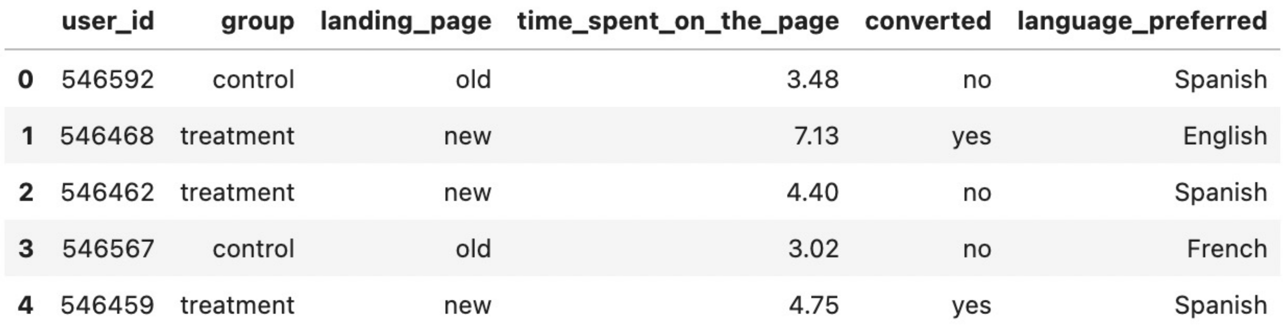
E-News Express Project

By: Nihal Kala

Problem Statement

- An online news portal wants to expand its business by adding new subscribers.
- The company has created a new landing page with the hopes that it would be able to gain new subscribers for the strategy to be effective using a/b testing.
- 100 users are randomly selected and divided into 2 groups with 50 in the control group (old) and 50 in the treatment group (new).

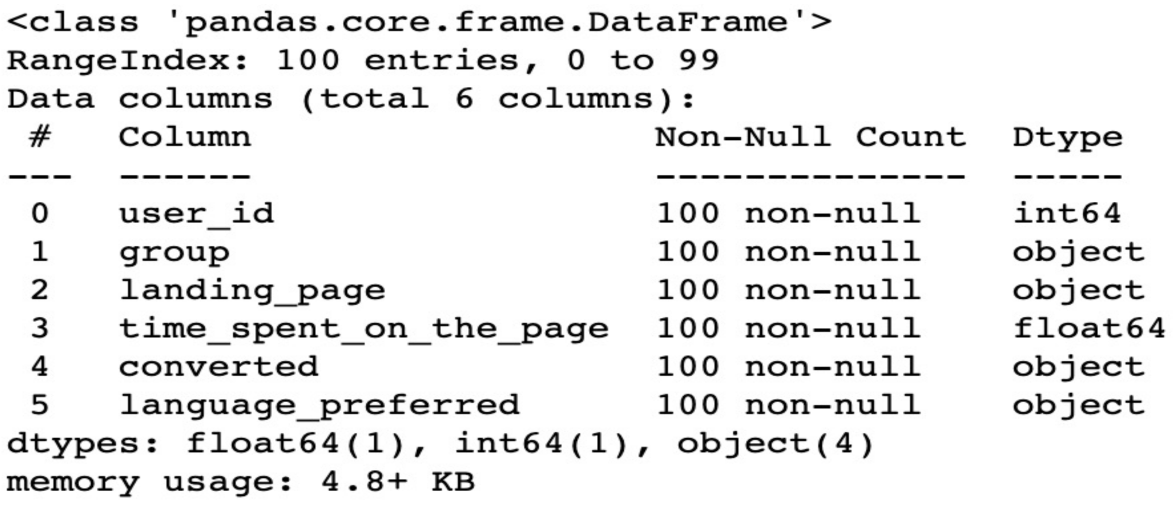
Structure of Data



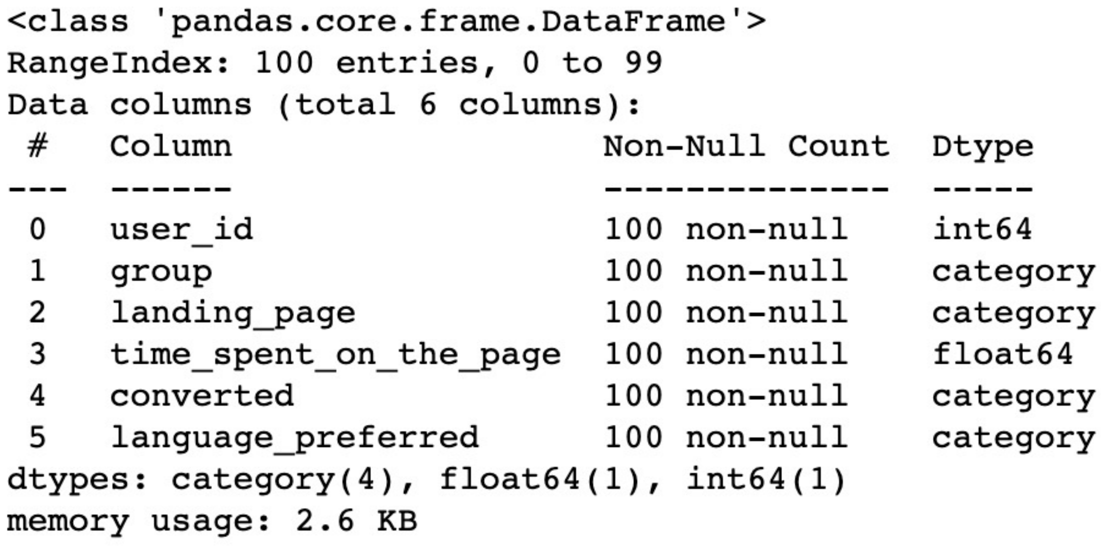
	user_id	group	landing_page	time_spent_on_the_page	converted	language_preferred
0	546592	control	old	3.48	no	Spanish
1	546468	treatment	new	7.13	yes	English
2	546462	treatment	new	4.40	no	Spanish
3	546567	control	old	3.02	no	French
4	546459	treatment	new	4.75	yes	Spanish

Shape of Data: 100 rows, 6 columns

Column Data Types Before & After (Conversion to category type)



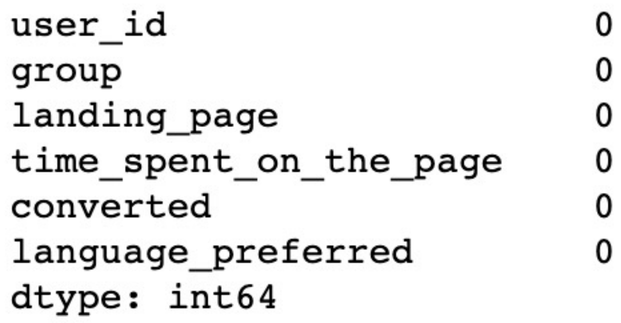
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 6 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   user_id                100 non-null    int64
 1   group                  100 non-null    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
```



```
<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    category
 2   landing_page           100 non-null    category
 3   time_spent_on_the_page 100 non-null    float64
 4   converted              100 non-null    category
 5   language_preferred     100 non-null    category
dtypes: category(4), float64(1), int64(1)
memory usage: 2.6 KB
```

Memory usage decreased by around 2.2 KB

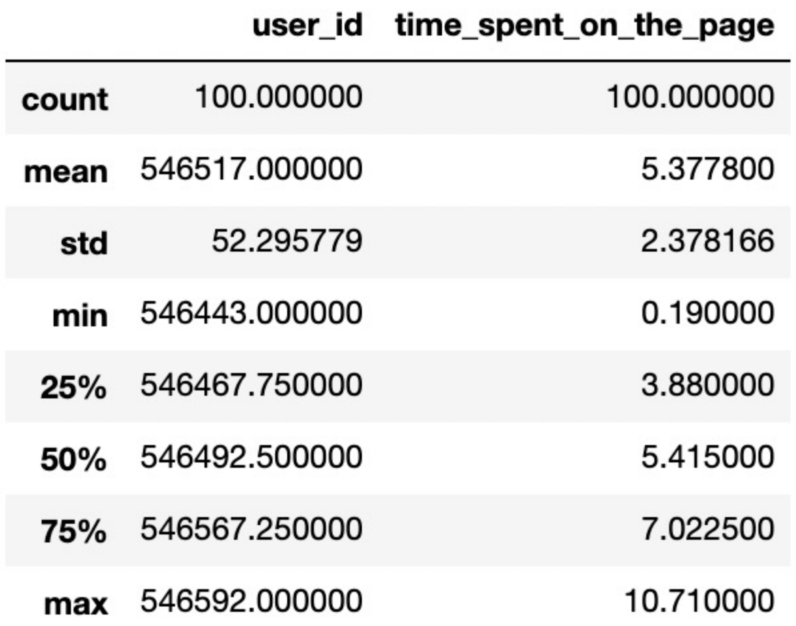
Null Values



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

No null values

Statistical Summary of Data

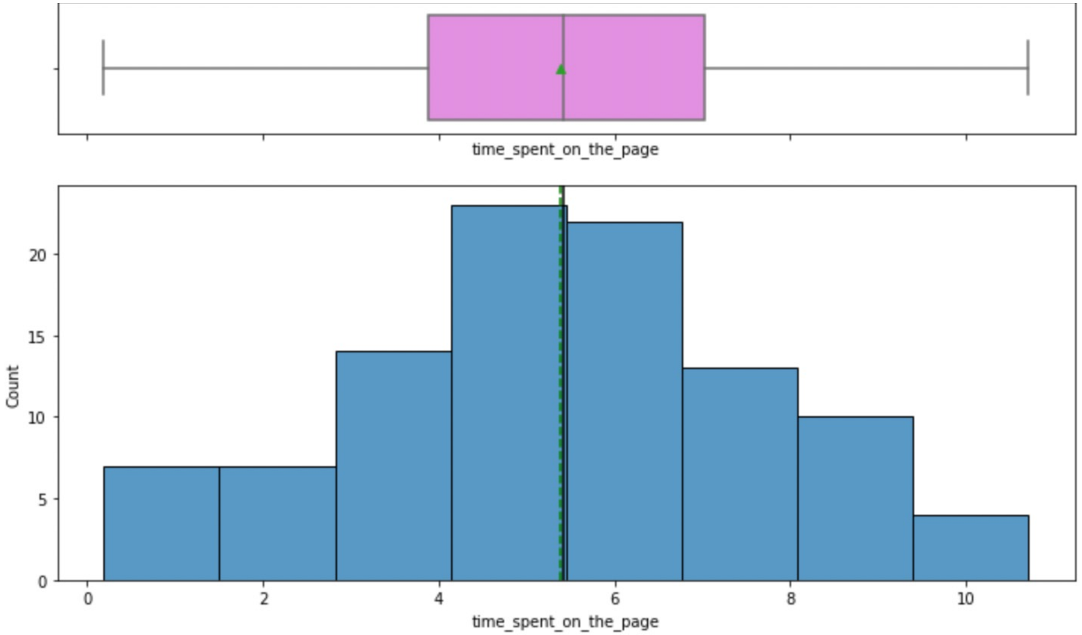


	user_id	time_spent_on_the_page
count	100.000000	100.000000
mean	546517.000000	5.377800
std	52.295779	2.378166
min	546443.000000	0.190000
25%	546467.750000	3.880000
50%	546492.500000	5.415000
75%	546567.250000	7.022500
max	546592.000000	10.710000

Five Point Summary 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, 55 users get converted and 45 users do not get converted after visiting the landing page.
- There are 3 unique preferred languages - English, French, and Spanish.

Time Spent on the Page Histogram and Box Plot

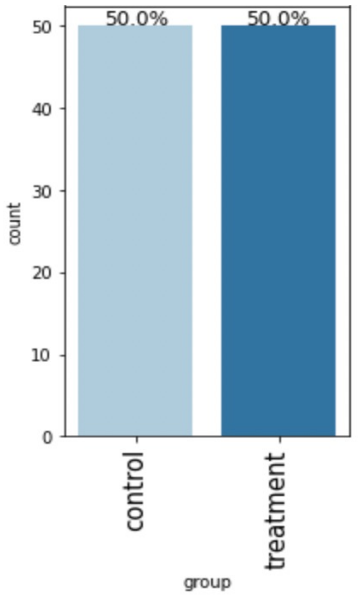


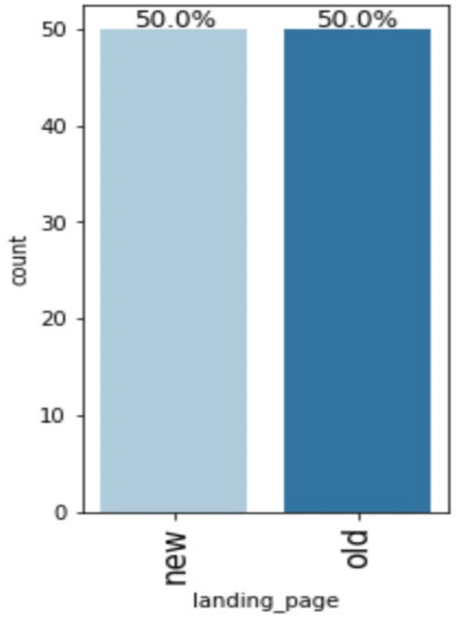
Histogram and Box Plot (Time Spent on the Page)

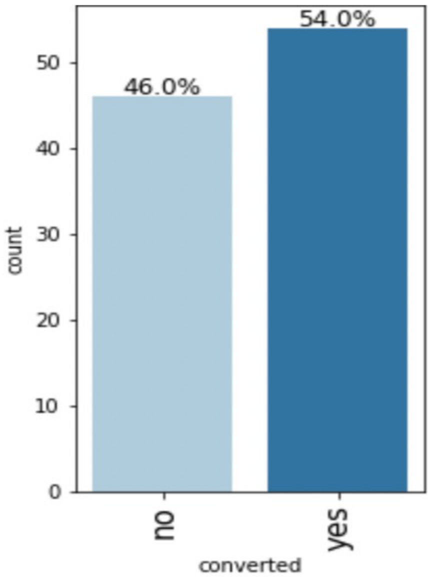
An average of 5.41 minutes was spent on the landing page with the interquartile range falling from 3.88 to 7.02 minutes.

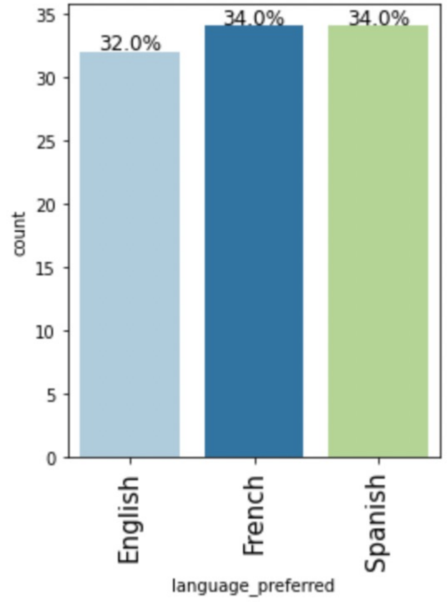
Univariate Analysis: Bar Graphs

More users subscribed and more users preferred Spanish and French over English.

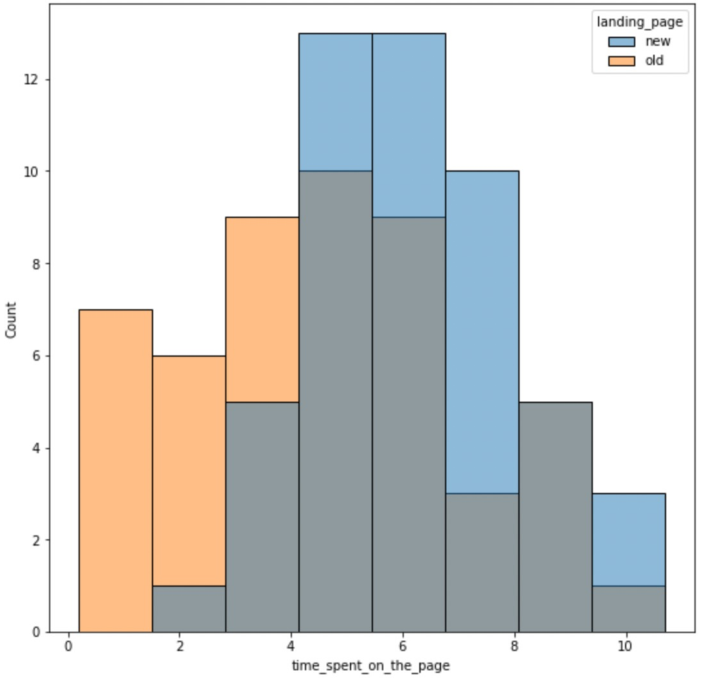


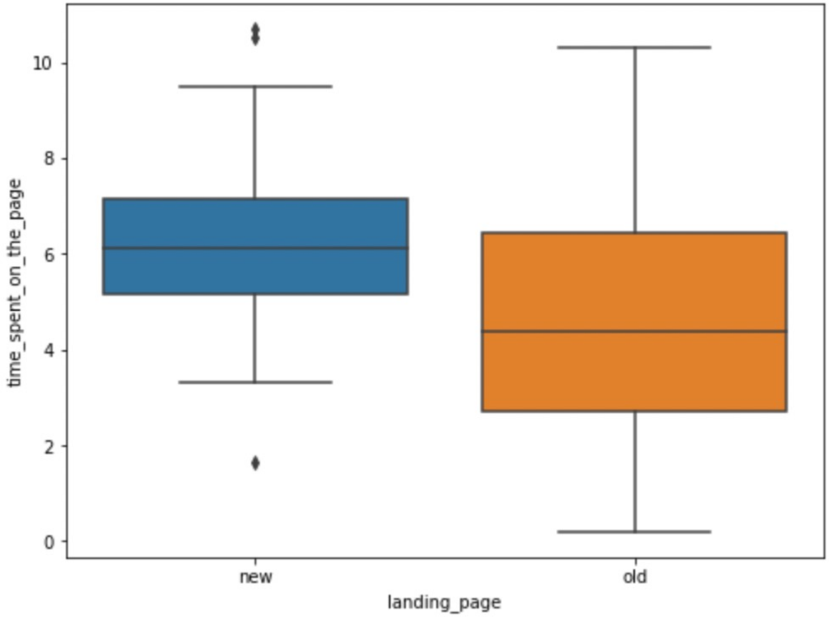






Bivariate Analysis: Time Spent on the Page vs. Landing Page Histogram and Box Plot

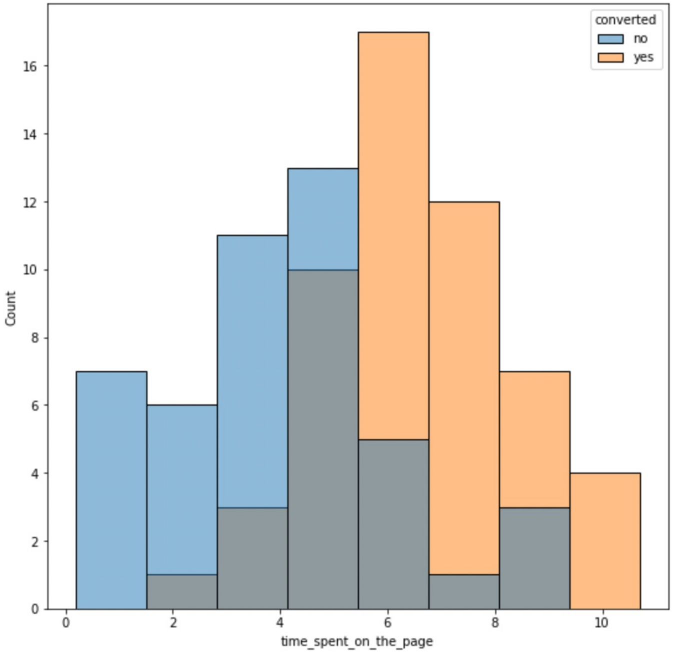


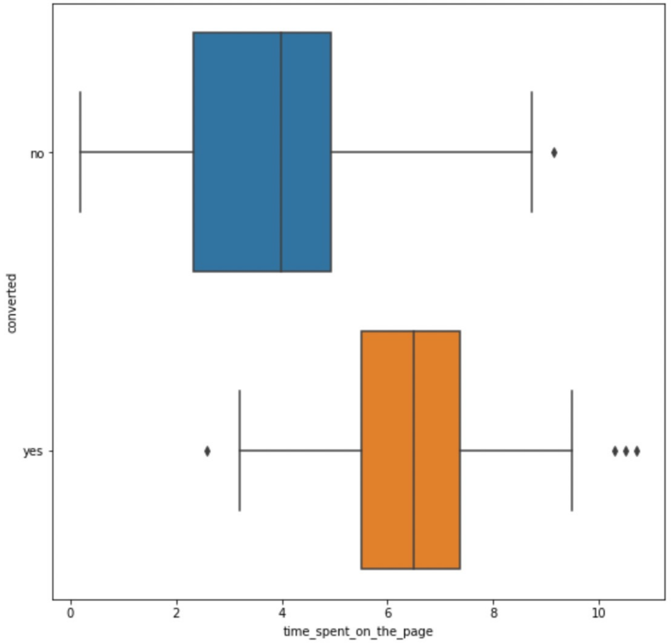


While there's a wider range of time spent on the older landing page, the average time spent on the new landing page is greater than the average time spent on the old landing page.

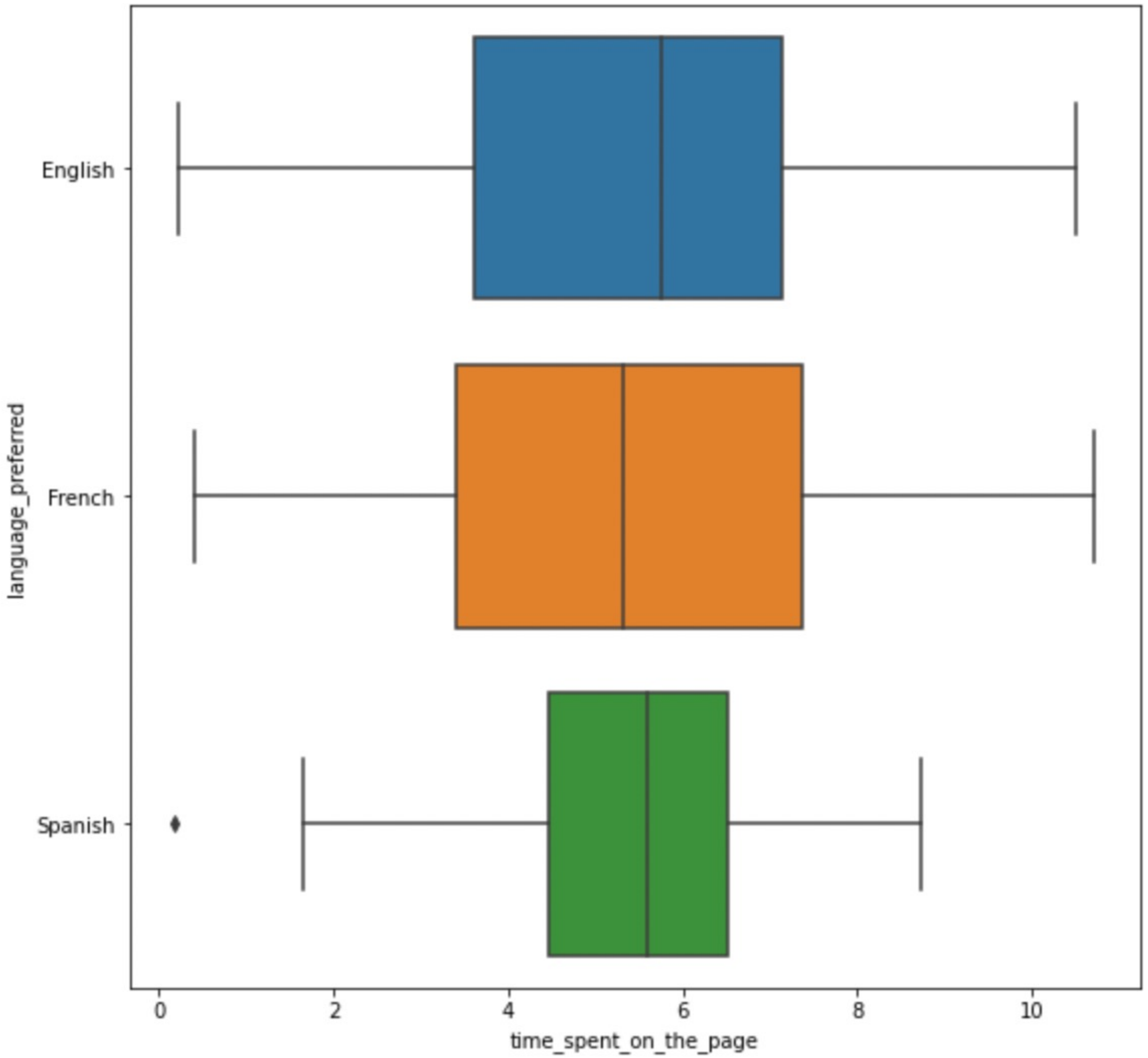
Bivariate Analysis: Time Spent on the page vs. Converted Histogram and Box Plot

While the time spent on the page range of people who didn’t subscribe is larger than the people who did subscribe, the people who did subscribe have larger average time spent on the page than people who didn’t subscribe.



