Activity_ Course 4 Automatidata project lab

July 4, 2023

1 Automatidata project

Course 4 - The Power of Statistics

You are a data professional in a data consulting firm, called Automatidata. The current project for their newest client, the New York City Taxi & Limousine Commission (New York City TLC) is reaching its midpoint, having completed a project proposal, Python coding work, and exploratory data analysis.

You receive a new email from Uli King, Automatidata's project manager. Uli tells your team about a new request from the New York City TLC: to analyze the relationship between fare amount and payment type. A follow-up email from Luana includes your specific assignment: to conduct an A/B test.

A notebook was structured and prepared to help you in this project. Please complete the following questions.

2 Course 4 End-of-course project: Statistical analysis

In this activity, you will practice using statistics to analyze and interpret data. The activity covers fundamental concepts such as descriptive statistics and hypothesis testing. You will explore the data provided and conduct A/B and hypothesis testing.

The purpose of this project is to demostrate knowledge of how to prepare, create, and analyze A/B tests. Your A/B test results should aim to find ways to generate more revenue for taxi cab drivers.

Note: For the purpose of this exercise, assume that the sample data comes from an experiment in which customers are randomly selected and divided into two groups: 1) customers who are required to pay with credit card, 2) customers who are required to pay with cash. Without this assumption, we cannot draw causal conclusions about how payment method affects fare amount.

The goal is to apply descriptive statistics and hypothesis testing in Python. The goal for this A/B test is to sample data and analyze whether there is a relationship between payment type and fare amount. For example: discover if customers who use credit cards pay higher fare amounts than customers who use cash.

This activity has four parts:

Part 1: Imports and data loading * What data packages will be necessary for hypothesis testing?

Part 2: Conduct EDA and hypothesis testing * How did computing descriptive statistics help you analyze your data?

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

Part 3: Communicate insights with stakeholders

- What key business insight(s) emerged from your A/B test?
- What business recommendations do you propose based on your results?

Follow the instructions and answer the questions below to complete the activity. Then, you will complete an Executive Summary using the questions listed on the PACE Strategy Document.

Be sure to complete this activity before moving on. The next course item will provide you with a completed exemplar to compare to your own work.

3 Conduct an A/B test

4 PACE stages

Throughout these project notebooks, you'll see references to the problem-solving framework PACE. The following notebook components are labeled with the respective PACE stage: Plan, Analyze, Construct, and Execute.

4.1 PACE: Plan

In this stage, consider the following questions where applicable to complete your code response: 1. What is your research question for this data project? Later on, you will need to formulate the null and alternative hypotheses as the first step of your hypothesis test. Consider your research question now, at the start of this task.

The research question for this data project is: "Is there a significant difference in the fare amounts between customers who pay with credit cards and customers who pay with cash?"

Complete the following steps to perform statistical analysis of your data:

4.1.1 Task 1. Imports and data loading

Import packages and libraries needed to compute descriptive statistics and conduct a hypothesis test.

Hint:

Before you begin, recall the following Python packages and functions that may be useful:

Main functions: stats.ttest_ind(a, b, equal_var)

Other functions: mean()

Packages: pandas, stats.scipy

[1]:

```
import pandas as pd
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
```

Note: As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

[2]:

```
# Load dataset into dataframe
taxi_data = pd.read_csv("2017_Yellow_Taxi_Trip_Data.csv", index_col = 0)
```

4.2 PACE: Analyze and Construct

In this stage, consider the following questions where applicable to complete your code response: 1. Data professionals use descriptive statistics for Exploratory Data Analysis. How can computing descriptive statistics help you learn more about your data in this stage of your analysis?

Computing descriptive statistics is an essential step in Exploratory Data Analysis (EDA) as it helps us understand and summarize the key characteristics of the dataset. In the context of the project scenario with the New York City TLC data, computing descriptive statistics can provide valuable insights into the fare amount and payment type variables.

By calculating measures such as mean, median, standard deviation, and quartiles, we can gain an understanding of the central tendency, spread, and distribution of the fare amounts. This information can help us identify any outliers or anomalies in the data, detect patterns or trends, and assess the variability in fare amounts across different payment types.

Descriptive statistics can also provide preliminary insights into the relationship between payment type and fare amount. For example, by comparing the descriptive statistics of fare amounts for credit card payments and cash payments, we can get an initial understanding of whether there are any noticeable differences in the average fare amounts or the variability between the two payment types.

Overall, computing descriptive statistics allows us to summarize and explore the key characteristics of the data, providing a foundation for further analysis and hypothesis testing.

4.2.1 Task 2. Data exploration

Use descriptive statistics to conduct Exploratory Data Analysis (EDA).

Hint:

Refer back to *Self Review Descriptive Statistics* for this step-by-step proccess.

