Project Business Statistics: E-news Express

Marks: 60

Define Problem Statement and Objectives

Business Context

The advent of e-news, or electronic news, portals has offered us a great opportunity to quickly get updates on the day-to-day events occurring globally. The information on these portals is retrieved electronically from online databases, processed using a variety of software, and then transmitted to the users. There are multiple advantages of transmitting new electronically, like faster access to the content and the ability to utilize different technologies such as audio, graphics, video, and other interactive elements that are either not being used or aren't common yet in traditional newspapers.

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 decision to subscribe.

[Companies often analyze user responses to two variants of a product to decide which of the two variants is more effective. This experimental technique, known as A/B testing, is used to determine whether a new feature attracts users based on a chosen metric.]

Objective 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. In order to test the effectiveness of the new landing page in gathering new subscribers, the Data Science team conducted an experiment by randomly selecting 100 users and dividing them equally into two groups. The existing landing page was served to the first group (control group) and the new landing page to the second group (treatment group). Data regarding the interaction of users in both groups with the two versions of the landing page was collected. Being a data scientist in E-news Express, you have been asked to explore the data and perform 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:

Do the users spend more time on the new landing page than on the existing 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 time spent on the new page the same for the different language users?

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 [174…

```
# Libraries to help with reading and manipulating data
import pandas as pd
import numpy as np

# Libraries to help with data visualization
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Library to help with statistical analysis
import scipy.stats as stats
```

Reading the Data into a DataFrame

In [175…

```
df = pd.read_csv("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

In [176…

```
df.head()
```

Out[176]:

	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 [177…

```
df.tail()
```

Out[177]:

	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

In [178…

```
print (df.shape[0],'rows and',df.shape[1],'columns');
```

```
100 rows and 6 columns
```

In [179…

```
data.describe()
```

Out[179]:

	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

Statistical summary of time spent on the page is mentioned above. The mean spent time is 5.37 minutes. 50% of the users spent less than 5.415 minutes on the page.

In [180…

```
data.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
```

Out of the six columns in the provided dataset, two are numeric whereas four are objects.

In [181…

```
data.isnull().sum()
```

Out[181]:

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

All columns have values populated in them. There are no nulls in the data.

In [182…

```
data['user_id'].unique()
```

Out[182]:

```
array([546592, 546468, 546462, 546567, 546459, 546558, 546448, 546581,
       546461, 546548, 546588, 546546, 546491, 546478, 546578, 546466,
       546443, 546555, 546493, 546549, 546560, 546584, 546450, 546475,
       546456, 546455, 546469, 546586, 546471, 546575, 546464, 546556,
       546585, 546577, 546587, 546552, 546551, 546557, 546487, 546589,
       546559, 546570, 546489, 546453, 546488, 546565, 546460, 546458,
       546492, 546473, 546554, 546457, 546479, 546576, 546482, 546563,
       546569, 546454, 546562, 546574, 546470, 546467, 546572, 546590,
       546553, 546445, 546545, 546582, 546484, 546579, 546568, 546476,
       546452, 546444, 546591, 546583, 546573, 546485, 546486, 546547,
       546490, 546449, 546463, 546580, 546571, 546564, 546465, 546480,
       546447, 546561, 546477, 546451, 546566, 546474, 546550, 546446,
       546544, 546472, 546481, 546483], dtype=int64)
```

In [183…

```
data['user_id'].duplicated().sum()
```

Out[183]:

```
0
```

Each User ID is unique. Each row corresponds to a different user.

In [184…

```
data['group'].unique()
```

Out[184]:

```
array(['control', 'treatment'], dtype=object)
```

There are only two types of user groups, control and treatment.

In [185…

