E-News Express Project

Business Statistics

May 11, 2023

Contents / Agenda

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

Executive Summary

- Summary of observations and conclusions

  - The mean for both control and treatment group are almost similar, 5.4 minutes after introducing the E-news type.

  - 3 language offered; English, French and Spanish is doing well and having similar proportions of customer.

  - Most reader/customer spending time reading the E-news type is ranging between 4 to 7 minutes.

  - The new landing page of E-news is doing quite well compared with the old page (based on P-value), with customers also spending time more on E-news compared with the old page (based on P-value).

  - This is also similar with new customer, with more customer has converted to a new page regardless language offered (based on P-value).

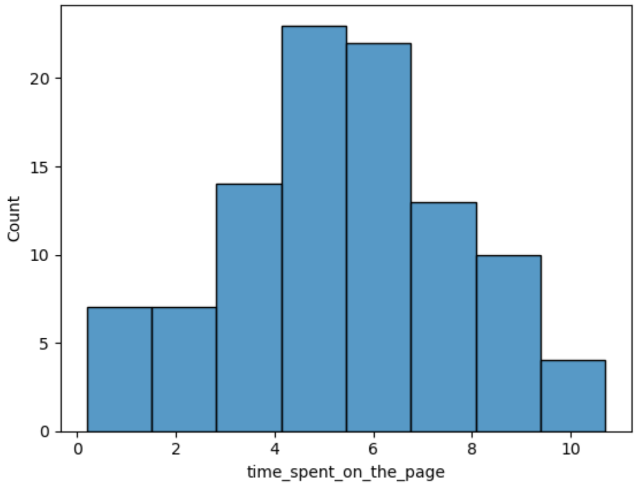
Business Problem Overview and Solution Approach

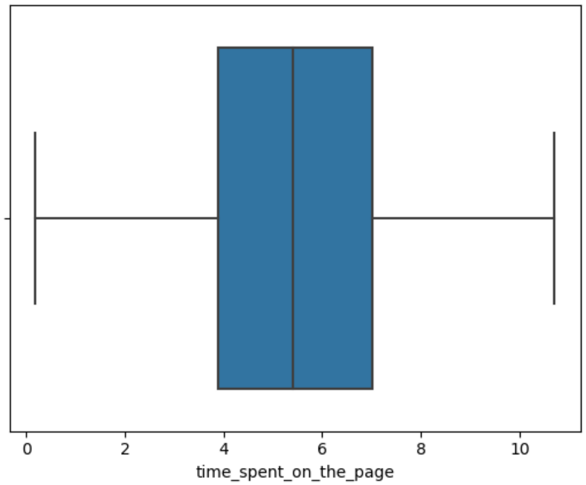
- Business problem overview:
  - The mean for both control and treatment group are almost similar, 5.4 minutes, therefore the company profit is not that high even after implementation of E-news type.
  - Only half of the user has been converted to subscribers. Although it is not bad for transformation from conventional newpapers to E-news, the company needs to improve more from time to time how to increase its subscribers.
  - New page is better than the old page, promising business ahead for electronic newspaper. The company needs to invest further for IT and other apps/software to maintain the current quality offerred.

Business Problem Overview and Solution Approach

- Solution approach/business improvement/recommendation
  - The company needs to look further analysis/improvement on the E-news surface to make it more friendly and reader able to spend more time reading and make a better profit.
  - 3 language offerred is promising, the company may need to look further to expand to other languages, thus attract new customers and expand business to other region.

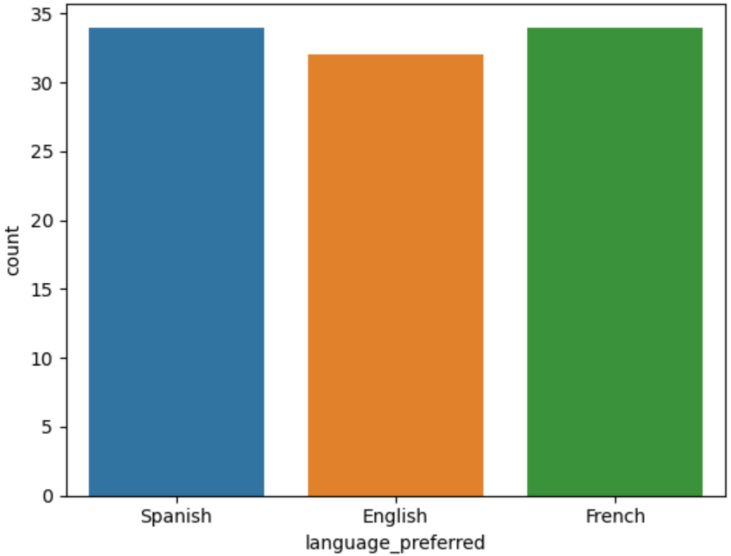
EDA Results – Business Overview after Implementation

