E-news Express Project

Business Statistics

Yair Brama – August 2024

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Hypotheses Tested and Results
- Appendix

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Executive Summary

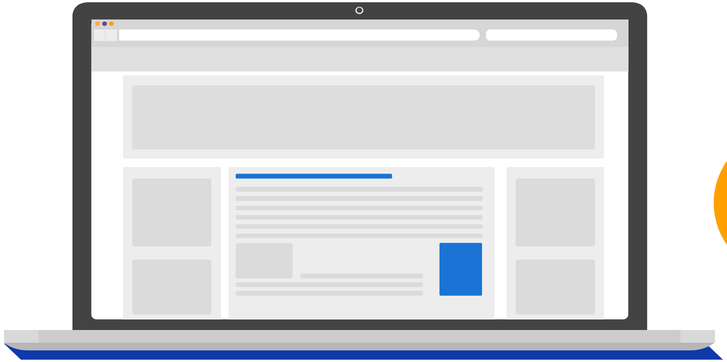
We have performed a statistical analysis (at a significance level of 5%) to determine the effectiveness of the new landing page in gathering new subscribers for the news portal by answering the following questions:

Question	Answer
Do the users spend more time on the new landing page than on the existing landing page?	Yes, 6.2 minutes vs. 4.5 minutes by average in our sample data.
Is the conversion rate (the proportion of users who visit the landing page and get converted) for the new page greater than the conversion rate for the old page?	Yes, ~60% in the new page vs. ~40% in the old page in our sample data
Does the converted status depend on the preferred language?	No, based on the data, we cannot assume that preferred language and conversion rates are dependent parameters
Is the time spent on the new page the same for the different language users?	Yes, based on the data, we cannot prove otherwise

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

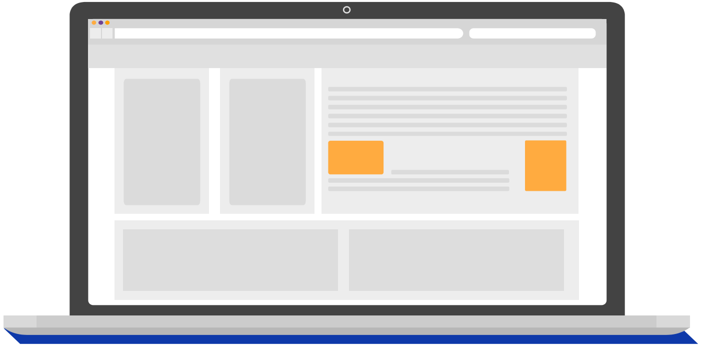
A/B Testing Executive Summary

Variation A – Old landing page



VS

Variation B – New landing page



Statistical test rejected H_0

H_0: Time spent on 2 pages is the same

Result - More time is spent in new page (6.2 minutes vs. 4.5 minutes)

H_0: Conversion rate is the same

Result - Higher conversion rate is found in new page (~60% vs. 40%)

Statistical test failed to reject H_0

H_0: Rate and language are independent

Result - Conversion rate and preferred language are independent

H_0: Time spent is similar for all languages

Result - Time spent on the page is similar by average for all languages

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Business Problem Overview and Solution Approach

BACKGROUND

E-news Express, an online news portal, aims to expand its business by acquiring new subscribers. With every visitor to the website taking certain actions based on their interest, the company plans to analyze these actions to understand user interests and determine how to drive better engagement. The executives at E-news Express are of the opinion that there has been a decline in new monthly subscribers compared to the past year because the current webpage is not designed well enough in terms of the outline & recommended content to keep customers engaged long enough to make a subscription

DECISION

The design team of the company has researched and created a new landing page that has a new outline & more relevant content shown compared to the old page.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

A/B Testing Process

01 Analyze data

100 unique users
2 landing pages
3 languages
Conversion rates

02 Form an hypothesis

1. New landing page is more effective in adding subscribers for the news portal
2. Preferred language plays a role in time spent and conversion rate

03 Experiment

perform a statistical analysis (at a significance level of 5%) to evaluate and compare the 2 landing pages

04 Evaluate results

- Are users spending more time in the new page?
- Are more users convert in the new page?
- Is there any relevance for the preferred language?

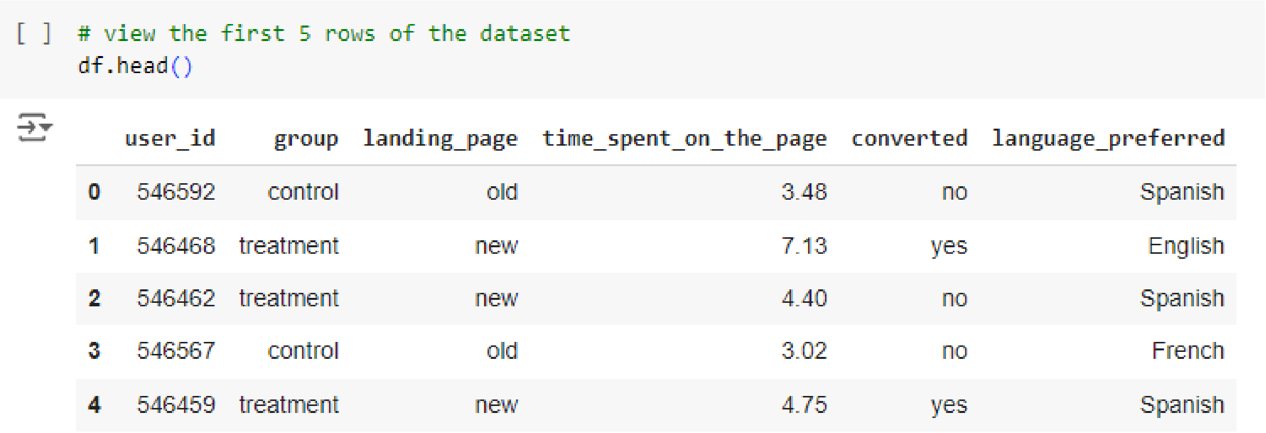
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results - Data Overview

The data includes 100 rows, 50 – 50 between the control group (old landing page) and treatment group (new landing page).

54 users converted to become subscribers (‘yes’ in converted field), and 46 have not (‘no’).

Language preferred – There are 34 Spanish, 34 French and 32 English speakers.

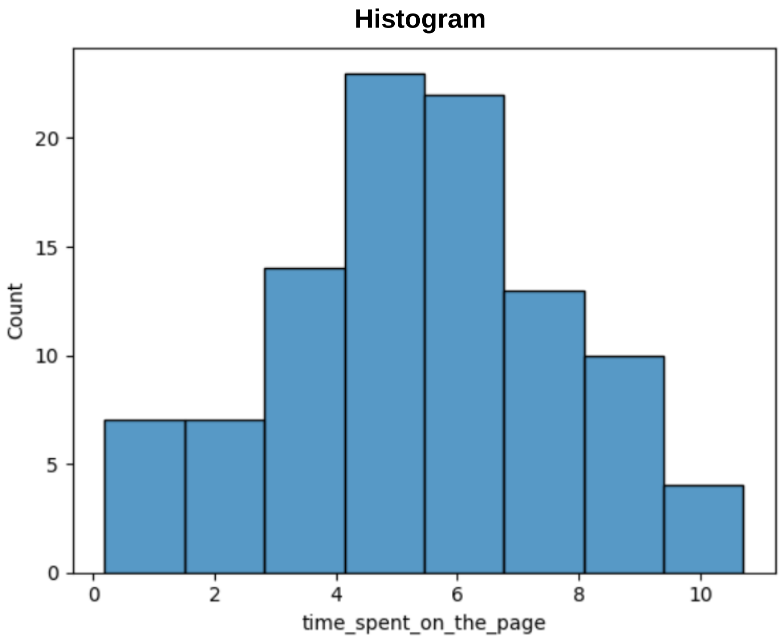


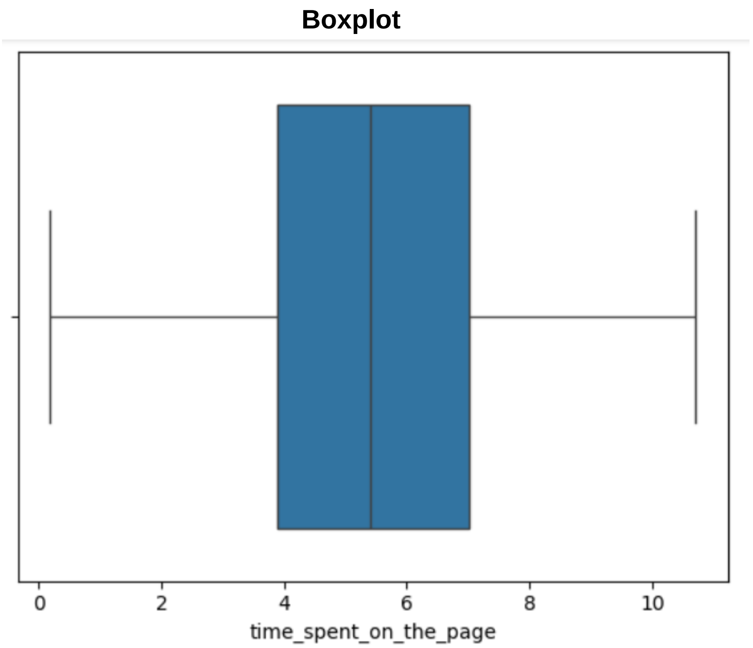
```
[ ] # view the first 5 rows of the dataset
    df.head()
```

	user_id	group	landing_page	time_spent_on_the_page	converted	language_preferred
0	546592	control	old	3.48	no	Spanish
1	546468	treatment	new	7.13	yes	English
2	546462	treatment	new	4.40	no	Spanish
3	546567	control	old	3.02	no	French
4	546459	treatment	new	4.75	yes	Spanish

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results - Univariate Analysis – Time Spent in Page