```
data['landing_page'].unique()
```

Out[185]:
```
array(['old', 'new'], dtype=object)
```

There are only two types of landing pages, the old one and the new one.

In [186…
```
data['converted'].unique()
```

Out[186]:
```
array(['no', 'yes'], dtype=object)
```

There are two distinct values in converted column, either a no or a yes signifying that the user became a customer or not.

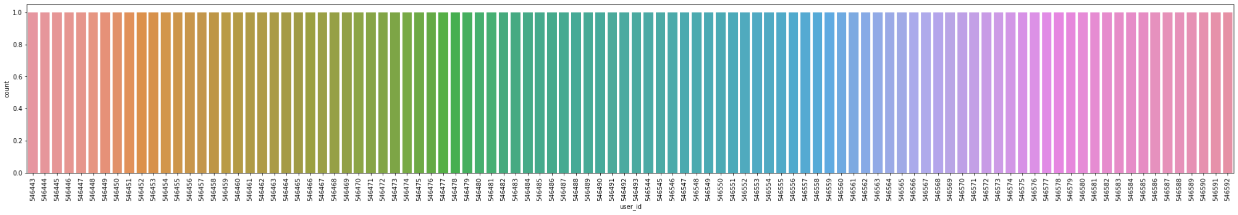
In [187…
```
data['language_preferred'].unique()
```

Out[187]:
```
array(['Spanish', 'English', 'French'], dtype=object)
```

There are three different languages in which the users can see the page.

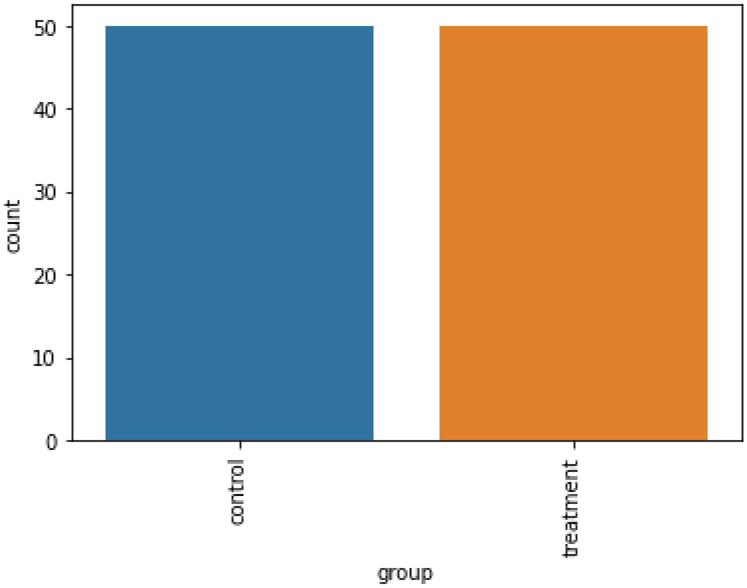
Univariate Analysis

In [188…
```
plt.figure(figsize = (35,5))
sns.countplot(data=df,x='user_id')
plt.xticks(rotation=90)
plt.show()
```



There are no duplicates in the user id, each user has a unique row and is mentioned only once.

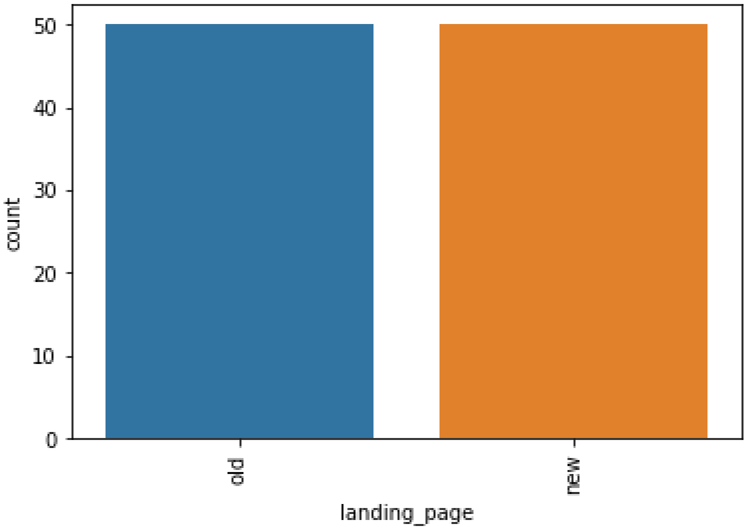
In [189…
```
sns.countplot(data=df,x='group')
plt.xticks(rotation=90)
plt.show()
```



The two groups control and treatment have an equal population.

In [190…

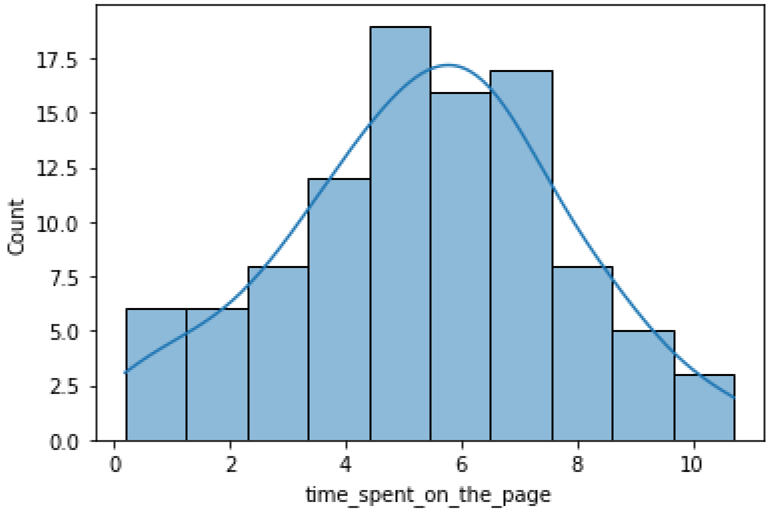
```
sns.countplot(data=df,x='landing_page')
plt.xticks(rotation=90)
plt.show()
```

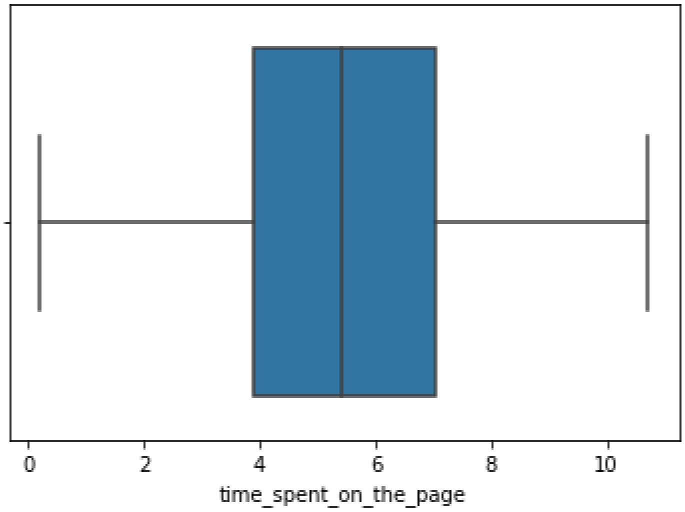


There's equal data of users on new and old landing page.

In [191…

```
sns.histplot(data=df,x='time_spent_on_the_page',bins = 10,kde=True);
plt.show();
sns.boxplot(data = df, x='time_spent_on_the_page')
plt.show();
```

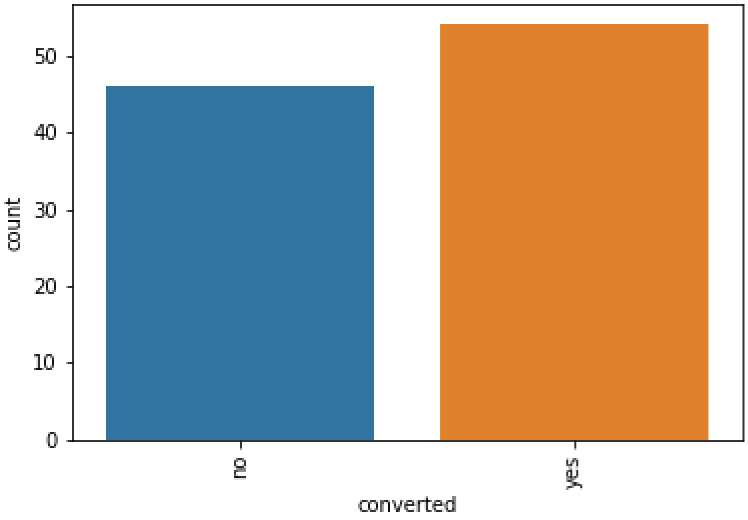




The time spent on the page follows a normal distribution. The minimum time spent on the page is 0.19 mins The maximum time spent on the page is 10.71 mins The average time spent on the page is 5.37 mins

In [192...

```
sns.countplot(data=df,x='converted')
plt.xticks(rotation=90)
plt.show()
```



In [193…

```
df.groupby('converted').count()
```

Out[193]:

converted	user_id	group	landing_page	time_spent_on_the_page	language_preferred
no	46	46	46	46	46
yes	54	54	54	54	54

54 users became subscriber whereas 46 did not.

In [194…

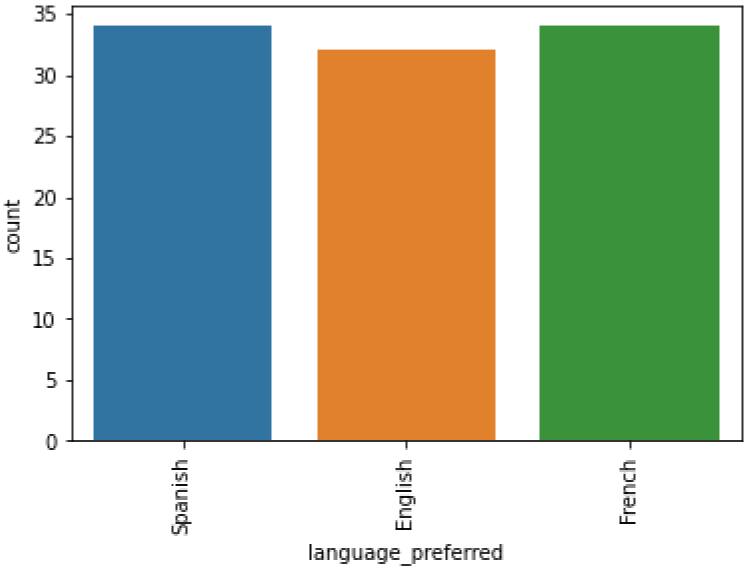
```
df.groupby('language_preferred').count()
```

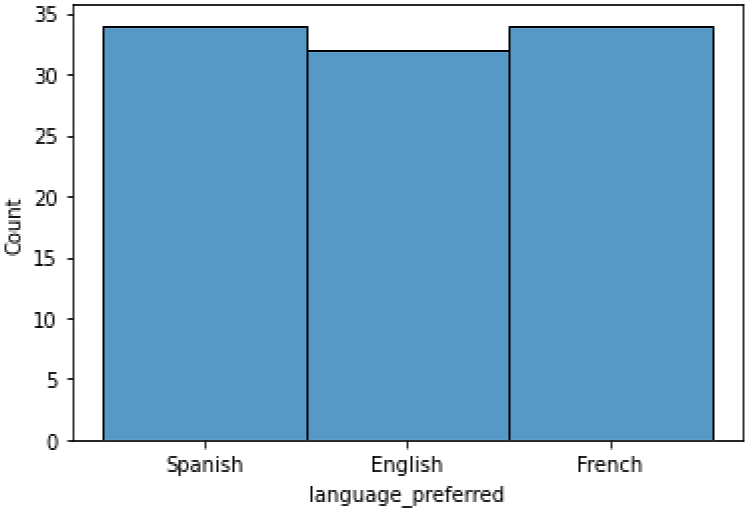
Out[194]:

language_preferred	user_id	group	landing_page	time_spent_on_the_page	converted
English	32	32	32	32	32
French	34	34	34	34	34
Spanish	34	34	34	34	34

In [55]:

```
sns.countplot(data=df,x='language_preferred')
plt.xticks(rotation=90)
plt.show()
sns.histplot(data=df,x='language_preferred');
plt.show();
```



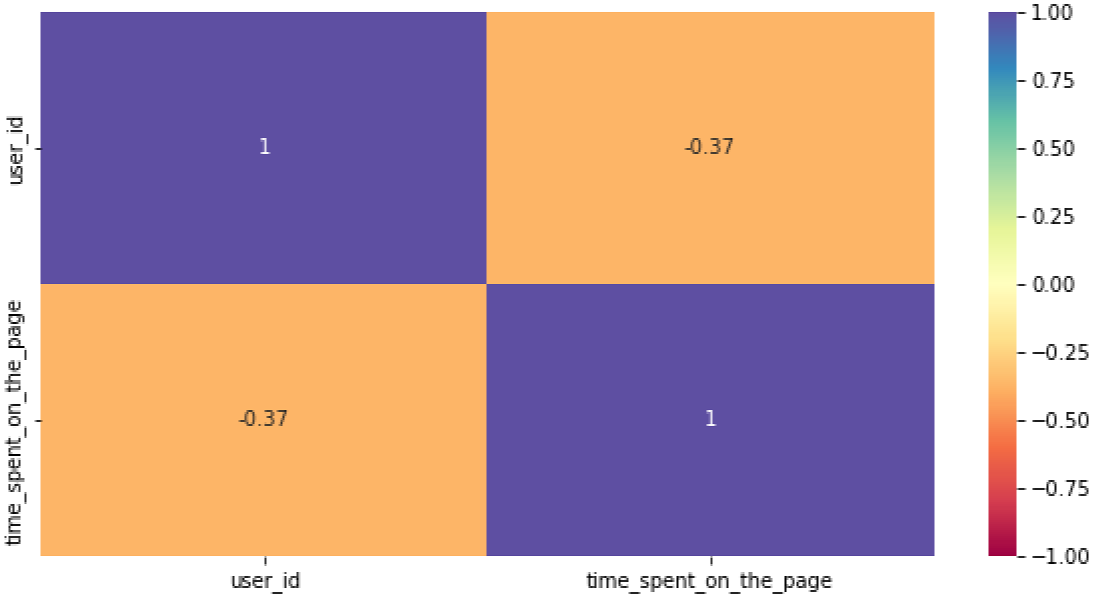


32 users preferred English whereas Spanish and French was preferred by 34 users each.

Bivariate Analysis

In [195…

```
# Write the code here
#df = df.drop('total_time', axis=1)
plt.figure(figsize=(10,5))
sns.heatmap(df.corr(),annot=True,cmap='Spectral',vmin=-1,vmax=1)
plt.show()
```



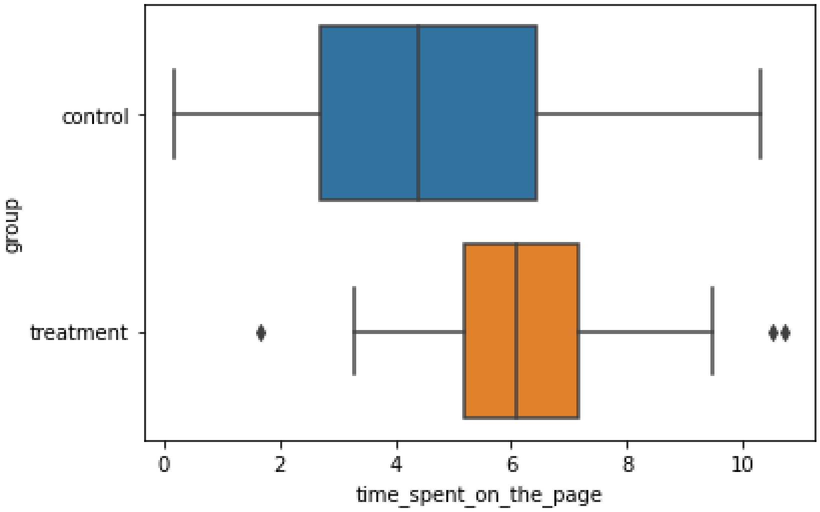
There is no strong correlation between user_id and time spent on the page.

In [196…

```
#plt.figure(figsize=(10,5))
#sns.scatterplot(data=df,x='time_spent_on_the_page',y='group')
#plt.show()

#plt.figure(figsize=(15,7))
#sns.lineplot(data=df,x='time_spent_on_the_page',y='group')
#plt.show()
```

```
sns.boxplot(data=df,x='time_spent_on_the_page',y='group')
plt.show()
```



The treatment group, i.e. users provided with the new page have a higher median in time spent on the landing page.

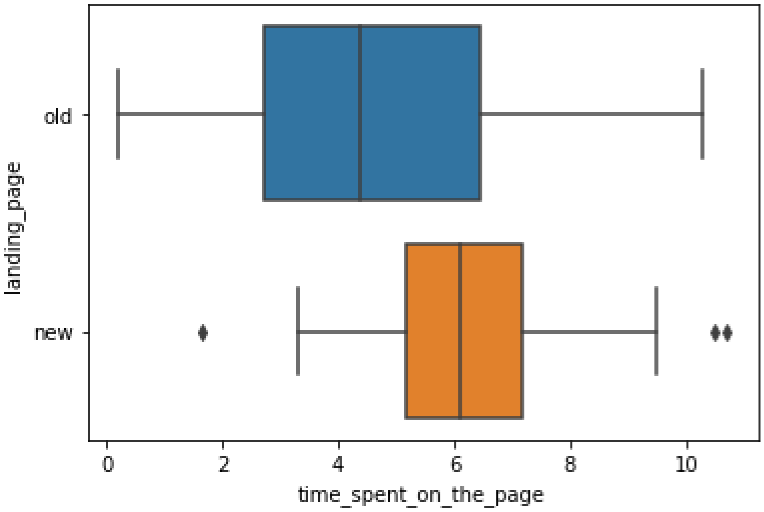
In [197…

```
#plt.figure(figsize=(10,5))
#sns.scatterplot(data=df,x='time_spent_on_the_page',y='group')

#plt.show()

#plt.figure(figsize=(15,7))
#sns.lineplot(data=df,x='time_spent_on_the_page',y='group')
#plt.show()

sns.boxplot(data=df,x='time_spent_on_the_page',y='landing_page')
plt.show()
```

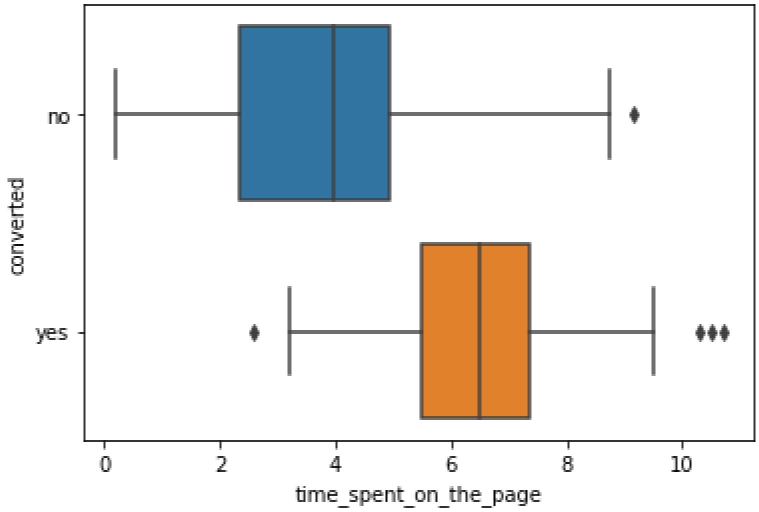


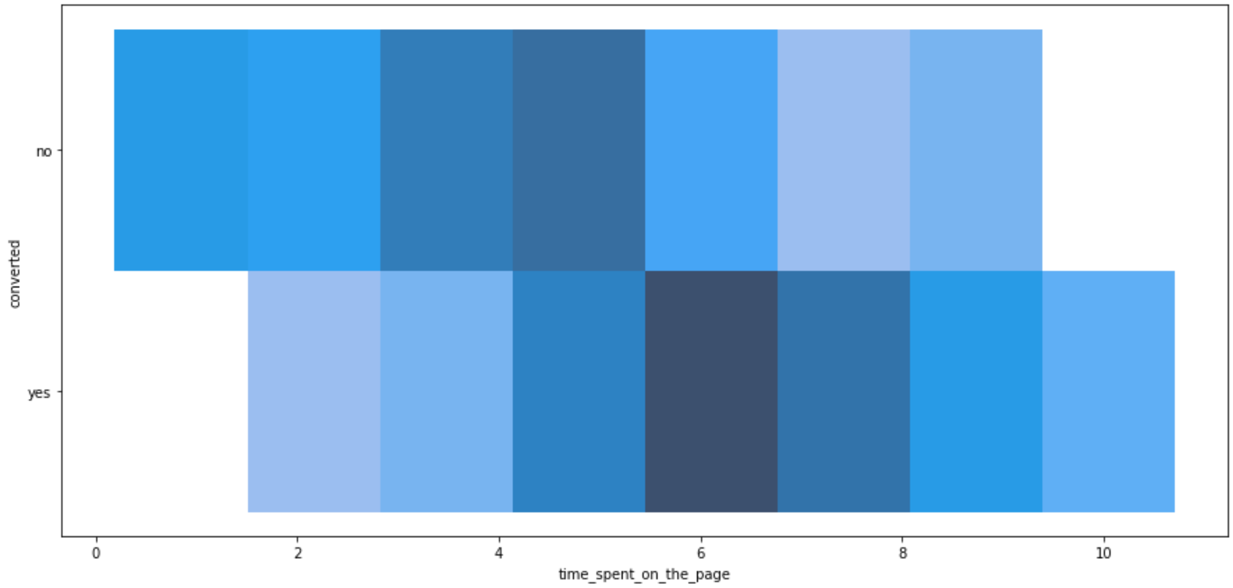
The users spent more time on the new landing page than the old landing page.

In [200…