Note: In the dataset, `payment_type` is encoded in integers: * 1: Credit card * 2: Cash * 3: No charge * 4: Dispute * 5: Unknown

[3]:
```
#==> ENTER YOUR CODE HERE

# Calculate descriptive statistics for all payment types
payment_type_stats = taxi_data.groupby('payment_type')['fare_amount'].describe()

# Display the descriptive statistics
payment_type_stats
```

[3]:
```
                count       mean        std    min  25%  50%     75%     max
payment_type
1             15265.0  13.429748  13.848964    0.0  7.0  9.5  15.000  999.99
2              7267.0  12.213546  11.689940    0.0  6.0  9.0  14.000  450.00
3               121.0  12.186116  14.894232   -4.5  2.5  7.0  15.000   65.50
4                46.0   9.913043  24.162943 -120.0  5.0  8.5  17.625   52.00
```

You are interested in the relationship between payment type and the total fare amount the customer pays. One approach is to look at the average total fare amount for each payment type.

[4]:
```
#==> ENTER YOUR CODE HERE
# Calculate the average fare amount for each payment type
average_fare_by_payment_type = taxi_data.groupby('payment_type')['fare_amount'].
 ↪mean()

# Display the average fare amounts
print(average_fare_by_payment_type)
```

```
payment_type
1    13.429748
2    12.213546
3    12.186116
4     9.913043
Name: fare_amount, dtype: float64
```

Based on the averages shown, it appears that customers who pay in credit card tend to pay a larger total fare amount than customers who pay in cash. However, this difference might arise from random sampling, rather than being a true difference in total fare amount. To assess whether the difference is statistically significant, you conduct a hypothesis test.

4.2.2 Task 3. Hypothesis testing

Before you conduct your hypothesis test, consider the following questions where applicable to complete your code response:

1. Recall the difference between the null hypothesis and the alternative hypotheses. Consider your hypotheses for this project as listed below.

H_0: There is no difference in the average total fare amount between customers who use credit cards and customers who use cash.

H_A: There is a difference in the average total fare amount between customers who use credit cards and customers who use cash.

Your goal in this step is to conduct a two-sample t-test. Recall the steps for conducting a hypothesis test:

1. State the null hypothesis and the alternative hypothesis
2. Choose a signficance level
3. Find the p-value
4. Reject or fail to reject the null hypothesis

Note: For the purpose of this exercise, your hypothesis test is the main component of your A/B test.

You choose 5% as the significance level and proceed with a two-sample t-test.

[ ]:

```
#==> ENTER YOUR CODE HERE

from scipy.stats import ttest_ind

#Significance Level = 5% (  = 0.05)
significance_level = 0.05

# Separate the total fare amounts for credit card users and cash users
credit_card_fare = taxi_data[taxi_data['payment_type'] == 1]['total_amount']
cash_fare = taxi_data[taxi_data['payment_type'] == 2]['total_amount']

# Perform the two-sample t-test
t_statistic, p_value = stats.ttest_ind(credit_card_fare, cash_fare,␣
 ↪alternative='two-sided', equal_var=True)



# Print the p-value
print("T-Statistic:", t_statistic, "P-value:", p_value)
```

Based on the provided p-value of (approximately 9.60e-73), which is significantly smaller than the chosen significance level of 0.05, we reject the null hypothesis. This indicates 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 t-statistic of 18.10468311143095 confirms that the difference between the two groups is substantial.

4.3 PACE: Execute

Consider the questions in your PACE Strategy Document to reflect on the Execute stage.

4.3.1 Task 4. Communicate insights with stakeholders

Ask yourself the following questions:

1. What business insight(s) can you draw from the result of your hypothesis test?
2. Consider why this A/B test project might not be realistic, and what assumptions had to be made for this educational project.

Q1: 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.

Q2: While this A/B test project serves as a valuable educational exercise, it may not be entirely realistic due to several factors:

1. Dataset Size and Representation: The dataset used in this project represents a specific sample of New York City Taxi and Limousine Commission (TLC) data, which may not fully capture the diverse range of trips and passengers in reality. The dataset may not include all possible variables that can influence fare amounts, such as traffic conditions, time of day, or specific trip details. In a real-world scenario, a much larger and more comprehensive dataset would be needed to draw more robust conclusions.

2. Simplified Hypothesis: The hypothesis tested in this project focuses on the difference in average total fare amount between credit card and cash users. While this provides valuable insights into payment method preferences, there are likely many other factors that contribute to fare amounts, such as distance traveled, additional charges, and discounts. Considering only the payment type as the differentiating factor oversimplifies the complexity of fare pricing in real-world scenarios.

3. Assumptions and Limitations: In this educational project, certain assumptions have been made, such as assuming the data is representative and unbiased, assuming equal variances for the t-test, and using a fixed significance level of 0.05. In reality, these assumptions may not hold true, and more thorough data preprocessing and analysis would be necessary to address these issues.

4. Practical Implementation Challenges: Implementing an A/B test in a real-world setting involves practical challenges such as ensuring proper randomization, controlling external factors, and managing the logistics of collecting and analyzing data in real-time. These challenges may not be fully addressed in this educational project due to its scope and limitations.

Considering these factors, it's important to recognize that this educational project provides a foundational understanding of hypothesis testing and statistical analysis. However, for real-world applications, a more comprehensive approach, including a larger dataset, consideration of multiple variables, and careful consideration of assumptions and implementation challenges, would be necessary to draw reliable and actionable insights.

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged.