E-news Express Project

By Minh Ngo
July 2021

Table of contents

Content
Project Background & Objective
Data analysis - EDA
Univariate analysis
Multivariate analysis
Hypothesis Testing
Key insights & Conclusion

Project Objective

Objective

- Determined whether the new landing page is more effective in gathering new subscribers via data analysis and collecting actionable insights from the A/B testing

- We will focus on key 2 main tasks:

  - Extract insights from the data

  - Answer 4 important question to assess the effectiveness of the new landing page by conducting hypothesis testing
    - Do the users spend more time on the new landing page than the old landing page?
    - 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?
    - Does the converted status depend on the preferred language?
    - Is the mean time spent on the new page same for the different language users?

Approach

- Approach
  - EDA is used to explore insights from the data set
    - Univariate analysis were conducted for each variable and then multivariate analysis was used to explore relationship among each pair of variables
  - Conduct various hypothesis tests for each prementioned question
  - Based on insights, provide recommendations for the management

greatlearning
Power Ahead

Data Analysis - EDA

Data information

- The dataset contains various data about the customers in 2 groups: control (seeing old page) and treatment (seeing new page) group

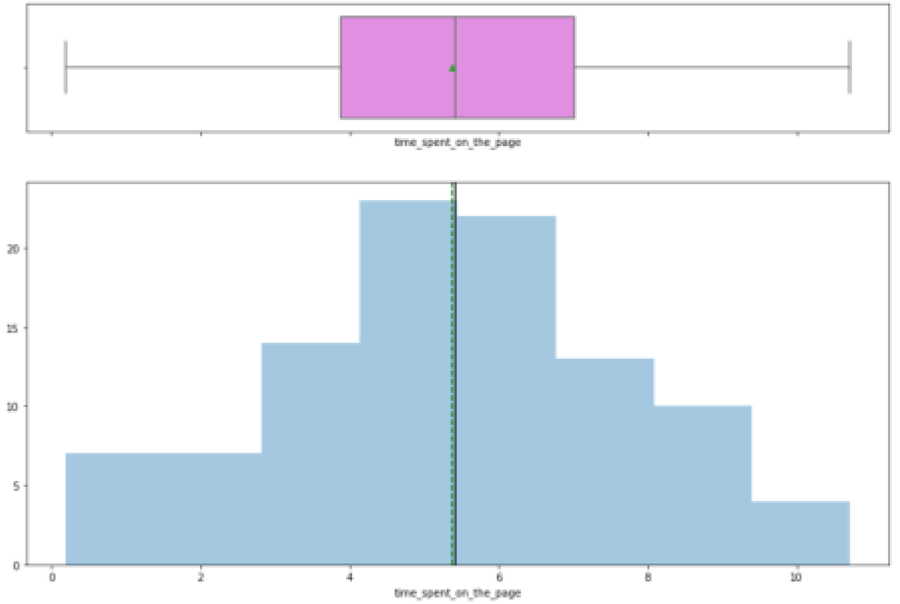
#	Variable	Description
1	user_id	Id of user visiting the website
2	group	Represent whether the user belongs to control or treatment group
3	landing_page	Represents whether the landing page is new or old
4	time_spent_on_the_page	The amount of time in mins spent by users on landing page
5	converted	Represent whether user gets converted into subscriber in the news portal
6	language_preferred	Average time the customer **wants** to use treadmill every week

Observation	Variable
100	5

Note:
- There is no missing value from the dataset
- There is only 1 integer variable (time_spent_on_the_page), there rest are object and were converted into categorical variables

Exploratory Data Analysis (EDA) – time spent on the page

time_spent_on_the_page

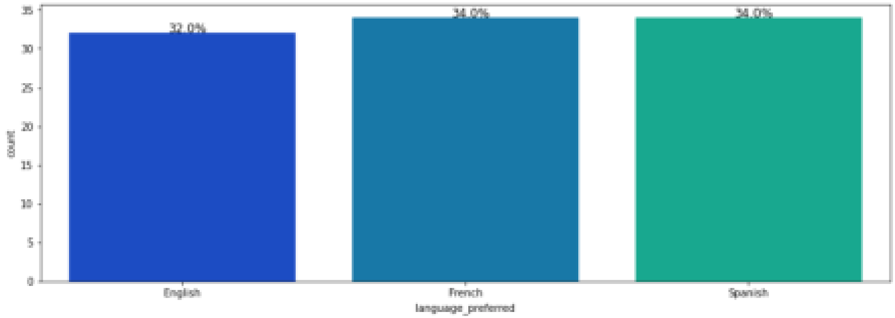


- The distribution of time spent on the page is very close to normal distribution (looking at the shape)
- There is no outlier in this variable (the only numerical variable in the dataset)
- From the boxplot we can see the mean of time is ~5.4 minutes

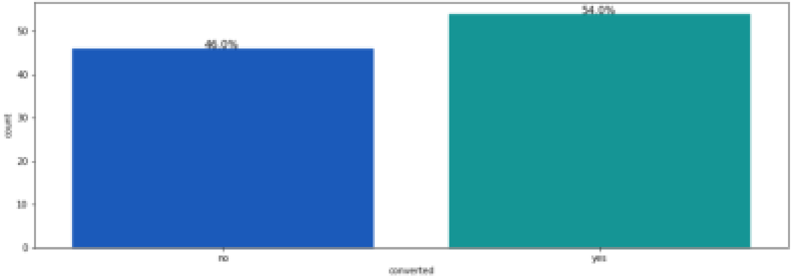
Univariate Analysis

EDA – language preferred & converted status

Language preferred



Converted status



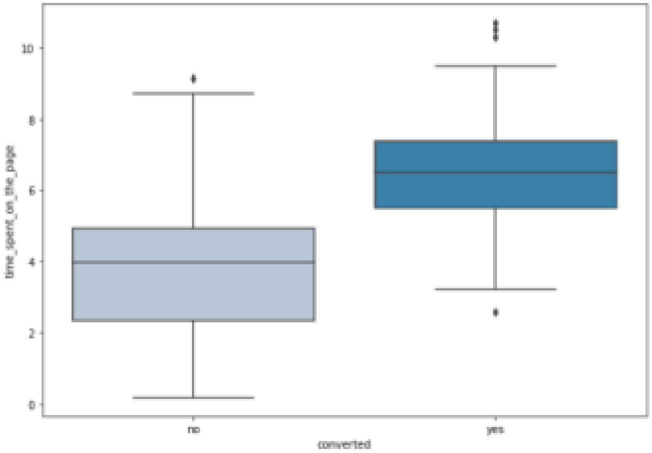
- Spanish and French are used more (by number of users) than English and have same proportion (34%)
- English users are slightly lower by a small margin (2%).

- Of the 100 users, the conversion rate is 54%, in other words, of 100 users there are 54 users who decided to subscribe

** The other variables (ID, landing page and group) are not analyzed in this project because: ID is randomly assigned and do not have any meaning behind. We know that landing page and group are equally divided into 2 groups, according to the design of this test, hence no need to further analyze*

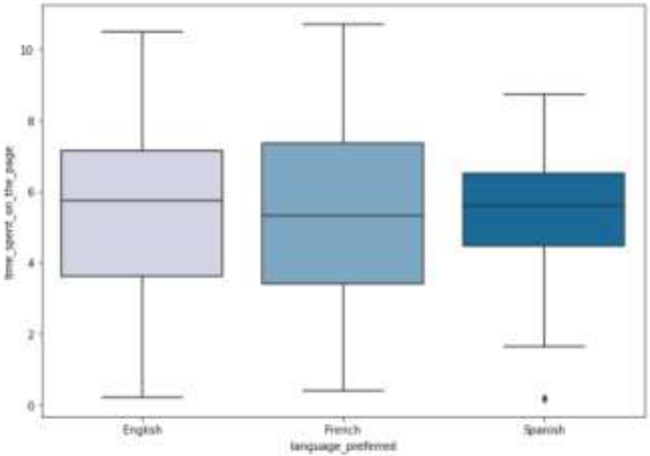
EDA – Time spent with converted status & language preferred

Time spent vs converted status



- For group of converted users, on average they spend more time on the page than non-converted group.
- The average time on the page of converted users is ~6.3mins, while average time on non-converted user is ~4mins
- Converted users also have some outliers in which users spend more than 10 mins on the site.

Time spent vs language preferred

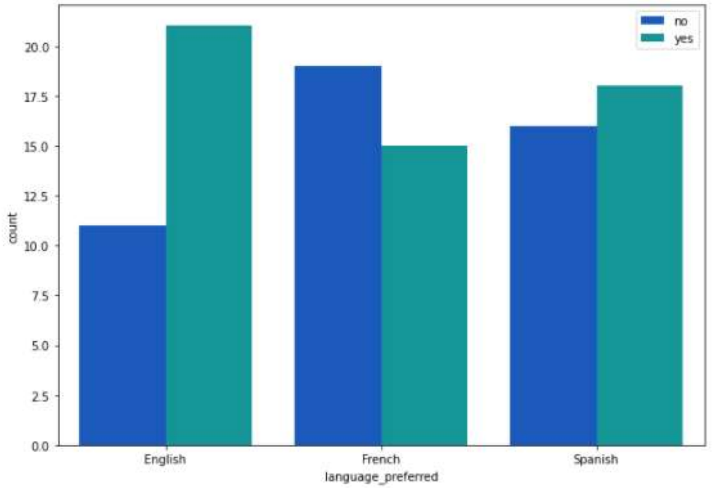


- Generally, on average, French users spend slightly less time (5.8 mins) than English and Spanish users.
- The range of time spent by Spanish user is considerably smaller than that of English and French.

Multivariate Analysis

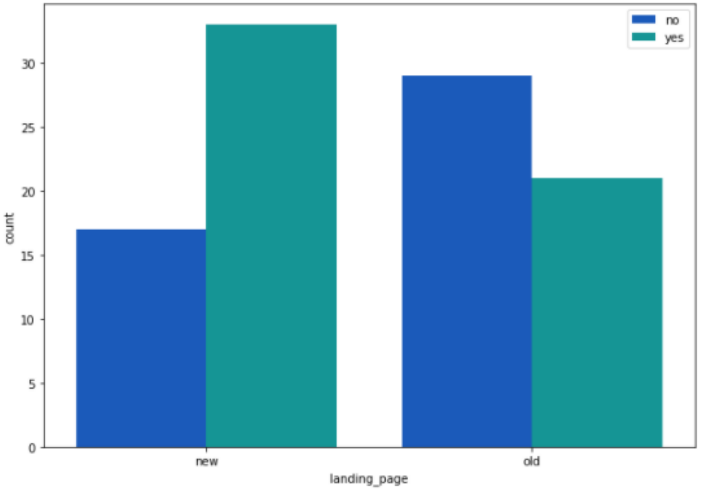
EDA –converted status with language preferred and landing page

Converted status vs Language preferred



- For group of converted users, on average they spend more time on the page than non-converted group.
- The average time on the page of converted users is ~6.3mins, while average time on non-converted user is ~4mins
- Converted users also have some outliers in which users spend more than 10 mins on the site.

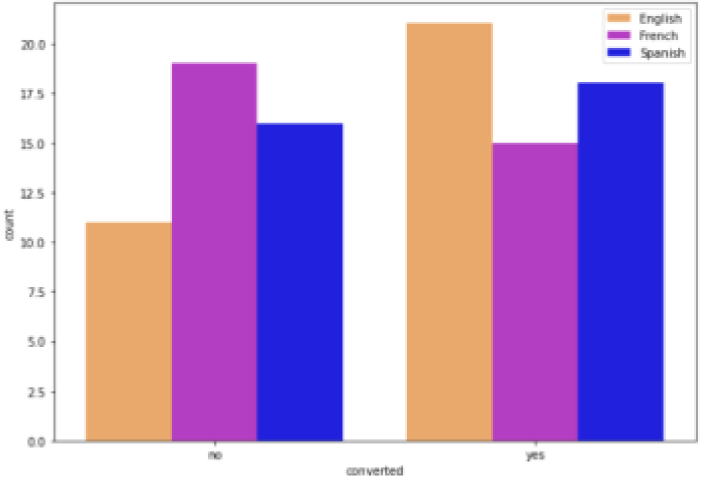
Converted status vs landing page



- New landing page has much higher conversion ratio than old landing
- Within the new landing page group, there are almost twice the number of converted user vs non-converted
- While within old landing page group, the number of converted users is smaller than non-converted

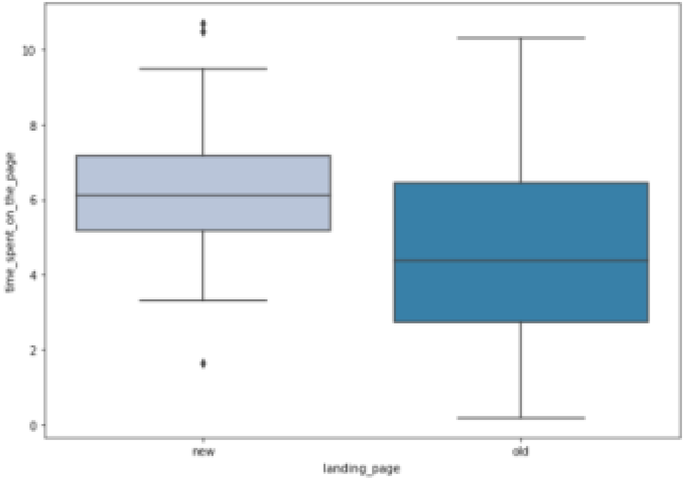
EDA –converted status with language preferred and landing page

Converted status vs Language preferred



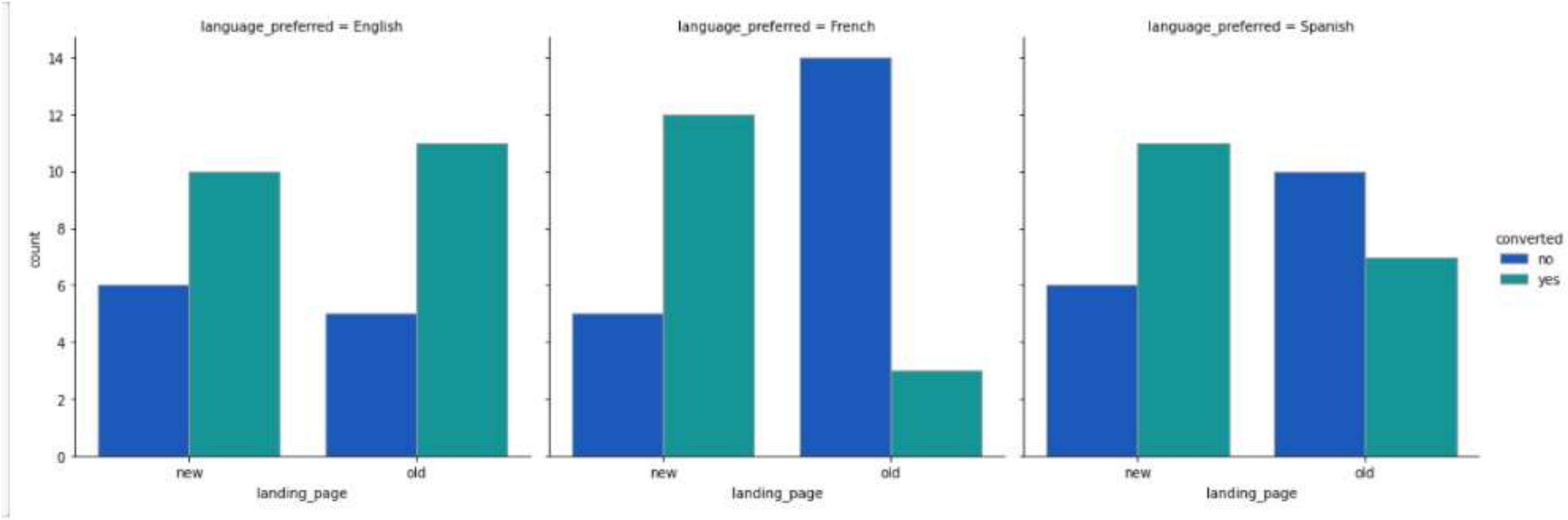
- Of the converted group, English is used the most, next is Spanish and French.
- Of the non-converted group, French is used the most, next are Spanish and English. This means that English seems to be a best language option to convert users, while French should not be used.

Landing page vs time spent on the page



- Users who use new landing page spend generally more time, ranging from 3.7-9.8 mins with some outliers where users spend more than 10 mins
- The time spent of old landing page has larger variance than new landing page

EDA: landing page and language and converted status



Observations from the sample:

- For language= French : new landing page has much higher conversion ratio than old landing page
- For language= English: new landing page has slightly lower conversion ratio (subscriber / total user) than old landing page
- For language= Spanish: new landing page help converts users much better than old landing page

HYPOTHESIS TESTING

To assess the effectiveness of the new landing page, we follow this process

Objective

Assess based on various metrics

Hypothesis tests

Conclusion

- Assess the effectiveness of the new landing page

4 key questions to answer:

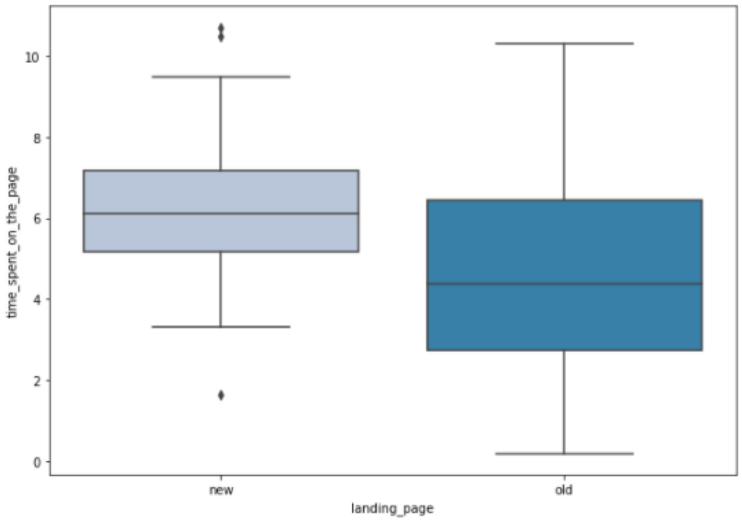
- Do the users spend more time on the new landing page than the old landing page?
- Is the conversion rate (the proportion of users who visit the landing page and get converted) for the new page greater than the conversion rate for the old page?
- Does the converted status depend on the preferred language?
- Is the mean time spent on the new page same for the different language users?

- We conduct different tests for different questions (and problem)
- There are 5 main steps in the hypothesis test
- Conclude based on the statistical evidence from the test

- Final conclusion about the effectiveness of the new page
- Gain additional insights

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

Landing page vs time spent on the page



Hypothesis

The hypothesis we want to test here is **customers spend more time on new landing page than old one**
Let $\mu1,\mu2$ be the mean of time_spent_on_the_page of control group and treatment groups respectively.
We will test the null hypothesis
Ho: $\mu1=\mu2$
against the alternative hypothesis:
Ha: $\mu1<\mu2$

Select Appropriate Test

For the problem, we have to compare the sample means from 2 independent populations (subscribers vs non-subscriber) when standard deviation are unknown.
Therefore , we will use **2-sample independent t-test** here
The significance level is 5%

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

Assumptions

We test to see if these assumptions are met:
•Continuous data - Yes, the usage time is measured on a continuous scale.
•Normally distributed populations - Yes, we can assume the population is normal and sample size n>30.
•Independent populations - As we are taking random samples for two different type of users, the two samples are from two independent populations.
•Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.
•Population standard deviation is known - No
•Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.

-> All assumptions are met so we can use this test

Test

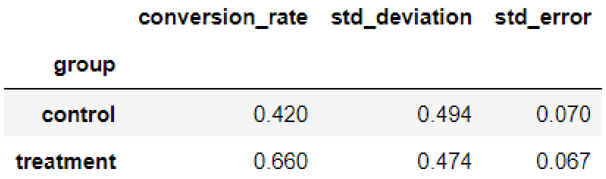
We prepared data and conducted the test.
After testing, we can find the p-value =0.000131 which is lower than 0.05 (significance level)

Conclusion

As the p-value (~0.0001) is less than the level of significance (0.05), we can reject the null hypothesis.
Thus, it could be concluded that there is **enough statistical evidence to conclude that the time spent on the new landing page (treatment group) is higher than old landing page (control group)**

2-Is the conversion rate of the new landing page higher than old landing page?

Conversion rate by group



	conversion_rate	std_deviation	std_error
group			
control	0.420	0.494	0.070
treatment	0.660	0.474	0.067

Hypothesis

The hypothesis we want to test here is **conversion rate of new landing page is higher than old landing page**

Let p1,p2 be the conversion ratio (number of converted user / total users) in control group and treatment groups respectively.
We will test the null hypothesis
Ho: $p1=p2$
against the alternative hypothesis:
Ha: $p1<p2$

Select Appropriate Test

For the problem, we need to compare the proportion (conversion rate) of two populations (subscribers vs non-subscriber) and conclude if conversion rate of new landing page is **higher** than old landing page, based on data of 2 sample groups.
Therefore we will use the **2 sample 1 tailed Z-test (proportion)**
The significance level is 5%

2-Is the conversion rate of the new landing page higher than old landing page?

Assumptions

We test to see if these assumptions are met:
•Binomally distributed population - Yes, a conversion option is either CONVERTED or NOT CONVERTED
•Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.
•Can the binomial distribution approximated to normal distribution - Yes.

-> All assumptions are met so we can use this test

Test

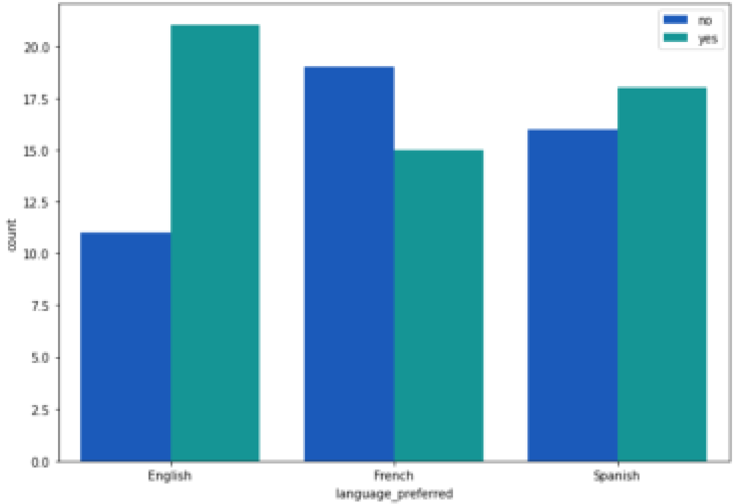
We prepared data and conducted the test.
After testing, we can find the p-value =0.00802 which is lower than 0.05 (significance level)

Conclusion

As the p-value (~0.008) is less than the level of significance (0.05), we can reject the null hypothesis.
Thus, it could be concluded that we **have enough statistical evidence to support that conversion rate of new landing page is higher than old landing page.**

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

Conversion rate by language users



Hypothesis

The hypothesis we want to test here is **conversion rate depends on the language preferred in the site**

Let p1,p2 be the conversion ratio (number of converted user / total users) in control group and treatment groups respectively.
We will test the null hypothesis
Ho: converted status does not depend on preferred language
against the alternative hypothesis:
Ha: converted status depends on preferred language

Select Appropriate Test

For the problem, basically we want to check whether categorical variables (converted status (yes/no)) from a population (language user) are independent.
Therefore we can use **Chi-Square Test for independence**
The significance level is 5%

3-Is the conversion rate of the new landing page higher than old landing page?

Assumptions

We test to see if these assumptions are met:
•Categorical variable: yes (yes/no answer is categorical)
•expected value of the number of sample observations in each level of the variable is at least 5: yes
•Random sampling from the population: yes, we know that the sample is randomly collected Hence we can use Chi-Square Test for independence

-> All assumptions are met so we can use this test

Test

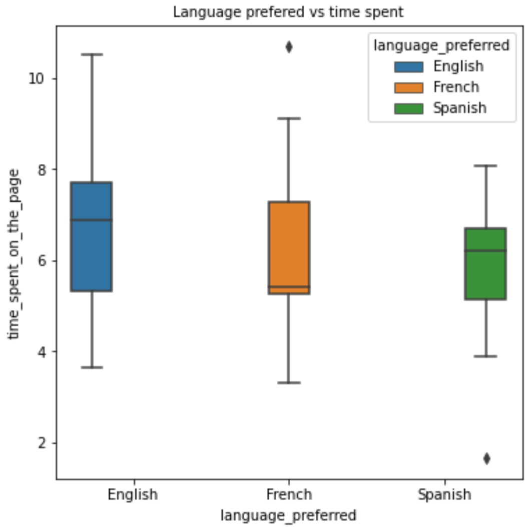
We prepared data and conducted the test.
After testing, we can find the p-value =0.21 which is higher than 0.05 (significance level)

Conclusion

As the p-value (~0.21) is higher than the level of significance (0.05), we **fail to reject the null hypothesis**.
Thus, it could be concluded that we **do not have enough statistical evidence to support that conversion status depends on the language preferred.**

4. Is the mean of time spent on the new page the same for different language user?

Language preferred vs time spent



Hypothesis

The hypothesis we want to test here is **conversion rate depends on the language preferred in the site**

Let $\mu1$,$\mu2$, $\mu3$ be the mean of time spent on the NEW page of group English, Spanish and French users

We will test the null hypothesis

Ho: $\mu1$=$\mu2$= $\mu3$

against the alternative hypothesis:

Ha: at least mean of the time spent on new page of one language group user is different from the rest (All μ are not equal)

Select Appropriate Test

The problem concerning 3 means of 3 populations : English, Spanish and French user and we want to see if one of them is different from other mean.

For this we will use the One-way (One factor) ANOVA test.

The significance level is 5%

4. Is the mean of time spent on the new page the same for different language user?

Assumptions

We test to see if these assumptions are met:

- Residuals (experimental error) are approximately normally distributed (Shapiro-Wilks Test)
- Homogeneity of variances (variances are equal between treatment groups) (Levene's or Bartlett's Test)
- Observations are sampled independently from each other (no relation in observations between the groups and within the groups): yes

We have tested these 2 assumptions using Shapiro-Wilks Levene's test (see the notebook) and these 2 assumptions are met.
For 3rd assumption we know that observations are sampled independently.

-> All assumptions are met so we can use this test

Test

We prepared data and conducted the test.
After testing, we can find the p-value =0.4 which is higher than 0.05 (significance level)

Conclusion

As the p-value (~0.21) is higher than the level of significance (0.05), we **fail to reject the null hypothesis**.
Thus, it could be concluded that we **do not have enough statistical evidence to support that mean of time spent is different for different language users.**

Key insights & Conclusion

Key insights from EDA

After EDA analysis, we have drawn the following insights:

- Spanish and French are used by slightly more users than English
- The conversion rate of the whole sample (100 users) is 56%
- On average converted users spend more time (2.3 minutes more) on the page than non-converted (non-subscriber). This implies that once users find the portal interesting, they want to linger around more and this increase the likelihood of conversion
- Generally average users who use French spend less time on the page than English and Spanish
- In terms of language:
    - English user has higher number of converted users than other languages. French has lowest number of converted users and highest number of non-converted users.
    - Of the converted group, English is used the most, next is Spanish and French. Of the non-converted group, French is used the most, next are Spanish and English. This means that English seems to be a best language option to convert users, while French should not be used.
- Landing page vs language: new landing page all perform better in French and Spanish. Only English has lower conversion ratio on new landing page than old
- New landing page vs old landing page:
    - New landing page has much higher conversion ratio than old landing page
    - People spend more time on new landing page than the old one

Key insights from Hypothesis testing

After conducting hypothesis testing on 4 questions, we have the following inference from different statistical test:

	Question	Inference
1	Do the users spend more time on the new landing page than the old landing page?	We have enough statistical evidence to conclude that the time spent on new page is higher than old landing page
2	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?	We have enough statistical evidence to conclude that conversion rate of new landing page is higher than old landing page
3	Does the converted status depend on the preferred language?	We do not have enough statistical evidence to conclude that converted status depend on preferred language
4	Is the mean time spent on the new page same for the different language users?	We do not have enough statistical evidence to conclude that mean of time spent on new page same for different language users

What does this all mean?

- **Conclusion**: New landing page is **more effective t**han the old one in the following ways:
    - in keeping users in the portal for longer period of time.
    - Users are more likely to subscribe using new landing page

  Therefore, it is recommended to use the new landing page instead of old one.

- **Other areas to examine further:**
    - Why does new landing page in English have lower conversion rate than old landing page?
    - If language is not the determining factor for conversion (in other words, if conversion rate doesn't depend on the language), what would it be?
    - Why do French users spend less time than other users? Is this because of the poorer content in French (i.e. less relevant news, poorer writing quality etc.)
- **Potential next steps:**
    - Conduct a focus group or survey on specific language users
    - Conduct other A/B test to answer above questions