```
sns.boxplot(data=df,x='time_spent_on_the_page',y='converted')
plt.show()
#sns.countplot(data=df,x='time_spent_on_the_page',hue='converted')
#plt.xticks(rotation=90)
```

```
#plt.show()
plt.figure(figsize=(15,7))
sns.histplot(data=df,x='time_spent_on_the_page',y='converted')
plt.show()
```

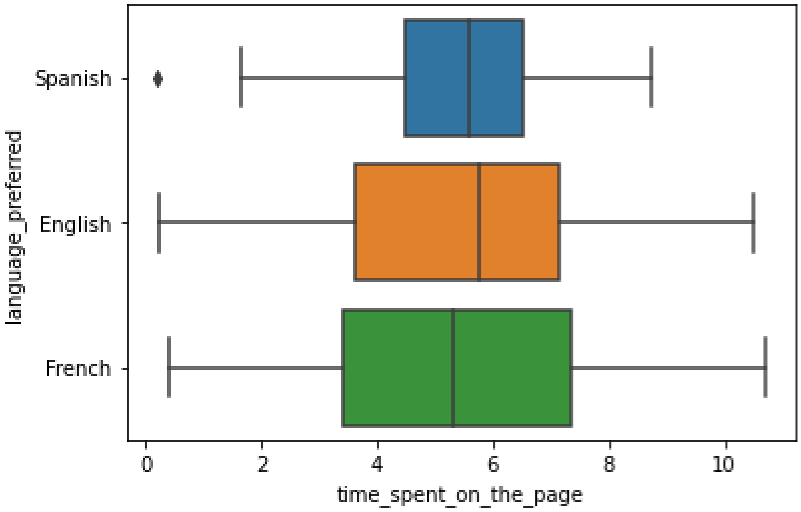


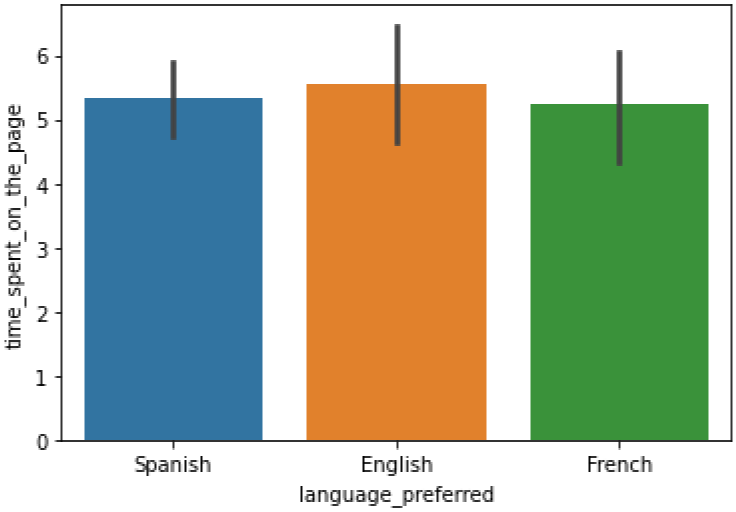


The users spending more time on the landing page, have a slightly higher conversion to subscriber.

In [201…

```
sns.boxplot(data=df,x='time_spent_on_the_page',y='language_preferred')
plt.show()
sns.barplot(data=df,x='language_preferred',y='time_spent_on_the_page')
plt.show()
```

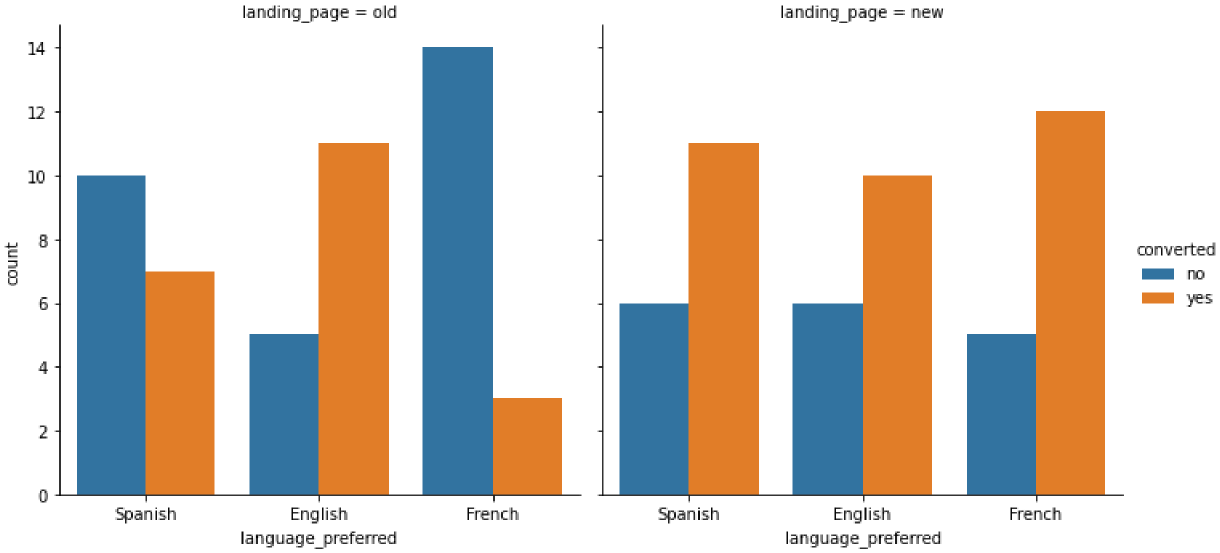




The time spent on the page based on the language preference is relatively similar.

In [202…

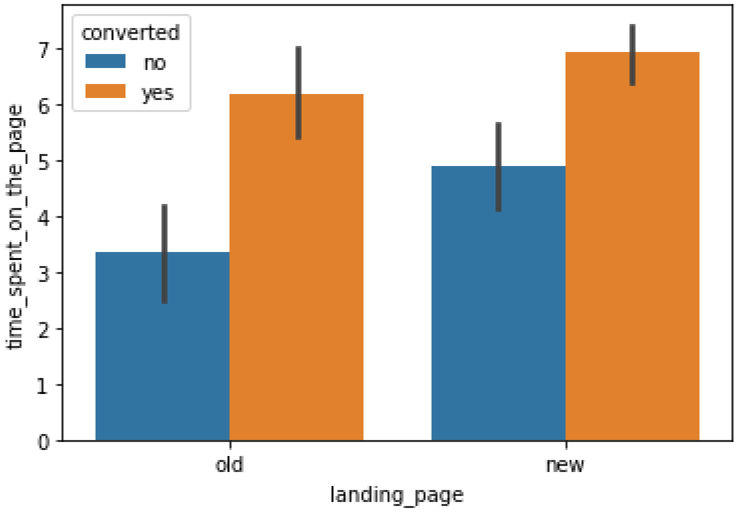
```
sns.catplot(data=df,x='language_preferred',hue='converted',col='landing_page',kind ='c
plt.show()
```



New landing page has been popular with users preferring French and has a higher probability of increasing the customer base in French.

In [203…

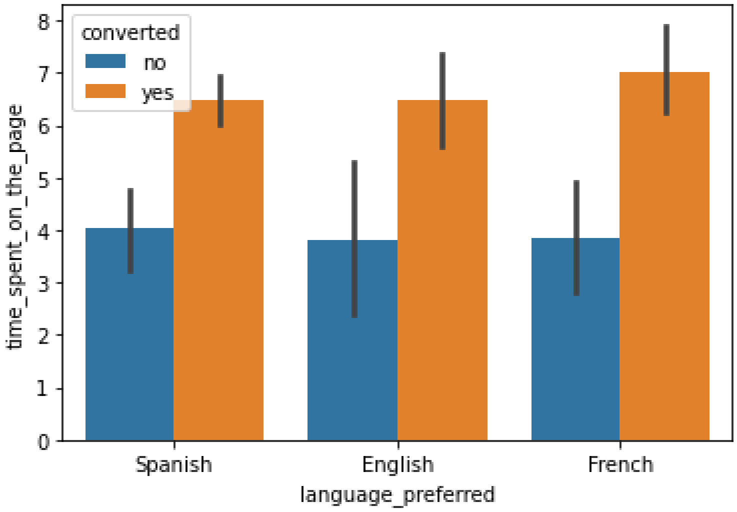
```
sns.barplot(data=df,x='landing_page',y='time_spent_on_the_page',hue='converted')
plt.show()
```



More users became subscribers when exposed to the new landing page.

In [204…