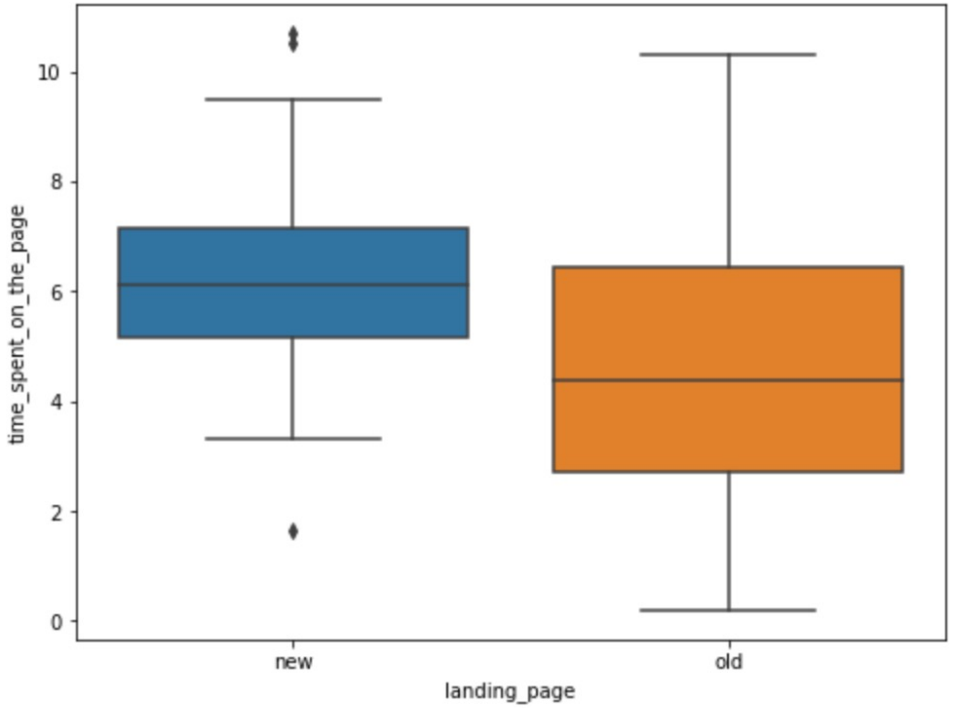


Bivariate Analysis: Time spent on the page vs. Language Preferred Box Plot



The ranges of time spent on the page for English and French are significantly larger than Spanish, while the average time spent on the page is greatest for English.

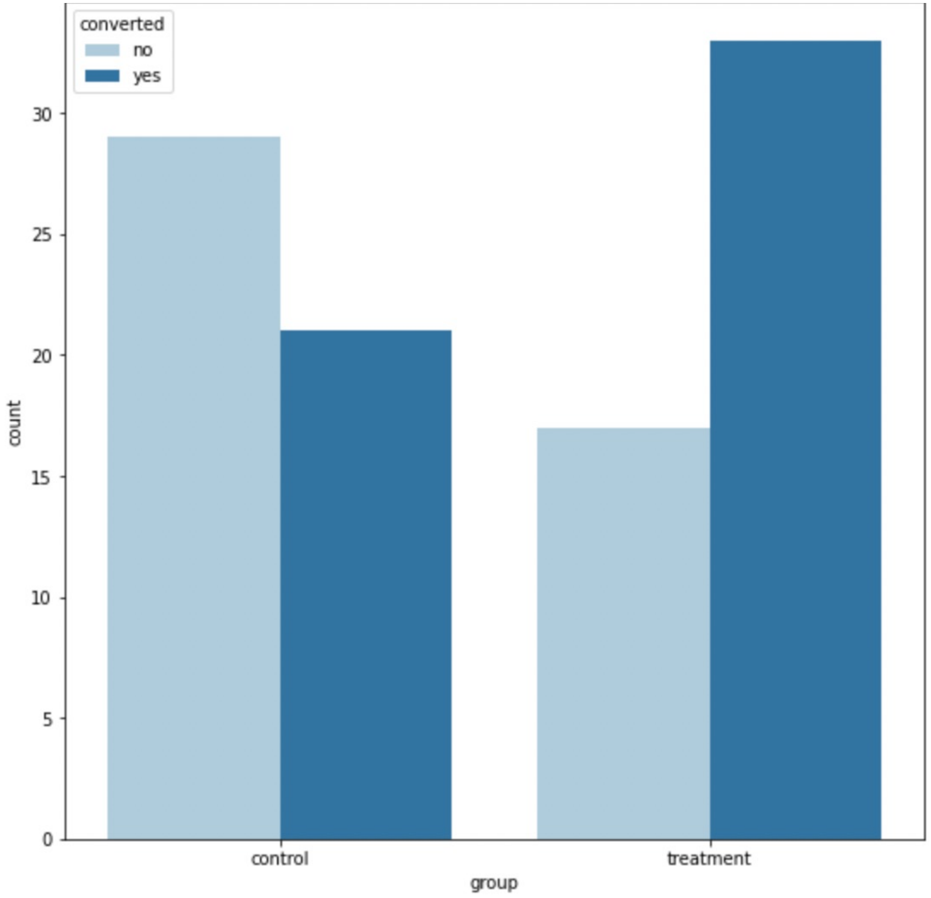
Q1: Do the users spend more time on the new landing page than the existing landing page?



