and preferences.

5. Enhance data collection and analysis capabilities: As the TLC continues to gather a large volume of data on taxi trips and fare transactions, investing in advanced data collection and analysis capabilities can provide valuable insights for decision-making. This may involve implementing automated data collection systems, leveraging data analytics tools, and employing data scientists or analysts to extract actionable insights from the data.

Overall, the recommendations aim to optimize revenue, improve customer satisfaction, and align the fare collection process with evolving payment preferences. By implementing these recommendations, the New York City TLC can leverage the findings from the hypothesis test to make informed decisions and enhance its operations in the taxi and limousine industry.