New_York_TLC_2_AB_testing

April 1, 2025

1 New York TLC project

Part 2 - A/B testing

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?

2 Conduct an A/B test

In this stage, consider the following questions where applicable to complete your code response: 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.

Possible response: The research question for this data project: "Is there a relationship between total fare amount and payment type?"

2.0.1 Task 1. Imports and data loading

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

```
[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)
```

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?

Possible response: 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.

2.0.2 Task 2. Data exploration

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

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

```
[3]: taxi_data.describe(include='all')
```

```
[3]:            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 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()
```

```
[4]: 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 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.

2.0.3   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)

Possible response: Since the p-value is significantly smaller than the significance level of 5%, you reject the null hypothesis.

Notice the 'e-12' at the end of the pvalue result.

You conclude that there is a statistically significant difference in the average fare amount between customers who use credit cards and customers who use cash.

2.0.4 Task 4. Communicate insights with stakeholders

Possible 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.

Possible responses: 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.