Q1: Do the users spend more time on the new landing page than the existing landing page?

- Let u1 and u2 be the mean time on the new landing page and mean time on the old landing page respectively.
- Null Hypothesis: Ho: u1=u2 (Mean time on new landing page = mean time on old landing page)
- Alternate Hypothesis: Ha: u1>u2 (Mean time on new landing page > mean time on old landing page)
- Test: One-tailed T test dealing with two population means from two independent populations at 5% significance level
- 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 old page is: 2.58.
- As the p-value 0.0001392381225166549 is less than the level of significance, we reject the null hypothesis.
- Inference: We have enough statistical evidence to say the mean time that users spent on the new landing page is greater than the mean time users spent on the existing landing page.

Q2: 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?



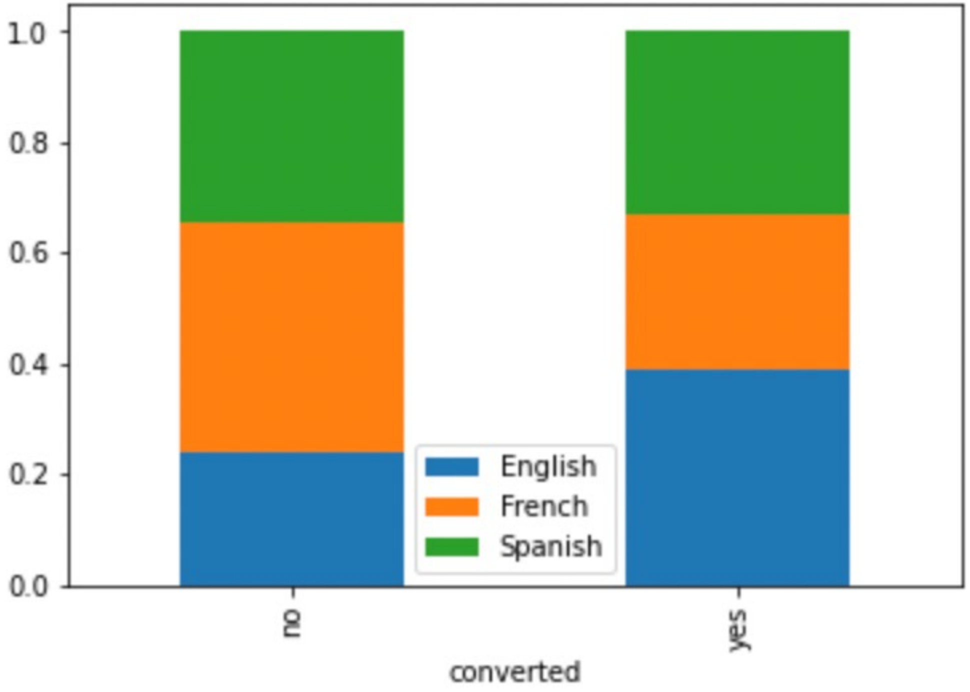
Conversion Rate old:
21/50 = 0.42
Conversion Rate new:
33/50 = 0.64

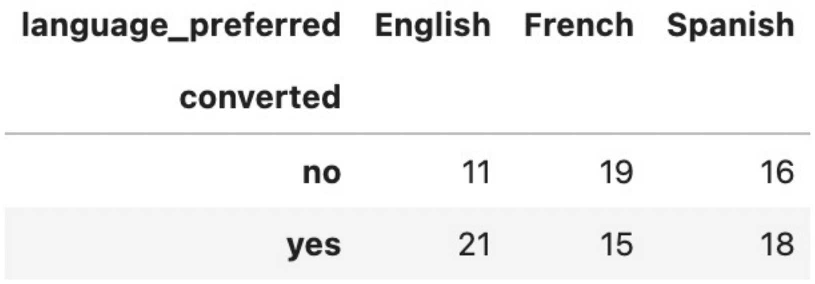
Q2: Is the conversion rate (the proportion of users who visit the landing page and get converted) for the new page greater than the conversion rate for the old page?

