Enews Express

Business Statistics

19th August 2022

Contents / Agenda

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

Executive Summary

E-news Express, an online news portal wants to expand its business by acquiring new subscribers and their design team has researched and created a new landing page which has a new outline and more relevant content shown compared to the old page. An experiment was further conducted on two randomly selected groups to monitor the interaction of these users with the new landing page and old existing landing page. I have been provided with data and after thorough exploration and statistical analysis of the data, the following are my deductions:

- From the sample data which contains 100 observations made by the control group (old page users) and treatment group (new page users) we observed that the users spent more time on the new landing page than the old landing page.

- We observed that from our 100 users that 54 users converted to subscribers and 46 did not.

- The users that converted to subscribers (yes) spent an average of about 6.2 minutes on the page while the users who didn't convert spent an average of around 4 minutes.

Executive Summary

- The conversion rate for users on the new landing page is greater than conversion rate of users to subscribers the old landing page.

- The conversion status from users to subscribers was not dependent on any preferred language.

- From our visual analysis it seems that users who preferred English language spent a longer time than users who preferred French and Spanish languages, however after carrying out the analysis of variance test, I had to conclude that the same average time was spent by all users regardless of language preferred, bearing in mind that the difference in time spent was not more than a minute difference.

Executive Summary

Conclusions and Recommendations

- It is glaring that the A/B technique experiment conducted by the Data Science team was quite successful. They have been able to test the effectiveness of the new landing page by seeing how their online users have responded to it, this is evidenced by the longer period spent on the new landing page and the rate at which users subscribed when compared to the old existing page.

- As this experiment has proved to be successful on a sample size, E-news Express can take bolder steps in designing a more improved and engaging landing page to drive better engagement amongst users such that more users will convert to subscribers and boost the business further.

Business Problem Overview and Solution Approach

Objectives

Enews Express wants to expand its business by acquiring new subscribers and to achieve this they have successfully carried out an experiment on their users' interaction on both their old landing page and their new landing page. As a data analyst, my responsibility is to explore the dataset, analyse it and address the following questions below:

- Whether users spend more time on the new landing page than on the old landing page?
- Comparing if conversion rate for the new page is greater than that for the old page?
- To ascertain the relationship between converted status and the preferred language?
- Analyse if time spent on the new page is same for the different language users?

After exploring, analysing the dataset and providing answers to the concerns raised by Enews Express. The company plans to act on my actionable insights and conclusion to further understand their user interests and determine how to drive a better engagement and expand their business successfully.

Business Problem Overview and Solution Approach

Methodology

To explore the dataset and extract insights using Exploratory Data Analysis, we carried out the following steps:

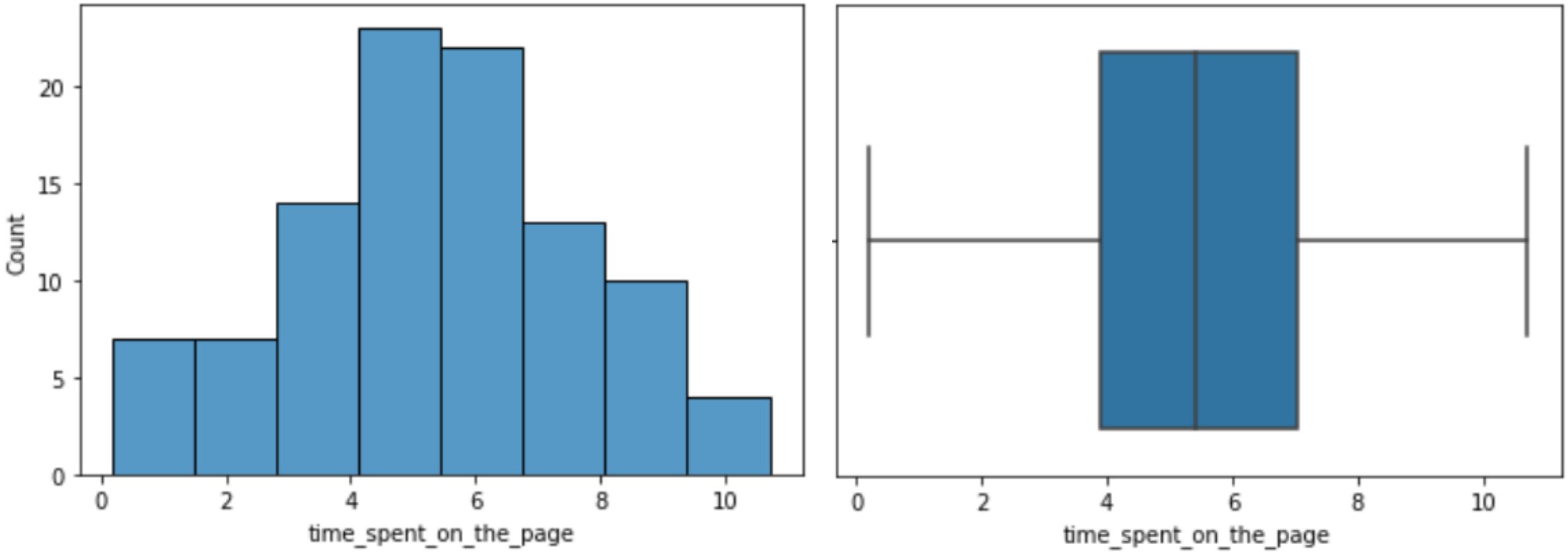
- Imported the necessary libraries into our Google Colab
- Uploaded the E-news Express dataset.
- Carried out sanity checks on our dataset to ensure data was loaded properly.
- Confirmed the absence of missing values.
- Confirmed the absence of duplicated values.
- Carried out Exploratory Data Analysis.
- Carried out hypothesis testing.
- Made inferences based on our hypothesis result.

EDA Results

- Display data head to view the first few rows to ascertain if data has been loaded properly or not.
- The dataset has 100 rows and 6 columns (4 categorical and 2 numerical variables).
- There are no missing values evidenced by presence of only non-null counts.
- There are no duplicates.
- The dataset set comprises 1 float, 1 integer and 4 object data types.
- Memory usage is at 4.8+ KB.
- Statistical summary of numerical variables shows that mean time spent on the webpage is about 5.4 minutes, maximum time spent at 10.71 minutes while least time spent was 0.2 minutes.
- Value counts of the categorical variable show:
  - Landing page is divided equally into old (50%) and new(50%) landing pages.
  - The user group is divided equally into Control (50%) and Treatment (50%)
  - Language preferred shows the following: Spanish-34 users, French-34 users and 32 users.
  - Users who converted to subscribers are 54 (Yes) and 46(No).

EDA Results

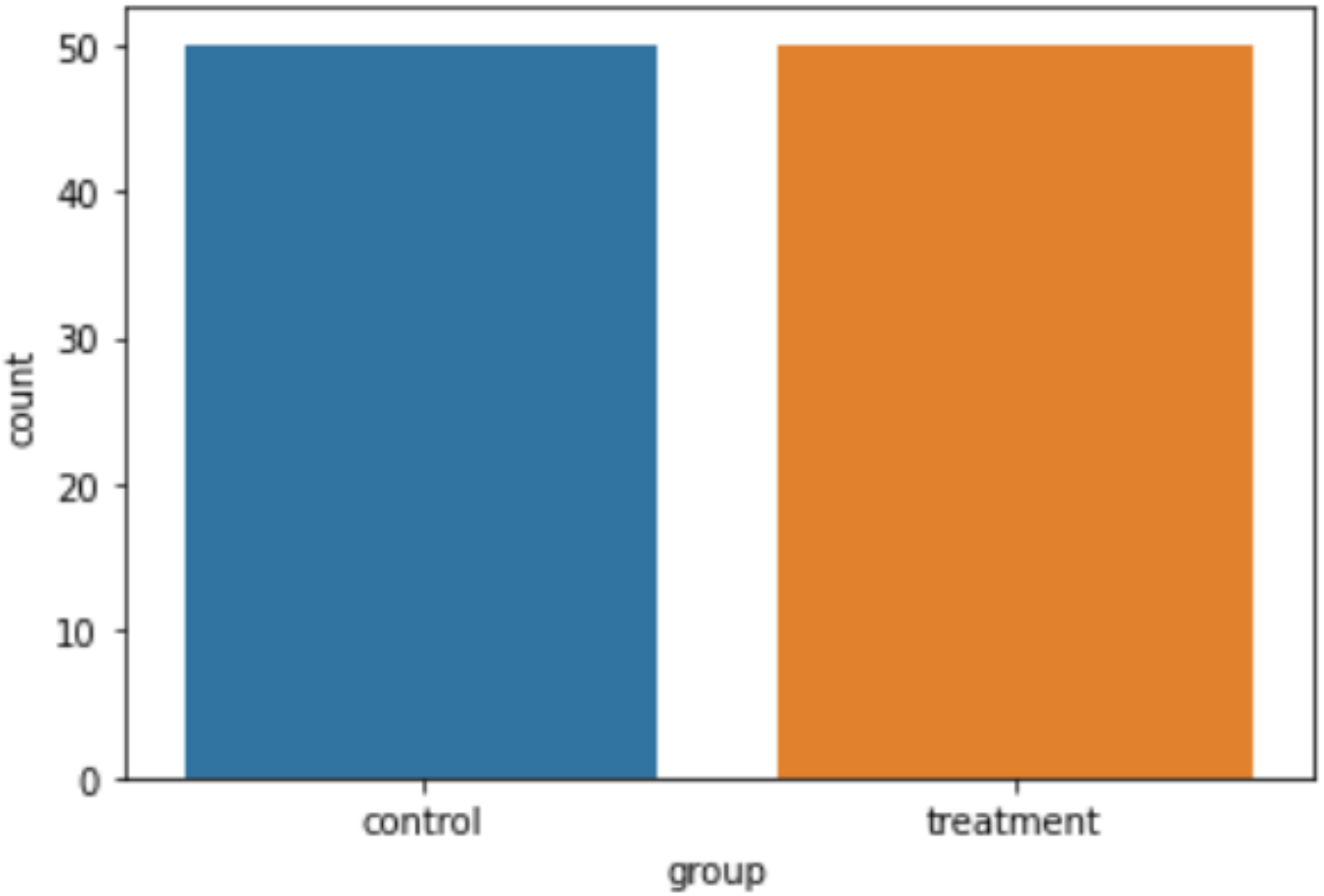
Univariate Analysis - **Time spent on the page**



- This is a normal distribution with no evidence of outliers,
- The longest time spent on the page is around 10.7 minutes while shortest time spent is around 0.2 minutes.

EDA Results

Univariate Analysis- Group



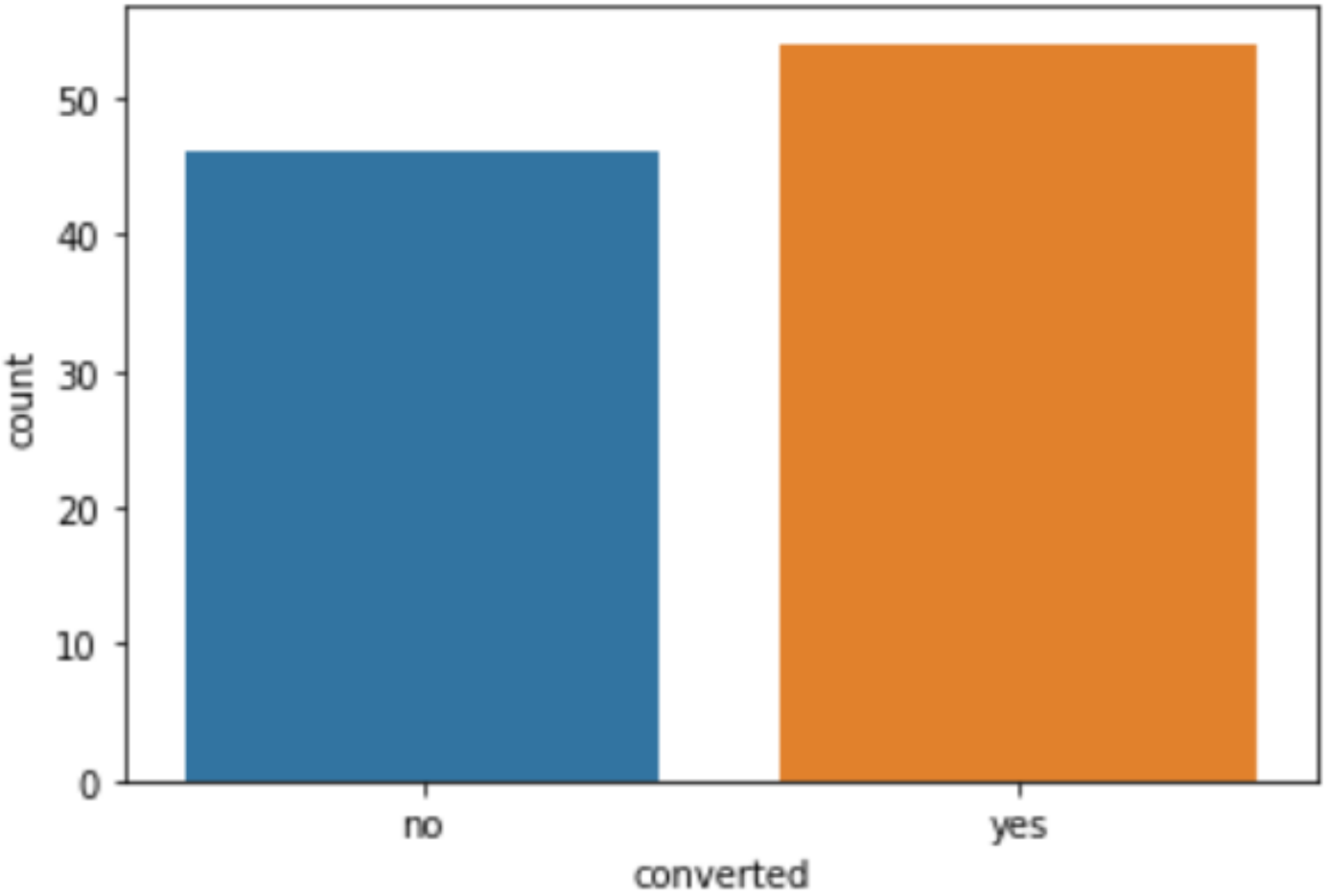
The user group is divided into

- Control
- Treatment

Link to Appendix slid on data background check

EDA Results

Univariate Analysis- Conversion status

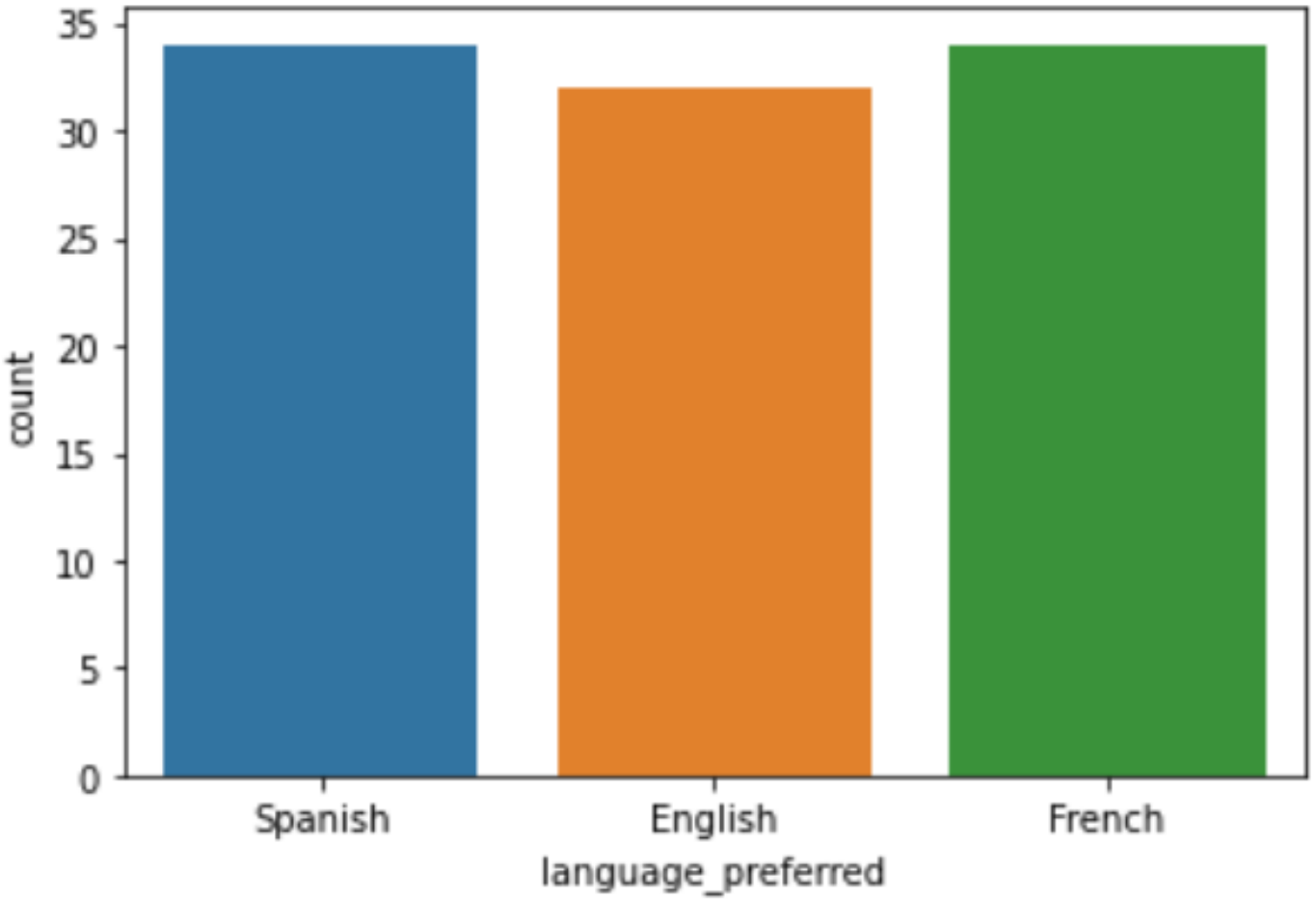


The users who converted to subscribers:

- No (did not convert)
- Yes (converted)

EDA Results

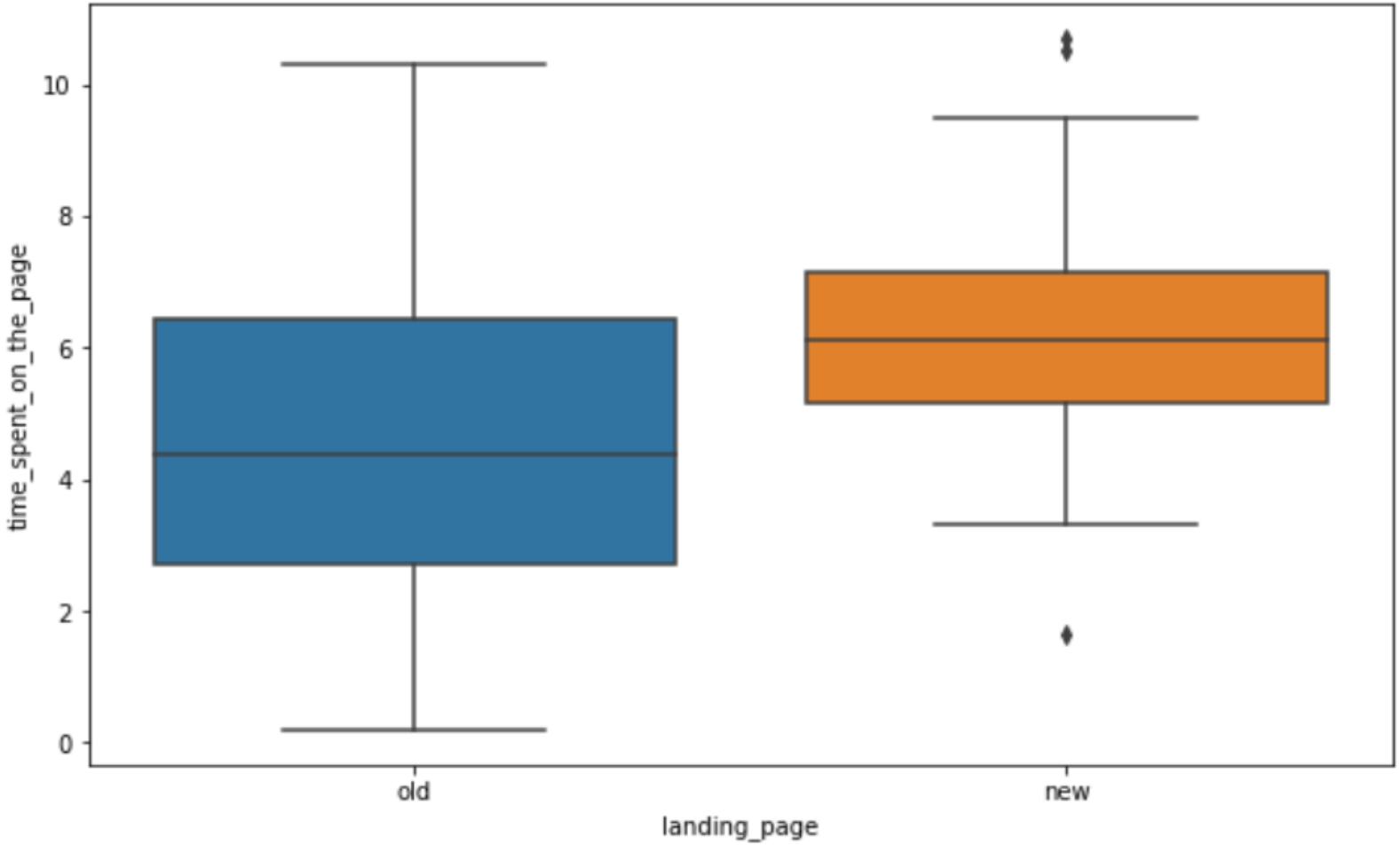
Univariate Analysis- Language preferred by users



- The distribution shows that French and Spanish are the more preferred languages.

EDA Results

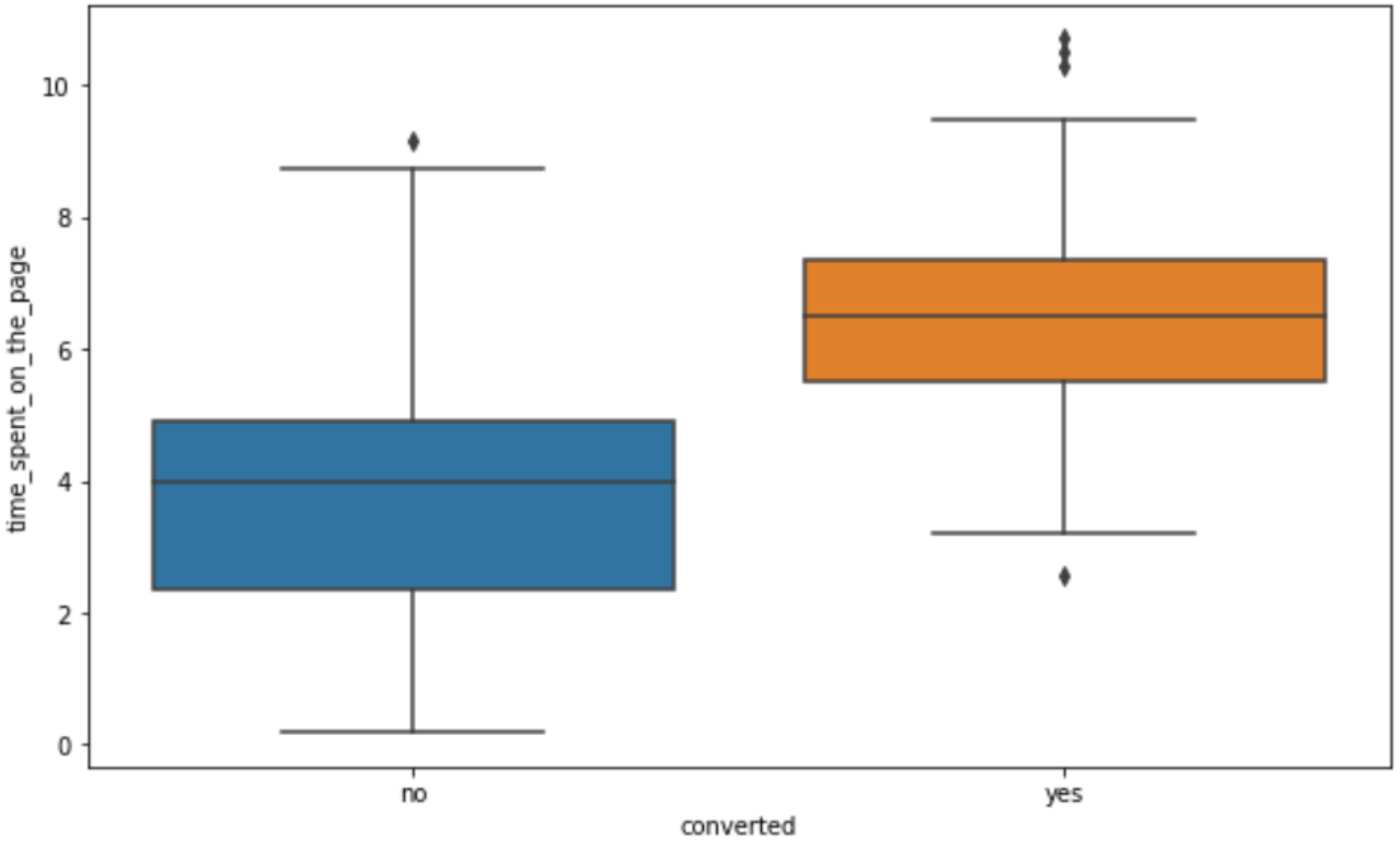
Bivariate Analysis- Time spent on the page vs. Landing page



- The plot shows that the new landing page has some outliers on both sides.
- The average time spent on the new landing page is about 6 minutes while average time spent on old landing page is around 4.5 minutes.

EDA Results

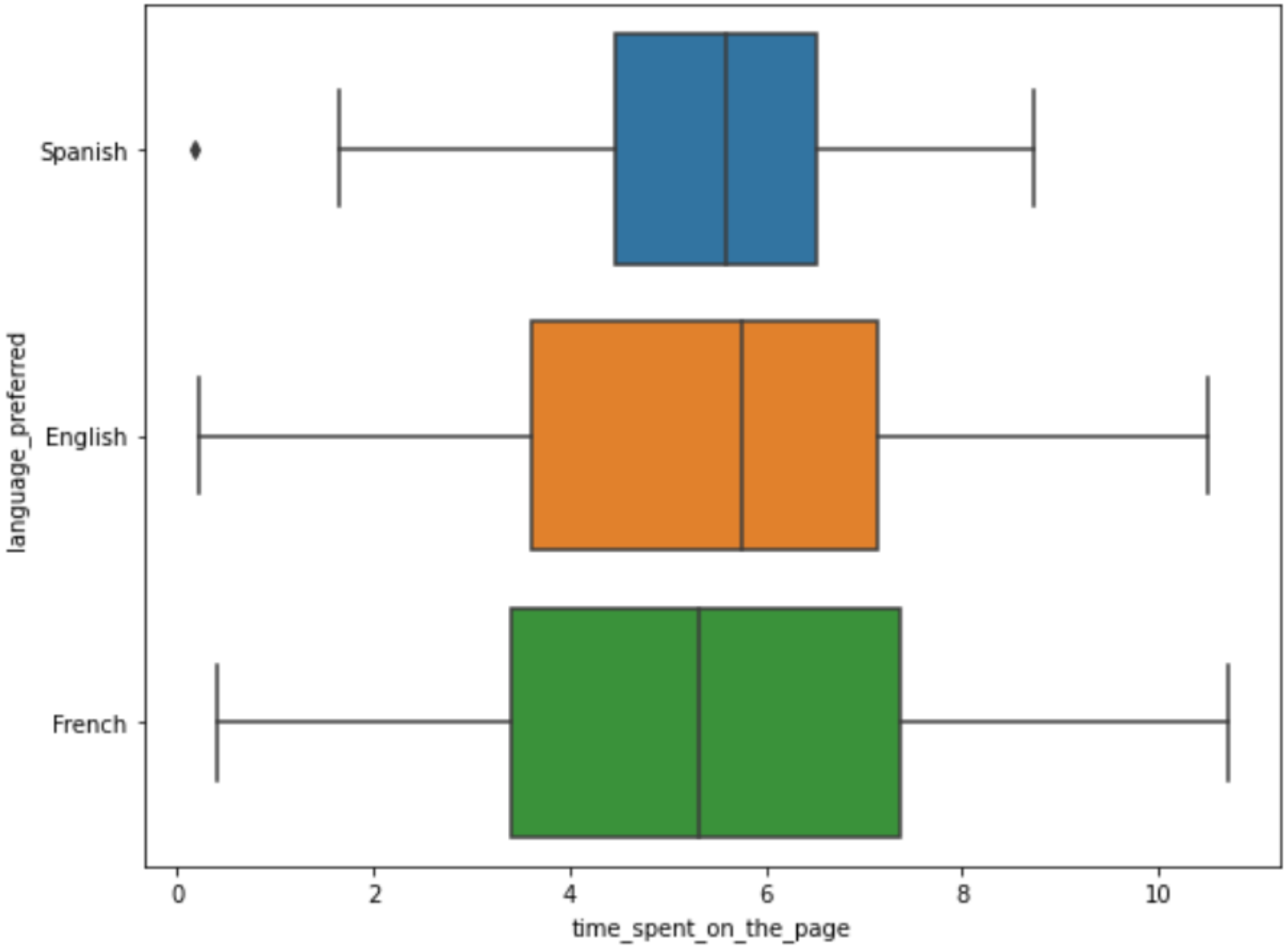
Bivariate Analysis- Conversion status vs. Time spent on page



- The users that converted to subscribers (yes) spent an average of about 6.2 minutes on the page while the users who didn't convert spent an average of around 4 minutes.
- The distribution shows that converted new subscribers spent a longer time than old subscribers.

EDA Results

Bivariate Analysis- Preferred Language vs. Time spent on page

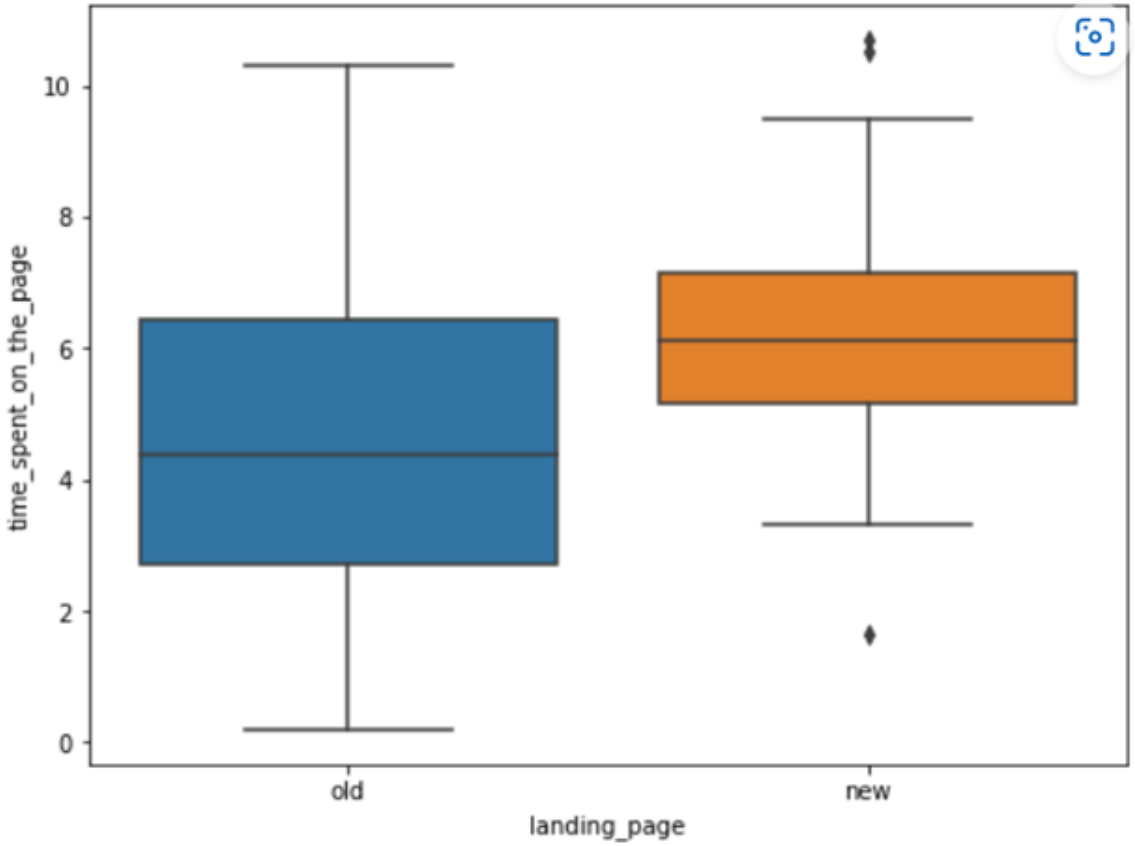


- From our distribution it seems that the users who prefer English language spent longer average time than both users who prefer French and Spanish languages.
- Users who prefer Spanish data has some outliers.

Hypotheses Tested and Results

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

Time spent on the page vs. Landing page



- Hypothesis Tested**:** Test to compare two sample means from two independent populations when standard deviation is unknown- **2 Sample Independent t-test.**
- The p-value is **0.00013923812251665 49** and is less than the level of significance (α),

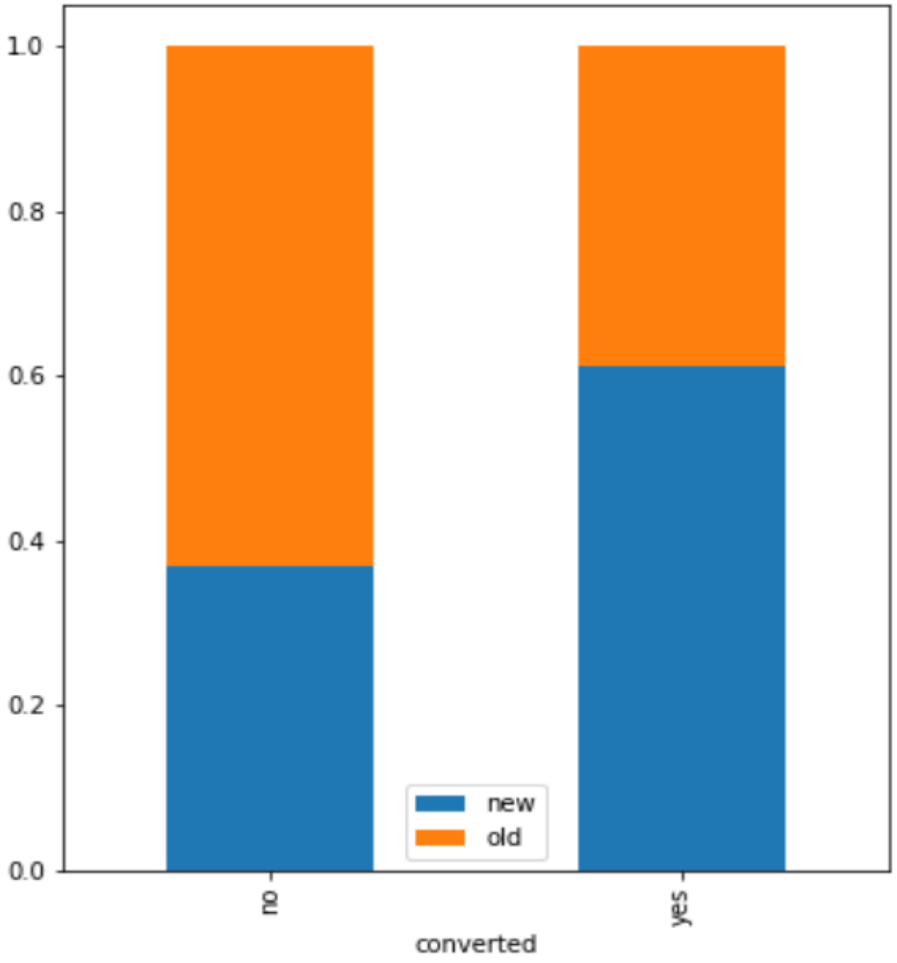
Inference

- Therefore, I have enough statistical evidence to reject the null hypothesis and support the claim that users spend more time on the new landing page than the old existing landing page.