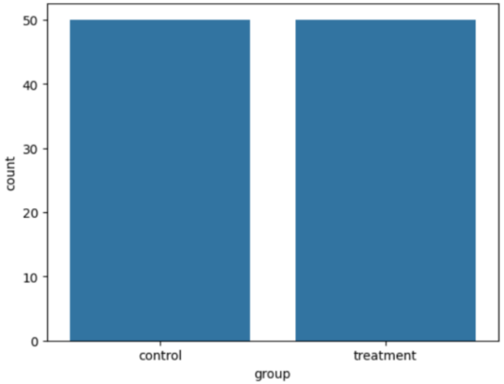




Observation – In this sampling observation, the time spent is distributed normally

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

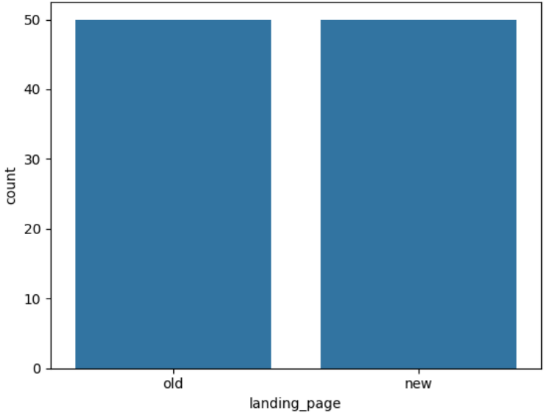
EDA Results - Univariate Analysis – Cont.

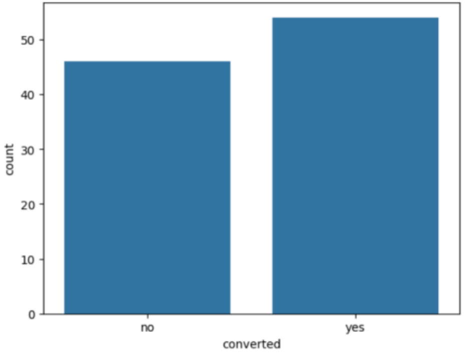


Observations:

Old vs. new landing page – evenly distributed, 50 – 50

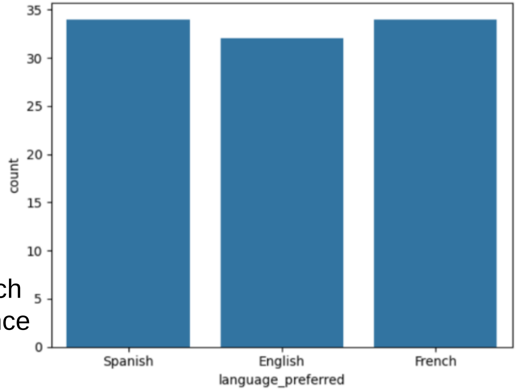
Control vs. treatment groups - evenly distributed, 50 – 50





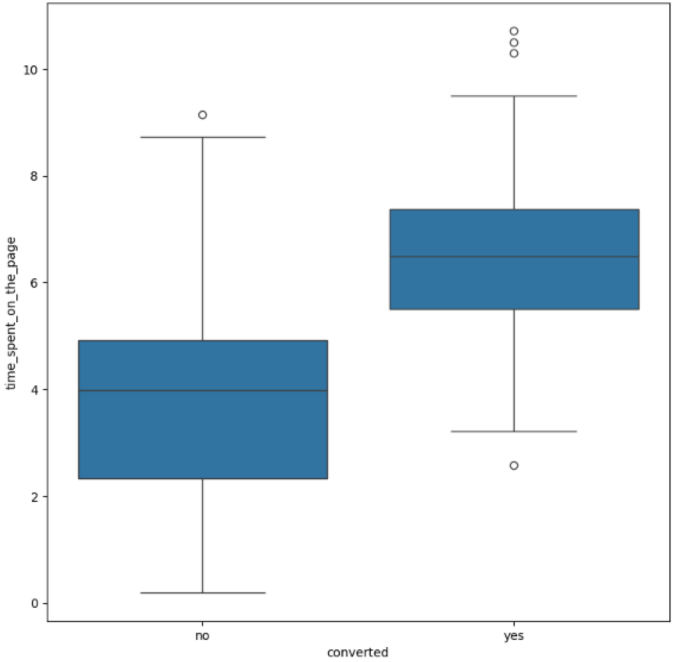
Conversion (rate) – NOT evenly distributed (54 – 46), which allows us to check if one page is more effective than the other in terms of conversion rate

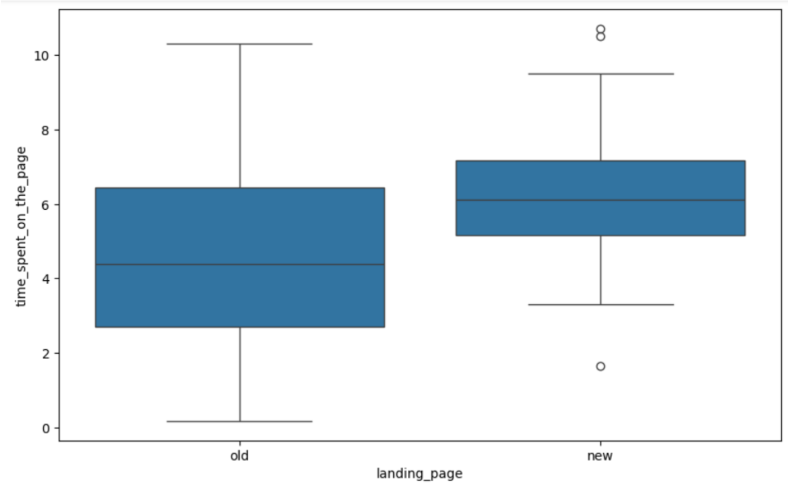
Language preferred – not evenly distributed, (34 French and Spanish, 32 English). We will check if this difference means anything to affect the conversion rate



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Multivariate Analysis – Time spent/conversion rate on each Landing Page

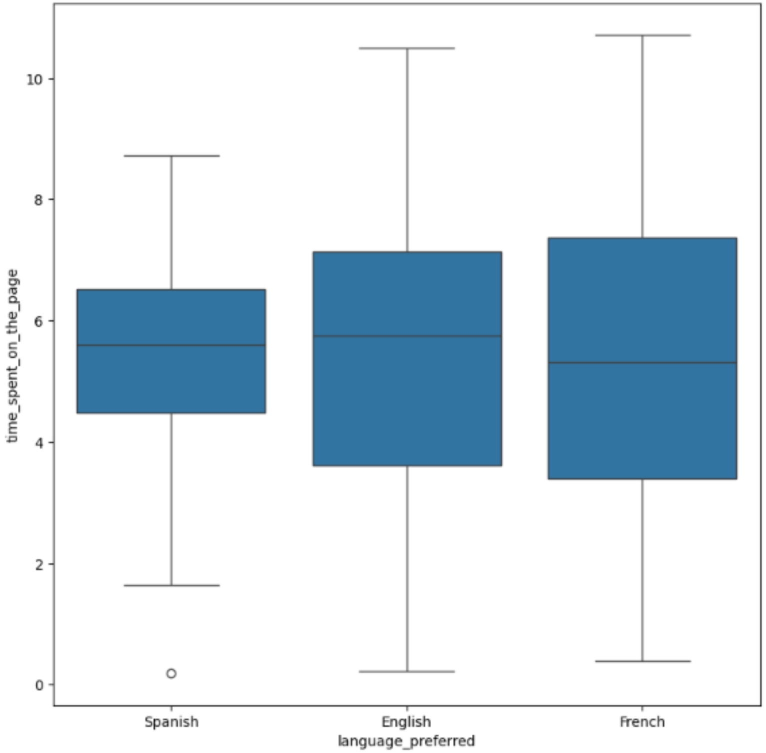




Observation – We can see a difference in the average time spent on each page (~2 minutes) and a very similar pattern regarding the conversion rate. It’s clear that people spend more time in the new landing page, and they spend more time when they convert to be subscribers

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Multivariate Analysis – Cont.



Observation – By average, all preferred languages spend the same amount of time on the website

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 1 - Do the users spend more time on the new landing page than the existing landing page?

01 Visual Analysis

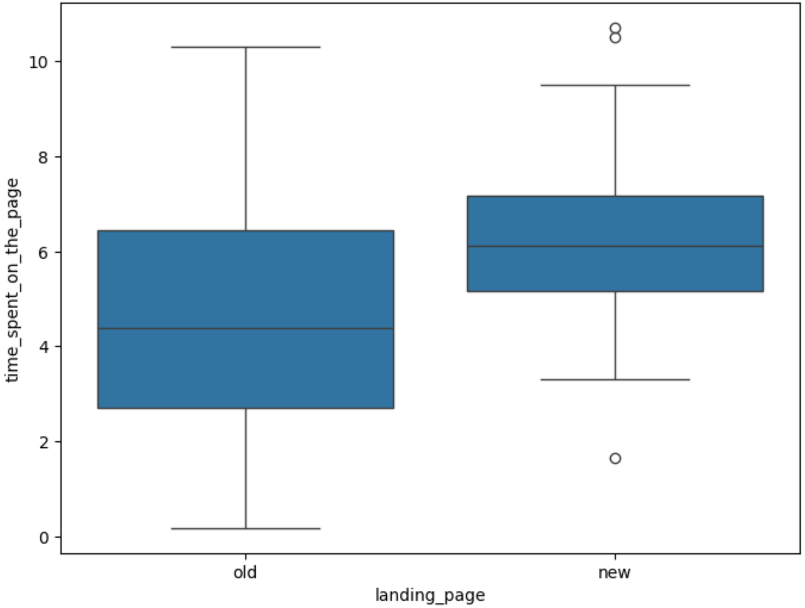
Shows longer time spent on the new page

03 Define appropriate test

Two independent sample t-test

05 Collect Data

The means and std of 2 samples, the time spent in each version of the page



02 Null and alt hypotheses

H0: μ1=μ2

Ha: μ1<μ2

04 Signifcance Level

α = 0.05

06 Calculate the p-value

The p-value is 0.00013923812251665490

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 1 Results and Inference

- As the p-value 0.0001392381225166549 is less than the level of significance, we reject the null hypothesis that both sites have the same time spent in average.

- The statistical inference allows us to conclude that the new landing page is more effective in keeping users for longer time than the old landing page, and the answer to the question is **yes, users spend more time on the new landing page than the existing landing page.**

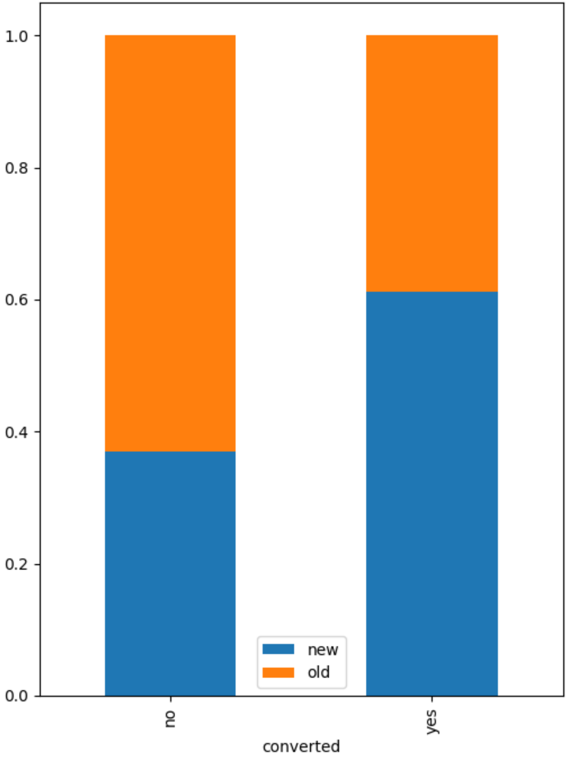
Link to Appendix slide on details of the test performed

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 2 - Is the conversion rate for the new page greater than the conversion rate for the old page?

01
Visual Analysis

Shows higher rate of conversion in the new page



02
Null and alt hypotheses

H0:p1=p2

Ha: (old) p1 < p2 (new)

03
Define appropriate test

2 proportions Z-test (proportions_ztest)

04
Signifcance Level

α = 0.05

05
Collect Data

The numbers of users served the new and old pages are 50 and 50 respectively

06
Calculate the p-value

The p-value is 0.008026308204056278

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 2 Results and Inference

- As the p-value 0.008026308204056278 is less than the level of significance, we reject the null hypothesis that both sites have the same proportion between users who converted and users who did NOT convert.

- The statistical inference allows us to conclude that the new landing page is more effective in converting users to subscribers than the old landing page and the answer to the question is **yes, the conversion rate for the new page is greater than the conversion rate for the old page**

Link to Appendix slide on details of the test performed

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 3 - Does the converted status depend on the preferred language?

01 Visual Analysis

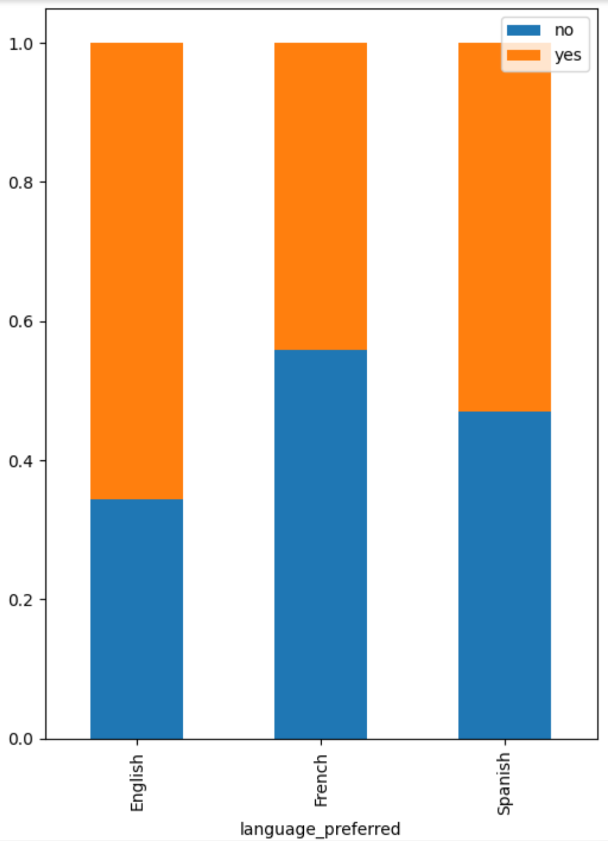
Shows some different proportions between languages

03 Define appropriate test

Chi Square Test for Independence(chi2_contingency)

05 Collect Data

Contingency table of how many converted in each language



02 Null and alt hypotheses

Ho:Rate is independent from language
Ha: Rate is not independent from language

04 Signifcance Level

α = 0.05

06 Calculate the p-value

The p-value is 0.2129888748754345

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 3 Results and Inference

- As the p-value 0.2129888748754345 is greater than the level of significance, we fail to reject the null hypothesis that language and conversion are independent.

- The statistical inference allows us to conclude that the preferred language and the conversion rate are independent from each other. The answer to the question is **no, we cannot assume that the converted status depends on the preferred language**

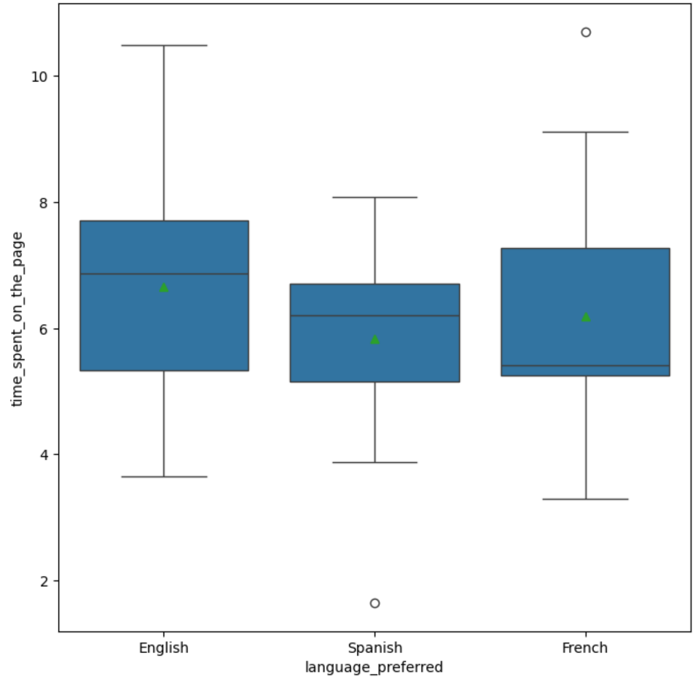
Link to Appendix slide on details of the test performed

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 4 - Is the time spent on the new page same for the different language users?

Visual Analysis

There is a slight difference between languages, is it significat?



Define appropriate tests

One-way Anova:

a. Shapiro-Wilk’s test

b. Levene’s test

c. ANOVA F-test (f_oneway)

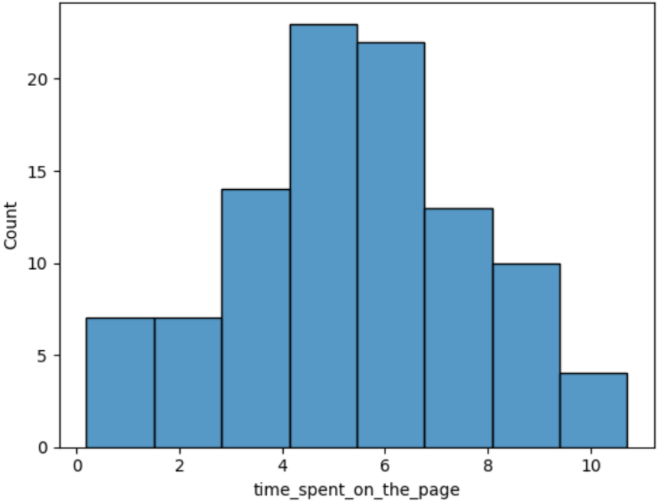
Tests planning

First, we will test that the data of ‘time spent’ is normally distributed (Shapiro-Wilk’s) and the variances are equal (Levene’s).

Then we will look at the Independent samples of time spent for each language and examine the hypothesis (F-test)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 4a, 4b - Verify Normal Distribution and equal variances



Null and alt hypotheses – Shapiro Wilk's

H0: The time spent is normally distributed

Ha: The time spent is NOT normally distributed

p-value – Shapiro Wilk's

The p-value is 0.5642956935237358 is greater than 0.05 – Failed to reject H0

Null and alt hypotheses – Levene's

H0: All the population variances are equal

Ha: At least one variance is different from the rest

p-value – Levene's

The p-value is 0.46711357711340173 is greater than 0.05 – Failed to reject H0

Signifcance Level

α = 0.05

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 4c - Is the time spent on the new page same for the different language users?

Define appropriate test

ANOVA F-test (f_oneway)

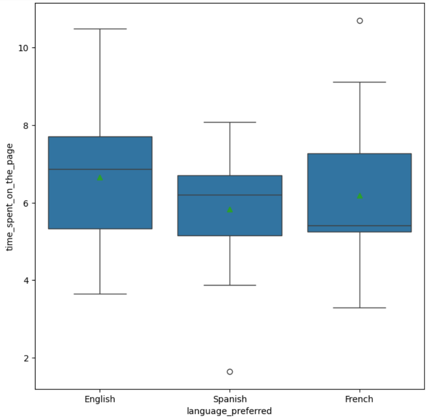
Null and alt hypotheses

H0: µ1=µ2=µ3

Ha: At least one preferred language is different from the rest.

Collect Data

Independent samples of time spent for each language



Signifcance Level

α = 0.05

Calculate the p-value

The p-value is 0.43204138694325955 is greater than 0.05 – Failed to reject Ho

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Test 4 Results and Inference

- We performed the pre-requisite tests to verify that the population is normally distributed, and the independent samples have the same variation (Shapiro-Wilk's test and Levene's test)

- In the F-test, as the p-value 0.43204138694325955 is greater than the level of significance, we fail to reject the null hypothesis that in a specific language, more time is spent on the website.

- The statistical inference allows us to conclude that **yes, the time spent on the new page is the same for the different language users**

Link to Appendix slide on details of the test performed

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

APPENDIX

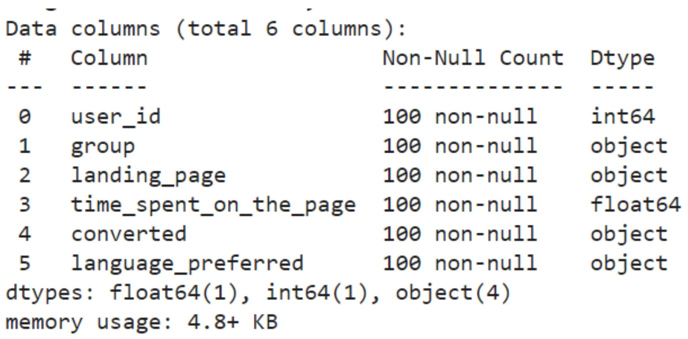
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Background and Contents

- How many rows and columns are present in the data?

Answer:: 100 rows, 6 columns

Datatypes of the different columns in the dataset

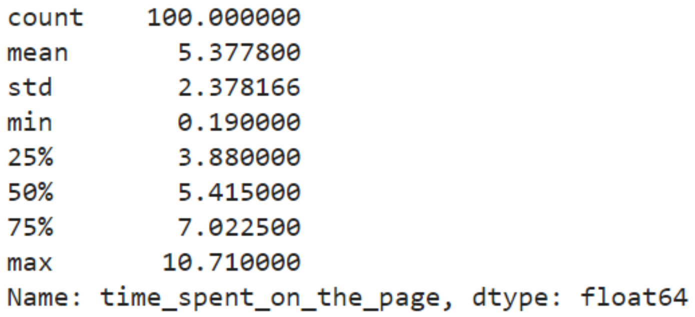


```
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   user_id                 100 non-null    int64
 1   group                   100 non-null    object
 2   landing_page            100 non-null    object
 3   time_spent_on_the_page  100 non-null    float64
 4   converted               100 non-null    object
 5   language_preferred      100 non-null    object
dtypes: float64(1), int64(1), object(4)
memory usage: 4.8+ KB
```

- Are there any missing values in the data? Are there any duplicates?

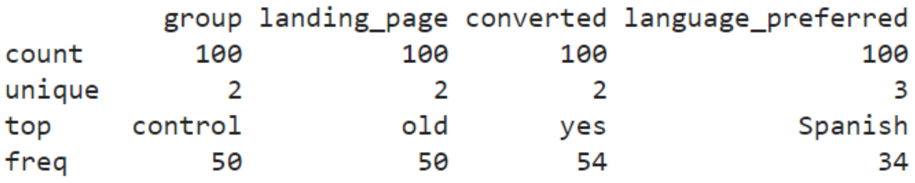
Answer: No. All rows/columns are filled, with no duplicates

Analysis of the data in the time_spent_on_the_page column:



```
count    100.000000
mean       5.377800
std        2.378166
min        0.190000
25%        3.880000
50%        5.415000
75%        7.022500
max       10.710000
Name: time_spent_on_the_page, dtype: float64
```

Analysis of the categorical columns:



```
          group landing_page converted language_preferred
count       100          100       100                100
unique        2            2         2                  3
top     control          old       yes            Spanish
freq         50           50        54                 34
```

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing Details – Test 1

- Null and alternate hypotheses:

μ1,μ2 are the old landing page mean time spent on the page and the new landing page time spent, respectively.

H0: μ1=μ2
Ha: μ1<μ2

- Hypothesis Test selected:

2-sample ind. t-test where:
The mean time spent for the old landing page is 4.532400000000001
The mean time spent for the new landing page is 6.2232
The standard deviation of time spent for the old landing page is 2.58
The standard deviation of time spent for the new landing page is 1.82

- p-value obtained:

```
test_stat, p_value = ttest_ind(df['time_spent_on_the_page'][df['landing_page']=='new'],
df['time_spent_on_the_page'][df['landing_page']=="old"], equal_var = False, alternative =
'greater’)
The p-value is 0.0001392381225166549
```

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing Details – Test 2

- Null and alternate hypotheses:

p_1,p_2 are the proportions of converted pages in the old landing page and the new landing page, respectively.

H_0:p1=p2

Ha:p1<p2

- Hypothesis Test selected:

2 proportions Z-test (proportions_ztest) where:

new_converted = number of converted users in the new landing page

old_converted = number of converted users in the old landing page

n_control = Size of control group = 50

n_treatment = Size of treatment group = 50

- p-value obtained:

```
test_stat, p_value = proportions_ztest([new_converted, old_converted] , [n_treatment,
n_control], alternative ='larger')
The p-value is 0.008026308204056278
```

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing Details – Test 3

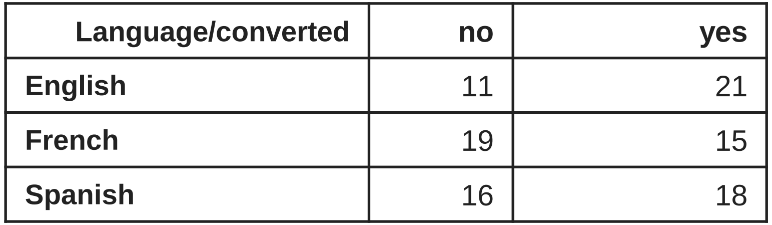
- Null and alternate hypotheses:

H_0: Conversion rate is independent of the language

Ha: Conversion rate depends on the language

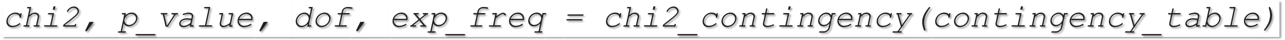
- Hypothesis Test selected:

Chi Square Test for Independence(chi2_contingency),
Where contingency_table is :

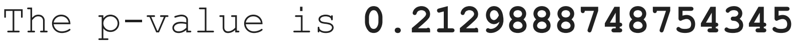


Language/converted	no	yes
English	11	21
French	19	15
Spanish	16	18

- p-value obtained:



```
chi2, p_value, dof, exp_freq = chi2_contingency(contingency_table)
```



```
The p-value is 0.2129888748754345
```

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing Details – Test 4a - Shapiro-Wilk's Test

- Null and alternate hypotheses:

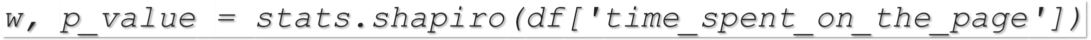
H_0: The time spent on the page distribution follows a normal distribution

Ha: The time spent on the page distribution does not not follow a normal distribution

- Hypothesis Test selected:

Shapiro-Wilk's test

- p-value obtained:



```
w, p_value = stats.shapiro(df['time_spent_on_the_page'])
```

```
The p-value is 0.5642956935237358
```

Which is greater than the level of significance, so we fail to reject the null hypothesis, and we can continue with the assumption that the population is normally distributed

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing Details – Test 4b – Levene’s Test

- Null and alternate hypotheses:

H_0: All the population variances are equal

Ha: At least one variance is different from the rest

- Hypothesis Test selected:

Levene’s test

Where:

time_spent_English = df_new[df_new['language_preferred']=="English"]['time_spent_on_the_page']

time_spent_French = df_new[df_new['language_preferred']=="French"]['time_spent_on_the_page']

time_spent_Spanish = df_new[df_new['language_preferred']=="Spanish"]['time_spent_on_the_page’]

- p-value obtained:

```
statistic, p_value = stats.levene(time_spent_English, time_spent_French,
time_spent_Spanish)
```

```
The p-value is 0.46711357711340173
```

Which is greater than the level of significance, so we fail to reject the null hypothesis, and we can continue with the assumption that the all the population variances are equal

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Hypothesis Testing Details – Test 4c – One-Way F-Test

- Null and alternate hypotheses:

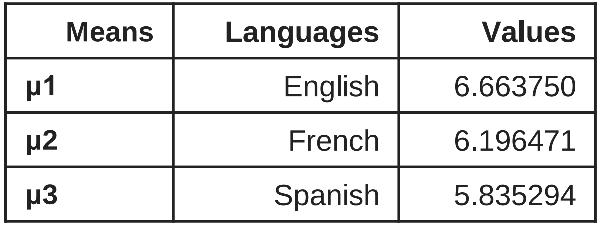
H_0: μ_1,μ_2,μ_3 are the means of time spent on the page for the preferred languages.

Ha: At least one preferred language is different from the rest.

- Hypothesis Test selected:

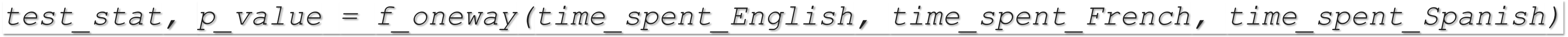
One-way ANOVA F-test (f_oneway)

Where μ_1,μ_2,μ_3 are :

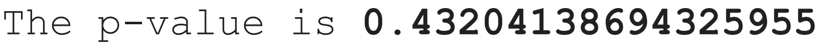


Means	Languages	Values
μ1	English	6.663750
μ2	French	6.196471
μ3	Spanish	5.835294

- p-value obtained:



```
test_stat, p_value = f_oneway(time_spent_English, time_spent_French, time_spent_Spanish)
```



```
The p-value is 0.43204138694325955
```

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Great Learning

Happy Learning !