Activity_ Course 4 Automatidata project lab

March 11, 2024

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.

==> ENTER YOUR RESPONSE HERE Is there a relationship between total fare amount and payment type?

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]:

```
#==> ENTER YOUR CODE HERE
import pandas as pd
from scipy import stats
```

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.

[3]:

```
# 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?

==> ENTER YOUR RESPONSE HERE In general, descriptive statistics are useful because they let you quickly explore and understand large amounts of data. In this case, computing descriptive statistics helps you quickly compare the average total fare amount among different payment types.

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

[4]:

```
#==> ENTER YOUR CODE HERE
taxi_data.head(10)
```

[4]:

```
           VendorID    tpep_pickup_datetime   tpep_dropoff_datetime  \
24870114          2   03/25/2017 8:55:43 AM   03/25/2017 9:09:47 AM
35634249          1   04/11/2017 2:53:28 PM   04/11/2017 3:19:58 PM
106203690         1   12/15/2017 7:26:56 AM   12/15/2017 7:34:08 AM
38942136          2   05/07/2017 1:17:59 PM   05/07/2017 1:48:14 PM
30841670          2  04/15/2017 11:32:20 PM  04/15/2017 11:49:03 PM
23345809          2   03/25/2017 8:34:11 PM   03/25/2017 8:42:11 PM
37660487          2   05/03/2017 7:04:09 PM   05/03/2017 8:03:47 PM
69059411          2   08/15/2017 5:41:06 PM   08/15/2017 6:03:05 PM
8433159           2   02/04/2017 4:17:07 PM   02/04/2017 4:29:14 PM
95294817          1   11/10/2017 3:20:29 PM   11/10/2017 3:40:55 PM
```

```
           passenger_count  trip_distance  RatecodeID store_and_fwd_flag  \
24870114                 6           3.34           1                  N
35634249                 1           1.80           1                  N
106203690                1           1.00           1                  N
38942136                 1           3.70           1                  N
30841670                 1           4.37           1                  N
23345809                 6           2.30           1                  N
37660487                 1          12.83           1                  N
69059411                 1           2.98           1                  N
8433159                  1           1.20           1                  N
95294817                 1           1.60           1                  N

           PULocationID  DOLocationID  payment_type  fare_amount  extra  \
24870114            100           231             1         13.0    0.0
35634249            186            43             1         16.0    0.0
106203690           262           236             1          6.5    0.0
38942136            188            97             1         20.5    0.0
30841670              4           112             2         16.5    0.5
23345809            161           236             1          9.0    0.5
37660487             79           241             1         47.5    1.0
69059411            237           114             1         16.0    1.0
8433159             234           249             2          9.0    0.0
95294817            239           237             1         13.0    0.0

           mta_tax  tip_amount  tolls_amount  improvement_surcharge  \
24870114       0.5        2.76           0.0                    0.3
35634249       0.5        4.00           0.0                    0.3
106203690      0.5        1.45           0.0                    0.3
38942136       0.5        6.39           0.0                    0.3
30841670       0.5        0.00           0.0                    0.3
23345809       0.5        2.06           0.0                    0.3
37660487       0.5        9.86           0.0                    0.3
69059411       0.5        1.78           0.0                    0.3
8433159        0.5        0.00           0.0                    0.3
95294817       0.5        2.75           0.0                    0.3

           total_amount
24870114          16.56
35634249          20.80
106203690          8.75
38942136          27.69
30841670          17.80
23345809          12.36
37660487          59.16
69059411          19.58
8433159            9.80
```

```
95294817          16.55
```

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

[5]:

```
#==> ENTER YOUR CODE HERE
taxi_data.groupby(['payment_type']).mean()[['fare_amount']]
```

[5]:

```
              fare_amount
payment_type
1               13.429748
2               12.213546
3               12.186116
4                9.913043
```

Based on the averages shown, it appears that customers who pay in credit card tend to pay a larger fare amount than customers who pay in cash. However, this difference might arise from random sampling, rather than being a true difference in 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 fare amount between customers who use credit cards and customers who use cash.

H_A: There is a difference in the average 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.

[9]:

```
#==> ENTER YOUR CODE HERE
credit_card = taxi_data[taxi_data['payment_type'] == 1]['fare_amount']
cash = taxi_data[taxi_data['payment_type'] == 2]['fare_amount']
stats.ttest_ind(a=credit_card, b=cash, equal_var=False)
```

```
[9]: Ttest_indResult(statistic=6.866800855655372, pvalue=6.797387473030518e-12)
```

==> ENTER YOUR DECISION TO ACCEPT OR REJECT THE NULL HYPOTHESIS Since the p-value is smaller than the significance level of 5%, you reject the null hypothesis. I conclude that there is a statistically significant difference in the average fare amount between customers who use credit cards and customers who use cash.

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.

==> ENTER YOUR RESPONSE HERE 1.The key business insight is that encouraging customers to pay with credit cards can generate more revenue for taxi cab drivers.

2.This project requires an assumption that passengers were forced to pay one way or the other, and that once informed of this requirement, they always complied with it. The data was not collected this way; so, an assumption had to be made to randomly group data entries to perform an A/B test. This dataset does not account for other likely explanations. For example, riders might not carry lots of cash, so it's easier to pay for longer/farther trips with a credit card. In other words, it's far more likely that fare amount determines payment type, rather than vice versa.

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.