Waze Project

Course 4 - The Power of Statistics

Your team is nearing the midpoint of their user churn project. So far, you've completed a project proposal, and used Python to explore and analyze Waze's user data. You've also used Python to create data visualizations. The next step is to use statistical methods to analyze and interpret your data.

You receive a new email from Sylvester Esperanza, your project manager. Sylvester tells your team about a new request from leadership: to analyze the relationship between mean amount of rides and device type. You also discover follow-up emails from three other team members: May Santner, Chidi Ga, and Harriet Hadzic. These emails discuss the details of the analysis. They would like a statistical analysis of ride data based on device type. In particular, leadership wants to know if there is a statistically significant difference in mean amount of rides between iPhone® users and Android™ users. A final email from Chidi includes your specific assignment: to conduct a two-sample hypothesis test (t-test) to analyze the difference in the mean amount of rides between iPhone users and Android users.

A notebook was structured and prepared to help you in this project. Please complete the following questions and prepare an executive summary.

Course 4 End-of-course project: Data exploration and hypothesis testing

In this activity, you will explore the data provided and conduct a hypothesis test.

The purpose of this project is to demostrate knowledge of how to conduct a two-sample hypothesis test.

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

This activity has three parts:

Part 1: Imports and data loading

- What data packages will be necessary for hypothesis testing?

Part 2: Conduct 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 hypothesis test?
- What business recommendations do you propose based on your results?

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.

Data exploration and hypothesis testing

Research question:

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

Complete the following tasks to perform statistical analysis of your data:

Task 1. Imports and data loading

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

In [1]:
```
# Import any relevant packages or libraries
import pandas as pd
from scipy import stats
```

Import the dataset.

In [2]:
```
# Load dataset into dataframe
df = pd.read_csv('waze_dataset.csv')
```

Answer:

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 amount of drives by device type.

Task 2. Data exploration

Use descriptive statistics to conduct exploratory eata analysis (EDA).

Note: In the dataset, `device` is a categorical variable with the labels `iPhone` and `Android`.

In order to perform this analysis, you must turn each label into an integer. The following code assigns a `1` for an `iPhone` user and a `2` for `Android`. It assigns this label back to the variable `device_type`.

Note: Creating a new variable is ideal so that you don't overwrite original data.

1. Create a dictionary called `map_dictionary` that contains the class labels (`'Android'` and `'iPhone'`) for keys and the values you want to convert them to (`2` and `1`) as values.
2. Create a new column called `device_type` that is a copy of the `device` column.
3. Use the `map()` method on the `device_type` series. Pass `map_dictionary` as its argument. Reassign the result back to the `device_type` series.

When you pass a dictionary to the `Series.map()` method, it will replace the data in the series where that data matches the dictionary's keys. The values that get imputed are the values of the dictionary.

```
Example:
df['column']
```

column
A
B
A
B

```
map_dictionary = {'A': 2, 'B': 1}
df['column'] = df['column'].map(map_dictionary)
df['column']
```

column
2
1

column
2
1

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
```

You are interested in the relationship between device type and the number of drives. One approach is to look at the average number of drives for each device type. Calculate these averages.

In [4]:
```
df.groupby('device_type')['drives'].mean()
```

Out[4]:
```
device_type
1    67.859078
2    66.231838
Name: drives, dtype: float64
```

Based on the averages shown, it appears that drivers who use an iPhone device to interact with the application have a higher number of drives on average. However, this difference might arise from random sampling, rather than being a true difference in the number of drives. To assess whether the difference is statistically significant, you can conduct a hypothesis test.

Task 3. Hypothesis testing

Your goal 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: 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).

Recall the difference between the null hypothesis (H_0) and the alternative hypothesis (H_A).

Question: What are your hypotheses for this data project?

Hypotheses:

H_0: There is no difference in average number of drives between drivers who use iPhone devices and drivers who use Androids.

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

Next, choose 5% as the significance level and proceed with a two-sample t-test.

You can use the `stats.ttest_ind()` function to perform the test.

Technical note: The default for the argument `equal_var` in `stats.ttest_ind()` is `True`, which assumes population variances are equal. This equal variance assumption might not hold in practice (that is, there is no strong reason to assume that the two groups have the same variance); you can relax this assumption by setting `equal_var` to `False`, and `stats.ttest_ind()` will perform the unequal variances t-test (known as Welch's `t`-test). Refer to the scipy t-test documentation for more information.

1. Isolate the `drives` column for iPhone users.
2. Isolate the `drives` column for Android users.
3. Perform the t-test

In [5]:
```
# 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)
```

Question: Based on the p-value you got above, do you reject or fail to reject the null hypothesis?

> *Since the p-value is larger than the chosen significance level (5%), you fail to reject the null hypothesis. You conclude that there is not** a statistically

> significant difference in the average number of drives between drivers who use iPhones and drivers who use Androids.*

Task 4: Communicate insights with stakeholders

Now that you've completed your hypothesis test, the next step is to share your findings with the Waze leadership team. Consider the following question as you prepare to write your executive summary:

- What business insight(s) can you draw from the result of your hypothesis test?

> *The key business insight is that drivers who use iPhone devices on average have a similar number of drives as those who use Androids.*

> *One potential next step is to explore what other factors influence the variation in the number of drives, and run additonal hypothesis tests to learn more about user behavior. Further, temporary changes in marketing or user interface for the Waze app may provide more data to investigate churn.*