Project 2

E-News Express

Business Statistics

28th October 2022

Name: Ann Mariya Jomon

Content /Agenda

01 Executive Summary

02 Business problem overview & Solution approach

03 EDA Results

04 Hypothesis tested & Results

05 Appendix

01

Executive Summary

Insights

EDA Results:

- The time spent on the page ranges from 0.2 mins to 10.7 mins.
- Most of the users spend around 4 to 6 mins.
- The total conversion rate is at 54%.
- 50% of the users spend more than 6 mins on new landing page.
- The users who have converted into subscribers spend more time on the page than the non-converted users.
- French and English language users spend more time on the page than Spanish language users. However, there are no major variances in the median time spent for the three languages.

Hypothesis test results:

- Users spend more time on the new landing page than the existing landing page.
- The conversion rate for the new page is greater than the conversion rate for the old page.
- The converted status does not depend on the preferred language.
- Time spent on the new page is same for different language users.

Recommendations

- ✓ The hypothesis tests conclude that the new landing page led to increased conversion rates and time spent by users. Hence, the company should **implement the new design**.
- ✓ **Incentives** such as discounts or coupons should be provided to new users to improve the conversion rates.
- ✓ E-news Express can **compare** their online portal with other similar portals with high reputation and demand. This can help the company to make necessary changes to increase their subscribers.
- ✓ Ensure to keep the landing page **designs up-to-date** and in line with the latest trends and user preferences.
- ✓ Subscribers can be provided with platforms to express their **concerns** or provide relevant **suggestions** to improve the online portal.

Recommendations

- ✓ **Social media marketing** can be adopted by the company to create brand awareness.
- ✓ Even though the language preferences don't have much impact on the time spent or the conversion rates, **introducing more language** options could help to increase the viewers.
- ✓ **Further analysis** will need to be performed on the time spent by users on the page. The focus here should be on what kind of news are they most interested in. This will help to make the portal more attractive.
- ✓ Comparing the time spent by users on other online news portals with the time spent by users on E-news Express can help to understand whether it is less or more than industry average.

02

Business problem overview & Solution approach

Problem Overview

- E-news Express is an online news portal wanting to acquire new subscribers.
- There has been a decline in new monthly subscribers compared to the past year due to the design of their webpage.
- Hence, the design team has created a new landing page.
- 100 users were selected at random and divided into 2 equal groups to analyze the interaction of users between the existing landing page and the new landing page.

Solution Approach

- EDA was performed to get a basic overview of the data along with univariate and bivariate analysis.
- Hypothesis test was conducted based on the questions. Following statistical tests were used for the same:
  - Two independent sample t-test
  - Two proportions z-test
  - Chi-square test of independence
  - One-way ANOVA test
- Insights and recommendations were derived from the above analysis.

03

EDA Results

Link to Appendix on data background check

Data Overview

Shape of dataset

Rows (No. of users)	100
Columns	6

Data types

integer	1
float	1
Object	4

Missing / Duplicated: None

Time spent on the page (statistical summary)

Mean	5.38 mins
Min.	0.19 mins
Median	5.42 mins
Max.	10.71 mins

Statistical Summary of categorical variables

	count	unique	top	freq
group	100	2	control	50
landing_page	100	2	old	50
converted	100	2	yes	54
language_preferred	100	3	Spanish	34

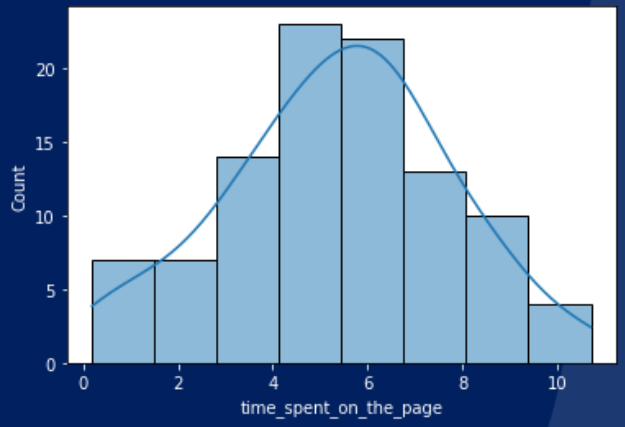
Link to Appendix on data background check

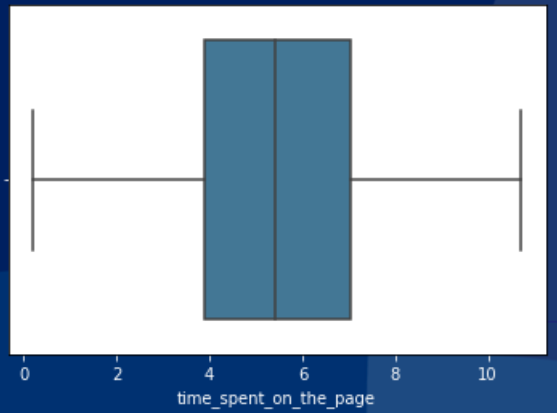
Univariate Analysis

Time spent on the page

Observations:

- The time spent on the page shows a normal distribution.
- It ranges from 0 to 11 mins approx.
- Boxplot has not outliers
- Most of the users spends around 4 to 6 mins.
- Median time is at 5.42 mins.
- 25% of the users spend more than 7 mins on the page.

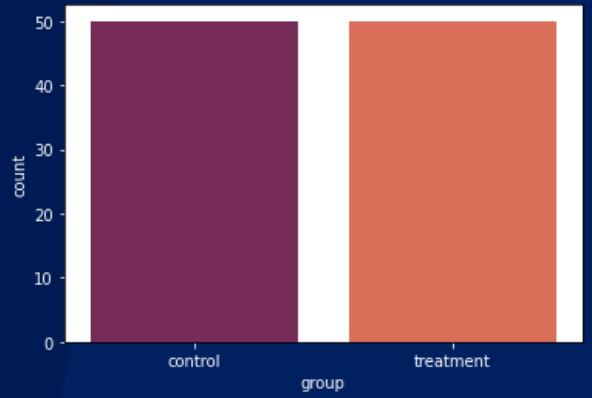




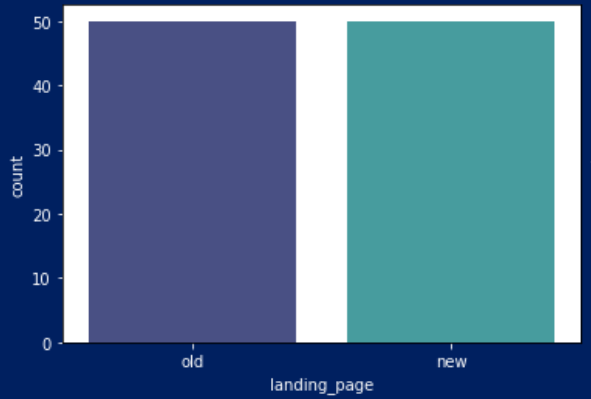
Link to Appendix on data background check

Univariate Analysis

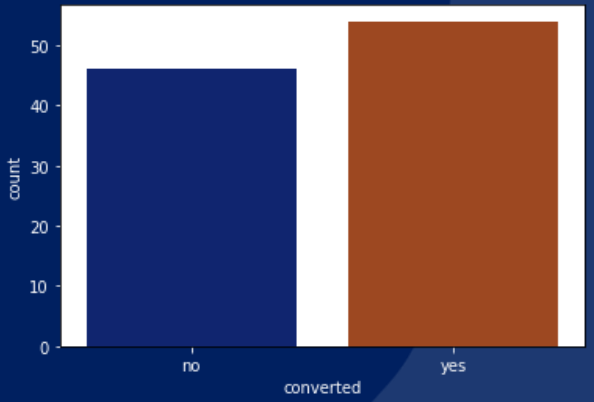
Group



Landing page



Converted



Observations:

- 100 users were divided into equal groups as ‘control’ & ‘treatment’ groups, assigned respectively to the old and new landing pages.
- There are more converted users (54) than non-converted users (46).
- Hence, the total conversion rate is at 54% which is just slightly more than half.

