E-news Express Data Analysis

Python Foundations : PGP-DSBA

Jan 27,2023

Contents / Agenda

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

Background

E - News Express is a media site.

The portal exhibits the contents in three different languages, depending on the reader's preference, namely English, French, and Spanish.

To consume the news content on a regular basis, users must subscribe to this portal. The company's design team has created a brand-new landing page with extra features in order to attract the attention of more new users.

The new landing page is expected to convert more users into new subscribers.

Business Problem Overview

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. They want to test the effectiveness of the new landing page in gathering new subscribers for the news portal by answering the following questions:

1. Do the users spend more time on the new landing page than on the existing 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?
3. Does the converted status depend on the preferred language?
4. Is the time spent on the new page the same for the different language users?Please mention the solution approach / methodology

Solution Approach

The goal of answering these questions is to determine whether the new page is effective in attracting new subscribers to the news portal.

To make a business decision, we will conduct statistical analysis on the collected data.

EDA - Data overview

```
# check the data types of the columns in the dataset
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
```

Observations

- There are 4 object datatype , 1 integer and 1 float.
- Every column has 100 observations

EDA - Data overview

```
# write your code here
df.isnull().sum()
```

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

Observations

- There are no missing values

EDA - Data overview

```
# write your code here to print the numeri
df.describe()
```

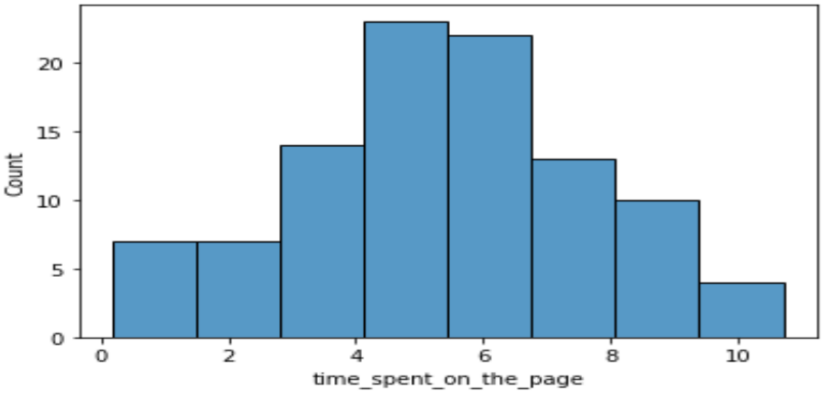
	user_id	time_spent_on_the_page
count	100.000000	100.000000
mean	546517.000000	5.377800
std	52.295779	2.378166
min	546443.000000	0.190000
25%	546467.750000	3.880000
50%	546492.500000	5.415000
75%	546567.250000	7.022500
max	546592.000000	10.710000

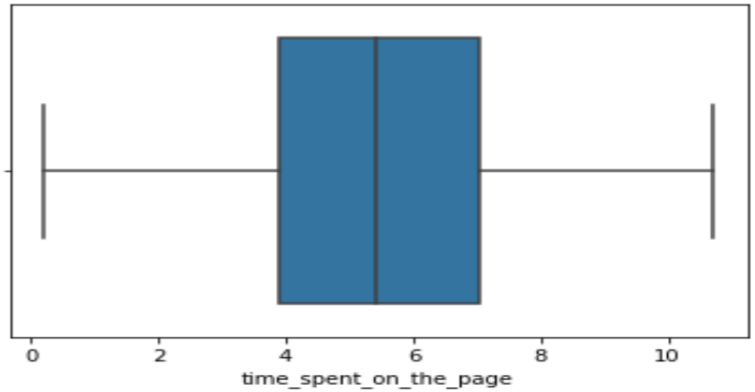
Observations

- Average time spent on the page is 5.37.
- Maximum time spent on the page is 10.71.
- Minimum time spent on the page is 0.19.

EDA – Univariate Analysis

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





Observations

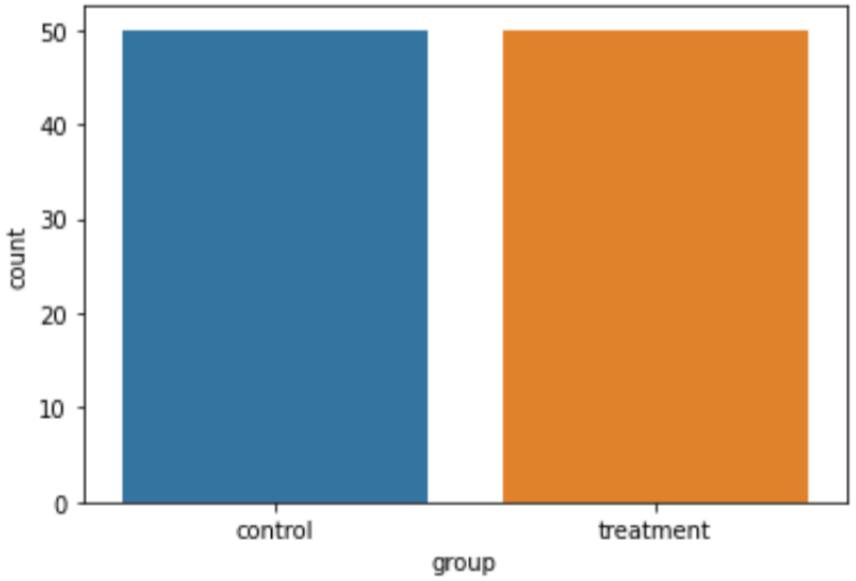
- The time spent on the page is approximately normally distributed
- There are no outliers in the time spent on page.
- The average or the mean is close to the median.

EDA – Univariate Analysis

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

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

```
sns.countplot(data=df,x='group')
plt.show()
```



Observations

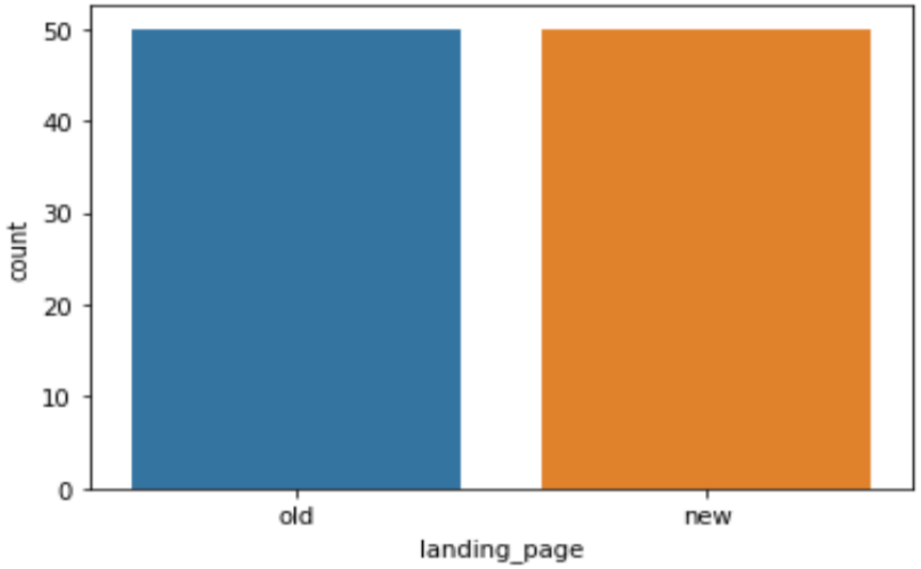
- Each group in control and tretment has 50 users.

EDA – Univariate Analysis

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

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

```
# complete the code to plot the countplot
sns.countplot(data=df,x="landing_page")
plt.show()
```



Observations

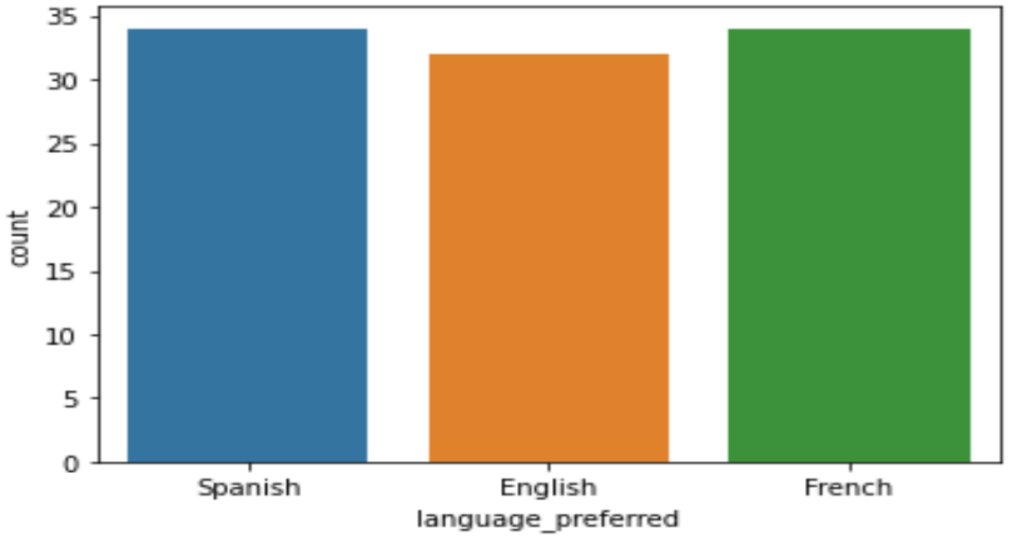
- There are two landing pages which are old and new.
- Both old and new has 50 users.

EDA – Univariate Analysis

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

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

```
# complete the code to plot the countplot
sns.countplot(data=df,x="language_preferred")
plt.show()
```



Observations

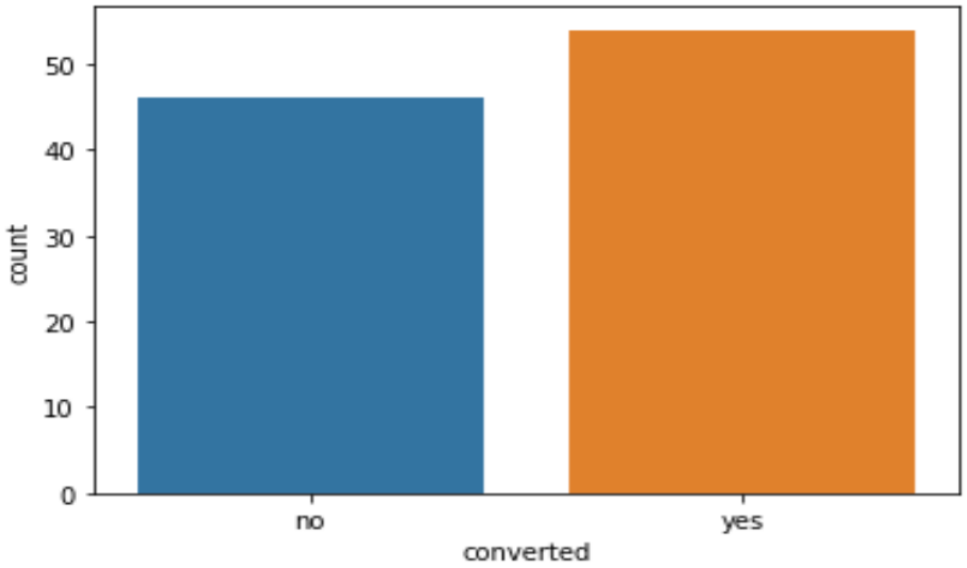
- The language preferred by the users are spanish, english and french.
- 34 users out of 100 sample prefer Spanish.
- 34 users out of 100 sample prefer French.
- 32 users out of 100 sample prefer English.

EDA – Univariate Analysis

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

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

```
# complete the code to plot the countplot
sns.countplot(data=df,x="converted")
plt.show()
```

