E-news Express Project

Problem Statement

The problem is to determine whether the new landing page is more effective than the old one in terms of user engagement and conversion rate. The following questions need to be addressed:

1. Do users spend more time on the new landing page than on the old one?
2. Is the conversion rate higher for the new landing page compared to the old one?
3. Does the converted status depend on the preferred language?
4. Is the mean time spent on the new page the same for different language users?

Data Dictionary

- 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

Let us start by importing the required libraries

In [32]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as stats
```

Understand the structure of data

In [31]:
```
df = pd.read_csv('/content/abtest.csv')
df
```

Out[31]:

	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
...	...	...	...	...	...	...
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

100 rows × 6 columns

In [33]:
```
df.head()
```

Out[33]:

	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

In [34]:
```
df.shape
```

Out[34]:
```
(100, 6)
```

Observation:

The data has 100 rows and 6 columns.

In [35]:
```
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
```

Observation:

This dataset 1 integer columns, 1 float column, and 4 object columns. All columns are non-null.

In [36]:
```
missing_values = df.isnull().sum()
print(missing_values)
```

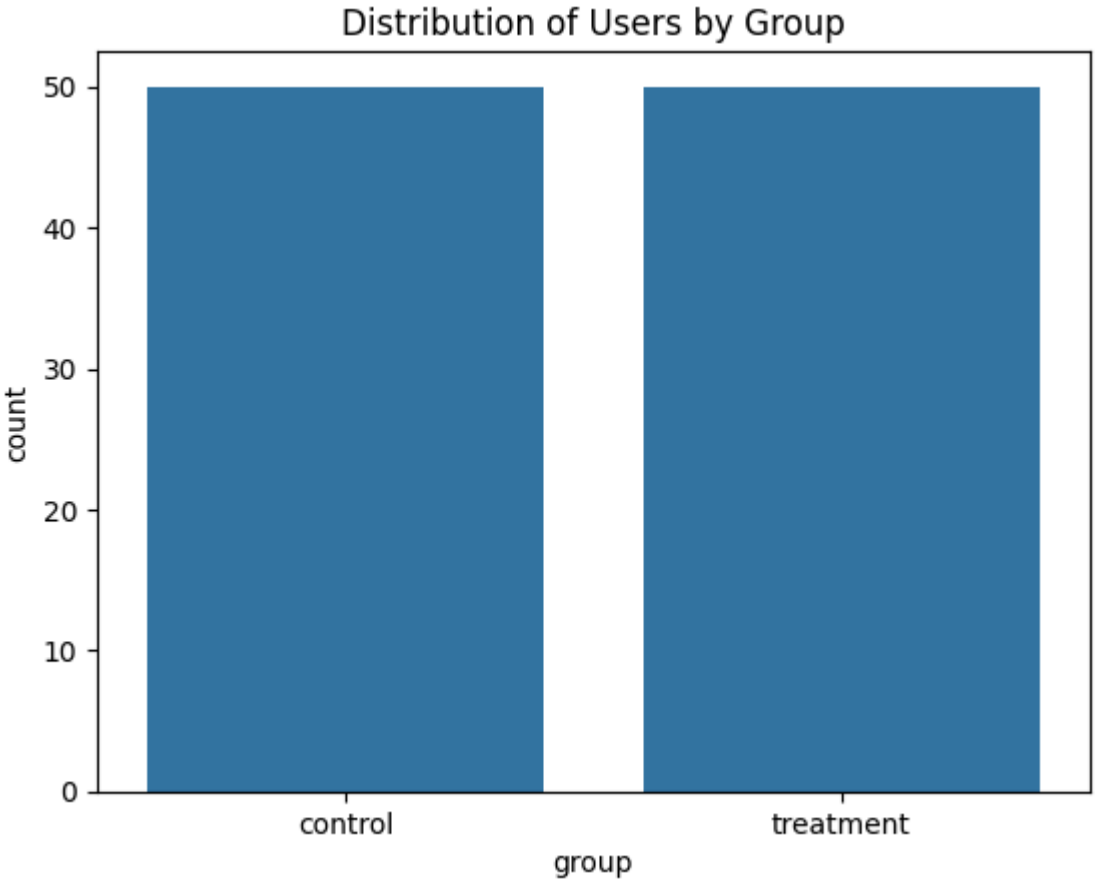
```
user_id                   0
group                     0
landing_page              0
time_spent_on_the_page    0
converted                 0
language_preferred        0
dtype: int64
```

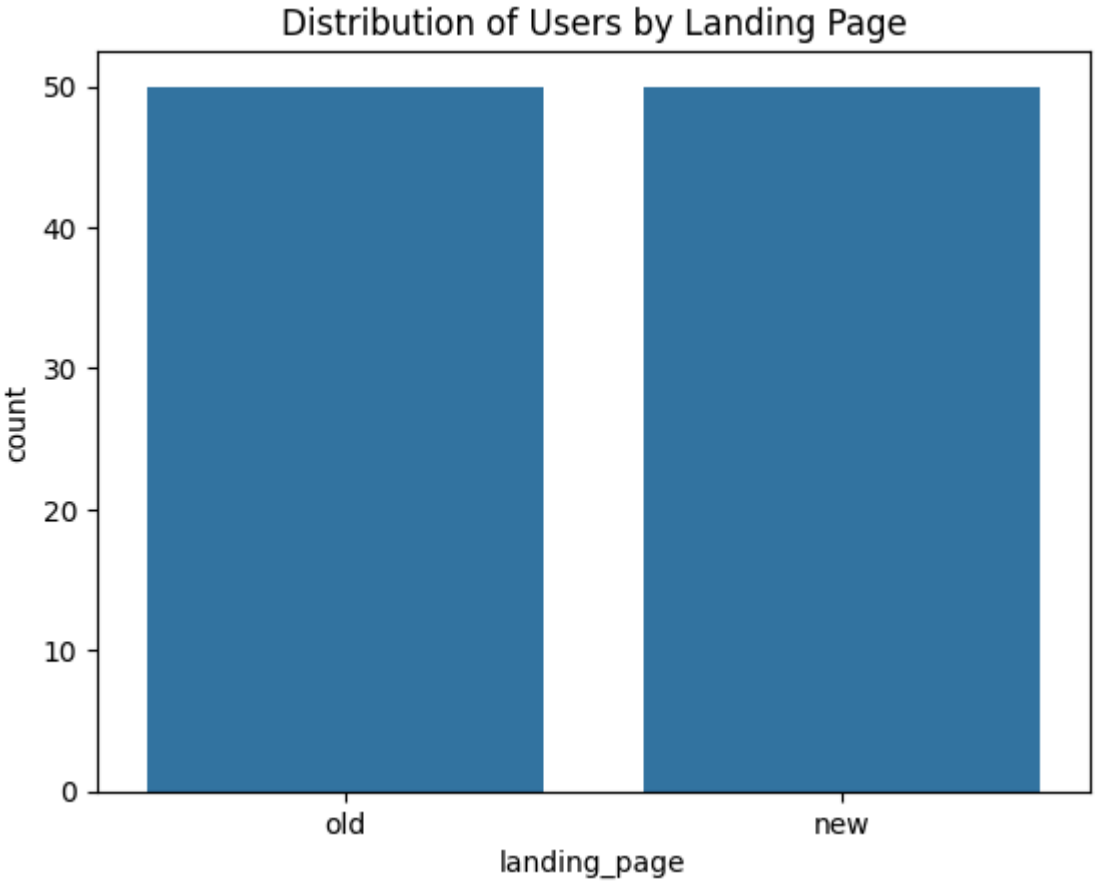
Observations:

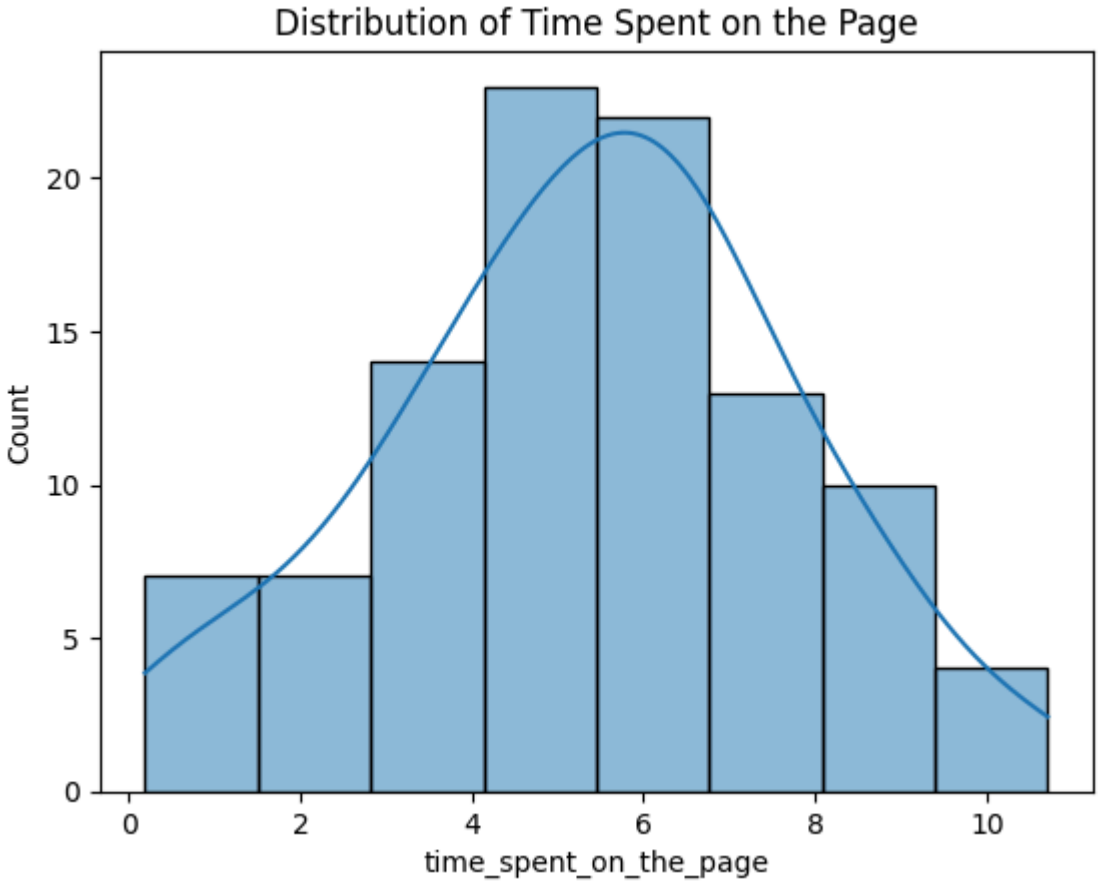
There is No missing value in dataset

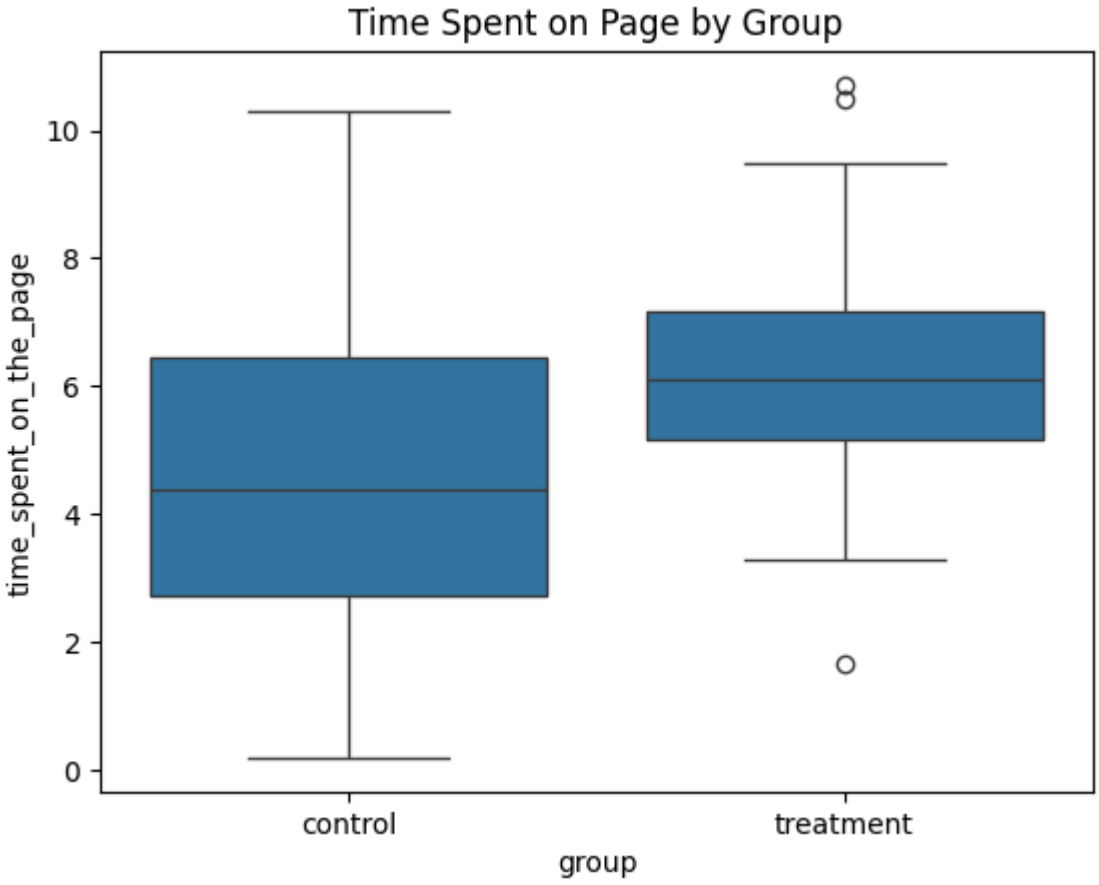
In [37]:

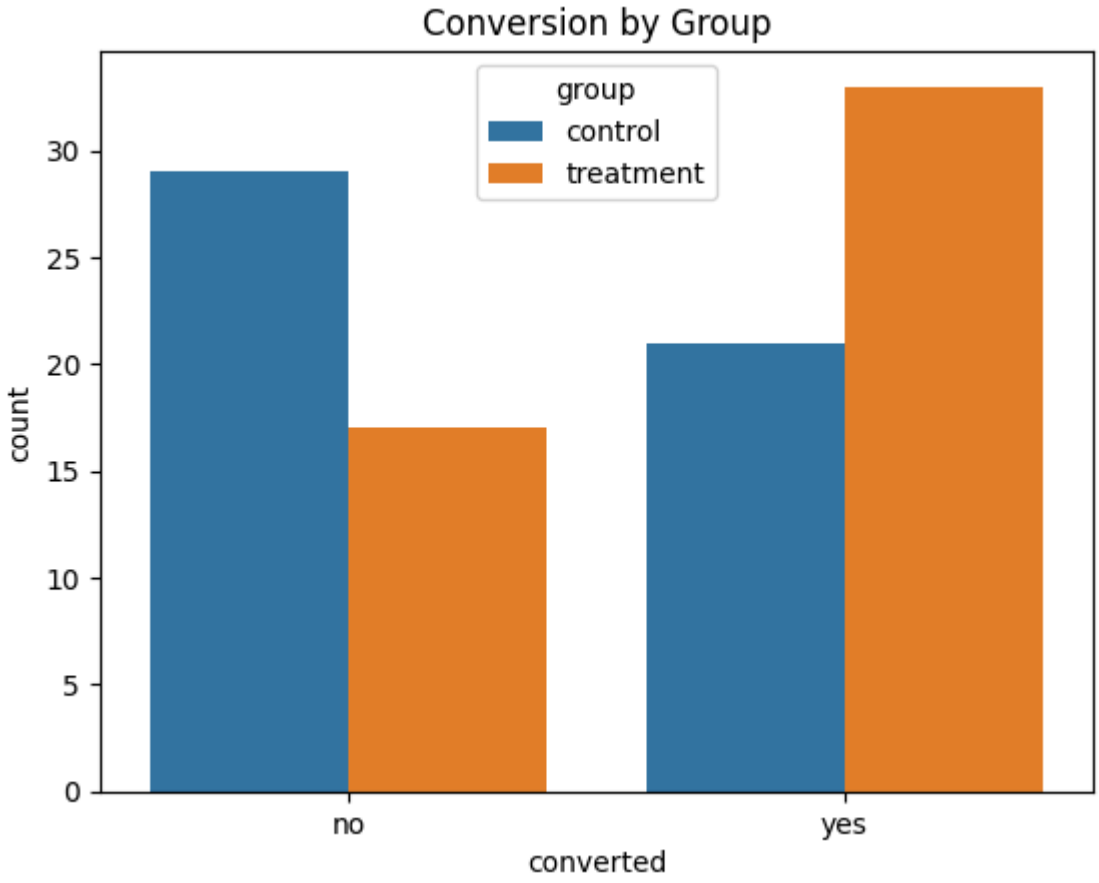
```
# Univariate analysis