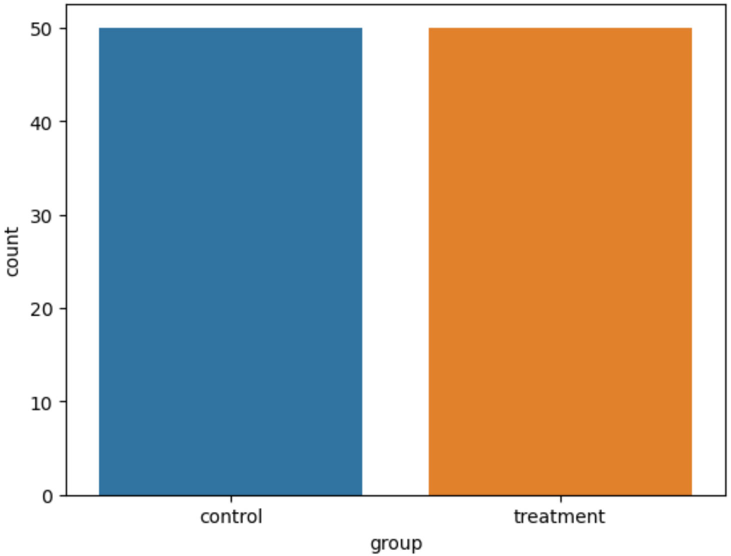




- Result shows treatment and control group spend time on the new page of E-news estimated ranging between 4-7 minutes

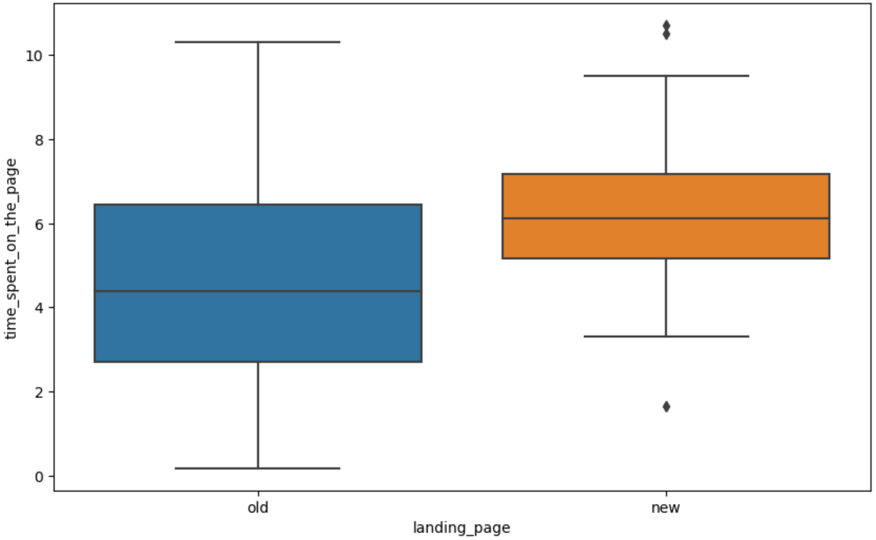
EDA Results – Business Overview after Implementation

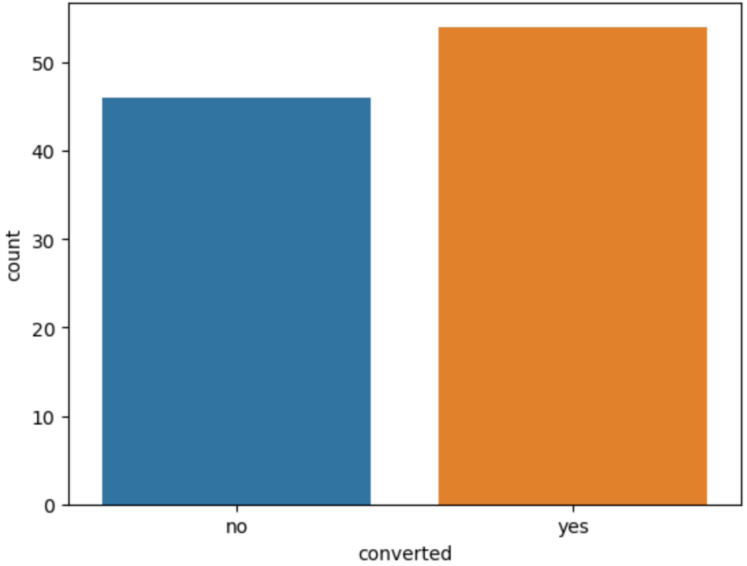




- Graph above shows both groups (treatment and control) do not have any significant impact on the chosen language. All language offered on the E-news; English, French and Spanish has similar proportions of reader/customer.

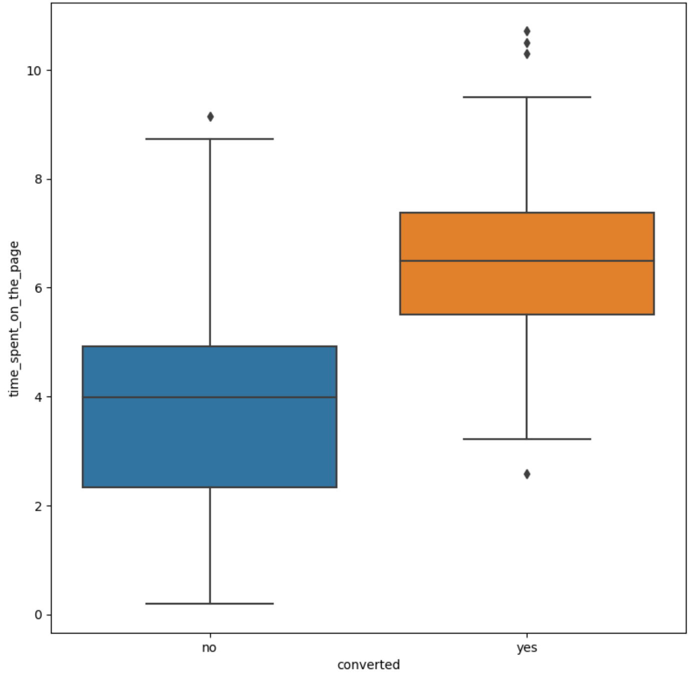
EDA Results – Business Overview after Implementation

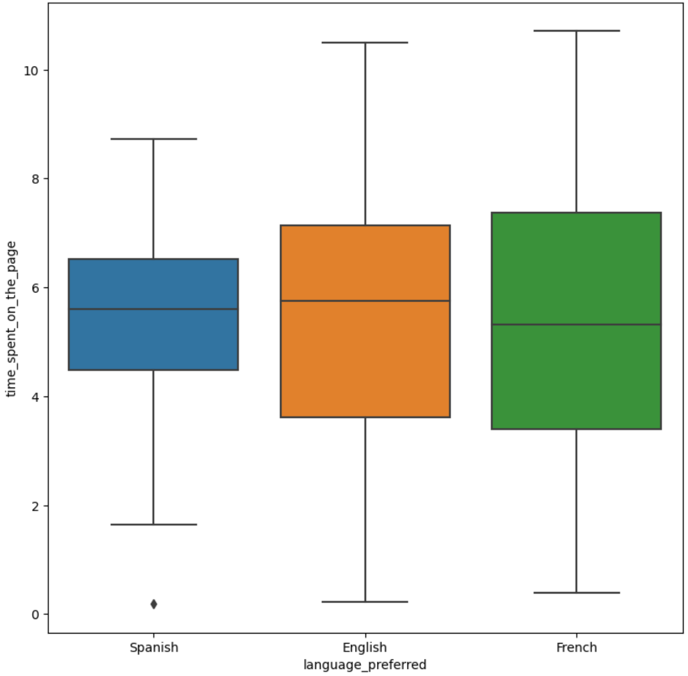




- Customers are comfortable and love the new page, and they have spent more amount of time on the new page, and there is a significant amount of them are converted to a new type of page.

EDA Results – Business Overview after Implementation

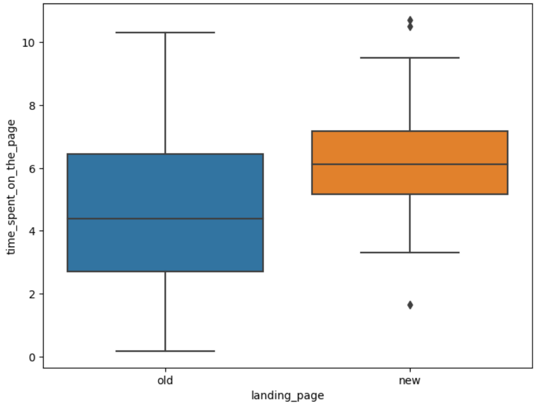




- The converted customer also observed spent more time on the new page compared with non-converted customer, and they spent similar amount of time regardless the language preferred.

Hypotheses Tested and Results – Users spend more time on the new landing page?

- Visual Analysis



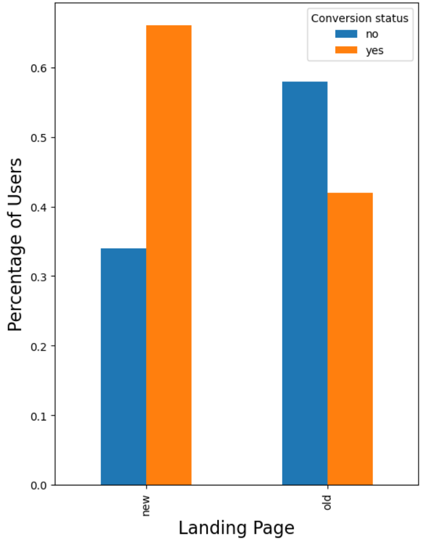
- Null and alternative hypotheses
  - H_0: Users spend more time on the new landing page than existing landing page, μ1 ≥ μ2
    H_a: Users spend less time on the new landing page than existing landing page, μ1 < μ2
  - Since this is two population means from two independent populations and the population standard deviations are known, the appropriate test is by using 2-sample t-test

Hypotheses Tested and Results – Users spend more time on the new landing page?

- Test result and interference
  - Based on the result, P-value is 0.9998683876471904 > α = 0.05, therefore we fail to reject the null hypothesis. The users spend more time on the new landing page compared with the old page.

Hypotheses Tested and Results – Conversion rate for the new page is greater than the conversion rate for the old page?

- Visual Analysis



- Null and alternative hypotheses
  - H_0: Conversion rate for new page is equal and greater than old page, μ1 ≥ μ2

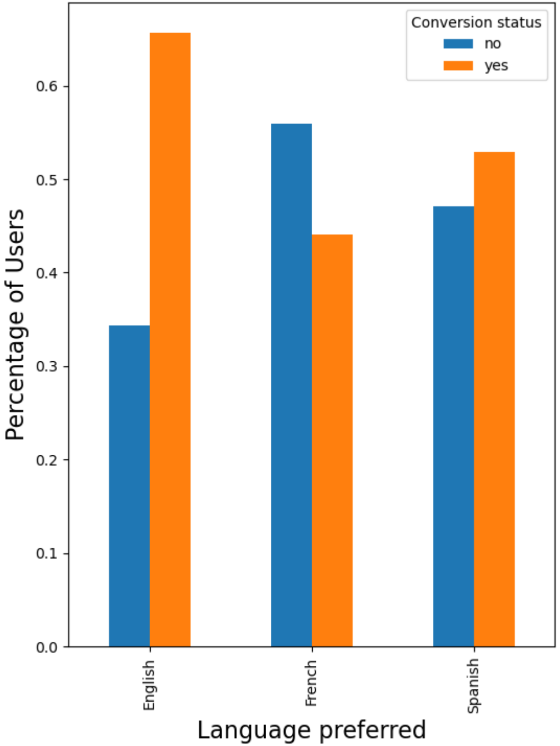
    H_a: Conversion rate for new page is lower than old page, μ1 < μ2

Hypotheses Tested and Results – Conversion rate for the new page greater than the conversion rate for the old page?

- Hypothesis test selected
  - This is a one-tailed test concerning two population proportions from two independent populations. It is random sampling from the population and the appropriate test would be 2-proportion z-test
- Test result and interference
  - Based on the result, P-value is 0.9919736917959437 > α = 0.05, therefore we fail to reject the null hypothesis. The conversion rate for the new page is equal and greater than the conversion rate for the old page.

Hypotheses Tested and Results – Converted status depend on the preferred language?

- Visual analysis and contingency table



Language preferred	Converted: No	Converted: Yes
English	11	21
French	19	15
Spanish	16	18

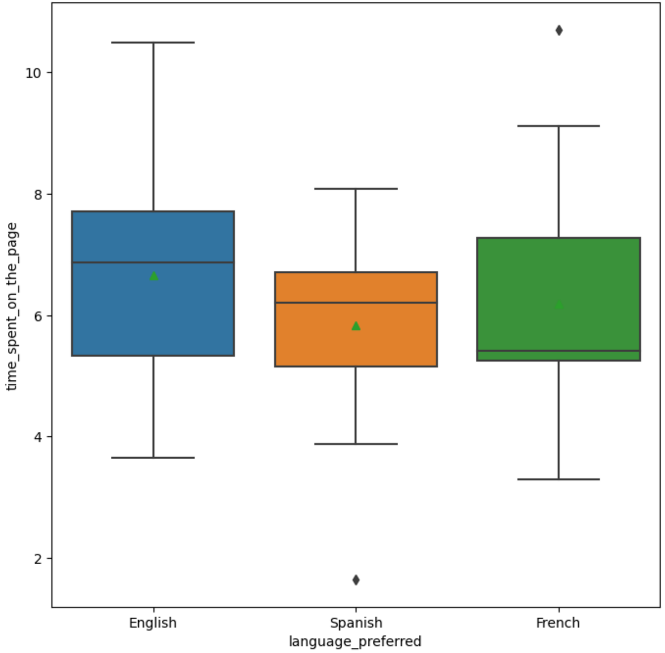
Hypotheses Tested and Results – Converted status depend on the preferred language?

- Null and alternative hypotheses
  - H_0: Conversion status is not depend on language preferred

    H_a: Conversion status depends on language preferred
- Hypothesis test selected
  - This is a problem of the test of independence, concerning two categorical variables - converted status and preferred language. We can try to use chi-square test.
- Test result and interference
  - Based on the result, P-value is 0.2129888748 > $\alpha = 0.05$, therefore we fail to reject the null hypothesis. The converted status is not depend on the preferred language.

Hypotheses Tested and Results – Time spent on the new page same for the different language users?

- Visual Analysis



Language preferred	Time_spent_on_the_page
English	6.663750
French	6.196471
Spanish	5.835294

Hypotheses Tested and Results – Time spent on the new page same for the different language users?

- Null and alternative hypotheses
  - H_0: All time spent on the new page is same for all language users, $\mu1 = \mu2 = \mu3$

    H_a: At least one group of language users is different
- Hypothesis test selected
  - This is about three population means. Therefore, ANOVA test is more appropriate. However, we need to ensure that it meets requirement for other test's first, which is normality testing, Shapiro-Wilk's test and equality of variance test, Levene's test
  - Shapiro_Wilk's test, P-value is 0.5643193125724792 > $\alpha = 0.05$, therefore it follows normal distribution
  - Levene's test, P-value is 0.46711357711340173 > $\alpha = 0.05$, therefore all population variance is equal

Hypotheses Tested and Results – Time spent on the new page same for the different language users?

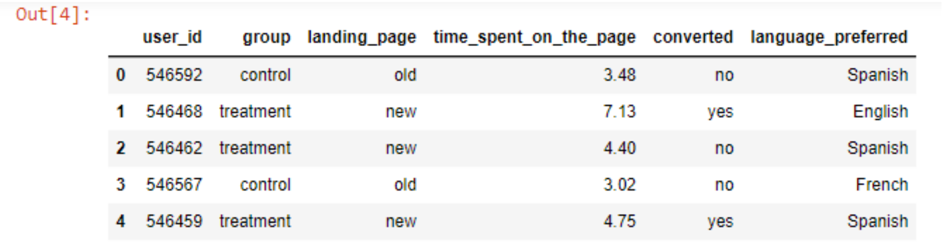
- Test result and interference
    - P-value is 0.43204138694325955 > α = 0.05, therefore we fail to reject the null hypothesis. Therefore, all time spent on the new page is same for all language users

APPENDIX

Data Background and Contents – Data Overview

- Displaying the first few rows and the last few rows of the dataset

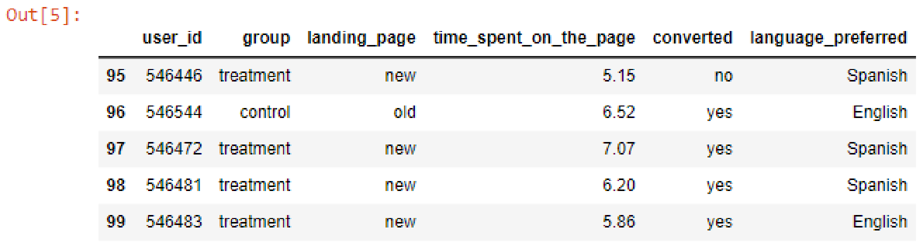
  - df.head()



Out[4]:

	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

  - df.tail()

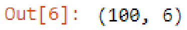


Out[5]:

	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

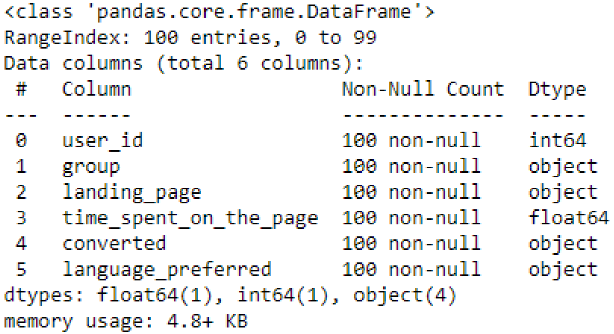
Data Background and Contents – Data Overview

- Checking shape of the the dataset
  - df.shape



```
Out[6]: (100, 6)
```

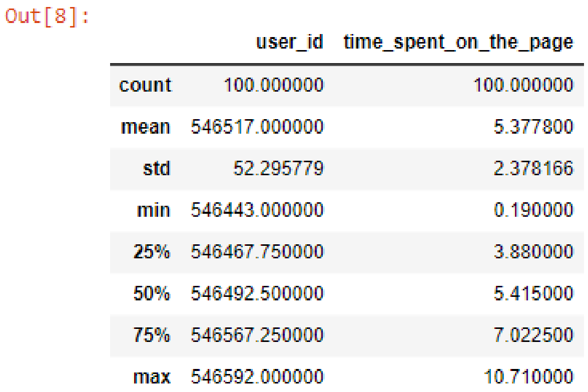
- Checking data types for the dataset
  - df.info()



```
<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
```

Data Background and Contents – Data Overview

- Displaying numerical statistical summary
  - df.describe()

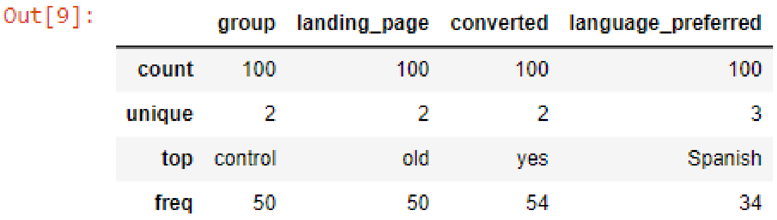


Out[8]:

	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

Data Background and Contents – Data Overview

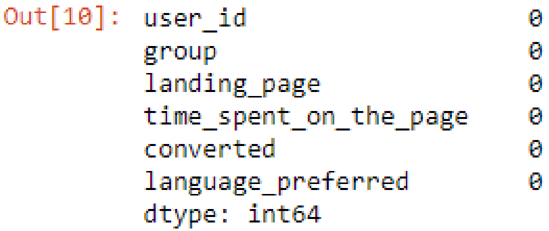
- Displaying categorical statistical summary



Out[9]:

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

- Checking for missing value

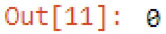


```
Out[10]: user_id                  0
         group                    0
         landing_page             0
         time_spent_on_the_page   0
         converted                0
         language_preferred       0
         dtype: int64
```