- Let u1 and u2 be the conversion rate for the new page proportion and conversion rate for the old page proportion respectively.
- Null Hypothesis: Ho: u1=u2 (Conversion rate for new page proportion = Conversion rate for old page proportion)
- Alternate Hypothesis: Ha: u1>u2 (Conversion rate for new page proportion > Conversion rate for old page proportion)
- Test: One-tailed Z test dealing with two population proportions from two independent populations at 5% significance level
- The number of users that serve the new and old pages are 50 and 50 respectively.
- As the p-value 0.016052616408112556 is less than the level of significance, we reject the null hypothesis.
- Inference: We have enough statistical evidence to say the conversion rate proportion for the new page is greater than the conversion rate proportion for the old page.

Q3: Is the conversion and preferred language are independent or related?

Crosstab and Contingency Table



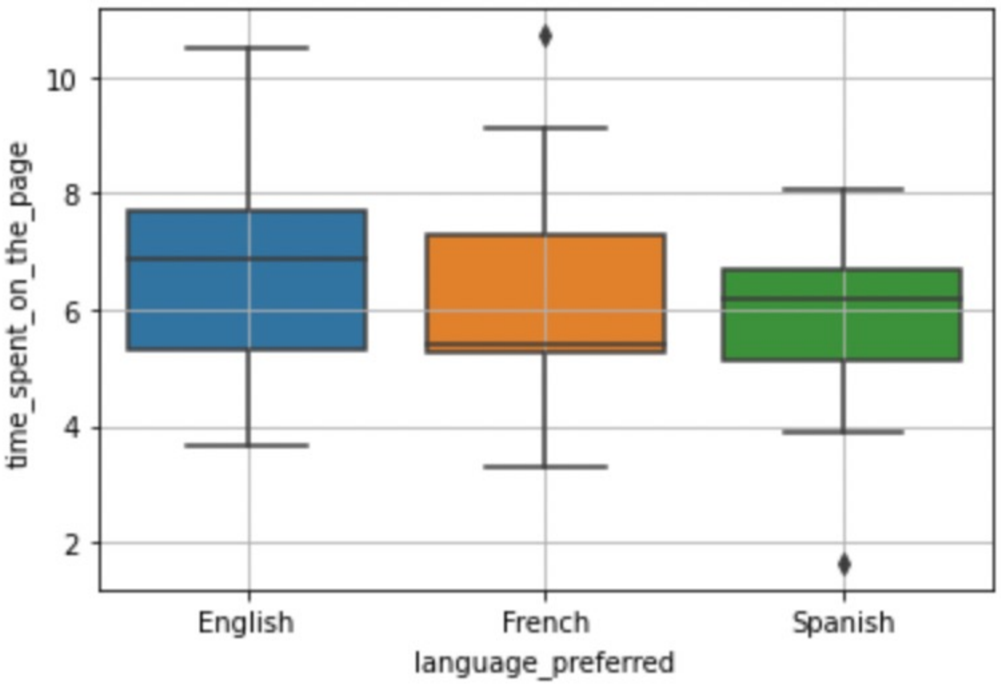


language_preferred	English	French	Spanish
converted			
no	11	19	16
yes	21	15	18

Q3: Is the conversion and preferred language are independent or related?

- Null Hypothesis: Conversion is independent of preferred language.
- Alternate Hypothesis: Conversion is not independent of preferred language.
- Test: Test of independence with 2 categorical variables: conversion status and preferred language at 5% significance level
- As the p-value 0.21298887487543447 is greater than the level of significance, we fail to reject the null hypothesis.
- Inference: We don't have enough statistical evidence to say that conversion is not independent of preferred language.

Q4: Is the time spent on the new page same for the different language users?



Mean time spent on new page by different language users
English 6.663750
French 6.196471
Spanish 5.835294

Q4: Is the time spent on the new page same for the different language users?

- Null Hypothesis: The mean times spent on the new page with respect to each language user is equal.
- Alternate Hypothesis: At least one of the mean time spent with respect to the 3 language users is different.
- Test: ANOVA Test at 5% significance level
- 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 test p-value: 0.8040016293525696
- Levene test p-value: 0.46711357711340173
- ANOVA one-way test p-value: 0.43204138694325955
- As the p-value 0.43204138694325955 is greater than the level of significance, we fail to reject the null hypothesis.
- Inference: We don't have enough statistical evidence to say that mean times spent on new page with respect to three language users are different.

Conclusions and Recommendations

- There is enough statistical evidence to say the mean time that users spent on the new landing page is greater than the mean time users spent on the existing landing page.
- There is enough statistical evidence to say the conversion rate proportion for the new page is greater than the conversion rate proportion for the old page.
- There is not enough statistical evidence to say that conversion is not independent of preferred language.
- There is not enough statistical evidence to say that mean times spent on new page with respect to three language users are different.
- Recommendations: Have a wider variety of preferred language options to see if more users would subscribe and spend more time on the new landing page. Have a translator function in case users want to view the pages in multiple languages to better assess how long users view the pages with the preferred languages.