sns.countplot(x='group', data=df)
plt.title('Distribution of Users by Group')
plt.show()

sns.countplot(x='landing_page', data=df)
plt.title('Distribution of Users by Landing Page')
plt.show()

sns.histplot(df['time_spent_on_the_page'], kde=True)
plt.title('Distribution of Time Spent on the Page')
plt.show()

# Bivariate analysis
sns.boxplot(x='group', y='time_spent_on_the_page', data=df)
plt.title('Time Spent on Page by Group')
plt.show()

sns.countplot(x='converted', hue='group', data=df)
plt.title('Conversion by Group')
plt.show()

sns.countplot(x='converted', hue='landing_page', data=df)
plt.title('Conversion by Landing Page')
```


Distribution of Users by Group
count
0
10
20
30
40
50
control
treatment
group

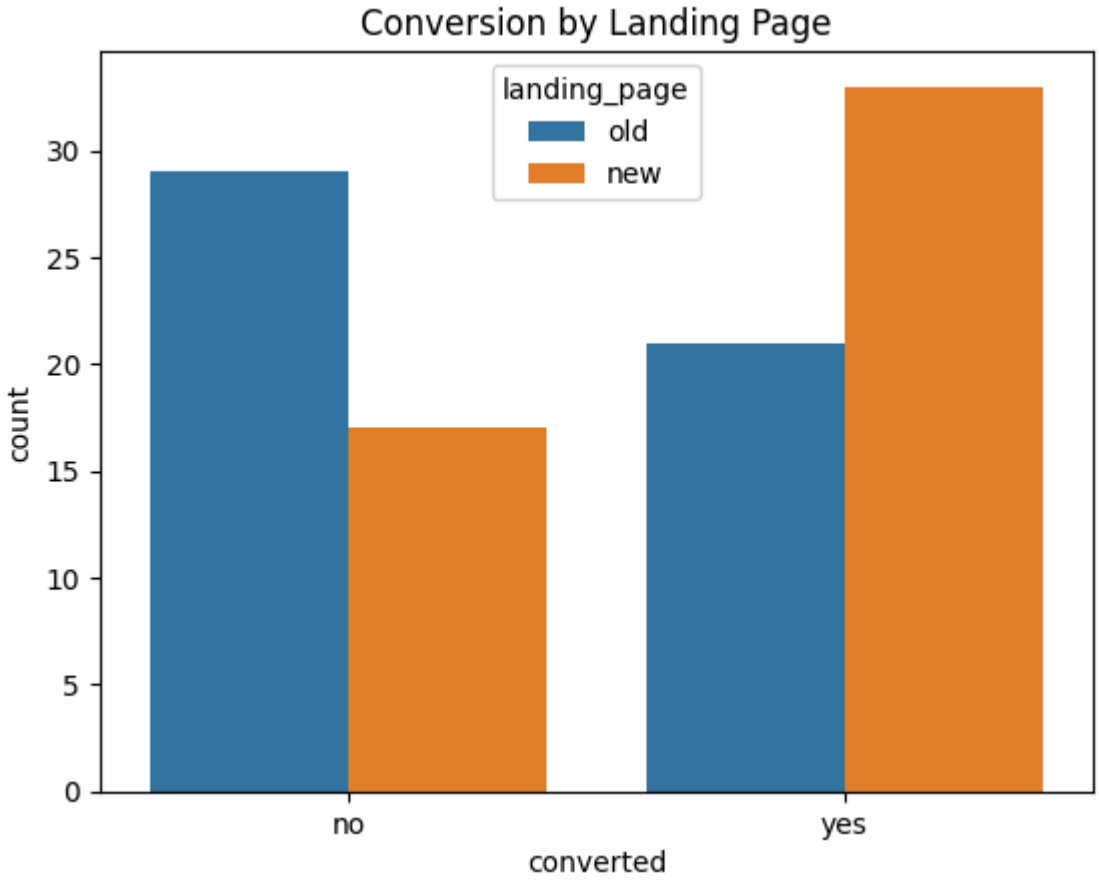

Distribution of Users by Landing Page
count
0
10
20
30
40
50
old
new
landing_page


Distribution of Time Spent on the Page
Count
time_spent_on_the_page
0
5
10
15
20
0
2
4
6
8
10


Time Spent on Page by Group
time_spent_on_the_page
group
control
treatment
0
2
4
6
8
10



Out[37]: Text(0.5, 1.0, 'Conversion by Landing Page')



Observation:

The visualizations help in understanding the distribution of users across groups, landing pages, and time spent on the pages. There are no obvious outliers or missing data issues.

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

- Null Hypothesis (H0): The mean time spent on the new landing page is equal to or less than the time spent on the old page.
- Alternative Hypothesis (H1): The mean time spent on the new landing page is greater than the time spent on the old page.

