Statistics: E-news Express

Objective

The design team of the company has created a new landing page. You have been assigned the task to decide whether the new landing page is more effective to gather new subscribers. Suppose you randomly selected 100 users and divided them equally into two groups. The old landing page is served to the first group (control group) and the new landing page is served to the second group (treatment group).

The object is to decide whether the new page is effective enough to gather new subscribers for the news

Questions

1. Explore the dataset and extract insights using Exploratory Data Analysis.

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

3. 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?

4. Does the converted status depend on the preferred language? [Hint: Create a contingency table using the pandas. crosstab() function]

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

*Consider a significance level of 0.05 for all tests.

Data Dictionary

1. user_id - This represents the user ID of the person visiting the website.

2. group - This represents whether the user belongs to the first group (control) or the second group (treatment).

3. landing_page - This represents whether the landing page is new or old.

4. time_spent_on_the_page - This represents the time (in minutes) spent by the user on the landing page.

5. converted - This represents whether the user gets converted to a subscriber of the news portal or not.

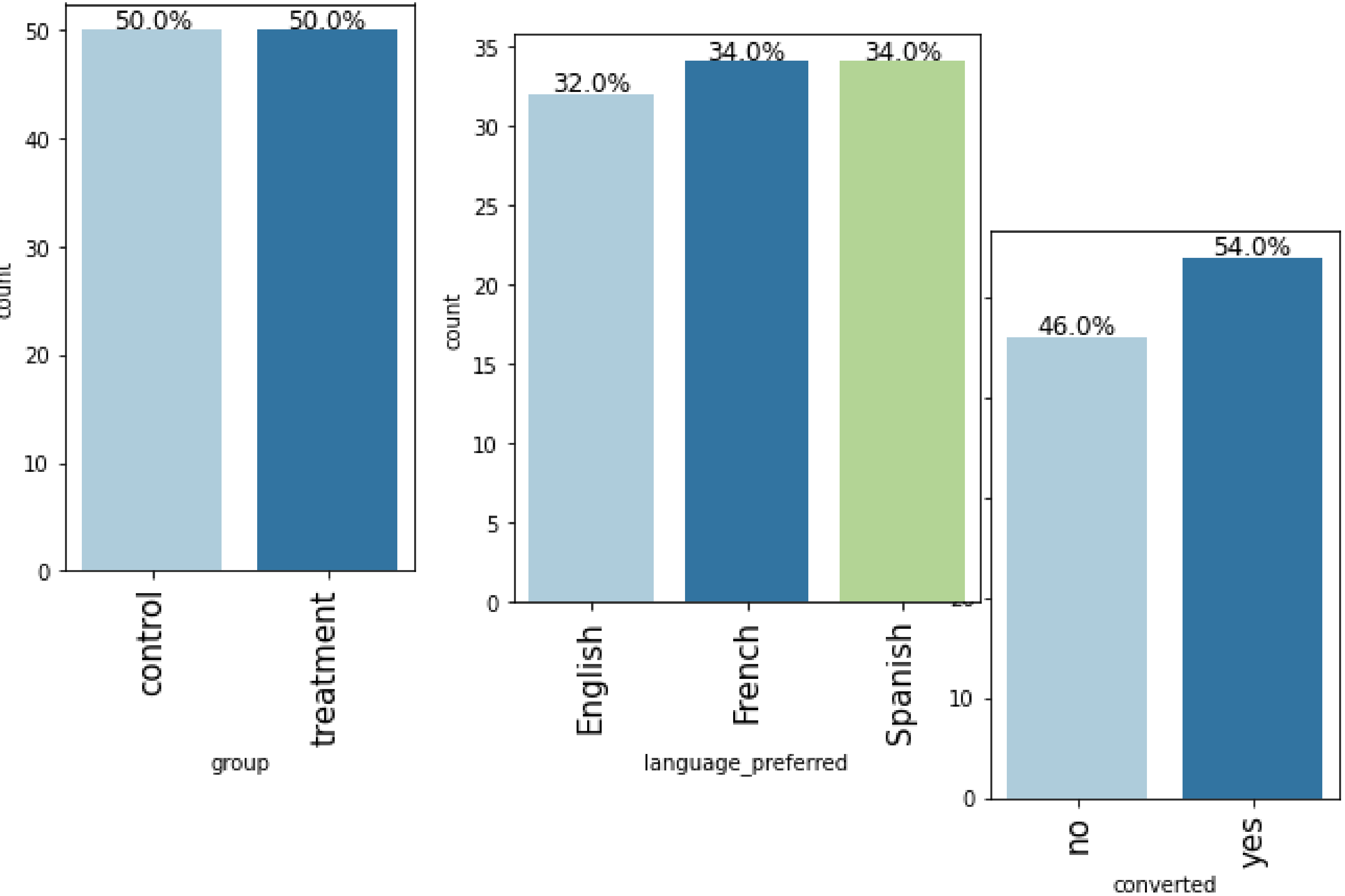
6. language_preferred - This represents the language chosen by the user to view the landing page.

