Waze Project

Milestone 4 / 4a - Compute descriptive statistics. Conduct hypothesis testing

Data exploration and hypothesis testing

The purpose of this project is to compute descriptive statistics and conduct a two-sample hypothesis test.

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

This notebook has four parts:

Part 1: Imports and data loading

Part 2: Data exploration

Part 3: Conduct hypothesis testing

Part 3: Communicate insights

Data exploration and hypothesis testing

"Do drivers who open the application using an iPhone have the same number of drives on average as drivers who use Android devices?"

Task 1. Imports and data loading

In [1]:

```
# Import any relevant packages or libraries
import pandas as pd
from scipy import stats
```

In [2]:

```
# Load dataset into dataframe
df = pd.read_csv('waze_dataset.csv')
```

Task 2. Data exploration

Using descriptive statistics to conduct exploratory data analysis (EDA).

In [3]:

```
# 1. Create `map_dictionary`
map_dictionary = {'Android': 2, 'iPhone': 1}

# 2. Create new `device_type` column
df['device_type'] = df['device']

# 3. Map the new column to the dictionary
df['device_type'] = df['device_type'].map(map_dictionary)

df['device_type'].head()
```

Out[3]:

```
0     2
```

```
1    1
2    2
3    1
4    2
Name: device_type, dtype: int64
```

Average number of drives for each device type

In [4]:

```python
df.groupby('device_type')['drives'].mean()
```

Out[4]:

```
device_type
1    67.859078
2    66.231838
Name: drives, dtype: float64
```

Given the displayed averages, it seems that iPhone device users tend to have a higher average number of drives when interacting with the application. However, it's important to consider that this disparity may be a result of random sampling rather than an actual difference in the number of drives. To determine if the distinction is statistically significant, we can perform a hypothesis test.

Task 3. Hypothesis testing

The goal is to conduct a two-sample t-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: This is a t-test for two independent samples. This is the appropriate test since the two groups are independent (Android users vs. iPhone users).

Hypotheses:

$H0$: There is no difference in average number of drives between drivers who use iPhone devices and drivers who use Androids.

HA : There is a difference in average number of drives between drivers who use iPhone devices and drivers who use Androids.

Two-sample test with 5% as the significance level with a two-sample t-test.

In [5]:

```python
# 1. Isolate the `drives` column for iPhone users.
iPhone = df[df['device_type'] == 1]['drives']

# 2. Isolate the `drives` column for Android users.
Android = df[df['device_type'] == 2]['drives']

# 3. Perform the t-test
stats.ttest_ind(a=iPhone, b=Android, equal_var=False)
```

Out[5]:

```
Ttest_indResult(statistic=1.4635232068852353, pvalue=0.1433519726802059)
```

p Value = 0.143...

As the p-value exceeds the selected significance level of 5%, we fail to reject the null hypothesis. This indicates that there is no statistically significant distinction in the average number of drives between iPhone users and Android users.

Task 4. Insights

The significant business insight is that, on average, drivers who utilize iPhone devices have a comparable number of drives to those using Androids.

One potential subsequent action is to investigate additional factors that influence the variation in the number of drives. Conducting additional hypothesis tests can help gain further insights into user behavior. Temporary alterations in marketing strategies or user interface for the Waze app could yield more data to examine churn patterns.