Activity_Course 4 Waze project lab

December 12, 2023

1 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.

2 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.

3 Data exploration and hypothesis testing

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

Consider the questions in your PACE Strategy Document and those below to craft your 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.

==> ENTER YOUR RESPONSE HERE

Complete the following tasks 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:

Main functions: stats.ttest_ind(a, b, equal_var)

Other functions: mean()

Packages: pandas, stats.scipy

[16]:
```
# Import any relevant packages or libraries
### YOUR CODE HERE ###
import pandas as pd                                          #read/manipulate␣
 ↪data
```

```
import numpy as np                                                    #maths,␣
 ↪statistics, multidimensional-array
from matplotlib import pyplot as plt          #normal EDA (non interactive)␣
 ↪(static) (low-level)
from scipy import stats
```

Import the dataset.

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.

[17]:
```
# Load dataset into dataframe
df = pd.read_csv('waze_dataset.csv')
df.head()
```

[17]:

	ID	label	sessions	drives	total_sessions	n_days_after_onboarding	\
0	0	retained	283	226	296.748273	2276	
1	1	retained	133	107	326.896596	1225	
2	2	retained	114	95	135.522926	2651	
3	3	retained	49	40	67.589221	15	
4	4	retained	84	68	168.247020	1562	

	total_navigations_fav1	total_navigations_fav2	driven_km_drives	\
0	208	0	2628.845068	
1	19	64	13715.920550	
2	0	0	3059.148818	
3	322	7	913.591123	
4	166	5	3950.202008	

	duration_minutes_drives	activity_days	driving_days	device
0	1985.775061	28	19	Android
1	3160.472914	13	11	iPhone
2	1610.735904	14	8	Android
3	587.196542	7	3	iPhone
4	1219.555924	27	18	Android

4.2 PACE: Analyze and Construct

Consider the questions in your PACE Strategy Document and those below to craft your response:
1. Data professionals use descriptive statistics for exploratory data analysis (EDA). How can computing descriptive statistics help you learn more about your data in this stage of your analysis?

1. Identifying the central tendency, spread, and outliers of the data.
2. Identifying patterns and relationships in the data.
3. Creating data visualizations to help you better understand the data.

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, `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_new`.

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
2
1

[18]:
```
# 1. Create `map_dictionary`
### YOUR CODE HERE ###
map_dictionary = {'Android': 2, 'iPhone': 1}

# 2. Create new `device_type` column
```

```
### YOUR CODE HERE ###
df['device_type'] = df['device']

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

[18]:

	ID	label	sessions	drives	total_sessions	\
0	0	retained	283	226	296.748273	
1	1	retained	133	107	326.896596	
2	2	retained	114	95	135.522926	
3	3	retained	49	40	67.589221	
4	4	retained	84	68	168.247020	
…	…	…	…	…	…	
14994	14994	retained	60	55	207.875622	
14995	14995	retained	42	35	187.670313	
14996	14996	retained	273	219	422.017241	
14997	14997	churned	149	120	180.524184	
14998	14998	retained	73	58	353.419797	

	n_days_after_onboarding	total_navigations_fav1	\
0	2276	208	
1	1225	19	
2	2651	0	
3	15	322	
4	1562	166	
…	…	…	
14994	140	317	
14995	2505	15	
14996	1873	17	
14997	3150	45	
14998	3383	13	

	total_navigations_fav2	driven_km_drives	duration_minutes_drives	\
0	0	2628.845068	1985.775061	
1	64	13715.920550	3160.472914	
2	0	3059.148818	1610.735904	
3	7	913.591123	587.196542	
4	5	3950.202008	1219.555924	
…	…	…	…	
14994	0	2890.496901	2186.155708	
14995	10	4062.575194	1208.583193	
14996	0	3097.825028	1031.278706	
14997	0	4051.758549	254.187763	
14998	51	6030.498773	3042.436423	

```
       activity_days  driving_days   device  device_type
0                 28            19  Android            2
1                 13            11   iPhone            1
2                 14             8  Android            2
3                  7             3   iPhone            1
4                 27            18  Android            2
…                 …             …       …            …
14994             25            17   iPhone            1
14995             25            20  Android            2
14996             18            17   iPhone            1
14997              6             6   iPhone            1
14998             14            13   iPhone            1

[14999 rows x 14 columns]
```

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.

[19]:
```
df.groupby('device_type')['drives'].mean()
```

[19]:
```
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.

4.2.2 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?

1. Null Hyp: No difference of number of drives between Androud & iPhone.
2. Alte Hyp: There's differences of number of drives between Androud & iPhone.

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

[24]:
```
# 1. Isolate the `drives` column for iPhone users.
### YOUR CODE HERE ###
drives_iPhone = df[df['device_type'] == 1]['drives']      #isolate column

# 2. Isolate the `drives` column for Android users.
### YOUR CODE HERE ###
drives_Android = df[df['device_type'] == 2]['drives']     #isolate column

# 3. Perform the t-test
### YOUR CODE HERE ###
stats.ttest_ind(a=drives_iPhone, b=drives_Android, equal_var=False)
```

[24]:
```
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?

significant level = 0.05, p-value =0.1433, fail to reject null hypothesis, Not statistically significant that there's differences of number of drives between Androud & iPhone.

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

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?

1. Both user groups exhibit similar driving behavior on average.
2. Consistency in user experience across both platforms is recommended.

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