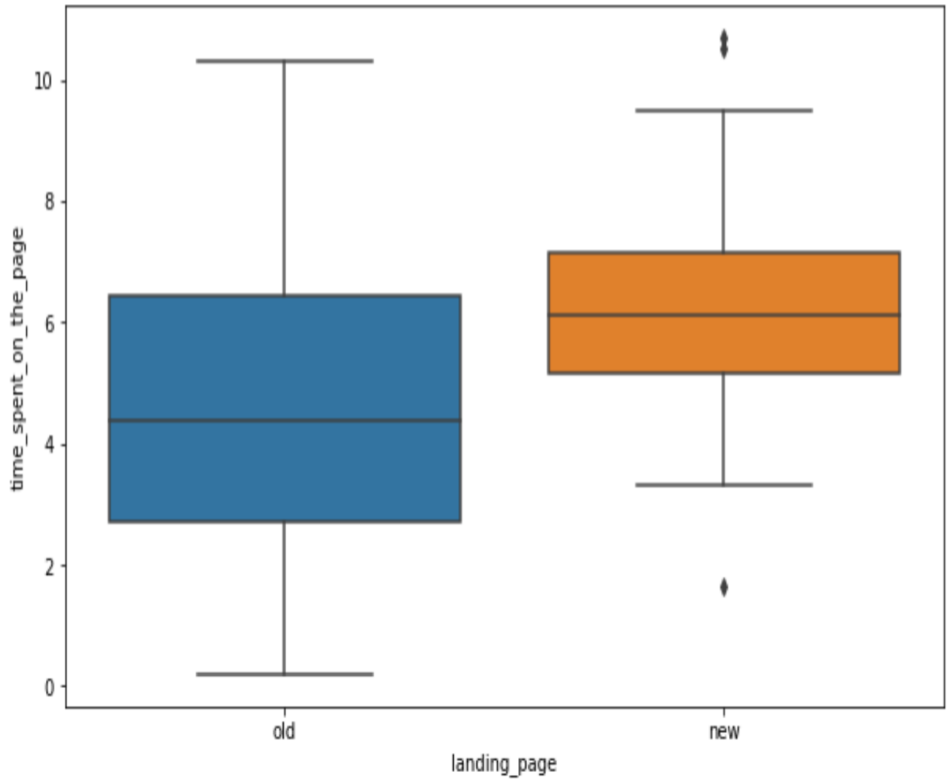


Observations

- from the graph , we can conclude that 54 users were converted as subscribers of the new landing page.
- Rest 46 were not converted.

EDA – Bivariate Analysis

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

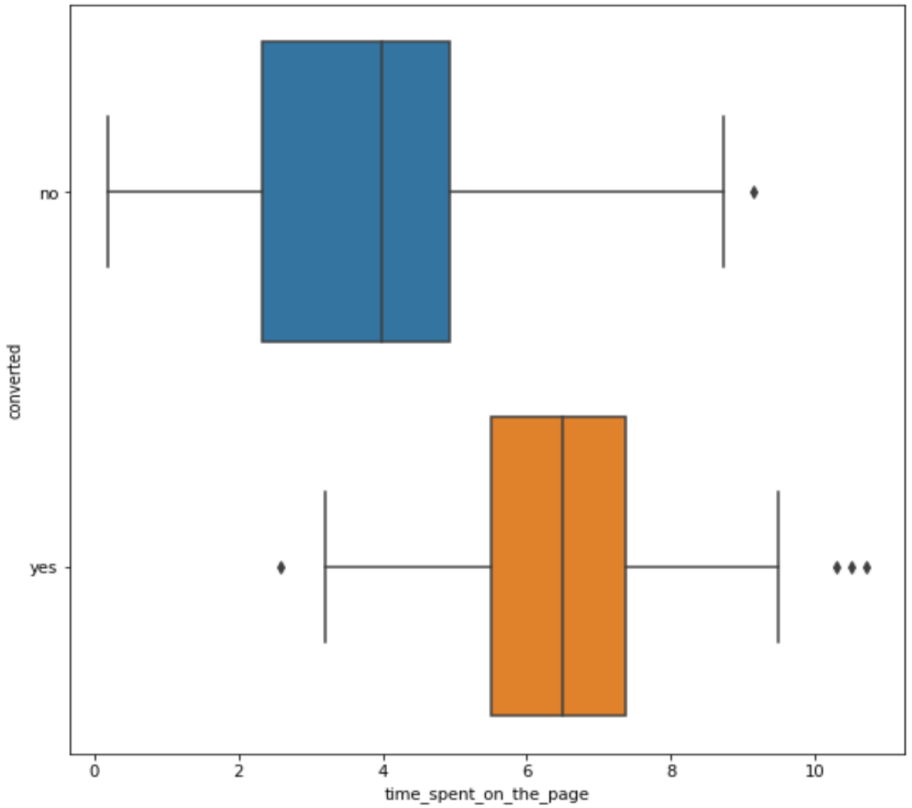


Observations

- Users on the new landing page spent average of 6 minutes on the page. There are outliers illustrating that some users spent more and less than quartiles.
- Users on the old landing page spent average of 4.22 minutes. There are no outliers.

EDA – Bivariate Analysis

```
# complete the code to plot a suitable graph to understand the relations
plt.figure(figsize=(9, 9))
sns.boxplot(data = df, x = "time_spent_on_the_page", y ='converted')
plt.show()
```