Data Background and Contents – Data Overview

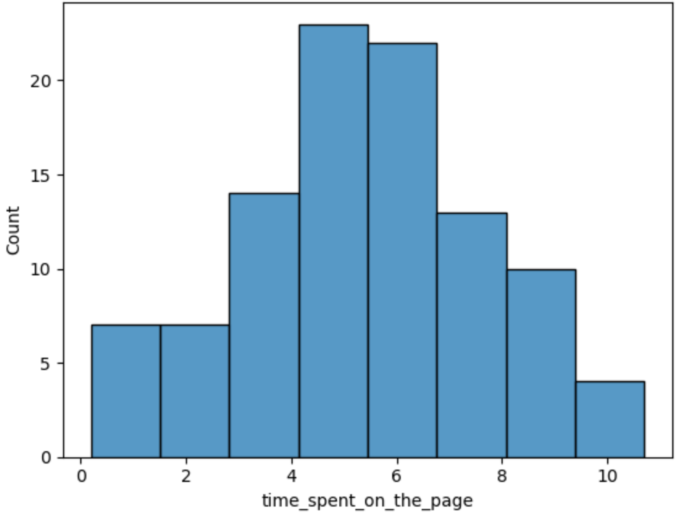
- Checking for duplicates

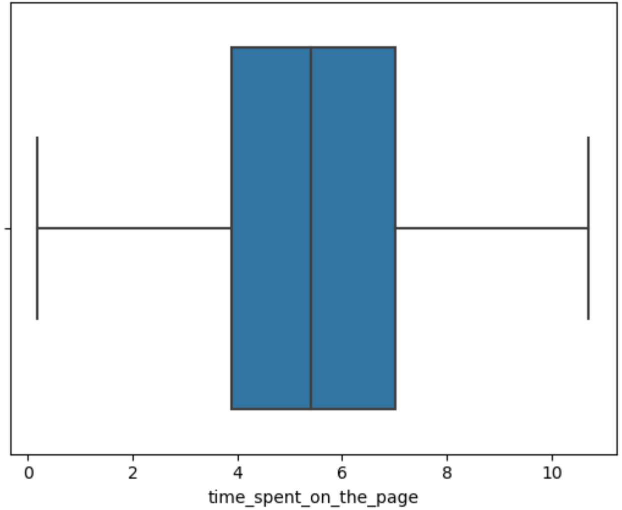
  - df.duplicated().sum()



```
Out[11]: 0
```

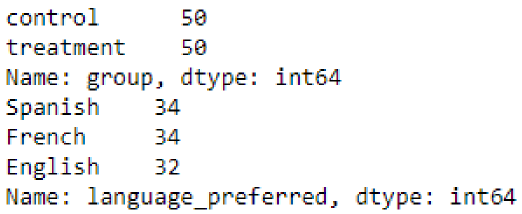
  - Time spent on the page



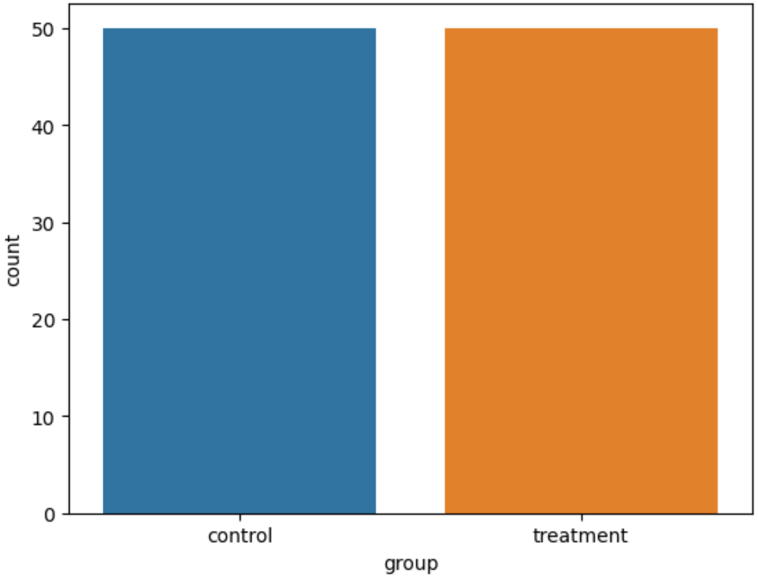


Data Background and Contents – Univariate Analysis

- Group

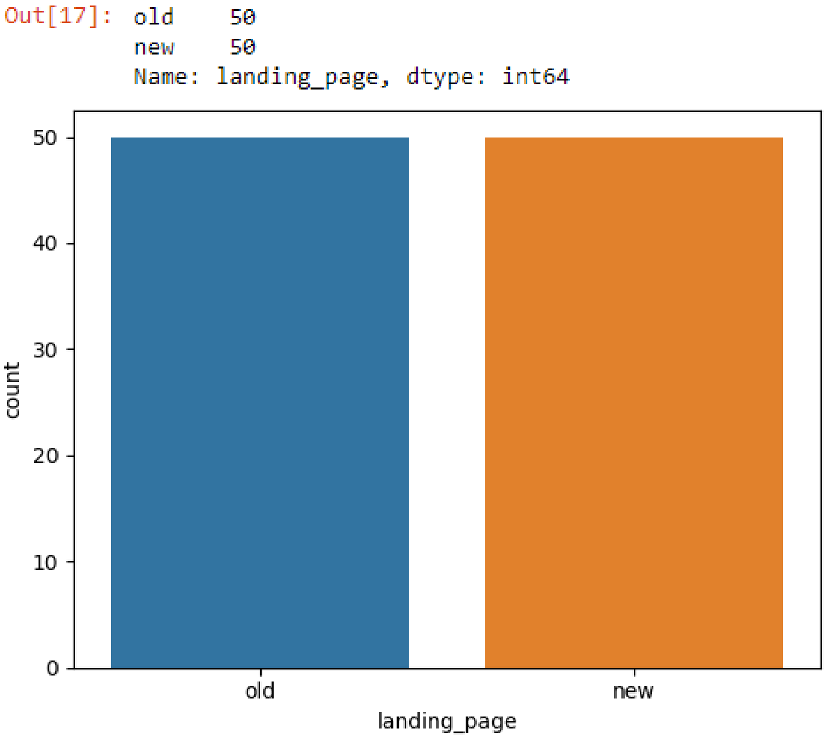


```
control      50
treatment    50
Name: group, dtype: int64
Spanish    34
French     34
English    32
Name: language_preferred, dtype: int64
```



Data Background and Contents – Univariate Analysis

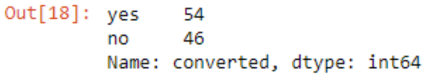
- Landing page



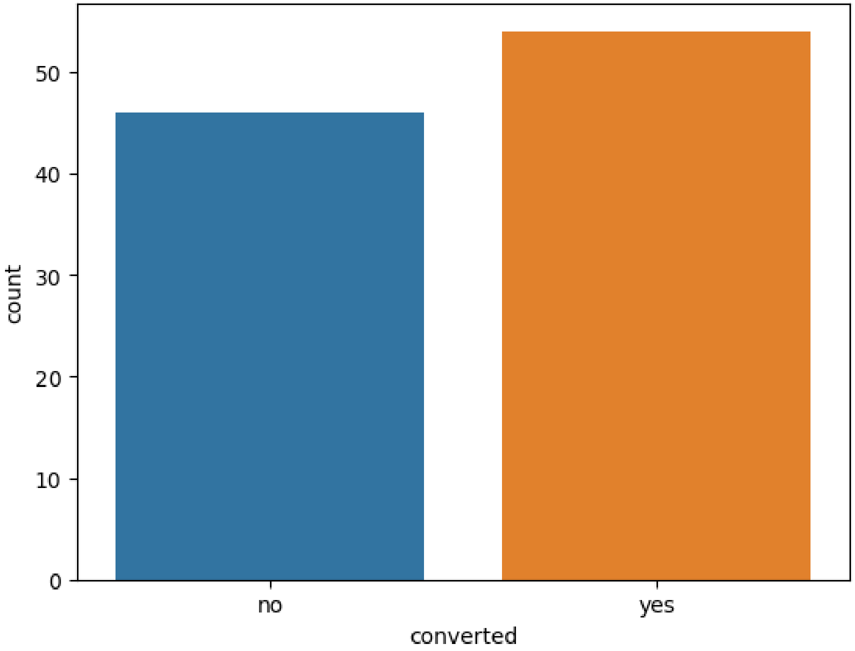
```
Out[17]: old    50
         new    50
         Name: landing_page, dtype: int64
```

Data Background and Contents – Univariate Analysis

- Converted

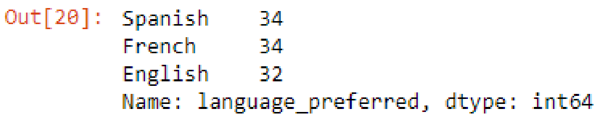


```
Out[18]: yes    54
         no     46
         Name: converted, dtype: int64
```

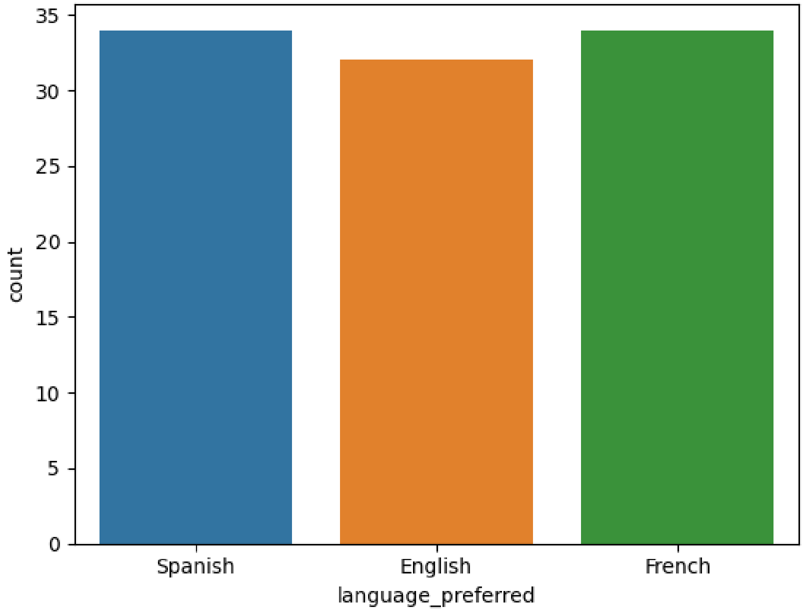


Data Background and Contents – Univariate Analysis

- Language preferred

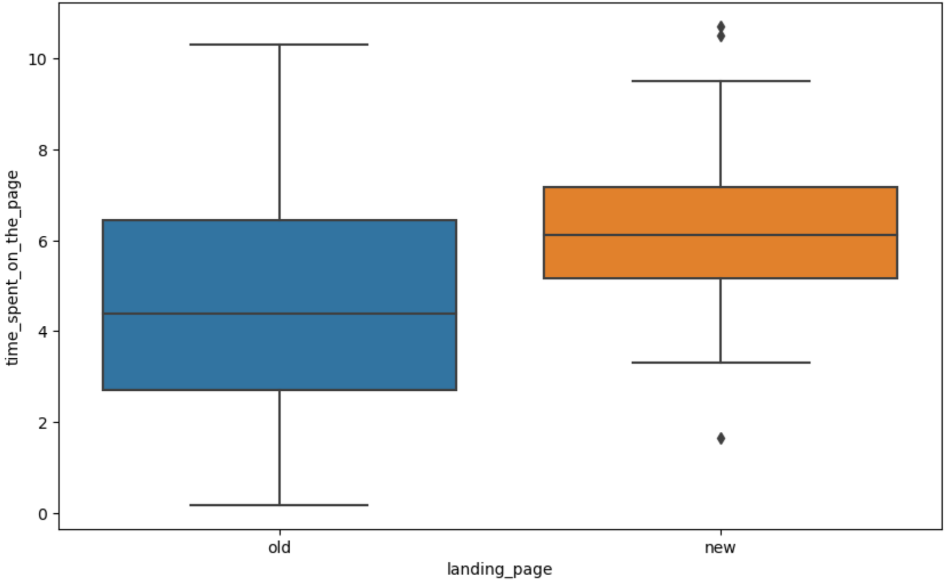


```
Out[20]: Spanish    34
         French     34
         English    32
         Name: language_preferred, dtype: int64
```