In [38]:
```
# Filter data by group
control_group = df[df['group'] == 'control']['time_spent_on_the_page']
treatment_group = df[df['group'] == 'treatment']['time_spent_on_the_page']

# Perform a t-test
t_stat, p_value = stats.ttest_ind(treatment_group, control_group, alternati
ve='greater')
print(f'T-statistic: {t_stat}, P-value: {p_value}')

if p_value < 0.05:
    print("Reject the null hypothesis: Users spend more time on the new lan
ding page.")
else:
    print("Fail to reject the null hypothesis: No significant difference in
time spent.")
```

```
T-statistic: 3.7867702694199856, P-value: 0.0001316123528095005
Reject the null hypothesis: Users spend more time on the new landing page.
```

Observation:

Depending on the p-value, we determine whether the new landing page results in significantly more time spent by users.

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

- Null Hypothesis (H0): The conversion rate for the new page is less than or equal to the conversion rate for the old page.
- Alternative Hypothesis (H1): The conversion rate for the new page is greater than the conversion rate for the old page.

In [39]:
```
# Convert 'converted' column to numeric values
df['converted'] = df['converted'].map({'yes': 1, 'no': 0})
print(df['converted'].head())
```

```
0    0
1    1
2    0
3    0
4    1
Name: converted, dtype: int64
```

In [40]:
```
# Conversion rates
conversion_rate_control = df[df['group'] == 'control']['converted'].mean()
conversion_rate_treatment = df[df['group'] == 'treatment']['converted'].mea
n()