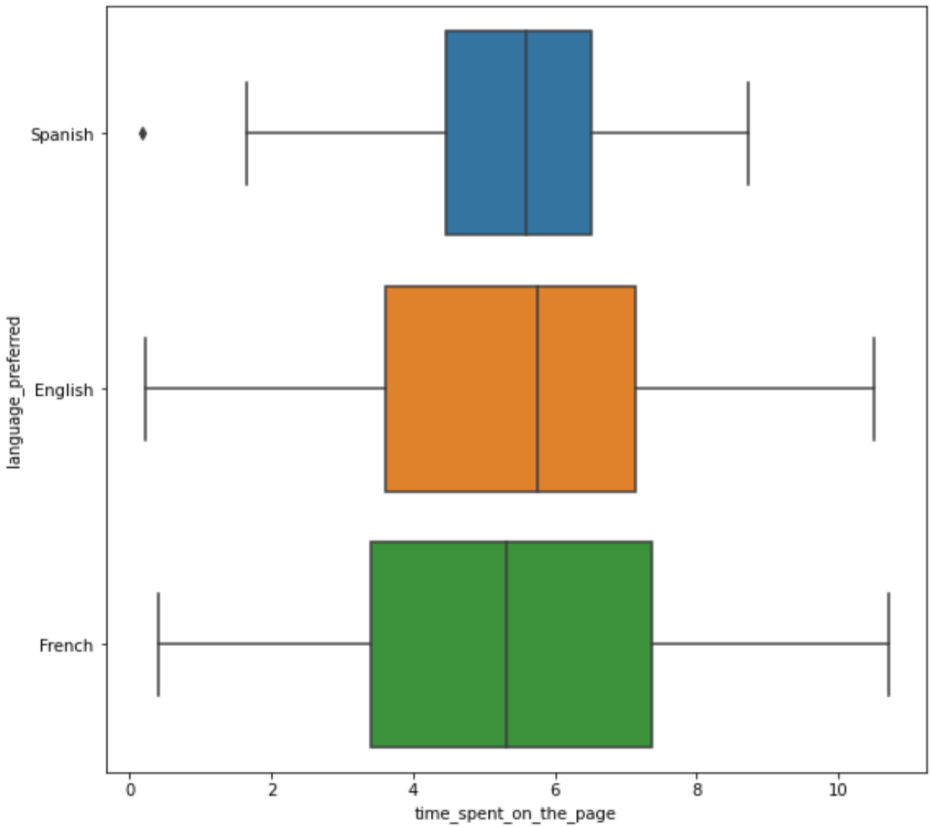
Observations

- On average the users who converted spend 6.22 minutes on the page. There are outliers illustrating that some users who converted spend more and less than the quartiles.

- On average the users who did not convert spend 4 minutes on the page. There are outliers illustrating that some users who did not convert spend more than the quartiles.

EDA – Bivariate Analysis

```
# write the code to plot a suitable graph to understand the distribution of
plt.figure(figsize=(9, 9))
sns.boxplot(data = df, x = 'time_spent_on_the_page', y = 'language_preferre
plt.show()
```

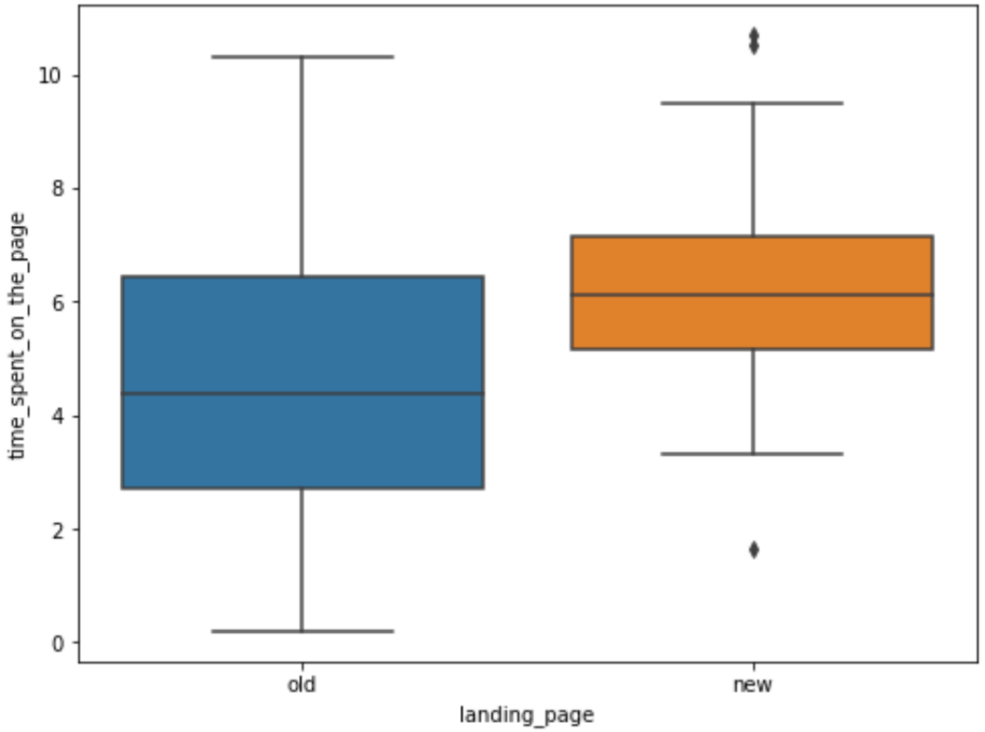


Observations

- The users who prefer French spend on average 5.3 mins. There are no outliers.
- The users who prefer English spend on average 5.9 mins. There are no outliers.
- The users who prefer Spanish spend on average 5.5 mins. There are outliers illustrating that some spanish users spend less than the quarlites.

Hypotheses Tested and Results

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



- The boxplot shows us that users spend more time on the new landing page then they do on the old one.

Step 1: Define the null and alternative hypotheses:

- Null Hypothesis: There is no difference in the time spent on the new and existing landing pages.
- Alternative Hypothesis: There is a difference with users spending more time on the new landing page.

Step 2: Select the appropriate test:

- Based on the information provided, the appropriate test to use would be a Student's t-test for independent samples, also known as a two-sample independent t-test. This test is used to compare the means of two independent populations when the population standard deviations are unknown. The t-test assumes that the two samples are independent and normally distributed, and that the variances of the two populations are equal.

Step 3: Decide the significance level:

- The problem statement state the $\alpha = 0.05$

Step 4: Collect and prepare data

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

Step 5: Calculate the p-value:

- The p-value is equal to 0.0001392381225166549

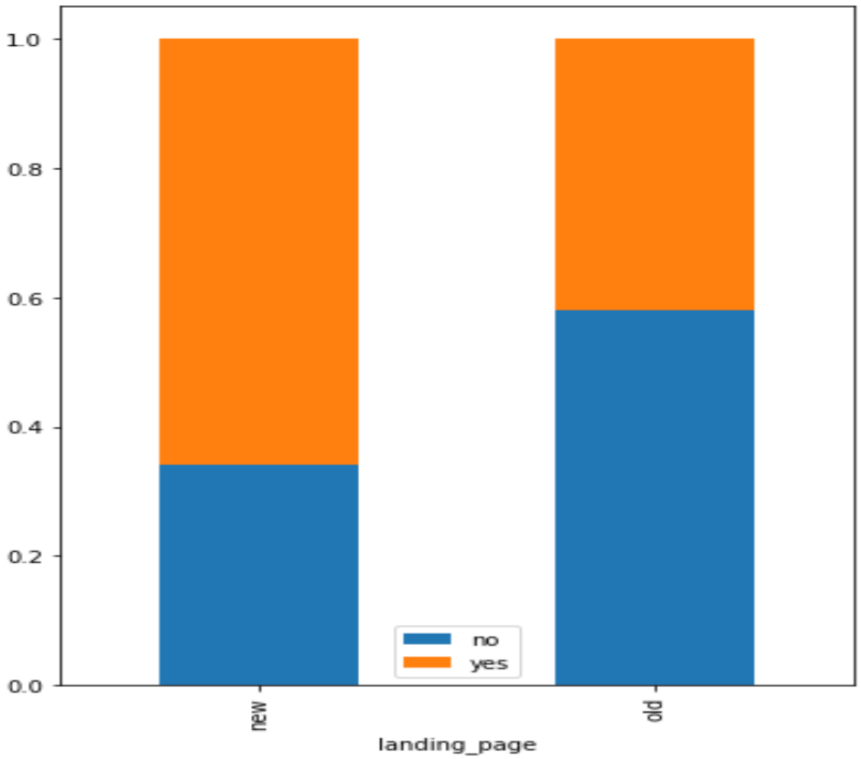
Step 6: Compare the p-value with α:

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

Step 7: Draw inference:

- Since the p-value is less than the significance, there is enough evidence to conclude that we reject the null hypothesis in favor of alternative hypothesis that the users spent 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?



- The bargraph shows that convertion rate is highest for users from the new landing page.

Step 1: Define the null and alternative hypotheses:

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

Step 2: Select the appropriate test:

- The appropriate test for this situation would be a z-test for the difference in two proportions. This test is used to compare the proportion of success of two independent populations. Since this is a one-tailed test and the population is normal, using a z-test would be appropriate.

