Activity_ Course 4 Automatidata project
lab_2023_05_16_02_26_39

May 16, 2023

0.1 Course 4 Automatidata project

Course 4 - The Power of Statistics

You are a data professional in a data analytics 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. You also discover follow-up emails from three other team members: Deshawn Washington, Luana Rodriguez, and Udo Bankole. These emails discuss the details of the analysis. A final email from Luana includes your specific assignment: to conduct an A/B test.

1 Course 4 end-of-course project: Statistical analysis

In this activity, 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.

The goal is to apply descriptive statistics and hypothesis testing in Python.

This activity has three parts:

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

Part 2: Conduct 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.

Recall that you have a helpful tool at your disposal! Refer to the PACE Strategy Document here to help apply your learnings, apply new problem-solving skills, and guide your approach to this project.

PACE strategy document

2 Conduct an A/B test

Now, you are trying to find ways to improve the work experience and compensation of taxi cab drivers.

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.

The purpose of this A/B test is 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 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 two parts:

Part 1: Exploratory data analysis Explore the NYC Taxi dataset with Python using a Jupyter notebook. This includes:

- Computing descriptive statistics

Part 2: Hypothesis testing with Python

- Conducting a two-sample hypothesis test

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.

2.1 PACE stages

- `[Plan](#scrollTo=psz51YkZVwtN&line=3&uniqifier=1)`
- `[Analyze](#scrollTo=mA7Mz_SnI8km&line=4&uniqifier=1)`

- [Construct](#scrollTo=Lca9c8XON8lc&line=2&uniqifier=1)
- [Execute](#scrollTo=401PgchTPr4E&line=2&uniqifier=1)

3 Pace: Plan Stage

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.

Is there a relationship between total fare amount and payment type?

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

3.1 Task 1. Imports and data loading

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

Hint 1

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

[3]:
```
import pandas as pd
import numpy as np
from scipy import stats
```

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.

[7]:
```
# RUN THIS CELL TO IMPORT YOUR DATA.

#==> ENTER YOUR CODE HERE
df = pd.read_csv("2017_Yellow_Taxi_Trip_Data.csv", index_col = 0)
df.head()
```

[7]:

	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	

```
           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

           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

           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

           total_amount
24870114          16.56
35634249          20.80
106203690          8.75
38942136          27.69
30841670          17.80
```

4 PACE: Analyze Stage and Construct Stage

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

4.1 Task 2. Data exploration

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

Hint 1

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

[6]:
```
df.describe(include='all')
```

[6]:

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

	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	\
count	22699.000000	22699.000000	22699.000000	22699	
unique	NaN	NaN	NaN	2	
top	NaN	NaN	NaN	N	
freq	NaN	NaN	NaN	22600	
mean	1.642319	2.913313	1.043394	NaN	
std	1.285231	3.653171	0.708391	NaN	
min	0.000000	0.000000	1.000000	NaN	
25%	1.000000	0.990000	1.000000	NaN	
50%	1.000000	1.610000	1.000000	NaN	
75%	2.000000	3.060000	1.000000	NaN	
max	6.000000	33.960000	99.000000	NaN	

	PULocationID	DOLocationID	payment_type	fare_amount	extra	\
count	22699.000000	22699.000000	22699.000000	22699.000000	22699.000000	
unique	NaN	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	NaN	
mean	162.412353	161.527997	1.336887	13.026629	0.333275	
std	66.633373	70.139691	0.496211	13.243791	0.463097	
min	1.000000	1.000000	1.000000	-120.000000	-1.000000	
25%	114.000000	112.000000	1.000000	6.500000	0.000000	
50%	162.000000	162.000000	1.000000	9.500000	0.000000	
75%	233.000000	233.000000	2.000000	14.500000	0.500000	
max	265.000000	265.000000	4.000000	999.990000	4.500000	

	mta_tax	tip_amount	tolls_amount	improvement_surcharge	\
count	22699.000000	22699.000000	22699.000000	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
25%          0.500000      0.000000      0.000000              0.300000
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 total fare amount the customer pays. One approach is to look at the average total fare amount for each payment type.

[8]:
```
cash= df[df['payment_type']==2]
card=df[df['payment_type']==1]
```

[9]:
```
print(cash['total_amount'].mean())
print(card['total_amount'].mean())
```

```
13.545820833908332
17.66357746478734
```

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 Task 3. Hypothesis testing

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

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.

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.

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

[12]:
```
stats.ttest_ind(a=card['total_amount'], b=cash['total_amount'], equal_var=False)
```

[12]:
```
Ttest_indResult(statistic=20.34644022783838, pvalue=4.5301445359736376e-91)
```

Since the p-value is extremely small (much smaller than the significance level of 5%), you reject the null hypothesis. You conclude that there is a statistically significant difference in the average total fare amount between customers who use credit cards and customers who use cash.

4.3 PACE: Execute Stage

Consider these questions to reflect on the Execute stage of this task.

4.4 Task 4: Communicate insights with stakeholders

In conclusion, 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 pedagogical project.

The key business insight is that encouraging customers to pay with credit cards will likely generate more revenue for taxi cab drivers.

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. The difference between average card payment fare and cash fare is inflated, because we use the total amount as the comparing variable. But cash fares all have tip values of $0, while card payments have non-zero values. A possible reason for this occurance is because cash tips aren't declared. In turn, this means that we capture tips in one group but not in the other. Instead, one should be comparing the fare_amount column.