Activity_ Course 4 Automatidata project lab

June 21, 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

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 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.

```
[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?

==> ENTER YOUR RESPONSE HERE

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 process.

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

```
[3]: # descriptive stats code for EDA taxi_data.describe(include='all')
```

```
[3]:
                   VendorID
                               tpep_pickup_datetime
                                                        tpep_dropoff_datetime
              22699.000000
     count
                                                22699
                                                                          22699
     unique
                                                22687
                                                                          22688
                        \mathtt{NaN}
                              07/03/2017 3:45:19 PM
                                                        10/18/2017 8:07:45 PM
     top
                        NaN
     freq
                        NaN
                                                                              2
     mean
                   1.556236
                                                  NaN
                                                                            NaN
                   0.496838
     std
                                                  NaN
                                                                            NaN
     min
                   1.000000
                                                  NaN
                                                                            NaN
     25%
                   1.000000
                                                  NaN
                                                                            NaN
     50%
                   2.000000
                                                  NaN
                                                                            NaN
     75%
                   2.000000
                                                  NaN
                                                                            NaN
                   2.000000
                                                  NaN
     max
                                                                            NaN
```

passenger_count trip_distance RatecodeID store_and_fwd_flag \

count	22699.0000	00 22699.000	0000	22699.000	0000	22699		
unique	NaN		NaN		NaN	2		
top	NaN		NaN		NaN	N		
freq	NaN		NaN		NaN	22600		
mean	1.642319 2.913		313	1.043	3394	NaN		
std	1.285231 3.653		3171	0.708	3391	NaN		
min	0.000000 0.000			1.000		NaN		
25%	1.000000 0.99			1.000		NaN		
50%	1.000000 1.6			1.000		NaN		
75%			0000	1.000		NaN		
max	6.0000			99.000		NaN		
mαx	0.000000 55.500000 55.000000				7000	Nan		
	PULocationID	DOLocationID	naur	ment_type	fare_amount	extr	· /	
aat	22699.000000			99.000000	22699.000000			
count		22699.000000	2203			22699.00000		
unique	NaN	NaN		NaN	NaN	Na		
top	NaN	NaN		NaN	NaN	Na		
freq	NaN	NaN		NaN	NaN	Na		
mean	162.412353	161.527997	1.336887		13.026629	0.33327		
std	66.633373	70.139691		0.496211	13.243791	0.46309	7	
min	1.000000	1.000000		1.000000	-120.000000	-1.00000	0	
25%	114.000000	112.000000		1.000000	6.500000	0.00000	0	
50%	162.000000	162.000000		1.000000	9.500000	0.00000	0	
75%	233.000000	233.000000		2.000000	14.500000	0.50000	0	
max	265.000000	265.000000		4.000000	999.990000	4.50000	0	
	mta_tax	tip_amount	tol	ls_amount	improvement s	provement_surcharge \		
count	22699.000000	22699.000000		99.000000	_	22699.000000		
unique	NaN	NaN		NaN		NaN		
top	NaN	NaN		NaN		NaN		
freq	NaN	NaN		NaN		NaN		
mean	0.497445	1.835781		0.312542		0.299551		
std	0.039465	2.800626		1.399212 0.015673				
min	-0.500000	0.000000		0.000000 -0.300000				
				0.000000 0.300000				
25%	0.500000	0.000000						
50%	0.500000	1.350000		0.000000		0.300000		
75%	0.500000	2.450000		0.000000		0.300000		
max	0.500000	200.000000	-	19.100000		0.300000		
	total_amount							
count	22699.000000							
unique	NaN							
top	NaN							
freq	NaN							
mean	16.310502							
std	16.097295							
min	-120.300000							
25%	8.750000							

```
50% 11.800000
75% 17.800000
max 1200.290000
```

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.

```
[4]: taxi_data.groupby('payment_type')['fare_amount'].mean()
```

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.

```
[5]: #hypothesis test, A/B test
#significance level
```

```
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)
```

[5]: Ttest indResult(statistic=6.866800855655372, pvalue=6.797387473030518e-12)

==> ENTER YOUR DECISION TO ACCEPT OR REJECT THE NULL HYPOTHESIS

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

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.