1. Exploratory Data Analysis

There is no none values.
There are 100 row and 6 (features or variable.
The variable Time_spend_in_the_page is float64 while other variable are objects

	user_id	group	landing_page	time_spent_on_the_page	converted	language_preferred
count	100.000000	100	100	100.000000	100	100
unique	NaN	2	2	NaN	2	3
top	NaN	treatment	old	NaN	yes	Spanish
freq	NaN	50	50	NaN	54	34
mean	546517.000000	NaN	NaN	5.377800	NaN	NaN
std	52.295779	NaN	NaN	2.378166	NaN	NaN
min	546443.000000	NaN	NaN	0.190000	NaN	NaN
25%	546467.750000	NaN	NaN	3.880000	NaN	NaN
50%	546492.500000	NaN	NaN	5.415000	NaN	NaN
75%	546567.250000	NaN	NaN	7.022500	NaN	NaN
max	546592.000000	NaN	NaN	10.710000	NaN	NaN

Univariate Analysis



Observations

* There are 100 unique users.

* There are 2 unique groups - control and treatment. Each group consists of 50 users.

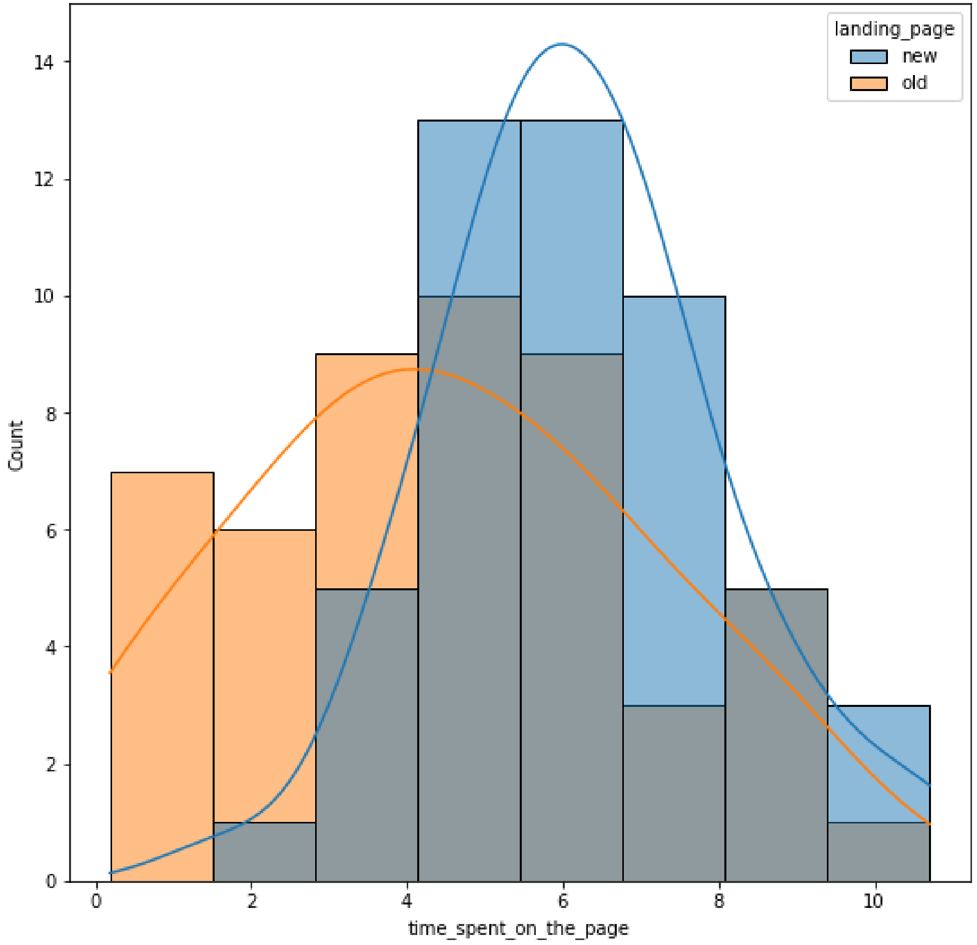
* There are 2 landing_pages - new and old.

* Overall, 54 users get converted and 46 users do not get converted after visiting the landing page.

* There are 3 unique preferred languages - English, French, and Spanish.

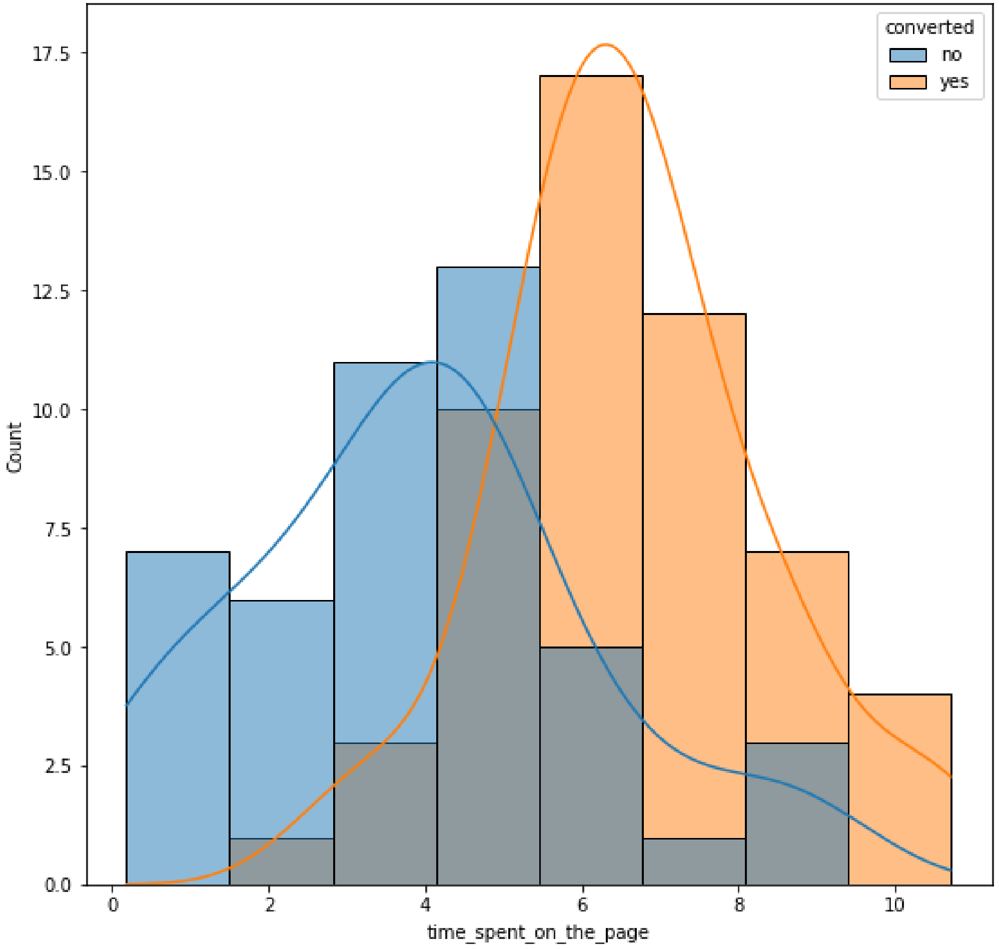
Bivariate Analysis

Landing page vs Time spent on the page



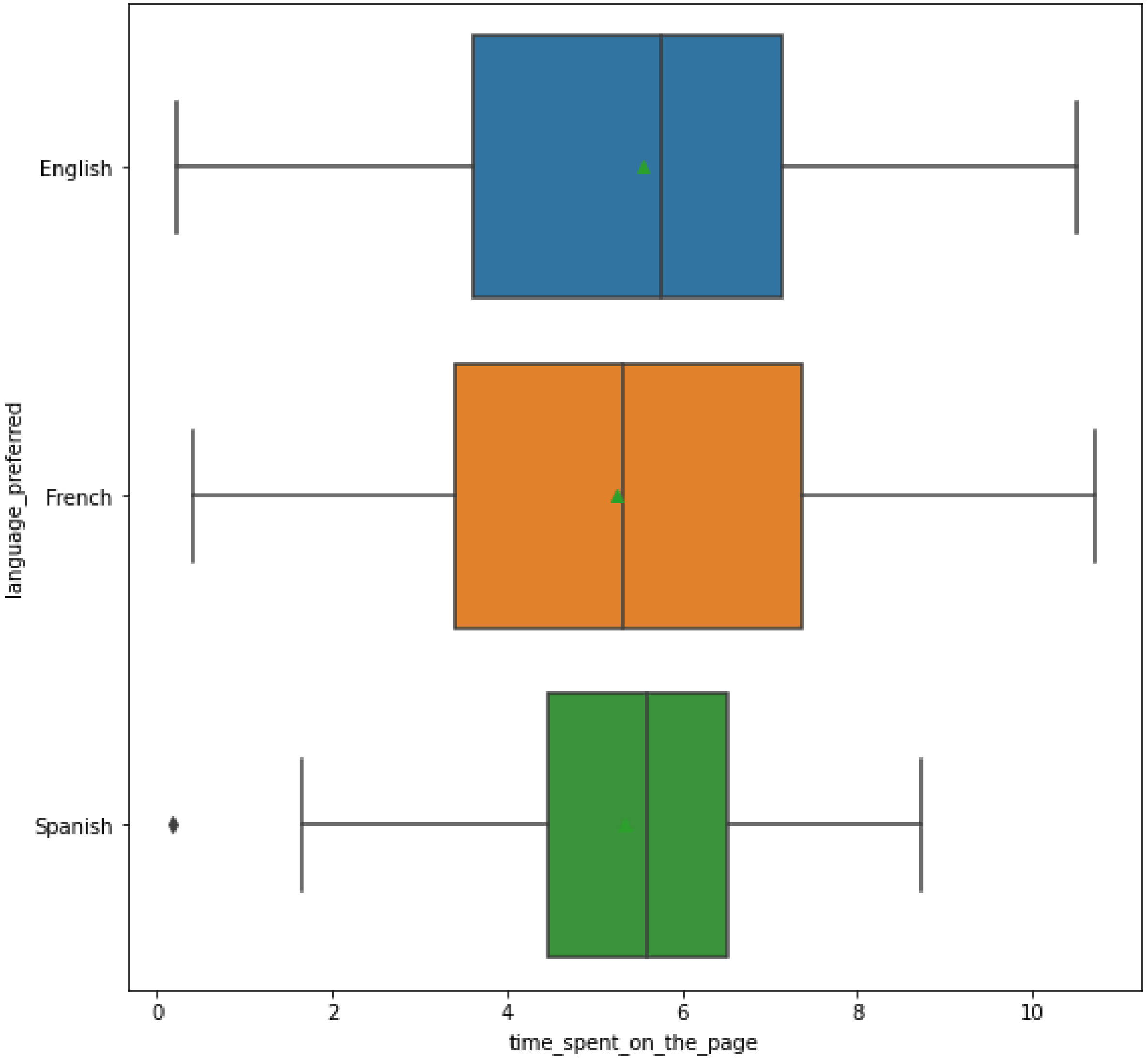
The mean time spend on the page are higher for new landing page than old landing page

Conversion status vs Time spent on the page



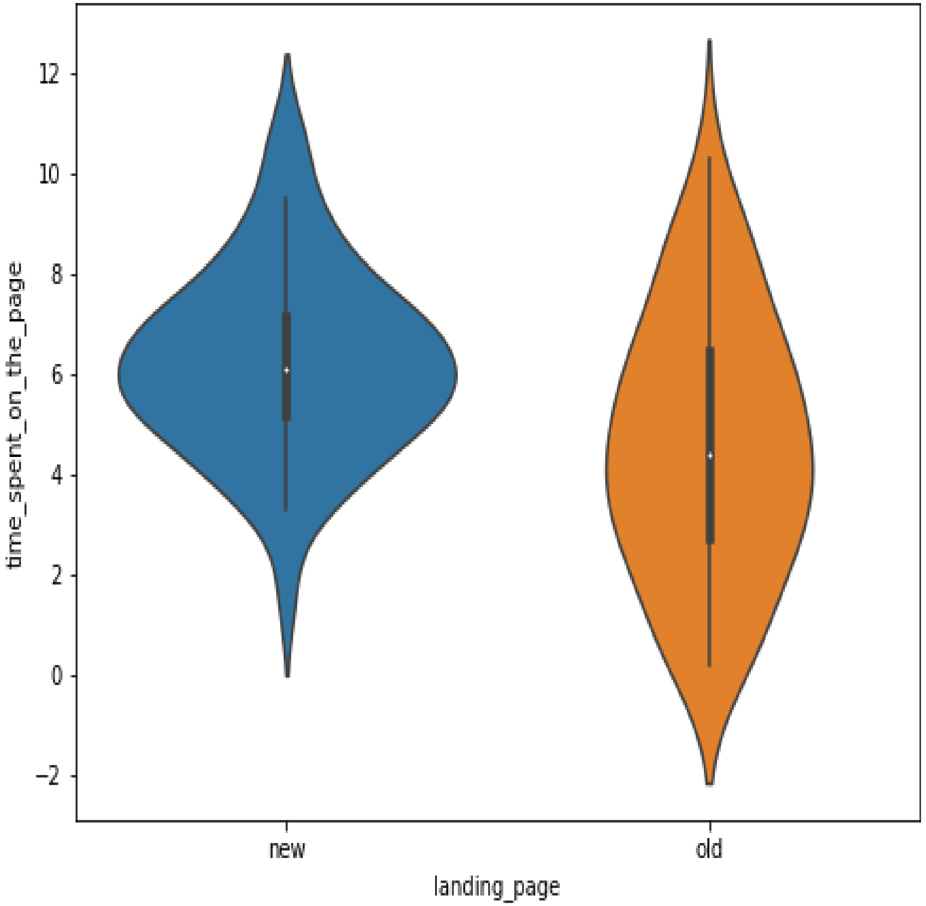
The mean time spend on the page are higher for converted page than not converted page

Time spent on the page *Vs* the language preferred



French has the larger variance and lower median.
Spanish and English have similar median, but English has bigger variance.

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



- the median time of new landing page is higher than old landing page.
- The time variance of the new landing page is small than the time variance of the old landing page

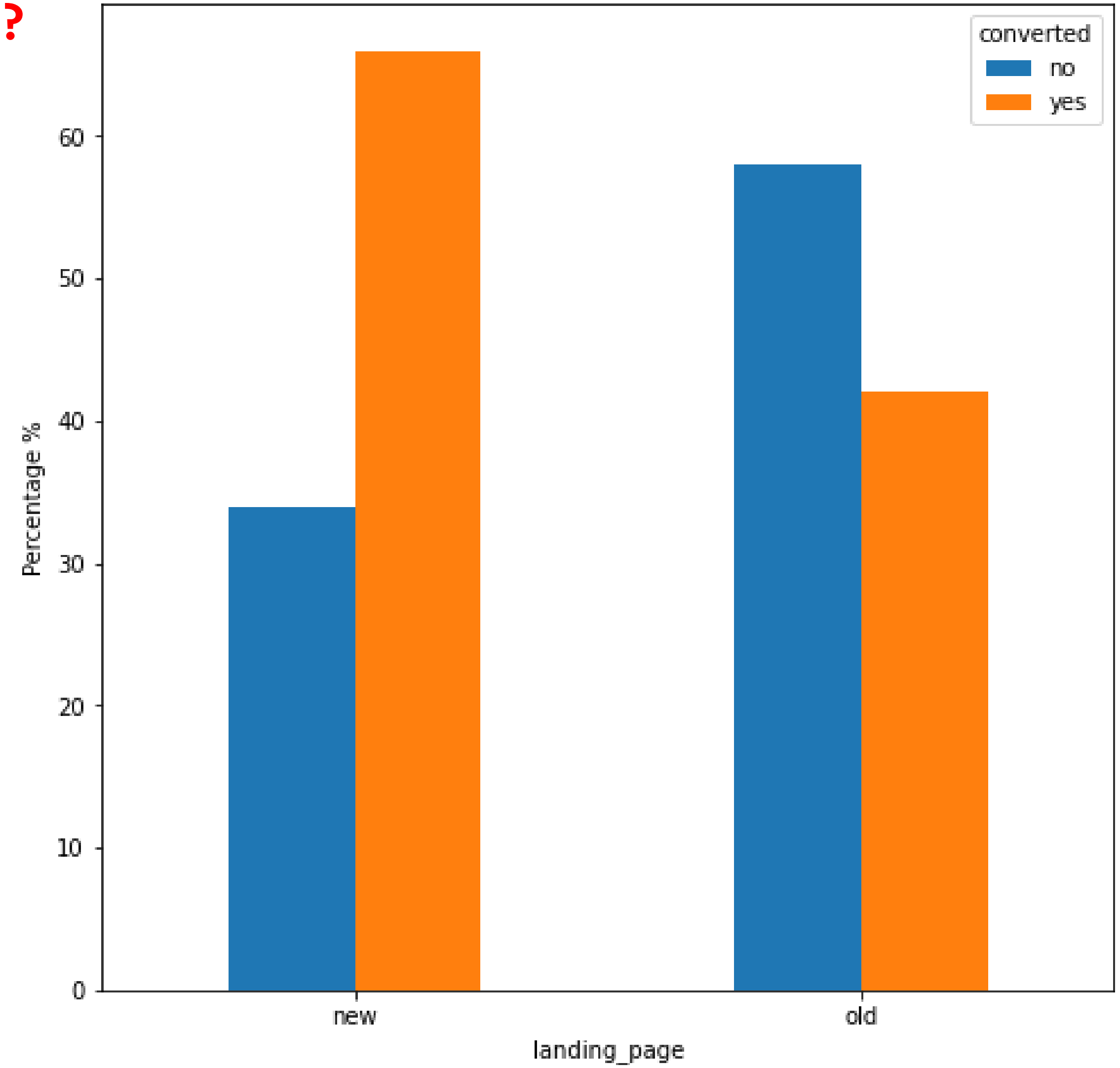
- the null and alternate hypotheses:

H0:u_new=u_old

H1:u_new>u_old

- This is a one-tailed test concerning two population means from two independent populations. The population standard deviations are unknown.
- I select t_test
- α = 0.05.
- test_stat, p_value = ttest_ind(time_spent_new,time_spent_old,alternative=‘greater’)
- Based on the sample standard deviations of the two groups(1.82,2.58), population standard deviations can be assumed to be equal.
-
- As the p-value 0.00013 is less than the level of significance, we reject the null hypothesis.
- The users spend more time on the new landing page.

3. 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?