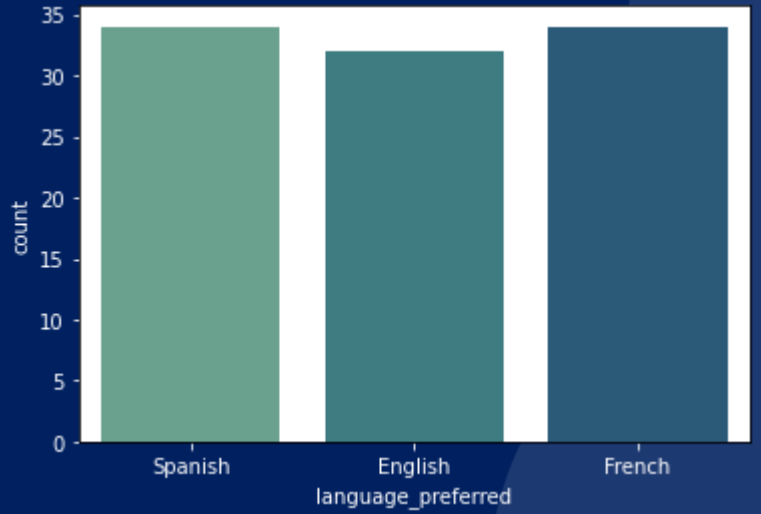
Link to Appendix on data background check

Univariate Analysis

Language preferred

Observations:

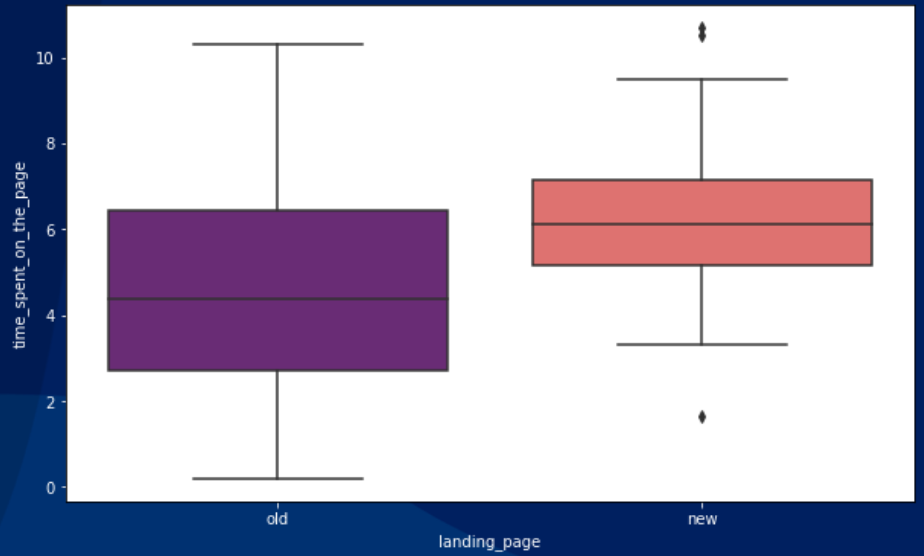
- Spanish and French have equal number of users (34 users) and is the highest.
- 32 users prefer English language.
- There is an almost equal distribution of users on the basis on language preferred.



Link to Appendix on data background check

Bivariate Analysis

Landing page vs Time spent on the page



Observations:

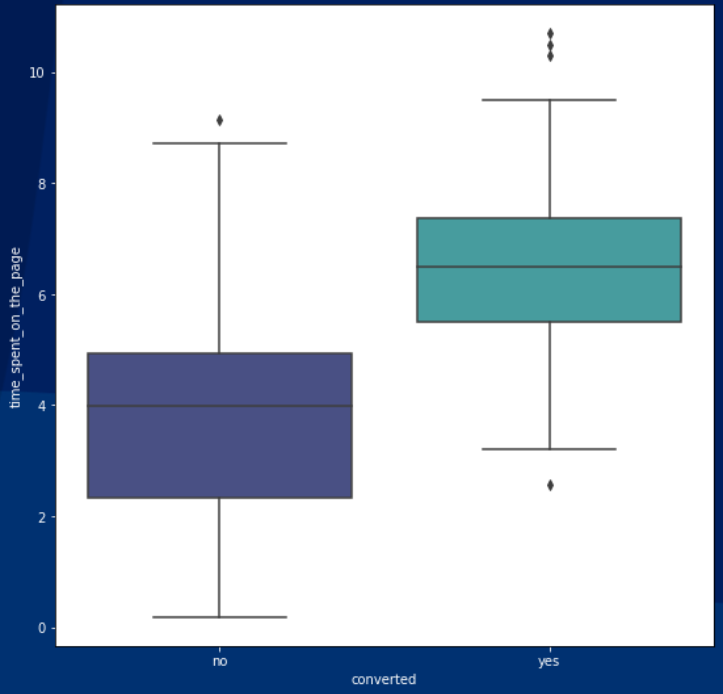
- Users spend more time on the new landing page.
- There are outliers for the new landing page.
- 50% of the users spend more than 6 mins on new landing page.