# Perform a proportion z-test
from statsmodels.stats.proportion import proportions_ztest
import numpy as np

count = np.array([df[df['group'] == 'control']['converted'].sum(),
                  df[df['group'] == 'treatment']['converted'].sum()])
nobs = np.array([df[df['group'] == 'control'].shape[0],
                 df[df['group'] == 'treatment'].shape[0]])

stat, p_value = proportions_ztest(count, nobs, alternative='larger')
print(f'Z-statistic: {stat}, P-value: {p_value}')

if p_value < 0.05:
    print("Reject the null hypothesis: Conversion rate is higher for the ne
w landing page.")
else:
    print("Fail to reject the null hypothesis: No significant difference in
conversion rates.")
```

```
Z-statistic: -2.4077170617153842, P-value: 0.9919736917959437
Fail to reject the null hypothesis: No significant difference in conversio
n rates.
```

Observation:

We infer if the new landing page has a significantly higher conversion rate.

1. Does the converted status depend on the preferred language?

- Null Hypothesis (H0): Conversion status is independent of the preferred language.
- Alternative Hypothesis (H1): Conversion status depends on the preferred language.

In [41]:
```
# Perform Chi-square test
contingency_table = pd.crosstab(df['language_preferred'], df['converted'])
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)
print(f'Chi2 Statistic: {chi2}, P-value: {p}')

if p < 0.05:
    print("Reject the null hypothesis: Conversion status depends on the pre
ferred language.")
else:
    print("Fail to reject the null hypothesis: Conversion status is indepen
dent of the preferred language.")
```

```
Chi2 Statistic: 3.0930306905370832, P-value: 0.2129888748754345
Fail to reject the null hypothesis: Conversion status is independent of th
e preferred language.
```

Observation:

The chi-square test determines if there's a significant relationship between the preferred language and conversion status.

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

- Null Hypothesis (H0): The mean time spent on the new page is the same for all languages.
- Alternative Hypothesis (H1): The mean time spent on the new page is different for at least one language.

In [42]:
```
# Filter data for treatment group
new_page_data = df[df['landing_page'] == 'new']

# Perform ANOVA test
anova_result = stats.f_oneway(new_page_data[new_page_data['language_preferr
ed'] == 'English']['time_spent_on_the_page'],
                              new_page_data[new_page_data['language_preferr
ed'] == 'Spanish']['time_spent_on_the_page'],
                              new_page_data[new_page_data['language_preferr
ed'] == 'French']['time_spent_on_the_page'])

print(f'ANOVA F-statistic: {anova_result.statistic}, P-value: {anova_resul
t.pvalue}')

# Inference
if anova_result.pvalue < 0.05:
    print("Reject the null hypothesis: Mean time spent on the new page is d
ifferent for at least one language.")
else:
    print("Fail to reject the null hypothesis: Mean time spent on the new p
age is the same for all languages.")
```

```
ANOVA F-statistic: 0.854399277000682, P-value: 0.43204138694325955
Fail to reject the null hypothesis: Mean time spent on the new page is the
same for all languages.
```

Observation:

ANOVA helps us understand if the mean time spent on the new landing page differs significantly across language preferences.

conclusion:-

- If users spend significantly more time on the new landing page, it indicates higher engagement, which could potentially lead to higher conversion rates.
- A higher conversion rate on the new page suggests that the new design and content are more effective in persuading users to subscribe.
- If conversion depends on the preferred language, it may indicate that content localization could further improve conversion rates.
- If the time spent on the new page varies significantly across languages, it could highlight the need for language-specific optimizations.

Buisness Reccomendation:-

- Consider rolling out the new landing page to all users if it shows significant improvement in engagement and conversion rates.
- Focus on optimizing content for different languages to maximize conversion rates across all user segments.
- Continue A/B testing for other features and content to ensure the landing page remains effective over time.
- Implement further segmentation analysis to target specific user groups more effectively.