```
sns.barplot(data=df,x='language_preferred',y='time_spent_on_the_page',hue='converted')
plt.show()
```



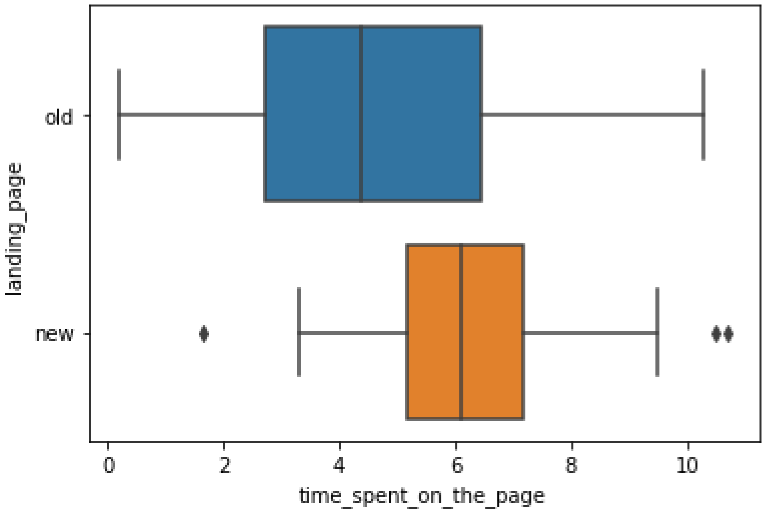
Users spending more time on the page irrespective of the language have a higher conversion rate.

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

Perform Visual Analysis

In [205…

```
sns.boxplot(data=df,x='time_spent_on_the_page',y='landing_page')
plt.show()
```



In [206…

```
df.groupby(['landing_page'])['time_spent_on_the_page'].mean()
```

Out[206]:

```
landing_page
new    6.2232
old    4.5324
Name: time_spent_on_the_page, dtype: float64
```

Users appear to be spending more time on the new landing page than the old landing page.

Step 1: Define the null and alternate hypotheses

H0 : Null Hypothesis - The mean time spent on the new landing page is the same as the mean time spent on the old landing page.

Ha : Alternate Hypothesis - The mean time spent on the new landing page is higher than the mean time spent on the old landing page.

Let µ1 and µ2 be the mean times spent on the new and old landing pages.

H0 : µ1 = µ2

Ha : µ1 > µ2

Step 2: Select Appropriate test

In [207…

```
df_new = df[df["landing_page"]=="new"]
df_old = df[df["landing_page"]=="old"]

µ1=df_new["time_spent_on_the_page"].mean()
µ2=df_old["time_spent_on_the_page"].mean()

std1=df_new["time_spent_on_the_page"].std()
std2=df_old["time_spent_on_the_page"].std()

print("µ1=",round(µ1,2))
print("μ2=",round(μ2,2))

print("std1=",round(std1,2))
print("std2=",round(std2,2))
```

```
μ1= 6.22
μ2= 4.53
std1= 1.82
std2= 2.58
```

Continuous data - Yes, the time spent on the page is measured on a continuous scale.

Normally distributed populations - Yes, the populations are assumed to be normal.

Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.

Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.

Random sampling from the population - Yes, the collected sample is a random sample.

We would use Two Independent Sample T-test for Equality of Means - Unequal Std Dev

Step 3: Decide the significance level

We can select a = 0.05.

Step 4: Collect and prepare data

This has already been done in Step 2

Step 5: Calculate the p-value

In [208…