Step 3: Decide the significance level:

- The problem statement state the $\alpha = 0.05$

Step 4: Collect and prepare data

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

Step 5: Calculate the p-value:

- The p-value is 0.008026308204056278

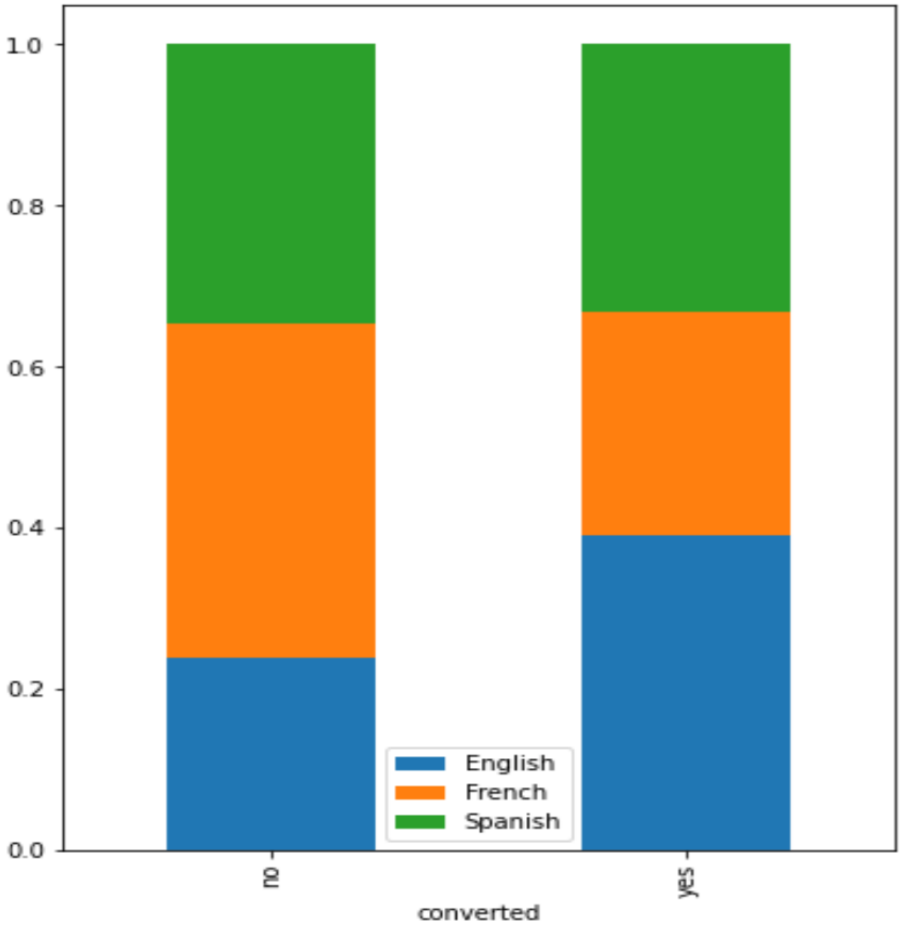
Step 6: Compare the p-value with α:

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

Step 7: Draw inference:

- Since the p-value is less than the significance, there is enough evidence to conclude that we reject the null hypothesis in favor of alternative hypothesis that the conversion rate for the new page is greater than the conversion rate for the old page.

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



- The bargraph here confirms the null hypothesis that the converted status is independent of the preferred language.

Step 1: Define the null and alternative hypotheses:

- Null Hypothesis: The converted status is independent of the preferred language.
- Alternative Hypothesis: The converted status depends on the preferred language.

Step 2: Select the appropriate test:

- The appropriate test for this situation would be chi-squared test for independence. This test is used to determine if there is a significant association between two categorical variables. The chi-squared test compares the observed frequencies of the two variables in a contingency table to the expected frequencies if the two variables were independent. Since you are trying to determine if there is a relationship between the converted status and the preferred language, the chi-squared test for independence would be the appropriate test to use.

Step 3: Decide the significance level:

- The problem statement state the $\alpha = 0.05$

Step 4: Collect and prepare data

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

Step 5: Calculate the p-value:

