A-B testing -New York City TLC

July 17, 2024

By Nitish Adhikari

Email id :nitishbuzzpro@gmail.com , +91-9650740295

Linkedin : https://www.linkedin.com/in/nitish-adhikari-6b2350248

1 Project

The project is for, the New York City Taxi & Limousine Commission (New York City TLC) which is reaching its midpoint on a project, having completed a project proposal, Python coding work, and exploratory data analysis.The New York City TLC: wants to analyze the relationship between fare amount and payment type.

2 Statistical analysis

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

Note: assumpion is 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 The goal for this A/B test is to sample data and analyze whether there is a relationship between payment type and fare amount.

Part 1: Imports and data loading

Part 2: Conduct EDA and hypothesis testing

Part 3: Communicate insights with stakeholders

3 Conduct an A/B test

4 PACE stage: Plan, Analyze, Construct, and Execute.

4.1 PACE: Plan

1. What is the research question for this data project?

Is there a difference in the average fare amount between customers who use credit cards and customers who use cash.

4.1.1 Task 1. Imports and data loading

[ ]:
```
import pandas as pd
from scipy.stats import ttest_ind
```

[18]:
```
import pandas as pd
from scipy.stats import ttest_ind
```

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

How can computing descriptive statistics help in this stage of the analysis?

1. Summarizing Data:. Measures of Central Tendency: Mean, median, and mode give a quick sense of where most data points lie. Measures of Dispersion: Range, variance, and standard deviation indicate the spread and variability in the data.
2. Understanding Distribution: Shape of Distribution: Skewness and kurtosis help understand the symmetry and peakedness of the data distribution. Histogram and Frequency Distribution: Visualizing how data points are distributed across different values.
3. Identifying Outliers: Interquartile Range (IQR): Helps detect outliers by looking at data points that fall significantly outside the central portion of the data. Box Plots: Visual tool to easily spot outliers.
4. Detecting Patterns and Relationships: Correlation Coefficients: Measures like Pearson's or Spearman's correlation coefficients show the relationship between variables. Scatter Plots: Visualize relationships and correlations between two continuous variables.
5. Assessing Data Quality: Missing Values: Count and percentage of missing values highlight data quality issues. Duplicate Values: Identifying and quantifying duplicates to ensure data integrity.
6. Comparing Subsets: Group Statistics: Mean, median, and other statistics computed for different subsets or groups within the data help compare these groups.
7. Informing Further Analysis: Feature Engineering: Insights from descriptive statistics can guide the creation of new features. Model Selection: Understanding data distribution and relationships helps in selecting appropriate modeling techniques.

4.2.1 Task 2. Data exploration

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

[21]:
```
taxi_data.isnull().sum()
```

[21]:
```
VendorID                 0
tpep_pickup_datetime     0
tpep_dropoff_datetime    0
passenger_count          0
trip_distance            0
RatecodeID               0
store_and_fwd_flag       0
PULocationID             0
DOLocationID             0
payment_type             0
fare_amount              0
extra                    0
mta_tax                  0
tip_amount               0
tolls_amount             0
improvement_surcharge    0
total_amount             0
dtype: int64
```

[22]:
```
taxi_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22699 entries, 24870114 to 17208911
Data columns (total 17 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   VendorID               22699 non-null  int64
 1   tpep_pickup_datetime   22699 non-null  object
 2   tpep_dropoff_datetime  22699 non-null  object
 3   passenger_count        22699 non-null  int64
 4   trip_distance          22699 non-null  float64
 5   RatecodeID             22699 non-null  int64
 6   store_and_fwd_flag     22699 non-null  object
 7   PULocationID           22699 non-null  int64
 8   DOLocationID           22699 non-null  int64
 9   payment_type           22699 non-null  int64
 10  fare_amount            22699 non-null  float64
 11  extra                  22699 non-null  float64
 12  mta_tax                22699 non-null  float64
 13  tip_amount             22699 non-null  float64
 14  tolls_amount           22699 non-null  float64
 15  improvement_surcharge  22699 non-null  float64
 16  total_amount           22699 non-null  float64
dtypes: float64(8), int64(6), object(3)
memory usage: 3.6+ MB
```

[12]:
```
taxi_data.describe(include='all')
```

[12]:

	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
```

We 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.

[11]:
```python
taxi_data.groupby('payment_type')['fare_amount'].mean()
```

[11]:
```
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, we need to conduct a hypothesis test.

4.2.2 Task 3. Hypothesis testing

The hypotheses for this project are 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

choosing 5% as the significance level and proceeding with a two-sample t-test.

[25]:
```
sample_creditcard = taxi_data[taxi_data['payment_type']==1]['fare_amount']
sample_cash = taxi_data[taxi_data['payment_type']==2]['fare_amount']

ttest_ind(sample_creditcard,sample_cash,equal_var=False)
```

[25]:
```
Ttest_indResult(statistic=6.866800855655372, pvalue=6.797387473030518e-12)
```

Based on the pvalue=6.797387473030518e-12 which is extremely less than the significance level of 0.05. We can reject the null hypothesis and accept the alternative hypothesis that There is a difference in the average fare amount between customers who use credit cards and customers who use cash.

4.3 PACE: Execute

4.3.1 Task 4. Communicate insights with stakeholders

1. What business insight(s) can be drawn from the result of the hypothesis test?
2. What assumptions had to be made for this project.

Business insights:

1. Higher Fare Amounts for Credit Card Payments: The significant difference in fare amounts indicates that customers who pay by credit card tend to have higher fare amounts compared to those who pay by cash. This could be due to various factors such as: Credit card users might be taking longer or more expensive trips.

2. Different Customer Profiles: The difference in fare amounts suggests that credit card users and cash users may belong to different customer segments with distinct behaviors and preferences. Understanding these segments can help in targeted marketing and personalized service offerings.

3. Encouraging Credit Card Payments: If higher fare amounts are associated with credit card payments, the business could benefit from encouraging more customers to use credit cards. This can be done through incentives such as discounts, loyalty points, or easier credit card processing options.

4. Streamlining Payment Processes: Understanding the preference and behavior of customers regarding payment methods can help in optimizing the payment process, reducing transaction times, and enhancing customer satisfaction.

5. Adjusting Pricing Models: If the fare amounts are consistently higher for credit card payments, the business might consider adjusting its pricing model to maximize revenue. This could include dynamic pricing strategies based on payment method or offering bundled services for higher fare trips.

what assumptions had to be made:

Independence: The test assumes that the samples are independent. In reality, there might be correlations or dependencies between the samples (e.g customers might use both payment methods).

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.