```
from scipy.stats import ttest_ind

# find the p-value
test_stat, p_value = ttest_ind(df_new["time_spent_on_the_page"],df_old["time_spent_on_
print('The p-value is ', p_value)
```

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

Step 6: Compare the p-value with α

The p-value of 0.00013 is significantly less than a which is 0.05.

Step 7: Draw inference

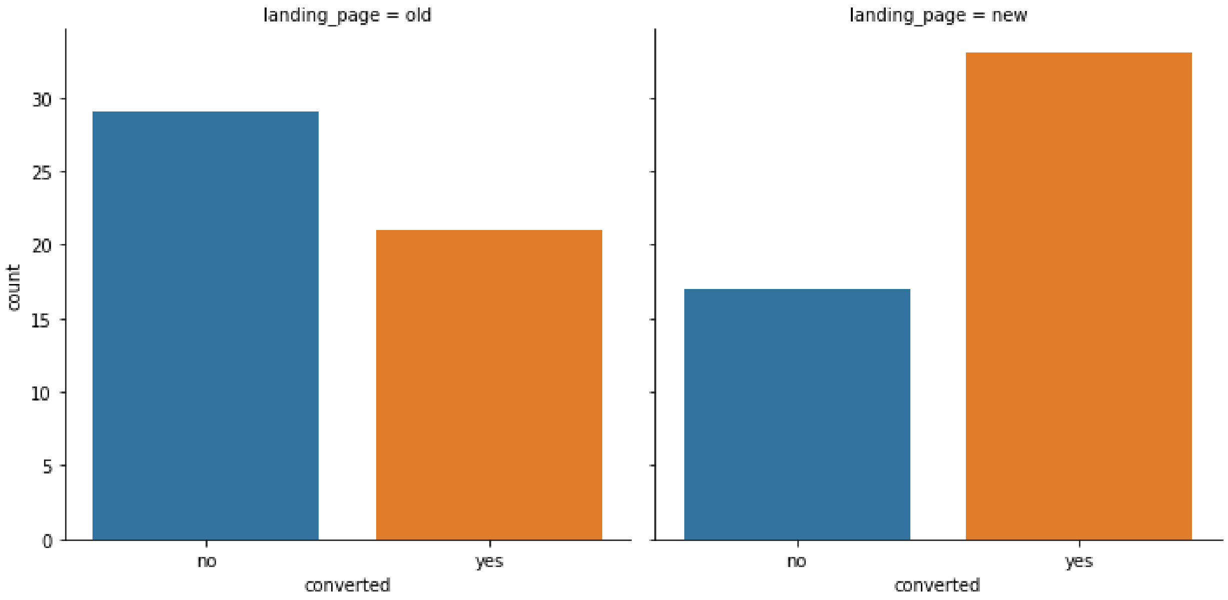
If p-value is less than the level of significance, we reject the null hypothesis.

There is enough evidence to support that users are spending more time on the new 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?

In [209…

```
sns.catplot(data=df,x='converted',col='landing_page',kind='count');
```



Users visiting the new landing page have a higher probability of subscribing.

In [210…

```
df_new_conv = df_new[df_new["converted"]=="yes"].count()
df_old_conv = df_old[df_old["converted"]=="yes"].count()

print(df_conv)
print(df_old_conv)
```

```
user_id                 33
group                   33
landing_page            33
time_spent_on_the_page  33
converted               33
language_preferred      33
dtype: int64
user_id                 21
group                   21
landing_page            21
time_spent_on_the_page  21
converted               21
language_preferred      21
dtype: int64
```

H0 : Null Hypothesis - The conversion rate (the proportion of users who visit the landing page and get converted) for the new page is the same as the conversion rate for the old page.

Ha : Alternate Hypothesis - The conversion rate for the new page is greater than the conversion rate for the old page.

Let p1 and p2 be the proportion of conversion on the new and old landing pages.

H0 : p1=p2

Ha : p1>p2

The level of significance is 0.05

Binomally distributed population - Yes, a user 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. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10. Here, n and p refer to the sample size and sample proportion respectively.

np1=50*(33/50) : 33>=10

n(1-p1) = 50 * ((50-33)/50) : 17 >=10

np2=50*(21/50) : 21>=10

n(1-p2) = 50 * ((50-21)/50) : 29 >=10

We would use proportions z-test for the problem

In [211…

```
from statsmodels.stats.proportion import proportions_ztest

# set the counts of defective items
converted_count = np.array([33,21])

# set the sample sizes
nobs = np.array([50,50])

# find the p-value
test_stat, p_value = proportions_ztest(converted_count, nobs)
print('The p-value is ' + str(p_value))
```

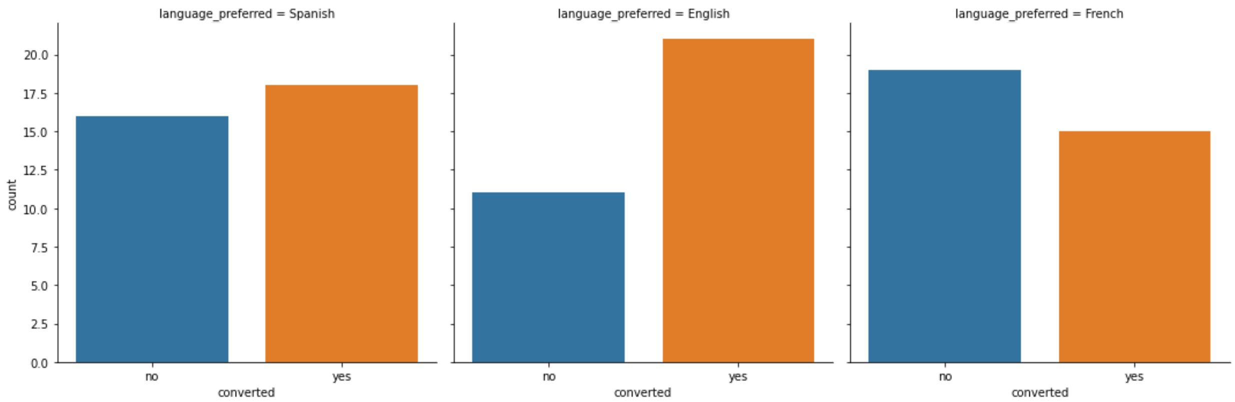
```
The p-value is 0.016052616408112556
```

As the p-value 0.016 is less than the level of significance, we reject the null hypothesis. We have enough evidence to suggest that conversion on new landing page is higher than the conversion on old landing page.

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

In [212…

```python
sns.catplot(data=df,x='converted',col='language_preferred',kind='count');
```



English users appear to have a higher conversion than French or Spanish users.

In [213…

```python
df1=pd.crosstab(df['converted'],df['language_preferred'])
df1
```

Out[213]:

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

H0 : Null Hypothesis - Converted status is independent of the preferred language.

Ha : Alternate Hypothesis - Converted status is dependent on the preferred language.

The level of significance is 0.05

- Categorical variables - Yes
- Expected value of the number of sample observations in each level of the variable is at least 5 - Yes, the number of observations in each level is greater than 5.
- Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.

We will perform chi square test of independence.

In [214…

```python
from scipy.stats import chi2_contingency

# find the p-value
chi, p_value, dof, expected = chi2_contingency(df1)
print('The p-value is', p_value)
```

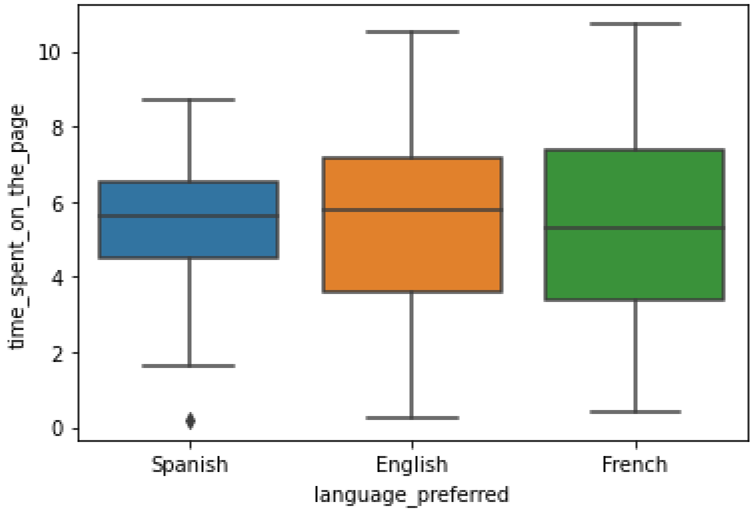
```
The p-value is 0.2129887487543447
```

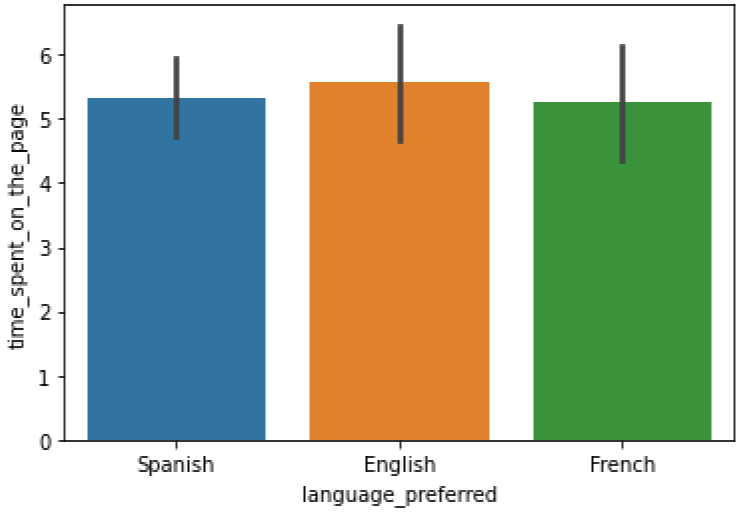
As the p-value is higher than level of significance we fail to reject the null hypotheses. Hence there is not enough evidence to suggest that the conversion status depends on the preferred language.

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

In [217…

```
sns.boxplot(data=df,x='language_preferred',y='time_spent_on_the_page')
plt.show()
sns.barplot(data=df,x='language_preferred',y='time_spent_on_the_page')
plt.show()
```





In [218…

```
df_new.groupby(['language_preferred'])['time_spent_on_the_page'].mean()
```

Out[218]:

```
language_preferred
English    6.663750
French     6.196471
Spanish    5.835294
Name: time_spent_on_the_page, dtype: float64
```

There is slight difference on the time spent for users of different language, but it is not very significant difference.

We would use ANOVA test as there are more than 2 samples.

Let μ1, μ2, μ3 be the mean time spent on the new landing page for the different language users.

μ1 - Spanish

μ2 - English

μ3 - French

H0: μ1 = μ2 = μ3

Ha: Atleast one of μ1, μ2, or μ3 is different

Now, the normality and equality of variance assumptions need to be checked.

```
For testing of normality, Shapiro-Wilk’s test is applied to the
response variable.

For equality of variance, Levene test is applied to the response
variable.
```

Shapiro-Wilk’s test

We will test the null hypothesis

> H_0 : Time Spent follows a normal distribution

against the alternative hypothesis

> H_a : Time Spent does not follow a normal distribution

In [219…

```python
w, p_value = stats.shapiro(df_new['time_spent_on_the_page'])
print('The p-value is', p_value)
```

```
The p-value is 0.8040016293525696
```

As p-value is significantly higher, we fail to reject the null hypotheses that a normal distribution is followed.

Levene’s test

We will test the null hypothesis

```
H0    : All the population variances are equal
```

against the alternative hypothesis

```
Ha    : At least one variance is different from the rest
```

In [220…

```python
from scipy.stats import levene
statistic, p_value = levene(df_new['time_spent_on_the_page'][df_new['language_preferre
                            df_new['time_spent_on_the_page'][df_new['language_preferre
                            df_new['time_spent_on_the_page'][df_new['language_preferre
```

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

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

As p-value is significantly higher, we fail to reject the null hypotheses that all population variances are equal.

- The populations are normally distributed - Yes, the normality assumption is verified using the Shapiro-Wilk’s test.
- Samples are independent simple random samples - Yes, we are informed that the collected sample is a simple random sample.
- Population variances are equal - Yes, the homogeneity of variance assumption is verified using the Levene's test.

In [221…

```
from scipy.stats import f_oneway

# perform one-way anova test
test_stat, p_value = f_oneway(df_new.loc[df_new['language_preferred']=='Spanish','time
                              df_new.loc[df_new['language_preferred']=='English','time
                              df_new.loc[df_new['language_preferred']=='French','time_
print('The p-value is ' + str(p_value))
```

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

As p-value is significantly higher than level of significance, we fail to reject the null hypotheses that all means are equal. We do not have enough evidence to suggest that atleast of the mean is different.

Conclusion

1) The newer landing page has a higher conversion rate than the older page

2) The newer page has a higher time spent than the older page.

3) There preference of language is not directly related to time spent on the page.

4) There preference of language is not directly related to the conversion rate.

Business Recommendations

1) The new landing page is working. Adding few more language options may increase the number of users

2) An advertisement strategy needs to be formulated to increase the customer base as the new landing page is quite engaging.

3) Some offers can be rolled out for new users that would help increase conversion/subscription.