Landing page	Median (in mins)	Min. (in mins)
Old	4.4	0.2
New	6.1	1.7

Link to Appendix on data background check

Bivariate Analysis

Conversion vs Time spent on the page



Observations:

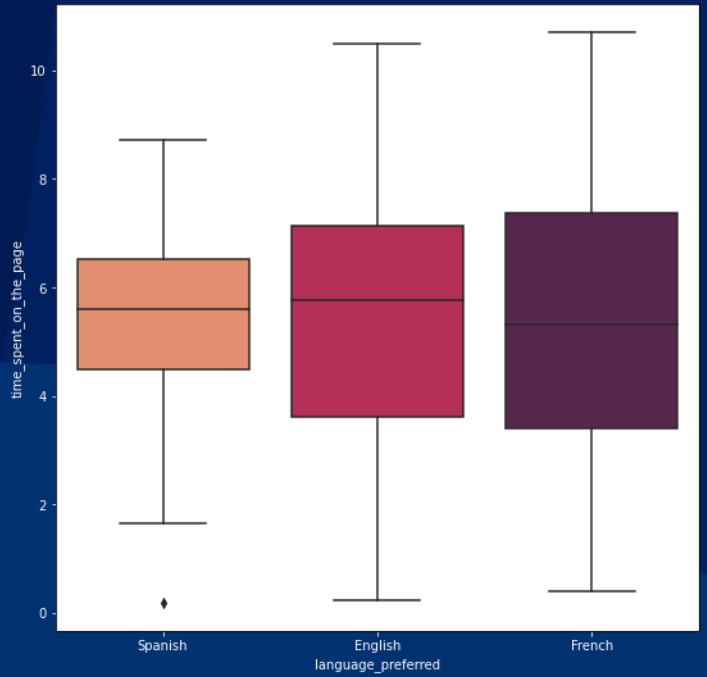
- The users who have converted into subscribers spend more time on the page than the non-converted users.
- There are outliers for both converted and non-converted users.

Converted	Median (in mins)	Min. (in mins)
Yes	6.5	2.6
No	4	0.2

Link to Appendix on data background check

Bivariate Analysis

Language preferred vs Time spent on the page



Observations:

- French and English language users spend more time on the page than Spanish language users.
- However, there are no major variances in the median time spent for the three languages.
- There are outliers for time spent by Spanish language users.

Language	Median (in mins)	Min. (in mins)	Max. (in mins)
Spanish	5.6	0.2	8.7
English	5.8	0.2	10.5
French	5.3	0.4	10.7

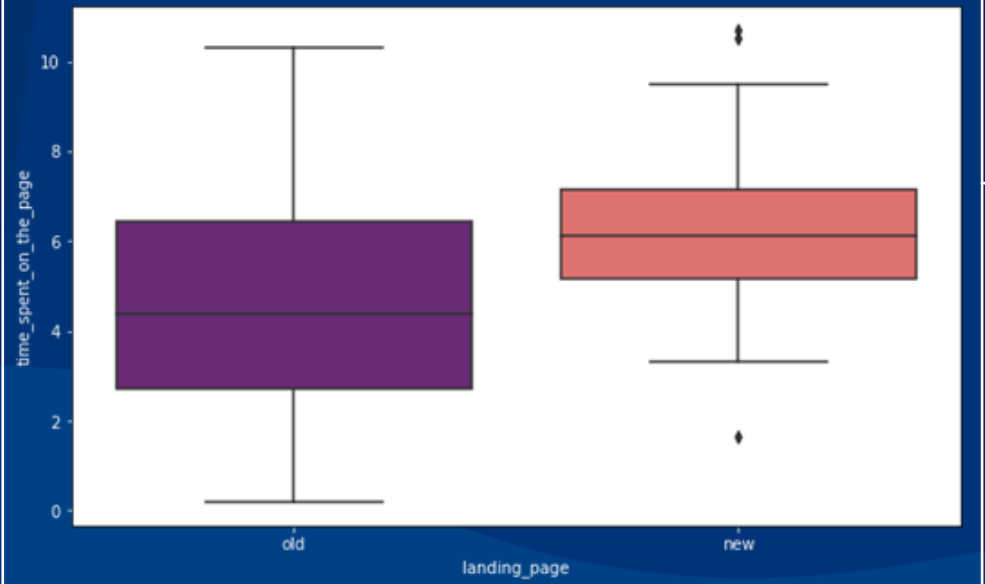
04

Hypothesis tested & Results

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

Link to Appendix slide on details of test performed

Visual Analysis



Hypothesis Tested

$H_0 : \mu_1 \leq \mu_2$

$H_a : \mu_1 > \mu_2$

μ_1: mean time spent on new landing page

μ_2: mean time spent on old landing page

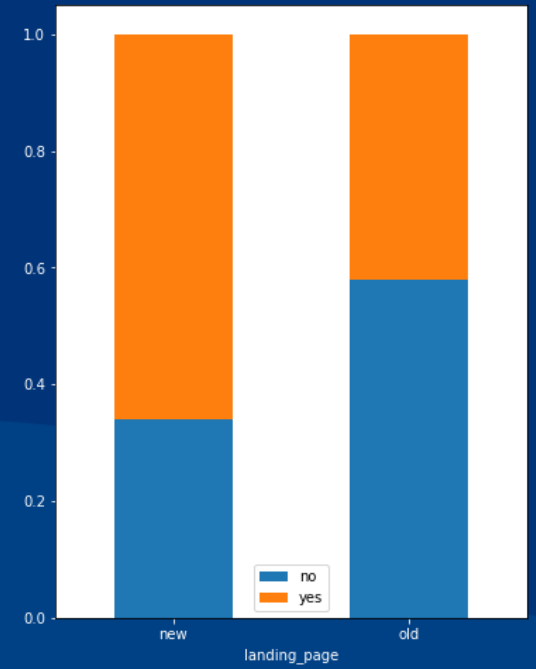
Test result & Inference

- α = 0.05
- p-value = 0
- p-value < 0.05
- Reject null hypothesis.
- Users spend more time on the new landing page than 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?

Link to Appendix slide on details of test performed

Visual Analysis



Hypothesis Tested

$H_0 : p_1 \leq p_2$

$H_a : p_1 > p_2$

p_1: conversion rate of new landing page

p_2: conversion rate of old landing page

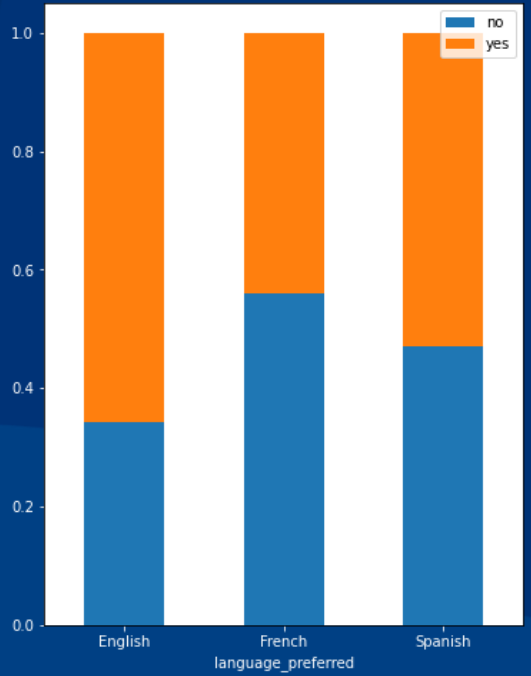
Test result & Inference

- α = 0.05
- p-value = 0.008
- p-value < 0.05
- Reject null hypothesis.
- The conversion rate for the new page is greater than the conversion rate for the old page.

Q3. Does the converted status depend on the preferred language?

Link to Appendix slide on details of test performed

Visual Analysis



Hypothesis Tested

H_0 : converted status is independent of preferred language

H_a : converted status depends on preferred language

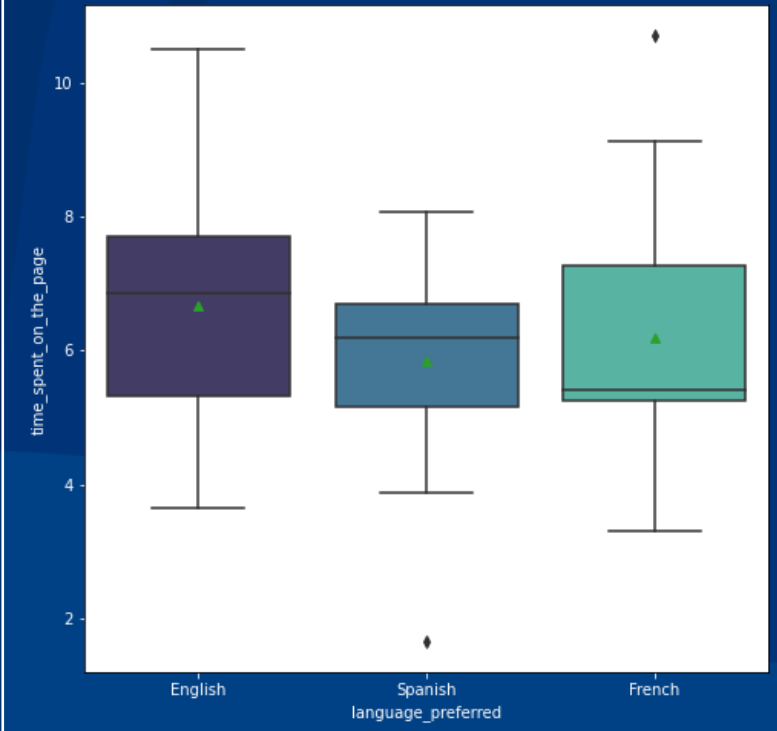
Test result & Inference

- α = 0.05
- p-value = 0.213
- p-value > 0.05
- Fail to reject null hypothesis.
- The converted status does not depend on the preferred language.
- English has the highest converted status.
- French has the lowest converted status.

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

Link to Appendix slide on details of test performed

Visual Analysis



Hypothesis Tested

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_a : At least one of the mean time spent is different

μ_1: Mean time spent by **English** language user on the new page

μ_2: Mean time spent by **Spanish** language user on the new page

μ_3: Mean time spent by **French** language user on the new page

Test result & Inference

- α = 0.05
- p-value = 0.432
- p-value > 0.05
- Fail to reject null hypothesis
- Time spent on the new page is same for different language users.
- The highest time spent is by English language user.

05

Appendix

Data background and Contents

The Dataset contains details of the interaction of users in both the groups with the two versions of the landing page. Following are the variables present in the dataset:

Variable	Description
user_id	Unique user ID of the person visiting the website
group	Group of the user: control / 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
language_preferred	Language chosen by the user to view the landing page (English / French / Spanish)

Data background and Contents

First 5 rows of the dataset

```
In [28]: # view the first 5 rows of the dataset
         df.head()
```

Out[28]:

	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

Last 5 rows of the dataset

```
In [29]: # view the last 5 rows of the dataset
         df.tail()
```

Out[29]:

	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

Hypothesis testing details

Test 1: Q1. Do the users spend more time on the new landing page than the existing landing page?

- Test used: **Two independent sample t-test**
- p-value: 0.000139
- Test statistic: 3.787
- Sample standard deviation of the time spent on the new page: 1.82
- Sample standard deviation of the time spent on the old page: 2.58

Assumptions:

- Continuous data
- Normal distribution
- Independent population
- Unequal population standard deviations
- Random sampling

Step 4: Collect and prepare data

```
In [26]:
# create subsetted data frame for new landing page users
time_spent_new = df[df['landing_page'] == 'new']['time_spent_on_the_page']

# create subsetted data frame for old landing page users
time_spent_old = df[df['landing_page'] == 'old']['time_spent_on_the_page'] ##Complete the code

In [27]:
print('The sample standard deviation of the time spent on the new page is:', round(time_spent_new.std(),2))
print('The sample standard deviation of the time spent on the old page is:', round(time_spent_old.std(),2))

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
```

Step 5: Calculate the p-value

```
In [28]:
# complete the code to import the required function
from scipy.stats import ttest_ind

# write the code to calculate the p-value
test_stat, p_value =  ttest_ind(time_spent_new, time_spent_old, equal_var = False, alternative = 'greater')

print('The p-value is', p_value)
test_stat

The p-value is 0.0001392381225166549
```

Hypothesis testing details

Test 2: 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?

