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?**

A: Pandas and scipy will be needed.

Part 2: Conduct EDA and hypothesis testing

- How did computing descriptive statistics help you analyze your data?

A: It allows me to see basic stats like mean, std, and the quartile distributions.

- How did you formulate your null hypothesis and alternative hypothesis?

**A: I used a significance level of 5% (0.05), comparing the sets of cash and credit card using a t-test. **

Part 3: Communicate insights with stakeholders

- What key business insight(s) emerged from your A/B test?

There is a significant difference between payment types.

- What business recommendations do you propose based on your results?

Leading customers to pay with a credit card would be beneficial in creating higher fares for the Taxi/Limo Industry in NYC. Future modeling or a study could be performed to see if this would indeed increase profitability.

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.

The research question: Is there a difference between fare amounts and the payment types?

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

In [80]:
```
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.

In [81]:
```
 # Load dataset into dataframe
taxi_data = pd.read_csv("../data/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?

Descriptive statistics can give you an idea of distrubution and averages amonst tabular data. This can lead you to make accurate assumptions to perform more analysis on certain data sets.

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 proccess.

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

In [82]:
```
taxi_data.describe()
```

Out[82]:

	VendorID	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocati
count	22699.000000	22699.000000	22699.000000	22699.000000	22699.000000	22699.0(
mean	1.556236	1.642319	2.913313	1.043394	162.412353	161.52
std	0.496838	1.285231	3.653171	0.708391	66.633373	70.13
min	1.000000	0.000000	0.000000	1.000000	1.000000	1.0(
25%	1.000000	1.000000	0.990000	1.000000	114.000000	112.0(
50%	2.000000	1.000000	1.610000	1.000000	162.000000	162.0(
75%	2.000000	2.000000	3.060000	1.000000	233.000000	233.0(
max	2.000000	6.000000	33.960000	99.000000	265.000000	265.0(

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.

In [83]:
```
payment_type_fare_amount = 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:

Recall the difference between the null hypothesis and the alternative hypotheses. Consider your hypotheses for this project as listed below.

H0: There is no difference in the average fare amount between customers who use credit cards and customers who use cash.

HA: 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.

In [84]:
```
sig_level = 0.05
credit = taxi_data[taxi_data['payment_type'] == 1]['fare_amount']
cash = taxi_data[taxi_data['payment_type'] == 2]['fare_amount']
p_value = stats.ttest_ind(a=credit, b=cash, equal_var=False)[1]
if p_value <= sig_level:
    print("Reject null hypothesis - p-value is " + str(p_value))
else:
    print("Fail to reject null hypothesis - p-value is " + str(p_value))
```

```
Reject null hypothesis - p-value is 6.797387473031004e-12
```

Since the p-value is less than the significance level, we 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.

1. The insights about this test are that there is a significance between credit card and cash payments being that credit cards see more favorable earnings for taxi cab drivers.

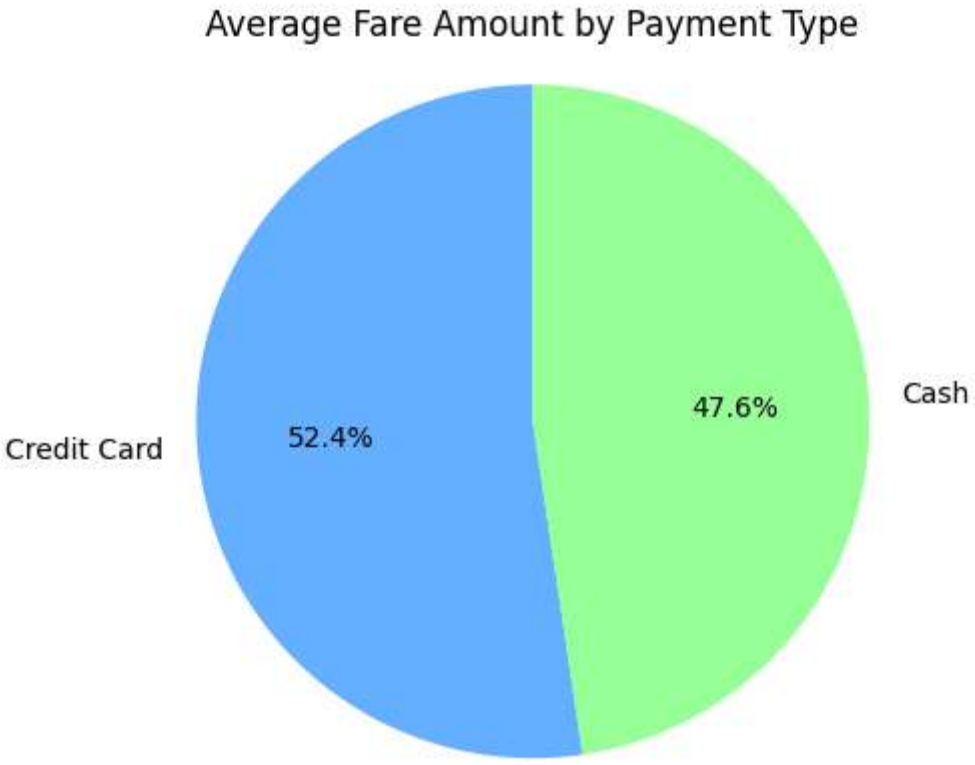
2. There may be additional factors that are outside of just the choice of credit/cash use. For instance, many people in the modern era do not carry large sums of cash, or any cash for that matter. This may affect their choices in payment types. Also, many people do not accept cash for services anymore, because of the likelyhood they may be targets of criminals, so credit cards are usually the go-to payment methods as they take up less space in one's wallet and can be discontinued immediately if stolen or lost.

EXTRASTEP ON MY OWN: Using a simple pie chart we can show the diversity of payment types. When comparing credit card and cash, the data shows that credit cards make up 4.8% more on average than cash payments.

In [87]:
```
import matplotlib.pyplot as plt
labels = ['Credit Card', 'Cash']
plt.pie(payment_type_fare_amount.loc[[1, 2]],
        labels=labels[:len(payment_type_fare_amount)],
        autopct='%1.1f%%',
        startangle=90,
        colors=['#66b3ff','#99ff99','#ff9999','#ffcc99'])

plt.title('Average Fare Amount by Payment Type')
plt.axis('equal')
#plt.savefig('payment_type_pie_chart.png')
```

Out[87]:
```
(np.float64(-1.0999998203277497),
 np.float64(1.0999989451037486),
 np.float64(-1.0999995079902272),
 np.float64(1.0999999765709632))
```



Congratulations! You've completed this lab. However, you may not notice a green check marknext 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.

In [ ]: