Project Business Statistics: E-news Express

Define Problem Statement and Objectives

The company's design team has researched and created a new landing page with a new structure and more relevant content than the previous page. To evaluate the efficacy of the new landing page in acquiring new subscribers, the Data Science team conducted an experiment in which 100 users were randomly divided into two equal groups. The existing landing page was served to the control group, while the new landing page was served to the treatment group. It was determined how users in both categories interacted with the two versions of the landing page. As a data scientist at E-news Express, you have been tasked with exploring the data and conducting a statistical analysis (at a significance level of 5%) to determine the efficacy of the new landing page in acquiring new news portal subscribers by answering the following questions.

Do consumers spend more time on the new landing page than on the previous one?

Is the new page's conversion rate (the proportion of users who visit the landing page and convert) higher than the previous page's conversion rate?

Does the status of conversion depend on the chosen language?

Is the time spent on the new page equivalent for consumers of various languages?

Data Dictionary

The data contains information regarding the interaction of users in both groups with the two versions of the landing page.

user_id - Unique user ID of the person visiting the website

group - Whether the user belongs to the first group (control) or the second group (treatment)

landing_page - Whether the landing page is new or old

time_spent_on_the_page - Time (in minutes) spent by the user on the landing page

converted - Whether the user gets converted to a subscriber of the news portal or not

language_preferred - Language chosen by the user to view the landing page

Import all the necessary libraries

In [1]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
```

Reading the Data into a DataFrame

In [2]:
```
df = pd.read_csv('/Users/camilaconyers/Downloads/abtest.csv')
```

Explore the dataset and extract insights using Exploratory Data Analysis

- Data Overview
  - Viewing the first and last few rows of the dataset
  - Checking the shape of the dataset
  - Getting the statistical summary for the variables
- Check for missing values
- Check for duplicates

Displaying the first 5 rows of the dataset

In [3]:
```
df.head()
```

Out[3]:

	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

Displaying the last 5 rows of the dataset

In [4]:
```
df.tail()
```

Out[4]:

	user_id	group	landing_page	time_spent_on_the_page	converted	language_preferred
95	546446	treatment	new	5.15	no	Spanish
96	546544	control	old	6.52	yes	English
97	546472	treatment	new	7.07	yes	Spanish
98	546481	treatment	new	6.20	yes	Spanish
99	546483	treatment	new	5.86	yes	English

Checking the shape of the dataset

In [5]:
```
df.shape
```

Out[5]:
```
(100, 6)
```

Checking the data types of the columns for the dataset

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
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
```

Getting the statistical summary for the numerical variables

In [7]:
```
df.describe().T
```

Out[7]:

	count	mean	std	min	25%	50%	75%	max
user_id	100.0	546517.0000	52.295779	546443.00	546467.75	546492.500	546567.2500	546592.00
time_spent_on_the_page	100.0	5.3778	2.378166	0.19	3.88	5.415	7.0225	10.71

In [8]:
```
print(df.select_dtypes(include=['int64' , 'float64']).describe())
```

```
             user_id  time_spent_on_the_page
count     100.000000              100.000000
mean   546517.000000                5.377800
std        52.295779                2.378166
min    546443.000000                0.190000
25%    546467.750000                3.880000
50%    546492.500000                5.415000
75%    546567.250000                7.022500
max    546592.000000               10.710000
```

Getting the statistical summary for the categorical variables

In [9]:
```
print(df.select_dtypes(include=['object']).describe())
```

```
          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
```

Check for missing values

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
user_id                   0
group                     0
landing_page              0
time_spent_on_the_page    0
converted                 0
language_preferred        0
dtype: int64
```

Check for duplicates

In [11]:
```
df.duplicated()
```

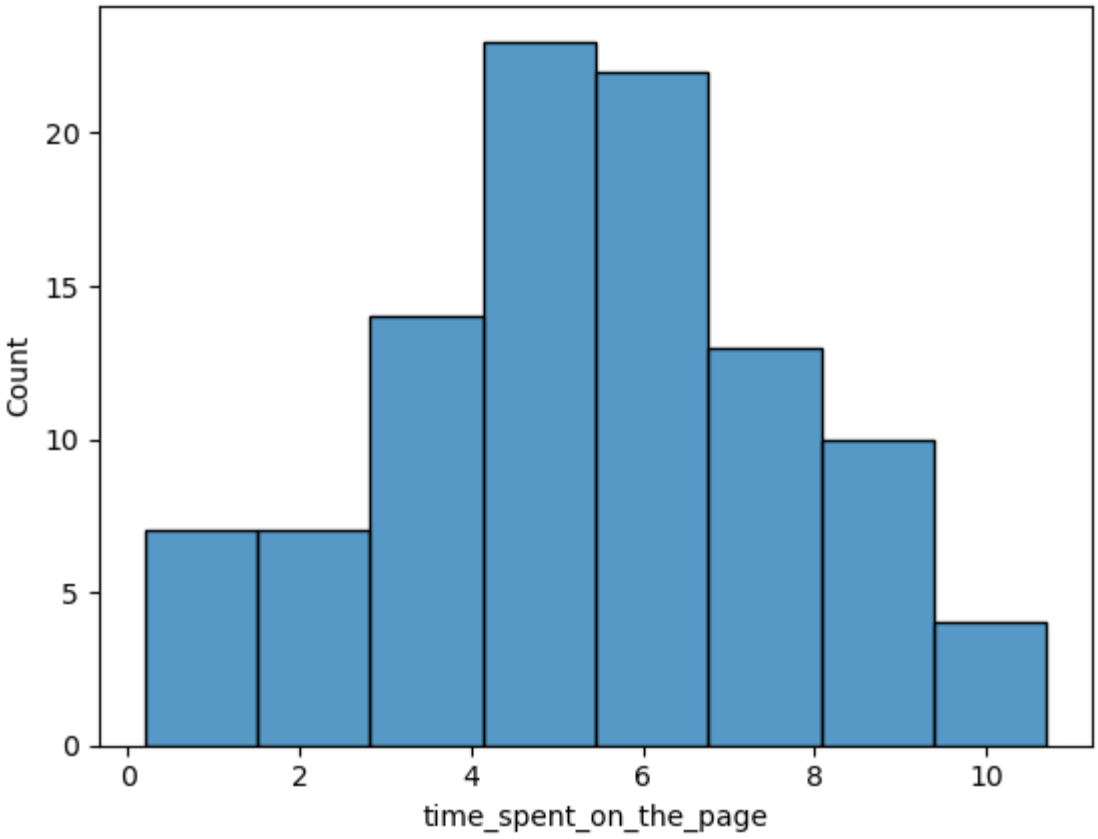
Out[11]:
```
0     False
1     False
2     False
```

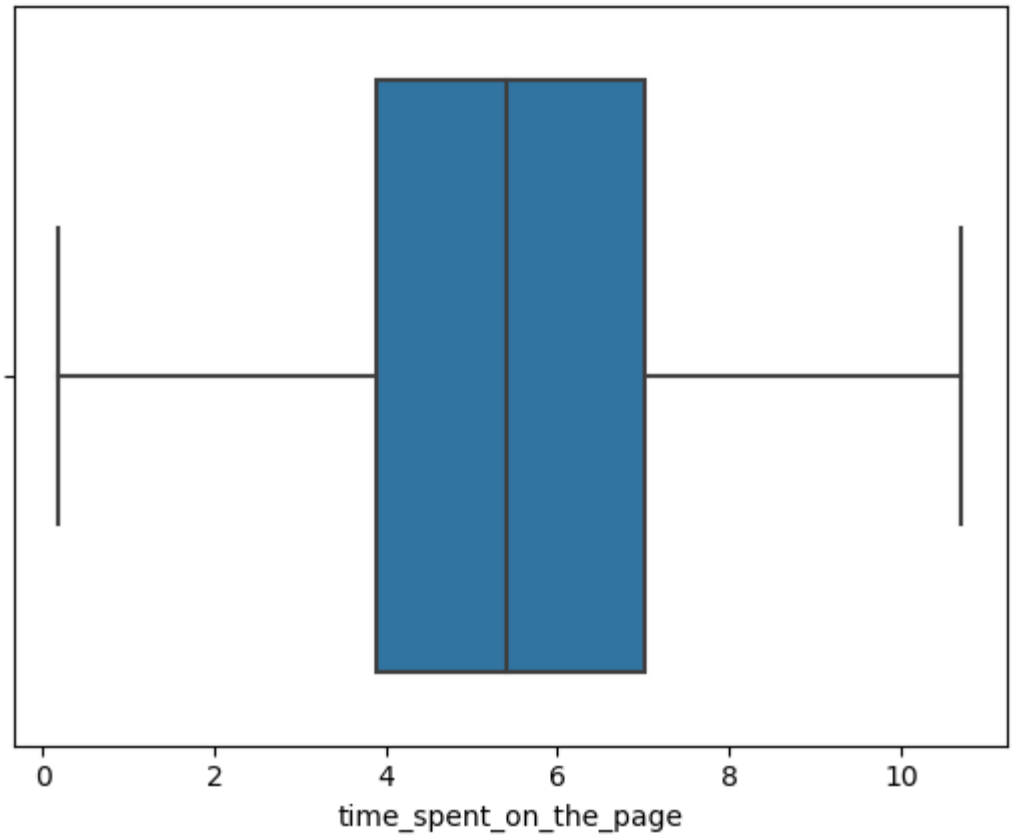
```
3     False
4     False
      ...
95    False
96    False
97    False
98    False
99    False
Length: 100, dtype: bool
```

Univariate Analysis

Time spent on the page

In [12]:
```
sns.histplot(data=df,x='time_spent_on_the_page')
plt.show()
sns.boxplot(data=df,x='time_spent_on_the_page')
plt.show()
```





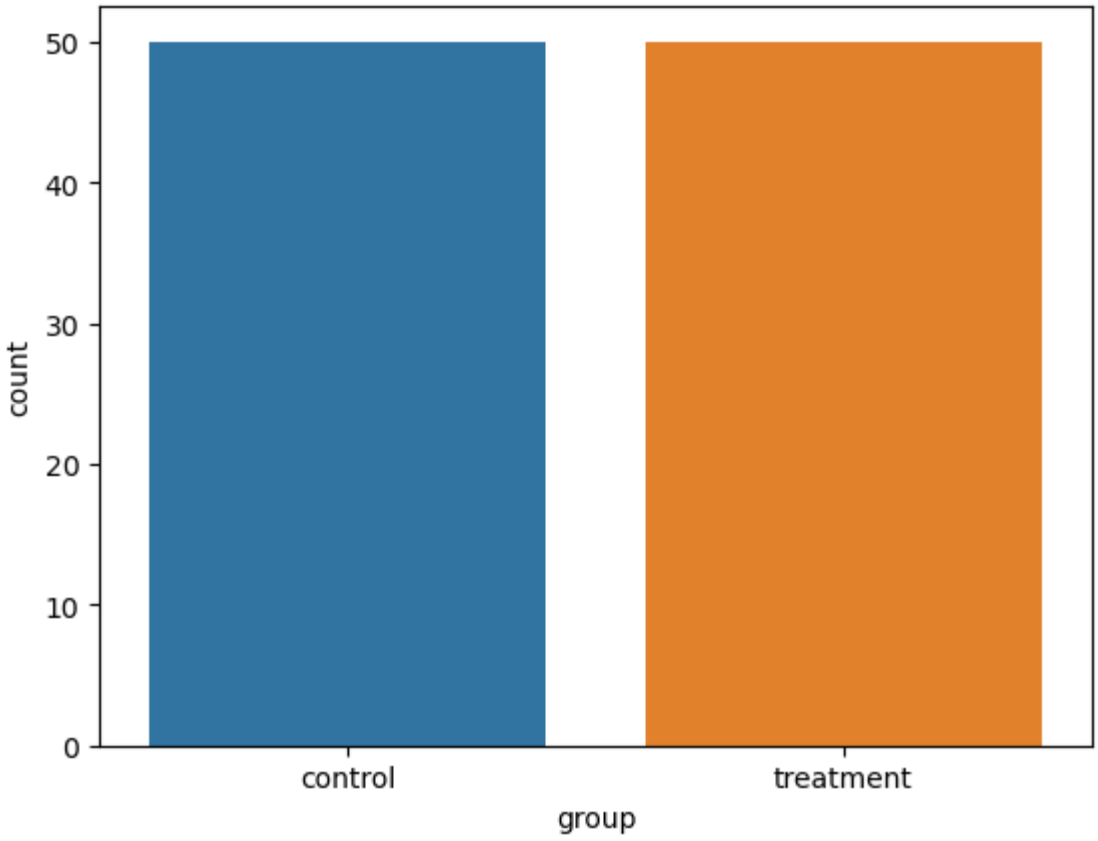
Insights: Time spend has a normal distrubution.

Group

In [13]:
```
df['group'].value_counts()
```

Out[13]:
```
control      50
treatment    50
Name: group, dtype: int64
```

In [14]:
```
sns.countplot(data=df,x='group')
plt.show()
```

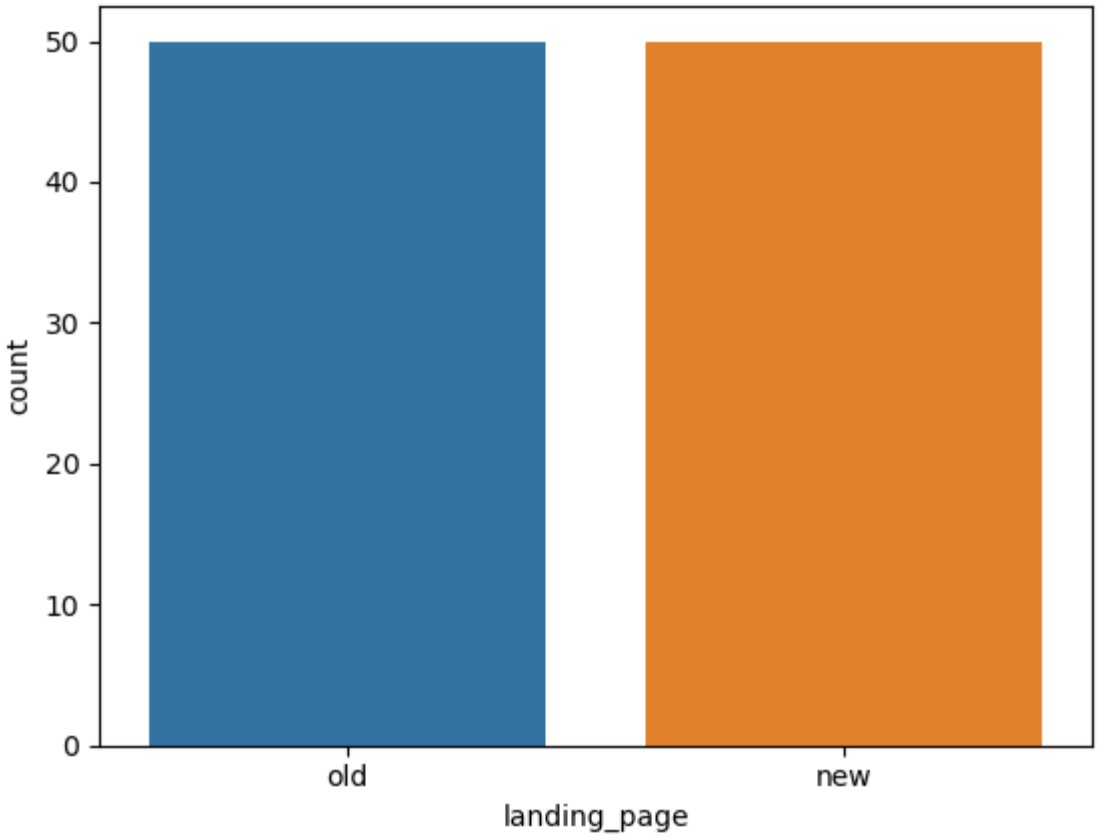


Landing page

In [15]:
```
df['landing_page'].value_counts()
```

Out[15]:
```
old    50
new    50
Name: landing_page, dtype: int64
```

In [16]:
```
sns.countplot(data=df,x='landing_page')
plt.show()
```

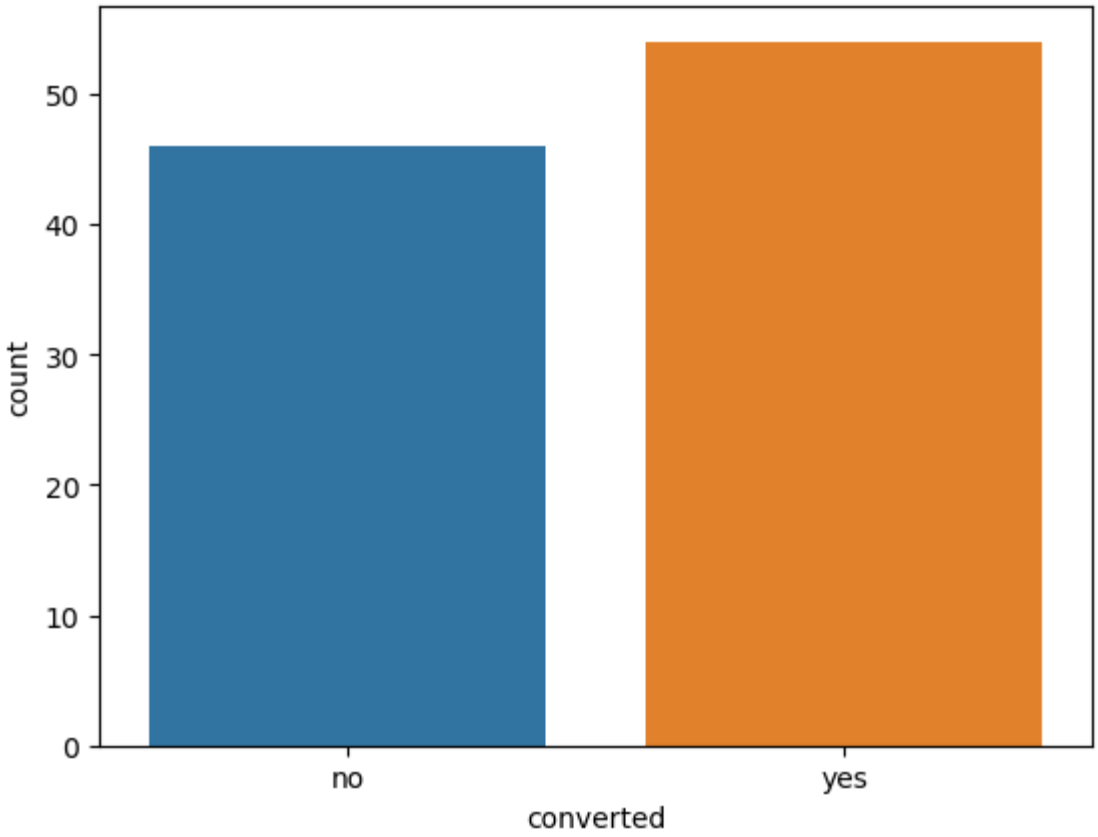


Converted

In [17]:
```
df['converted'].value_counts()
```

Out[17]:
```
yes     54
no      46
Name: converted, dtype: int64
```

In [18]:
```
sns.countplot(data=df,x='converted')
plt.show()
```

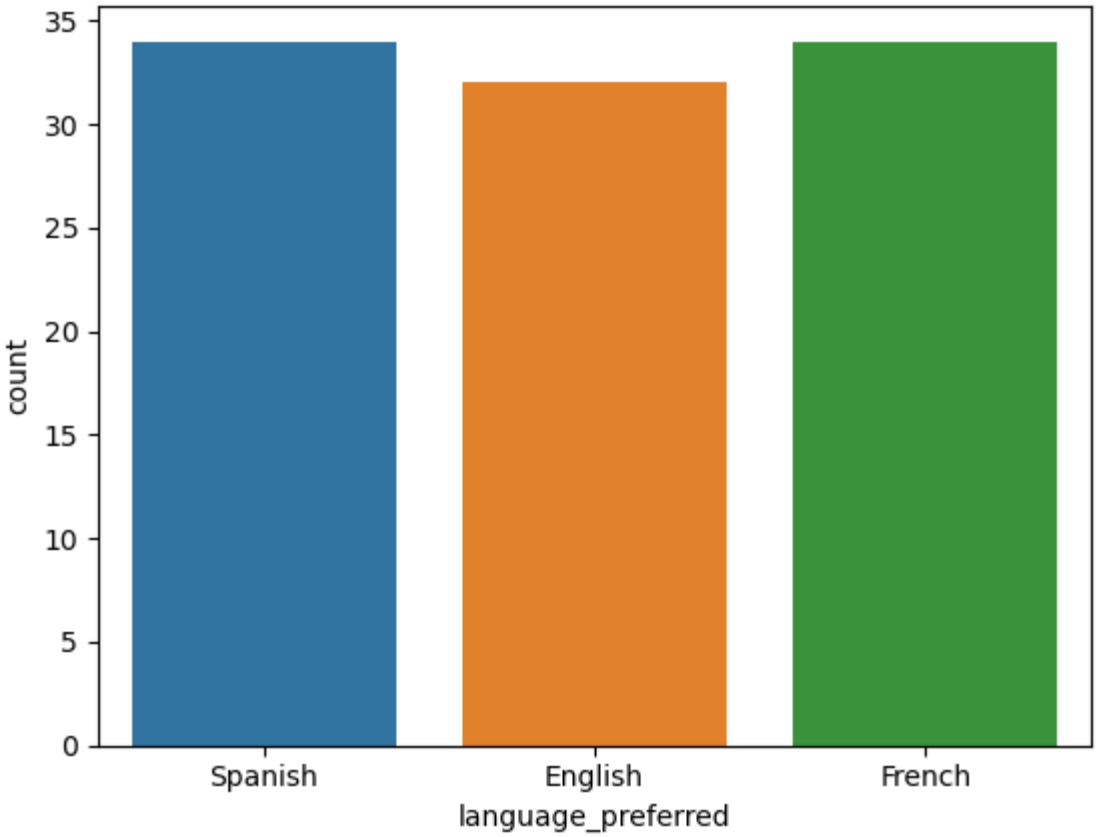


Language preferred

In [19]:
```
df['language_preferred'].value_counts()
```

Out[19]:
```
Spanish     34
French      34
English     32
Name: language_preferred, dtype: int64
```

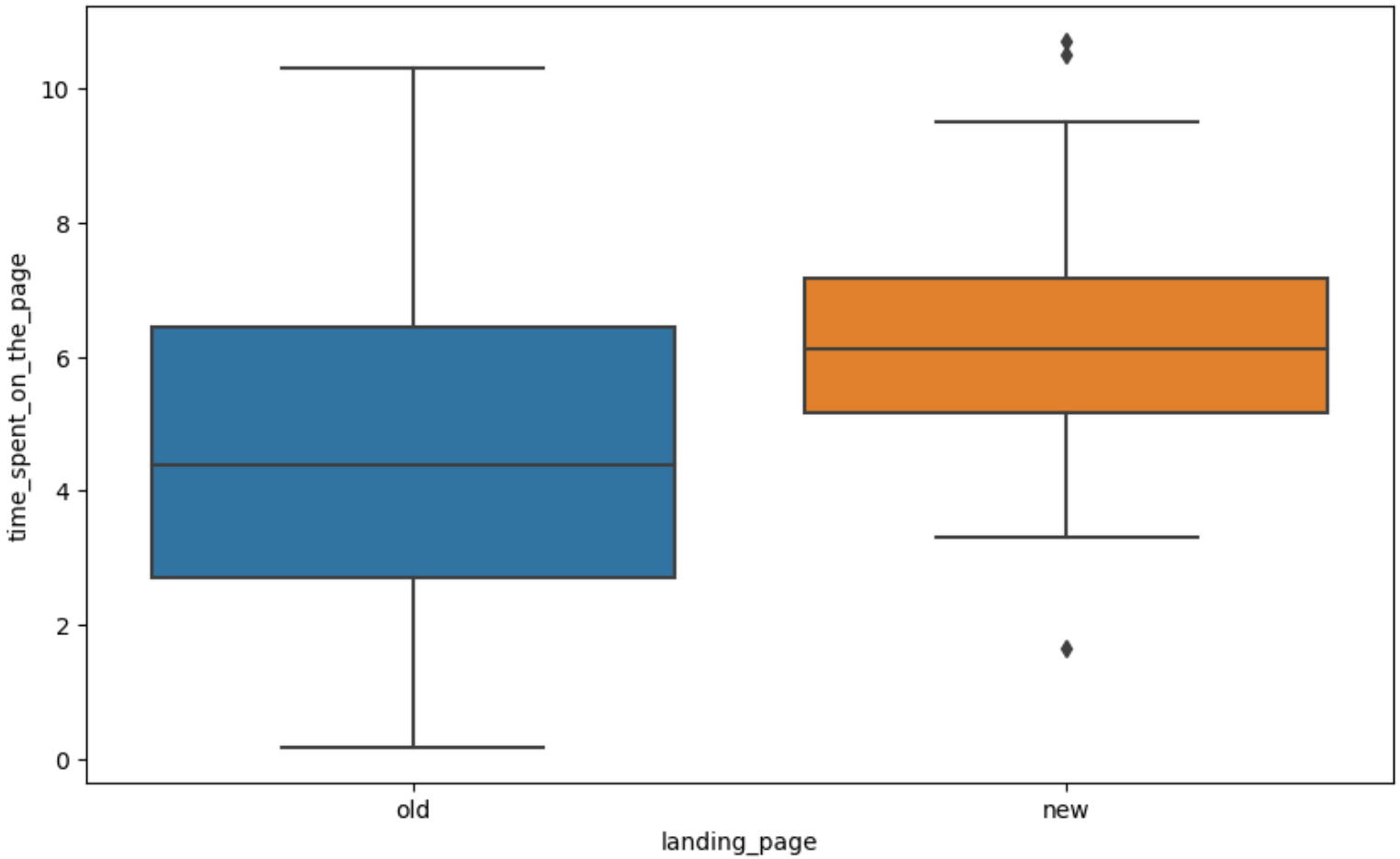
In [20]:
```
sns.countplot(data=df,x='language_preferred')
plt.show()
```



Bivariate Analysis

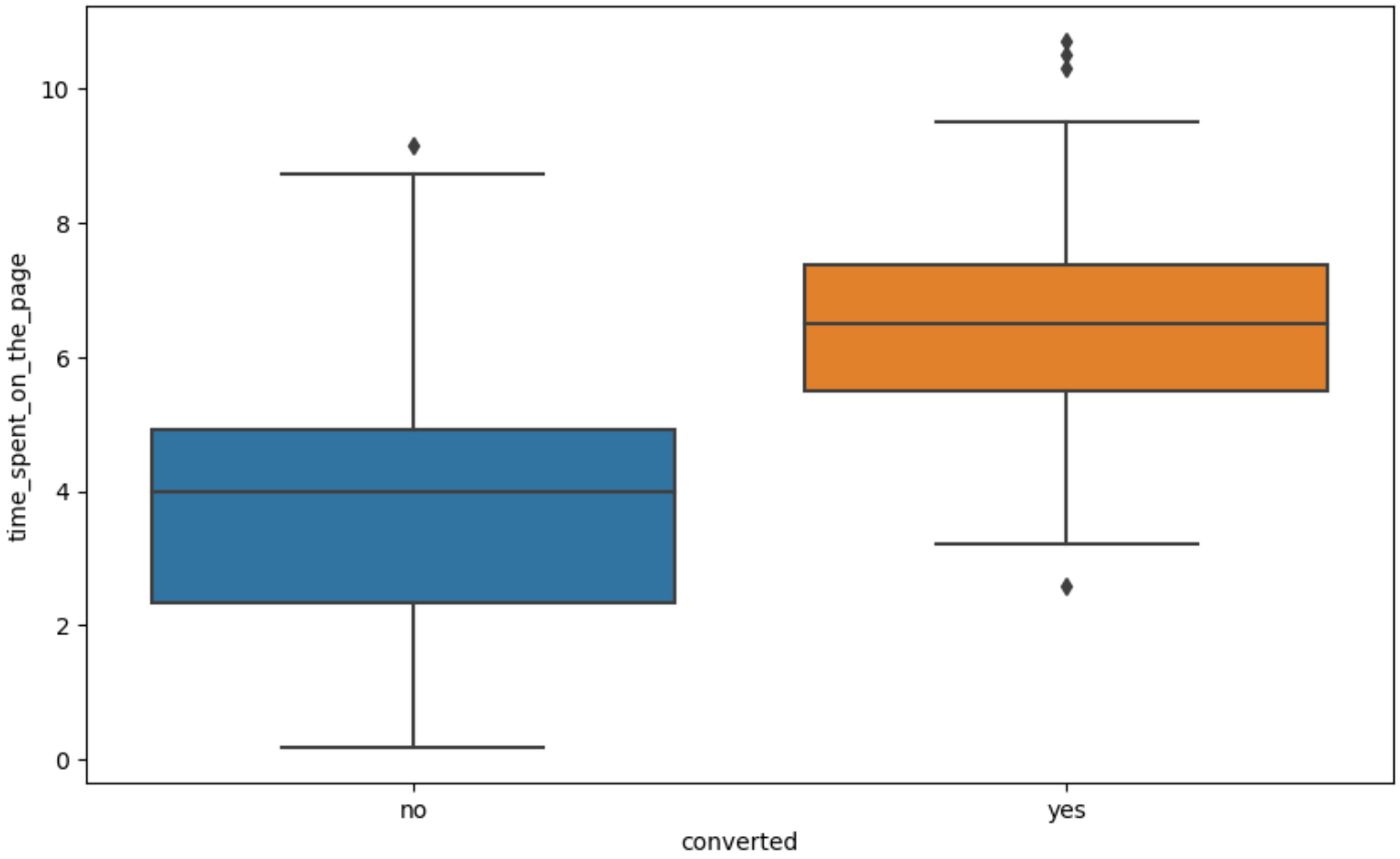
Landing page vs Time spent on the page

In [21]:
```
plt.figure(figsize=(10,6))
sns.boxplot(data=df,x='landing_page',y='time_spent_on_the_page')
plt.show()
```



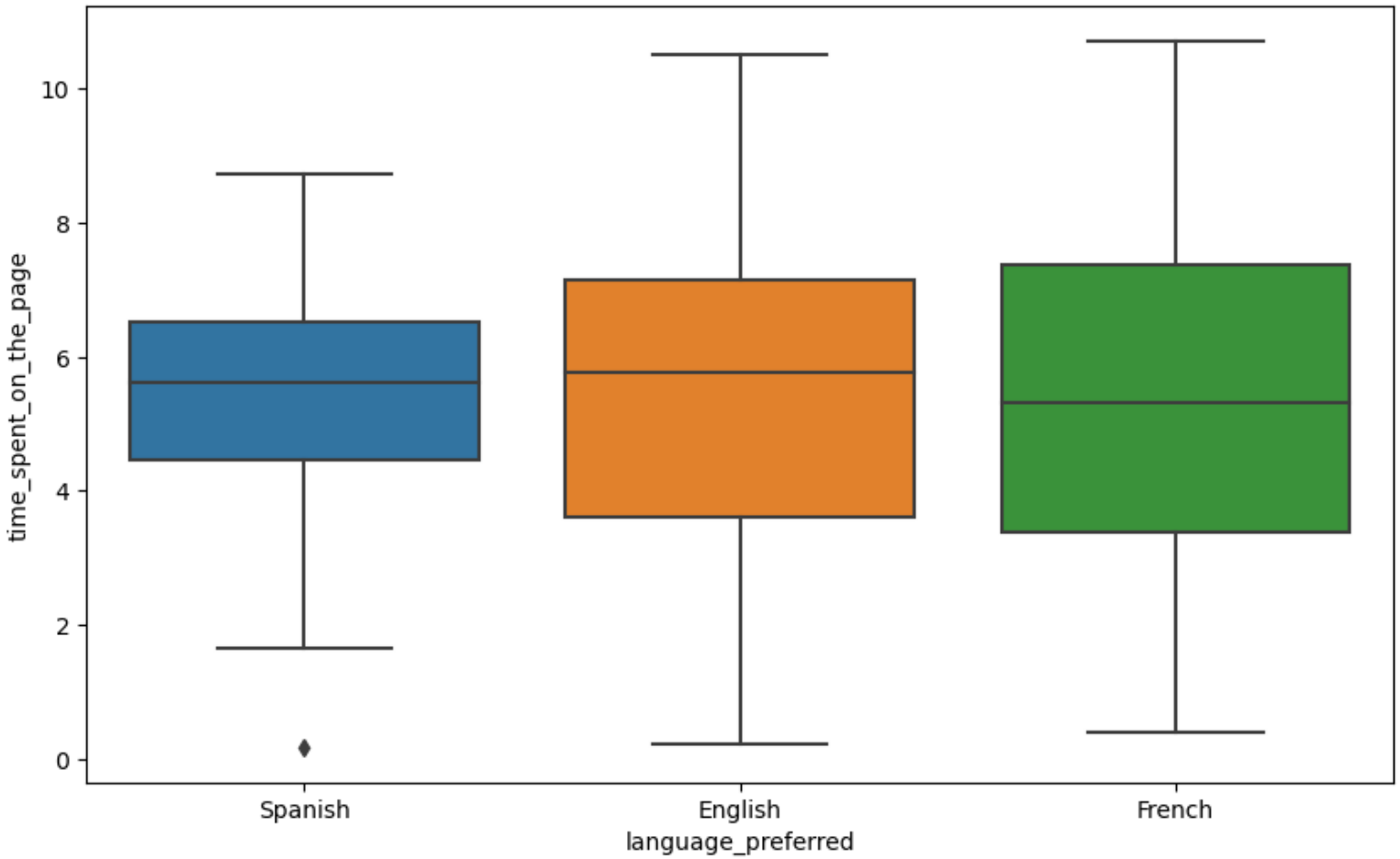
Conversion status vs Time spent on the page

In [22]:
```
# code to plot a suitable graph to understand the relationship between
'time_spent_on_the_page' and 'converted' columns
plt.figure(figsize=(10,6))
sns.boxplot(data=df,x='converted',y='time_spent_on_the_page')
plt.show()
```



Language preferred vs Time spent on the page

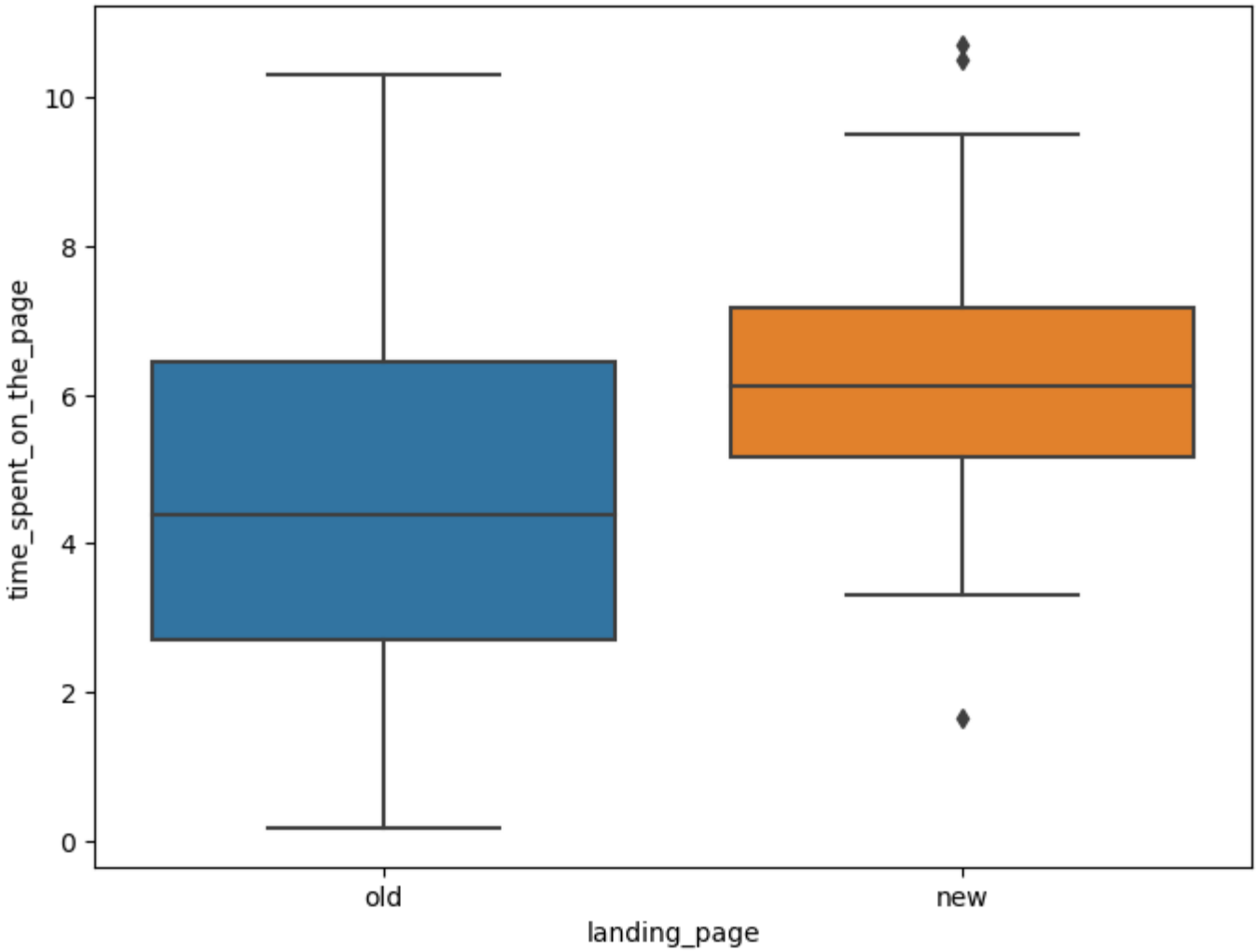
In [23]:
```
# code to plot a suitable graph to understand the distribution of
'time_spent_on_the_page' among the 'language_preferred'
plt.figure(figsize=(10,6))
sns.boxplot(data=df,x='language_preferred',y='time_spent_on_the_page')
plt.show()
```



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

Perform Visual Analysis

In [24]:
```
plt.figure(figsize=(8,6))
sns.boxplot(x = 'landing_page', y = 'time_spent_on_the_page', data = df)
plt.show()
```



Step 1: Define the null and alternate hypotheses

$H_0:$ Null Hypothesis The mean time spent by users on the new landing page is equal to the mean time spent on the existing landing page.

$H_a:$ Alternative Hypothesis The mean time spent by users on the new landing page is greater than the mean time spent on the existing landing page.

Step 2: Select Appropriate test

This is a two-sample independent t-test concerning two population means from two independent populations. The population standard deviations are unknown

H0: μ_new = μ_old

Ha: μ_new > μ_old

Step 3: Decide the significance level

As given in the problem statement, we select =0.05

Step 4: Collect and prepare data

In [25]:
```
# subsetted data frame for new landing page users
time_spent_new = df[df['landing_page'] == 'new']['time_spent_on_the_page']
```

```
# subsetted data frame for old landing page users
time_spent_old = df[df['landing_page'] == 'old']['time_spent_on_the_page']
```

In [26]:
```
print('The sample standard deviation of the time spent on the new page is:',
round(time_spent_new.std(),2))
print('The sample standard deviation of the time spent on the new page is:',
round(time_spent_old.std(),2))
```

```
The sample standard deviation of the time spent on the new page is: 1.82
The sample standard deviation of the time spent on the new page is: 2.58
```

Step 5: Calculate the p-value

In [27]:
```
from scipy.stats import ttest_ind

# assuming we're looking for whether time spent on the new page is greater than the
old one.
test_stat, p_value = ttest_ind(time_spent_new, time_spent_old, equal_var = False,
alternative = 'greater')

print('The p-value is', p_value)
```

```
The p-value is 0.0001392381225166549
```

Step 6: Compare the p-value with α

In [28]:
```
# print the conclusion based on p-value
if p_value < 0.05:
    print(f'As the p-value {p_value} is less than the level of significance, we reject
the null hypothesis.')
else:
    print(f'As the p-value {p_value} is greater than the level of significance, we
fail to reject the null hypothesis.')
```

```
As the p-value 0.0001392381225166549 is less than the level of significance, we reject the n
ull hypothesis.
```

Step 7: Draw inference

The p-value you've obtained (0.0001392381225166549) is indeed less than the common significance level (0.05). **Therefore, we reject the null hypothesis.**

Rejecting the null hypothesis means that there is statistically significant evidence to suggest that users spend more time on the new landing page than the existing one.

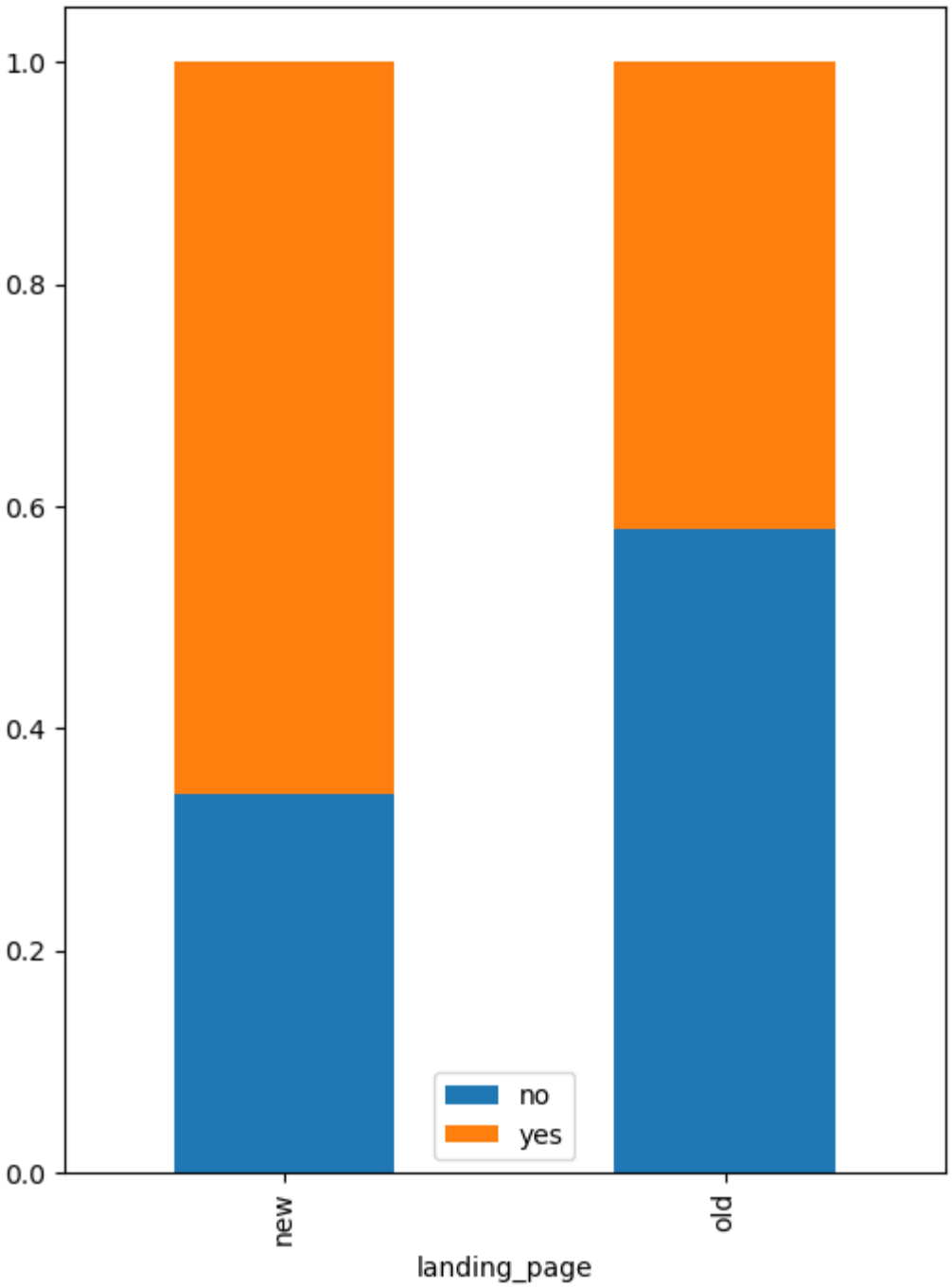
The alternative hypothesis was that the average time spent on the new page is greater than the average time spent on the old page, and the test has provided strong evidence in favor of this hypothesis.

A similar approach can be followed to answer the other questions.

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?

Perform Visual Analysis

In [29]:
```
pd.crosstab(df['landing_page'],df['converted'],normalize='index').plot(kind="bar",
figsize=(6,8),stacked=True)
plt.legend()
plt.show()
```



Step 1: Define the null and alternate hypotheses

$H_0:$ Null Hypothesis The conversion rate for the new page is less than or equal to the conversion rate for the old page. This means that the new page does not result in a higher conversion rate.

$H_a:$ Alternative Hypothesis The conversion rate for the new page is greater than the conversion rate for the old page. This implies that the new page is more effective at converting users.

Step 2: Select Appropriate test

This is a two-sample Z-test concerning two population means from two independent populations.

H0: μ_new <= μ_old

Ha: μ_new > μ_old.

Step 3: Decide the significance level

As given in the problem statement, we select α = 0.05.

Step 4: Collect and prepare data

In [30]:
```
# calculate the number of converted users in the treatment group
new_converted = df[df['group'] == 'treatment']['converted'].value_counts()['yes']

# calculate the number of converted users in the control group
old_converted = df[df['group'] == 'control']['converted'].value_counts()['yes']

n_control = df.group.value_counts()['control'] # total number of users in the control group
n_treatment = df.group.value_counts()['treatment'] # total number of users in the treatment group

print('The numbers of users served the new and old pages are {0} and {1} respectively'.format(n_control, n_treatment ))
```

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

Step 5: Calculate the p-value

In [31]:
```
# code to import the required function
from statsmodels.stats.proportion import proportions_ztest

# code to calculate the p-value
test_stat, p_value = proportions_ztest([new_converted, old_converted], [n_treatment, n_control], alternative = 'larger')

print('The p-value is', p_value)
```

```
The p-value is 0.008026308204056278
```

Step 6: Compare the p-value with α

In [32]:
```
# print the conclusion based on p-value
if p_value < 0.05:
    print(f'As the p-value {p_value} is less than the level of significance, we reject the null hypothesis.')
else:
    print(f'As the p-value {p_value} is greater than the level of significance, we fail to reject the null hypothesis.')
```

```
As the p-value 0.008026308204056278 is less than the level of significance, we reject the nu
ll hypothesis.
```

Step 7: Draw inference

The p-value in your output (0.008026308204056278) is less than the level of significance (0.05). **This indicates that we have strong evidence to reject the null hypothesis in favor of the alternative hypothesis.**

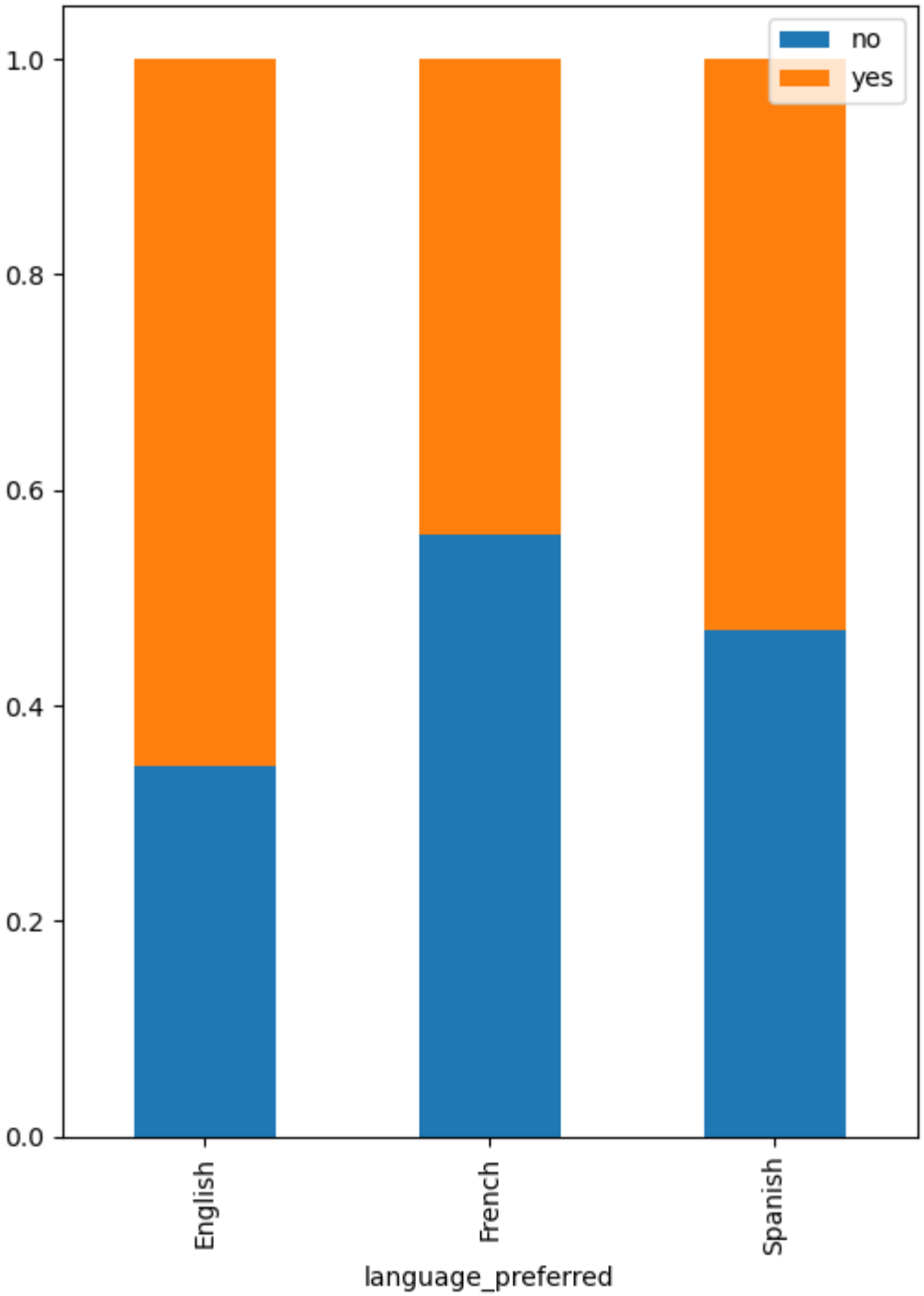
Rejecting the null hypothesis means that the conversion rate for the new landing page is significantly greater than the conversion rate for the old landing page. In other words, more users who visit the new landing page are likely to subscribe to E-news Express, compared to those who visit the old landing page.

This finding could inform business decisions at E-news Express, such as potentially implementing the new landing page design for all users, as it appears to be more effective at converting visitors into subscribers. However, other factors should also be considered, such as the cost of implementing the new design and whether the increase in conversion rate is significant enough to justify this cost.

3. Is the conversion and preferred language are independent or related?

Perform Visual Analysis

In [33]:
```
# code to visually plot the dependency between conversion status and preferred
langauge
pd.crosstab(df['language_preferred'], df['converted'],
normalize='index').plot(kind="bar", figsize=(6,8), stacked=True)
plt.legend()
plt.show()
```



Step 1: Define the null and alternate hypotheses

$H_0:$ The conversion status is independent of the preferred language. There is no relationship between the conversion status and the preferred language.

$H_a:$ The conversion status is not independent of the preferred language. There is a relationship between the conversion status and the preferred language.

Step 2: Select Appropriate test

Chi-square test. This is a problem of the test of independence, concerning two categorical variables - converted status and preferred language

Step 3: Decide the significance level

As given in the problem statement, we select α = 0.05.

Step 4: Collect and prepare data

In [34]:
```
#code to create a contingency table showing the distribution of the two categorical variables
contingency_table = pd.crosstab(df['converted'], df['language_preferred'])

contingency_table
```

Out[34]:

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

Step 5: Calculate the p-value

In [35]:
```
#code to import the required function
from scipy.stats import chi2_contingency

#code to calculate the p-value
chi2, p_value, dof, exp_freq = chi2_contingency(contingency_table)   # #complete the code by filling appropriate parameters in the blanks

print('The p-value is', p_value)
```

```
The p-value is 0.21298887487543447
```

Step 6: Compare the p-value with α

In [36]:
```
# print the conclusion based on p-value
if p_value < 0.05:
    print(f'As the p-value {p_value} is less than the level of significance, we reject the null hypothesis.')
else:
    print(f'As the p-value {p_value} is greater than the level of significance, we fail to reject the null hypothesis.')
```

```
As the p-value 0.21298887487543447 is greater than the level of significance, we fail to rej
ect the null hypothesis.
```

Step 7: Draw inference

The p-value of 0.21298887487543447 is indeed greater than the commonly used significance level of 0.05.

In this context, failing to reject the null hypothesis means that we do not have enough statistical evidence to conclude that there is a relationship between the conversion status and preferred language.

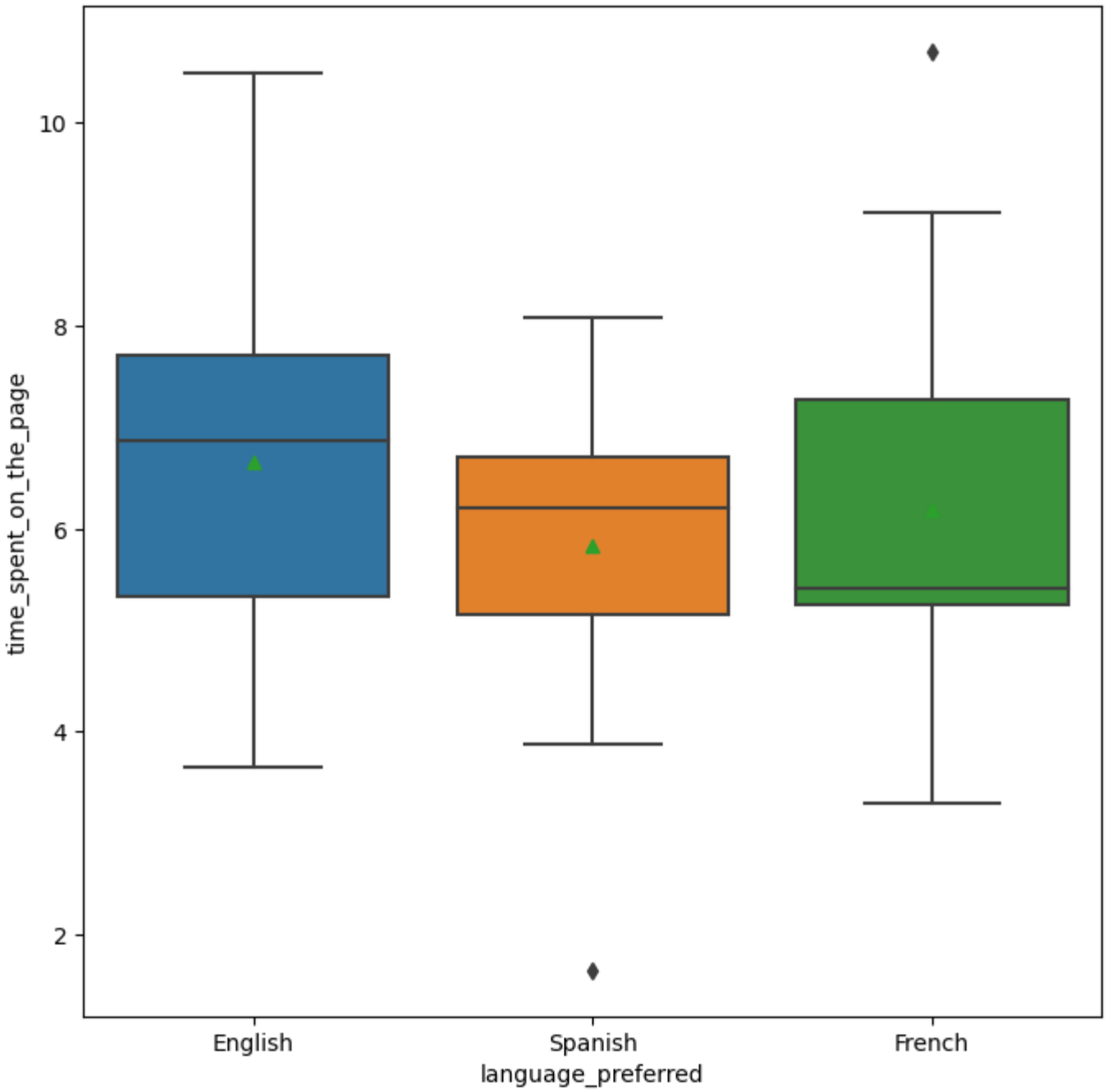
In other words, based on the data you have and the Chi-square test of independence, it seems that the conversion status (whether a user gets converted to a subscriber or not) is independent of the preferred language chosen by the user to view the landing page.

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

Perform Visual Analysis

In [37]:
```
# create a new DataFrame for users who got served the new page
df_new = df[df['landing_page'] == 'new']
```

In [38]:
```
# code to visually plot the time spent on the new page for different language users
plt.figure(figsize=(8,8))
sns.boxplot(x = 'language_preferred', y = 'time_spent_on_the_page', showmeans = True,
data = df_new)
plt.show()
```



In [39]:
```
#code to calculate the mean time spent on the new page for different language users
df_new.groupby(['language_preferred'])['time_spent_on_the_page'].mean()
```

Out[39]:
```
language_preferred
 English    6.663750
```

```
French     6.196471
Spanish    5.835294
Name: time_spent_on_the_page, dtype: float64
```

Step 1: Define the null and alternate hypotheses

$H_0:$ The mean time spent on the new page is the same for all language groups.

$H_a:$ At least one language group has a different mean time spent on the new page.

Step 2: Select Appropriate test

The appropriate test is an ANOVA Test. This is a problem, concerning three population means.

Step 3: Decide the significance level

As given in the problem statement, we select α = 0.05.

Step 4: Collect and prepare data

In [40]:
```
# subsetted data frame of the time spent on the new page by English language users
time_spent_English = df_new[df_new['language_preferred']=="English"]
['time_spent_on_the_page']

# subsetted data frames of the time spent on the new page by French and Spanish
language users
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']
```

Step 5: Calculate the p-value

In [41]:
```
# code to import the required function
from scipy.stats import f_oneway

# code to calculate the p-value
test_stat, p_value = f_oneway(time_spent_English, time_spent_French,
time_spent_Spanish)

print('The p-value is', p_value)
```

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

Step 6: Compare the p-value with α

In [42]:
```
# print the conclusion based on p-value
if p_value < 0.05:
    print(f'As the p-value {p_value} is less than the level of significance, we
reject the null hypothesis.')
else:
    print(f'As the p-value {p_value} is greater than the level of significance, we
fail to reject the null hypothesis.')
```

```
As the p-value 0.43204138694325955 is greater than the level of significance, we fail to rej
ect the null hypothesis.
```

Step 7: Draw inference

The p-value in this case is greater than the standard significance level of 0.05. So, would fail to reject the null hypothesis.

The null hypothesis was that the mean time spent on the new page is the same for users who prefer English, French, and Spanish. Failing to reject this null hypothesis means that I did not find sufficient evidence in my sample to conclude that there is a significant difference in the mean time spent on the new page among these three groups.

In other words, **the language preference of users does not seem to significantly affect the time they spend on the new page.**

Conclusion and Business Recommendations

Key Conclusions:

Overall, the new landing page seems to be performing well in terms of both user engagement (time spent on the page) and conversion rates. Future work should focus on identifying the specific features of the new page that contribute to these outcomes, and applying this knowledge to further optimize the website. Also, conducting similar analyses with more extensive and diverse data could help further validate these findings.

Users spend more time on the new landing page than the old one: We have significant statistical evidence (p-value = 0.0001392381225166549, less than significance level of 0.05) to support this. Thus, the new landing page appears to be more engaging for users. This could be due to a better design, more relevant content, improved navigation, etc.

Is the conversion rate for the new page greater than the conversion rate for the old page?

Yes, the conversion rate for the new page is greater than that for the old page.

Does the converted status depend on the preferred language?

No, there seems to be no significant dependence of converted status on the preferred language.

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

There is no statistically significant difference in the time spent on the new page among different language users.

Key Business Recommendations:

Adopt the New Landing Page: The analysis shows that users spend more time on the new landing page than the old one, and that the conversion rate for the new page is higher. Therefore, it is recommended to replace the old landing page with the new one. The longer time spent might be due to better design, user-friendly layout, more appealing content, etc. Also, a higher conversion rate indicates that the new landing page is more successful at encouraging users to subscribe to the e-newsletter.

Language Preference Doesn't Affect Conversion Rate: We found that the preferred language does not significantly impact the conversion rate. Therefore, while it is essential to cater to the language preferences of the users, it is equally important to focus on other areas (like the content of the landing page or the e-newsletter itself) that could have a more significant impact on conversion rates.

Continual Testing: While these tests have provided valuable insights, continually testing different designs, layouts, and content could help further optimize user experience and conversion rates. For example, you might want to experiment with different forms of media (like video), different calls-to-action, or different incentives for subscribing (like discounts or exclusive content).

In-depth Analysis of Users: Although the current analysis doesn't show the impact of user demographics on conversion rates, in-depth user analysis might provide some valuable insights. User information like their interests, age, time of activity, can be useful for targeted marketing and personalizing the user experience.

Customer Engagement: Encourage the users to spend more time on the page by making the content more engaging. More

time spent on the page likely increases the chance of conversion. Therefore, continually improve the content quality based on user feedback and performance metrics.