Link to Appendix slide on details of the test performed

Hypotheses Tested and Results

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?



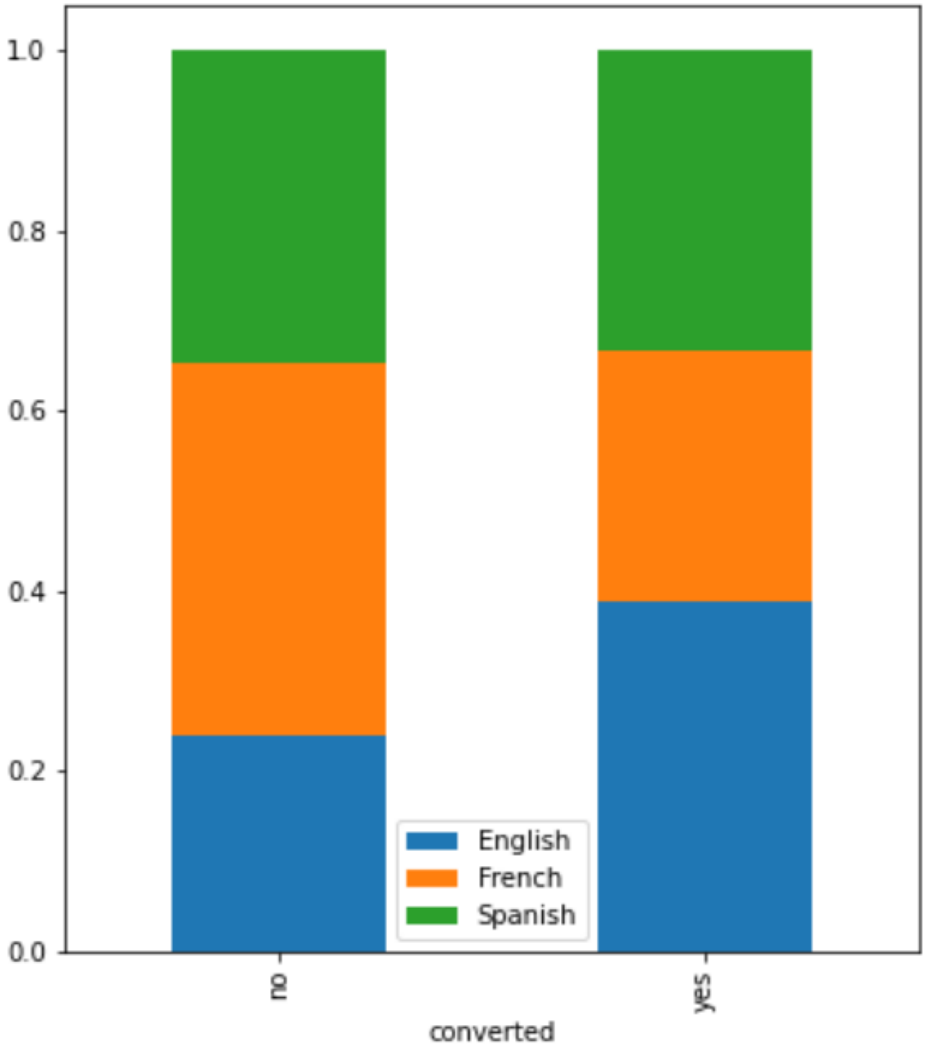
- Visually comparing the conversion rate for the new page and the conversion rate for the old page.
- Hypothesis Tested: Test to compare two sample proportions from two populations-**2 Sample Proportion Z test**
- The numbers of users served the new and old pages are 50 and 50 respectively.
- The p-value is **0.008026308204056278** and is less than level of significance (α).

Inference

- Therefore, I have enough statistical evidence to reject the null hypothesis and support the claim that conversion rate for the new page is greater than conversion rate for the old page.

Hypotheses Tested and Results

Does the converted status depend on the preferred language?



- View of the plot which shows the dependency between conversion status and preferred language.
- Visually it seems that about 40% of users who speak French did not convert (no category), while 40% of users who speak English converted (yes category).
- Hypothesis Tested: Test to check dependence relationship between two categorical variables - **Chi-square Test for Independence**
- The p-value is **0.21298887487543447**

