Hypothesis Testing with Python · Taxi Sales Data

Date: 17 April 2024

Project completed by: Anton Starshev
linkedin.com/in/starshev

Source

This is an end-of-course workplace scenario project **Automatidata, featuring a fictional data consulting firm** proposed within the syllabus of *Google Advanced Data Analytics Professional Certificate* on Coursera

Purpose

The purpose of this portfolio project is to demonstrate my knowledge of how to prepare, create and conduct hypothesis testing, as well as my ability to draw valuable insights for the benefit of business development.

Context

According to the fictional project scenario, I am working as a data professional in a data consulting firm, called Automatidata.

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles.

A new request from the New York City TLC arises: **to analyze the relationship between fare amounts and payment type**. The team agrees to perform a hypothesis test using the data.

Data

This project uses a dataset called **2017_Yellow_Taxi_Trip_Data.csv** gathered by the New York City Taxi & Limousine Commission and published by the city of New York as part of their NYC Open Data program.

In order to improve the learning experience and shorten runtimes, a sample was drawn from the 113 million rows in the 2017 Yellow Taxi Trip Data table. The dataset contains 22,699 rows (each row represents a different trip) and 17 columns.

Project Goal

The practical goal is to apply descriptive statistics and Hypothesis testing in Python, analyzing 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.

Solution

Starting my project, I broke down the execution process into four key phases in order to carry them out step by step:

1. Importing necessary Python packages and dataset loading
2. Performing Exploratory Data Analysis (EDA) and computing descriptive statistics
3. Conducting Hypothesis Testing
4. Formulating business insights and recommendations

1 · Data Loading

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

In [2]:
```
import pandas as pd
from scipy import stats
```

Loaded the scenario dataset into a DataFrame.

In [3]:
```
taxi_data = pd.read_csv("2017_Yellow_Taxi_Trip_Data.csv", index_col = 0)
```

2 · Data Exploration

Previewed the loaded data.

In [23]:
```
taxi_data.head()
```

Out[23]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	Rate
24870114	2	03/25/2017 8:55:43 AM	03/25/2017 9:09:47 AM	6	3.34	
35634249	1	04/11/2017 2:53:28 PM	04/11/2017 3:19:58 PM	1	1.80	
106203690	1	12/15/2017 7:26:56 AM	12/15/2017 7:34:08 AM	1	1.00	
38942136	2	05/07/2017 1:17:59 PM	05/07/2017 1:48:14 PM	1	3.70	
30841670	2	04/15/2017 11:32:20 PM	04/15/2017 11:49:03 PM	1	4.37	

Checked the data size.

In [29]:
```
taxi_data.shape
```

Out[29]:
```
(22699, 17)
```

Verified the data types and names of columns.

In [6]:
```
taxi_data.dtypes
```

Out[6]:
```
VendorID                  int64
tpep_pickup_datetime     object
tpep_dropoff_datetime    object
passenger_count           int64
trip_distance           float64
RatecodeID                int64
store_and_fwd_flag       object
PULocationID              int64
DOLocationID              int64
payment_type              int64
fare_amount             float64
extra                   float64
mta_tax                 float64
tip_amount              float64
tolls_amount            float64
improvement_surcharge   float64
total_amount            float64
dtype: object
```

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

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

Out[12]:

	fare_amount
count	22699.000000
mean	13.026629
std	13.243791
min	-120.000000
25%	6.500000
50%	9.500000
75%	14.500000
max	999.990000

Being interested in the relationship between payment type and the fare amount the customer pays, one approach within the EDA was to look at the average fare amount for each payment type.

Note: In the dataset, column *payment_type* is encoded in the following integers:

1 · Credit card
2 · Cash
3 · No charge
4 · Dispute
5 · Unknown

In [17]:
```
taxi_data.groupby('payment_type')[['fare_amount']].mean()
```

Out[17]:

payment_type	fare_amount
1	13.429748
2	12.213546
3	12.186116
4	9.913043

Observation: Based on preliminary research analysis, taxi customers who pay by card tend to spend more on average than those who pay with cash. However, this difference could be due to sample variability. So, the next step was to check the statistical significance of this difference through hypothesis testing.

3 · Hypothesis Test

Stated the null hypothesis and the alternative hypothesis:

H_0: There is no difference in the average fare amount between customers who use credit cards and customers who use cash.

H_1: There is a difference in the average fare amount between customers who use credit cards and customers who use cash.

Assigned a **5% significance level** to the Hypothesis Test.

Determined the type of Hypothesis Testing: **two-sample two-tailed t-test**.

Filtered the data into two groups based on the payment method: cash or credit card.

In [20]:
```
taxi_data_card = taxi_data[taxi_data['payment_type'] == 1]

taxi_data_cash = taxi_data[taxi_data['payment_type'] == 2]
```

Conducted the Hypothesis Test using the SciPy Stats module.

In [25]:
```
stats.ttest_ind(a = taxi_data_card.fare_amount, b = taxi_data_cash.fare_amount,
equal_var = False, alternative = 'two-sided')
```

Out[25]:
```
TtestResult(statistic=6.866800855655372, pvalue=6.797387473030518e-12, df=16675.48547403
633)
```

Test Result: Given the p-value is significantly smaller than the 5% significance level, I rejected the null hypothesis.

4 · Insight and Recommendation

Business Insight: Based on the conducted test, I concluded that there is a statistically significant difference in the average fare amount between customers who use credit cards and those who use cash. Specifically, customers who use credit cards exhibit a higher total amount compared to cash-paying customers. Therefore, encouraging customers to pay with credit cards can lead to increased revenue for taxi cab drivers, as evidenced by the statistical analysis.

Business Recommendation: *Based on our research, the Automatidata data team suggests that the New York City TLC promotes credit card payments among customers and develops strategies to incentivize their usage. For instance, implementing signage within cabs indicating a preference for credit card payments and requiring cab drivers to verbally communicate this preference to customers could be effective measures.*

Skills

Throughout this project, I showcased the following professional competencies:

- Working with DataFrame and conducting exploratory data analysis using the Pandas library
- Preparing and conducting Hypothesis testing using the Scipy Stats library
- Evaluating test results and formulating data-driven recommendations

Acknowledgment

I would like to express gratitude to Google and Coursera for supporting the educational process and providing the opportunity to refine and showcase skills acquired during the courses by completing real-life scenario portfolio projects, such as this.