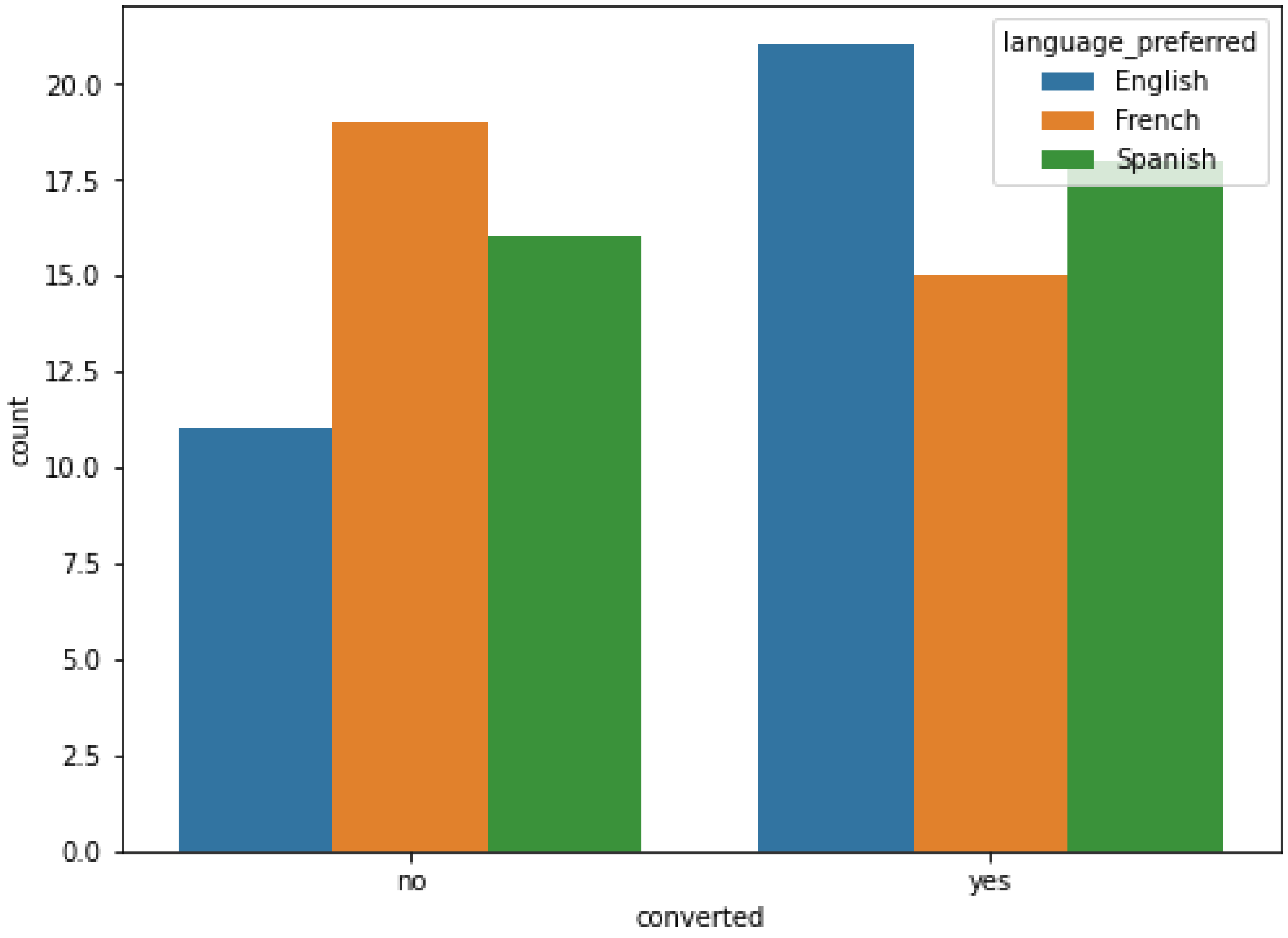
- The conversion rate is high for new pagethan the conversion rate for the old page
- The numbers of converted users in treatment group and numbers of converted users in control group are 33 and 21 respectively. The numbers of users served the new and old pages are 50 and 50 respectively

H0:conversion rate of new page=conversion rate of old page

H1:conversion rate of new page>conversion rate of old page

- This is a one-tailed test concerning two population proportions from two independent populations we choose proportions_ztest
- α = 0.05.
- The numbers of converted users in treatment group and numbers of converted users in control group are 33 and 21 respectively The numbers of users served the new and old pages are 50 and 50 respectively
- test_stat, p_value = proportions_ztest([new_converted,old_converted],[n_control,n_treatment],alternative='larger’)
- the p-value 0.008 is less than the level of significance, we reject the null hypothesis.
- Since we reject the null hypotheses. The conversation rate of new page is larger than conversation rate of the old page.