- Test used: **Two proportions z-test**
- p-value: 0.0080
- Test statistic: 2.408
- Converted users in treatment group: 33/50
- Converted users in control group: 21/50

Step 4: Collect and prepare data

In [18]:
```
# calculate the number of converted users in the treatment group
new_converted = df[df['group'] == 'treatment']['converted'].value_counts()['yes']
# calculate the number of converted users in the control group
old_converted = df[df['group'] == 'control']['converted'].value_counts()['yes'] # complete your code here

n_control = df.group.value_counts()['control'] # total number of users in the control group
n_treatment = df.group.value_counts()['treatment'] # total number of users in the treatment group

print('The numbers of users served the new and old pages are {0} and {1} respectively'.format(n_control, n_treatment ))
```

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

Step 5: Calculate the p-value

In [19]:
```
# complete the code to import the required function
from statsmodels.stats.proportion import proportions_ztest

# write the code to calculate the p-value
test_stat, p_value = proportions_ztest([new_converted, old_converted] , [n_treatment, n_control], alternative = 'larger')


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

```
The p-value is 0.008026308204056278
```

Assumptions:

- Binomially distributed populations: converted or not converted
- Binomial distribution approximated to normal distribution – Yes

$$np1 = 50 \times \frac{33}{50} = \mathbf{33} \geq \mathbf{10}$$
$$n(1-p1) = 50 \times \frac{50-33}{50} = \mathbf{17} \geq \mathbf{10}$$
$$np2 = 50 \times \frac{21}{50} = \mathbf{21} \geq \mathbf{10}$$
$$n(1-p2) = 50 \times \frac{50-21}{50} = \mathbf{29} \geq \mathbf{10}$$

- Independent populations
- Random sampling

Hypothesis testing details

Test 3: Q3. Does the converted status depend on the preferred language?

- Test used: Chi-square test of independence
- p-value: 0.213
- Test statistic: 3.093
- Degrees of freedom: 2

Assumptions:

- Categorical variables
- Observations in each level is greater than 5
- Random sampling

Step 4: Collect and prepare data

In [18]:
```
# complete the code to create a contingency table showing the distribution of
contingency_table = pd.crosstab(df['converted'], df['language_preferred'])

contingency_table
```

Out[18]:

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

Step 5: Calculate the p-value

In [21]:
```
# complete the code to import the required function
from scipy.stats import chi2_contingency

# write the code to calculate the p-value
chi2, p_value, dof, exp_freq = chi2_contingency(contingency_table)

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

```
The p-value is 0.21298887487543447
```

Hypothesis testing details

Test 4: Q4. Is the time spent on the new page same for the different language users?

- Test used: One-way ANOVA Test
- p-value: 0.432
- Test statistic: 0.854

Assumptions:

- Normal distribution (proven with Shapiro-Wilk's test)
- Independent simple random samples
- Population variances are equal (proven with Levene's test)

Step 4: Collect and prepare data

```
In [25]:
# create a subsetted data frame of the time spent on the new page by English language users
time_spent_English = df_new[df_new['language_preferred']=="English"]['time_spent_on_the_page']
# create subsetted data frames of the time spent on the new page by French and Spanish language use
time_spent_French = df_new[df_new['language_preferred']=='French']['time_spent_on_the_page']   # c
time_spent_Spanish = df_new[df_new['language_preferred']=='Spanish']['time_spent_on_the_page']   #
```

Step 5: Calculate the p-value

```
In [16]:
# complete the code to import the required function
from scipy.stats import f_oneway

# write the code to calculate the p-value
test_stat, p_value = f_oneway(time_spent_English, time_spent_French, time_spent_Spanish)

print('The p-value is', p_value)
test_stat

The p-value is 0.43204138694325955

Out[16]: 0.8543992770006822
```