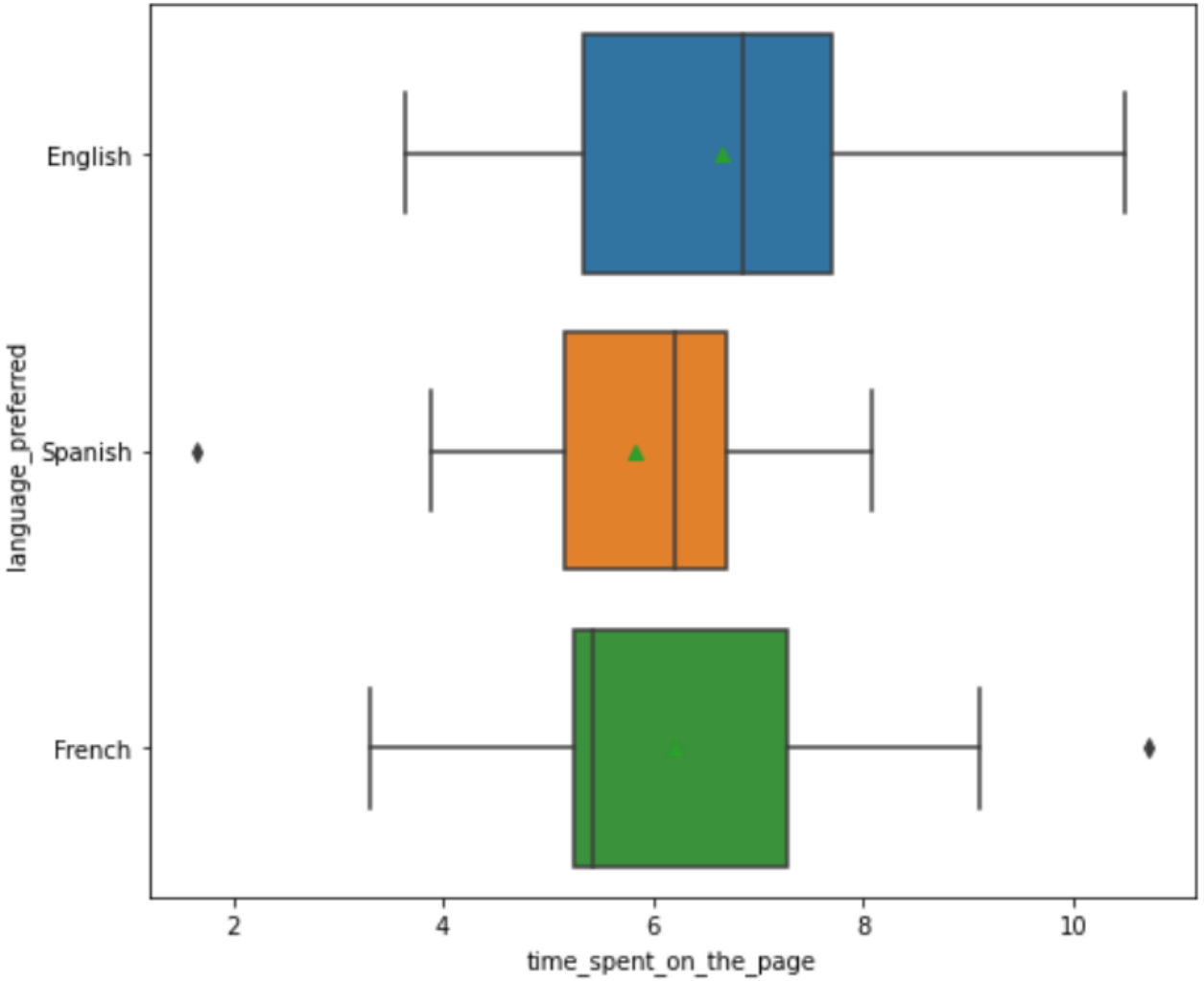
Inference:

- Since p-value is greater than the level of significance (α), I fail to reject the null hypothesis because I don't have enough statistical evidence to say that converted status is dependent on the preferred language.

Link to Appendix slide on details of the test performed

Hypotheses Tested and Results

Is the time spent on the new page same for the different



- View of the plot which shows time spent on the new page by different language users.
- Hypothesis tested- Test to compare sample means from more than two independent populations-**One way Anova Test**
- The p-value is **0.4320413869432595 5**

Inference

- P-value is greater than our level of significance (α), I fail to reject the null hypothesis because I do not have enough statistical evidence to conclude otherwise, therefore we accept that the mean time spent on the new landing page is the same for all three preferred language users.

Link to Appendix slide on details of the test performed

APPENDIX

Data Background and Contents

```
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
```

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

Numerical Statistical Summary

df.describe().T

	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

Data Background and Contents

```
df['group'].value_counts()
```

```
control      50
treatment    50
Name: group, dtype: int64
```

```
df['converted'].value_counts()
```

```
yes    54
no     46
Name: converted, dtype: int64
```

```
df['landing_page'].value_counts()
```

```
old    50
new    50
Name: landing_page, dtype: int64
```

```
df['language_preferred'].value_counts()
```

```
Spanish    34
French     34
English    32
Name: language_preferred, dtype: int64
```

To find missing values

```
df.isnull().sum()
```

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

To find duplicated values

```
df.duplicated().sum
```

```
<bound method NDFrame._add_
1      False
2      False
3      False
4      False
       ...
95     False
96     False
97     False
98     False
99     False
Length: 100, dtype: bool>
```

Hypothesis Testing Details

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

- Hypothesis Test: **2 Sample Independent t-test**
- Null and Alternative Hypotheses:

μ_1:average time spent on new landing page

μ_2: average time spent on old landing page

H_0 : $\mu_1 = \mu_2$

Ha : $\mu_1 > \mu_2$

- p-value obtained: **0.00013923812251665 49**
- 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

Hypothesis Testing Details

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?

- Hypothesis Test: **2 Sample Proportion Z test**
- Null and Alternative Hypotheses:

p1: proportion of users who visit the landing page and get converted (conversion rate) for new page

p2: proportion of users who visit the landing page and get converted (conversion rate) for old page

H0 : p1 = p2

Ha : $\mu 1 > \mu 2$

- p-value obtained: **0.008026308204056278**
- The numbers of users served the new and old pages are 50 and 50 respectively

Hypothesis Testing Details

Does the converted status depend on the preferred language?

- Hypothesis Test: **Chi Square Test of independence**
- Null and Alternative Hypotheses:

H0: Converted status is independent of the preferred language

Ha: Converted status is dependent of the preferred language

- p-value obtained: **0.21298887487543447**

```
contingency_table
```

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

Hypothesis Testing Details

Does the converted status depend on the preferred language?

- Hypothesis Test: **One way Anova Test**
- Null and Alternative Hypotheses:

Let μ be the mean time spent on the new landing page

H_0: The mean time spent on the new landing page is the same for all three preferred language users.

Ha: At least one of the mean time spent on the new landing page differs from the other two preferred language users.

- p-value obtained: **0.4320413869432595**5

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

Hypothesis Testing Details

I carried out **Shapiro Wilk** and **Levene tests** to be absolutely sure that our distribution was normal and variance was homogenous.

```
# Assumption 1: Normality
# import the required function
from scipy import stats

# find the p-value
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
```

e p-value of the test is very large, we fail to reject the null hypothesis that the response follows the normal distribution.

```
#Assumption 2: Homogeneity of Variance
# levene function from scipy.stats library for this test

# find the p-value
statistic, p_value = stats.levene(df[df['diet']=='A']['weightloss'],
                                  df[df['diet']=='B']['weightloss'],
                                  df[df['diet']=='C']['weightloss'])
print('The p-value is', p_value)
```

```
The p-value is 0.5376731304274011
```

the p-value is large than the 5% significance level, we fail to reject the null hypothesis of homogeneity of variances.