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



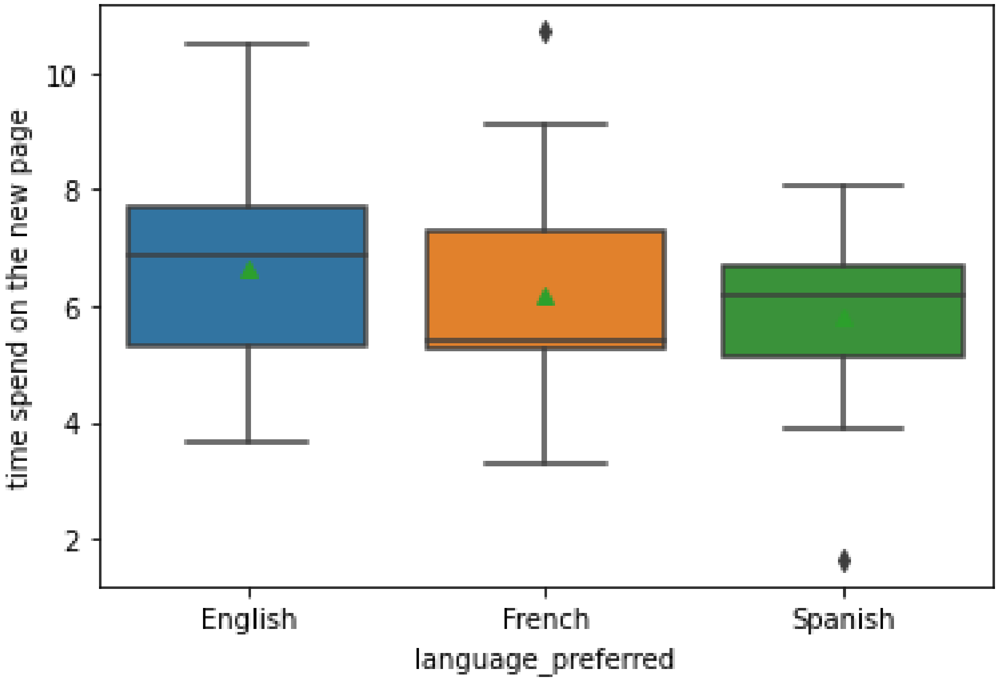
Does the converted status depend on the preferred language?

H0: Converted statue is independent of preferred languages
H1: Converted statue is not independent of preferred languages

- This is a problem of the test of independence, concerning two categorical variables - converted status and preferred language. I choose Chi2 contingency test
- α = 0.05.
- chi2, p_value, dof, exp_freq = chi2_contingency(pd.crosstab(df['converted'],df['language_preferred']))
- As the p-value 0.21 is greater than the level of significance, we fail to reject the null hypothesis.
- Since we failed to reject the null hypotheses. We find that The converted status is independent of languages.

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

the time spent on the new page for different language users



H0: U_english=U_spanish=U_french

H1: U_English,U_Spanish andt U_french
Are not equal

- This is a problem, concerning three population means so we choose ANOVA
- First, we need to check the assumptions of normality and equality of variance for the three groups
- 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
H0: Time spent on the new page follows a normal distribution
Ha: Time spent on the new page does not follow a normal distribution

```
w, p_value = shapiro(df_new['time_spent_on_the_page'])
```

Since p-value (0.80) of the test is very large than the 5% significance level, we fail to reject the null hypothesis that the response follows the normal distribution.

Levene’s test

We will test the null hypothesis
H: All the population variances are equal
Ha: At least one variance is different from the rest

```
statistic, p_value = levene( df_new[df_new['language_preferred']=="English"]['time_spent_on_the_page'],
                             df_new[df_new['language_preferred']=="French"]['time_spent_on_the_page'],
                             df_new[df_new['language_preferred']=="Spanish"]['time_spent_on_the_page'])
```

Since the p-value (0.467) is large than the 5% significance level, we fail to reject the null hypothesis of homogeneity of variances.

ANOVA Test

```
test_stat, p_value
=  f_oneway(time_spent_English,time_spent_French,time_spent_S
panish
```

As the p-value 0.43 is greater than the level of significance, we fail to reject the null hypothesis and the mean time spend in the new page are equal to all languages

Conclusion and Business Recommendations

- The users spend more time on the new landing page
- The conversion rate is higher of the new page than the conversion rate of the old page
- The mean time spend in the new page are equal to all languages.
- The converted status is independent of languages.
- We find the new page is effective in attracting new subscribers