- The p-value is 0.2129888748754345

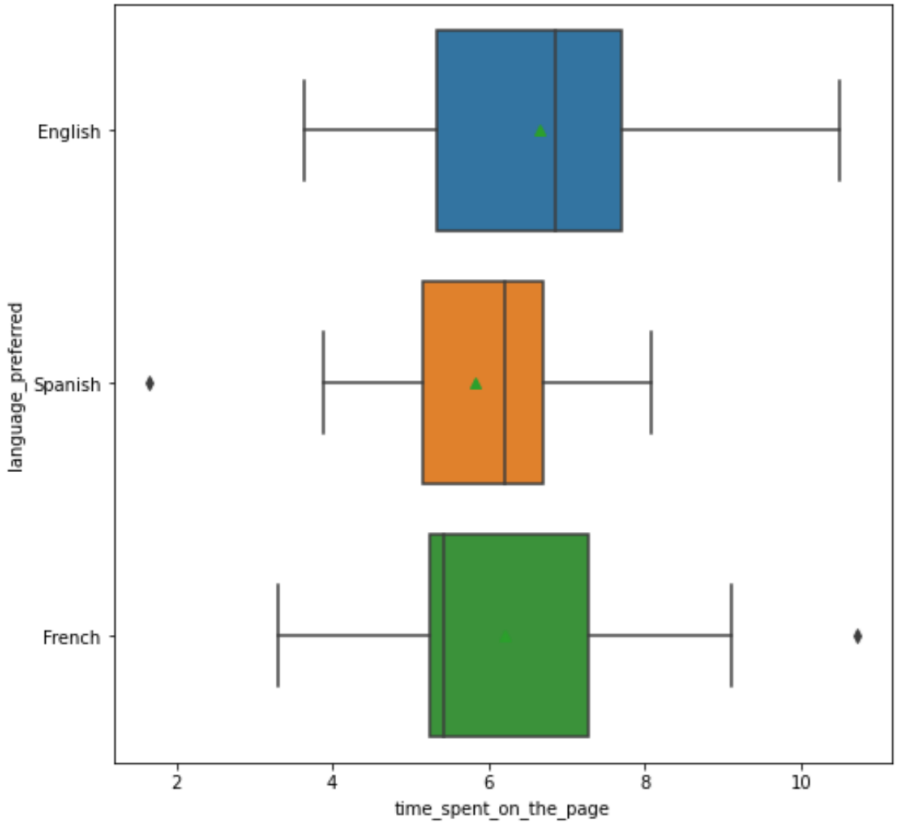
Step 6: Compare the p-value with α:

- As the p-value 0.2129888748754345 is greater than the level of significance, we fail to reject the null hypothesis.

Step 7: Draw inference:

- Since the p-value is greater than significance level, we fail to reject the null hypothesis. There is no enough evidence to support the alternative hypothesis that the conversion rate depends on users preferred language. hence, the language preference of the users and the conversion rate are independent.

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



- From the box plot we can say that mean of all language preferred by the users are Equal almost. This support our null hypothesis that the mean time spent by Spanish, French and English language users is equal.

Step 1: Define the null and alternative hypotheses:

- Null Hypothesis: The mean time spent by Spanish, French and English language users is equal.
- Alternative Hypothesis: At least one mean is unequal.

Step 2: Select the appropriate test:

- The appropriate test for this situation would be ANOVA (Analysis of Variance) test, which is used to compare the means of three or more independent groups. This test can be used to determine if there is a significant difference in the meantime spent on the new page among users of different languages. The ANOVA test can also help to determine which group or groups are responsible for the observed difference, if any. It is a parametric test, and it assumes that the data is normally distributed, independent and have equal variances among the groups.

Step 3: Decide the significance level:

- The problem statement state the $\alpha = 0.05$

Step 4: Collect and prepare data

```
# create a subsetted data frame of the time spent on the new page by  English, French and Spanish language users
time_spent_English = df_new[df_new['language_preferred']=="English"]['time_spent_on_the_page']
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:

- The p-value is equal to 0.0001392381225166549

Step 6: Compare the p-value with α:

- As the p-value 0.46711357711340173 is greater than the level of significance, we fail to reject the null hypothesis.

Step 7: Draw inference:

- As the p-value is greater than the level of significance, we cannot reject the null hypothesis. Hence, there is enough evidence to support the claim that the mean time spent on the new page is equal across the different language users.

Executive Summary

- Compared to the old landing page, users spent more time on the new one.
- Compared to users on the old landing page, users on the new page appear to have converted at a higher rate.
- Users of all of the chosen languages are largely evenly distributed on both the old and new pages.
- The conversion rate and the amount of time spent on the page do not appear to be impacted by the user's preferred language.
- The conversion status is independent of the preferred language
- The mean time spent on the new page is equal for the different language users.
- According to the statistical analysis, I believe that the new landing page was a success because it has been shown to be successful in luring more subscribers.

Recommendations

- Extend the study by including the age group of users to see if there is a specific age group that is more likely to subscribe. This could allow E-News Express to custom content to different age groups' interests.

- Perform a larger, non-stratified random sample to obtain a more accurate picture of population proportions and habits and examine the language groups in greater depth to see where additional content might drive more subscribers in each language group.

- Extend the study to include what types of news content people of different ages and languages are interested in, such as local news, global news, politics, and so on. By letting E-news express to custom content to various language groups, this might help drive subscribers.

Great Learning

Happy Learning !