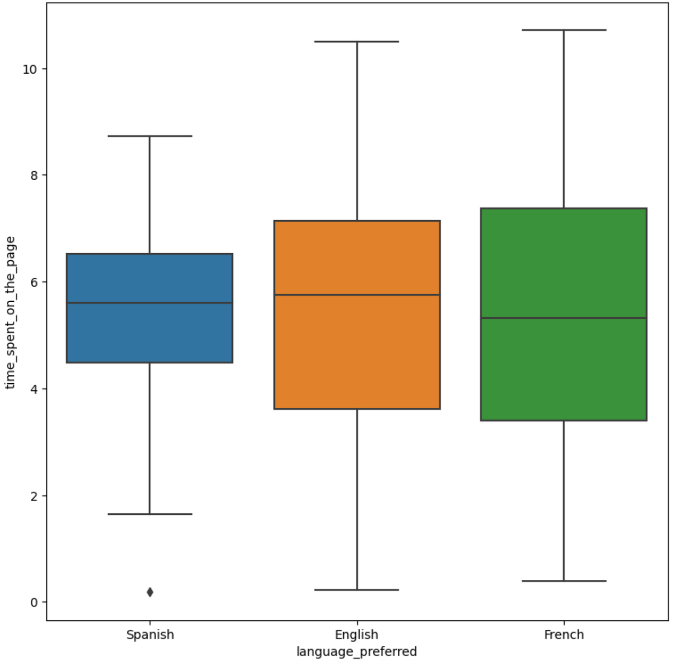
Data Background and Contents – Bivariate Analysis

- Landing page vs Time spent on the page



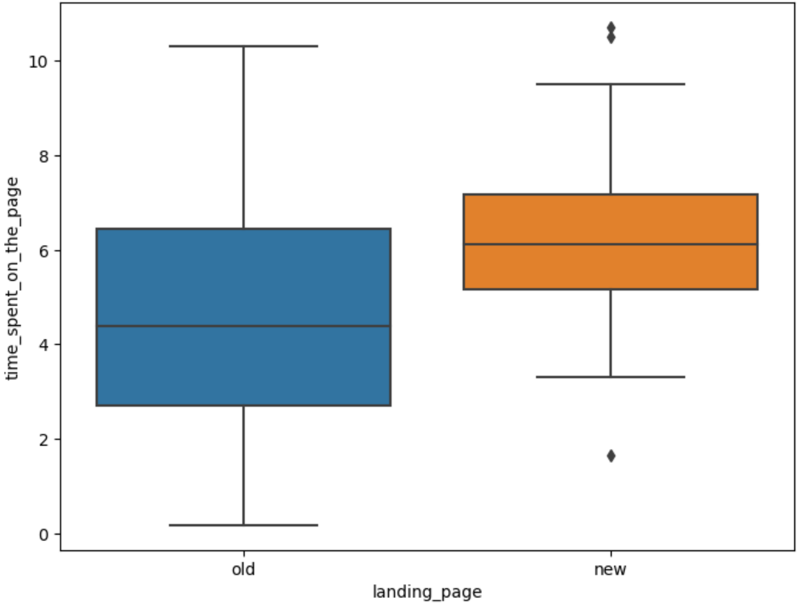
Data Background and Contents – Bivariate Analysis

- Language preferred vs Time spent on the page



Hypothesis Testing Details

- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - Visual analysis



Hypothesis Testing Details

- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - Null and alternative hypotheses
    - H_0: Users spend more time on the new landing page than existing landing page, $\mu1 \geq \mu2$
      H_a: Users spend less time on the new landing page than existing landing page, $\mu1 < \mu2$
  - Hypothesis test selected
    - This is a one-tailed test concerning two population means from two independent populations. The population standard deviations are known. The appropriate test is by using 2-sample t-test
    - Significance level, $\alpha = 0.05$

Hypothesis Testing Details

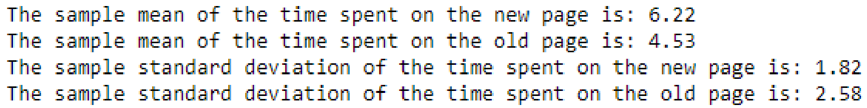
- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - Collect and analyze data (mean, std dev, z-scores)

```
# 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’]

#mean
print('The sample mean of the time spent on the new page is:', round(time_spent_new.mean(),2))
print('The sample mean of the time spent on the old page is:', round(time_spent_old.mean(),2))

#standard deviation
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))
```

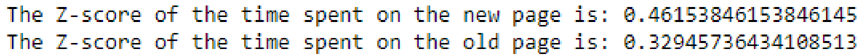
Hypothesis Testing Details

- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - Collect and analyze data (mean, std dev, z-scores)

```
#find the z-score
new_page = (6.22-5.38) / 1.82
print('The Z-score of the time spent on the new page is:', new_page)
old_page = (5.38-4.53) / 2.58
print('The Z-score of the time spent on the old page is:', old_page)
```



```
The sample mean of the time spent on the new page is: 6.22
The sample mean of the time spent on the old page is: 4.53
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
```



```
The Z-score of the time spent on the new page is: 0.46153846153846145
The Z-score of the time spent on the old page is: 0.32945736434108513
```

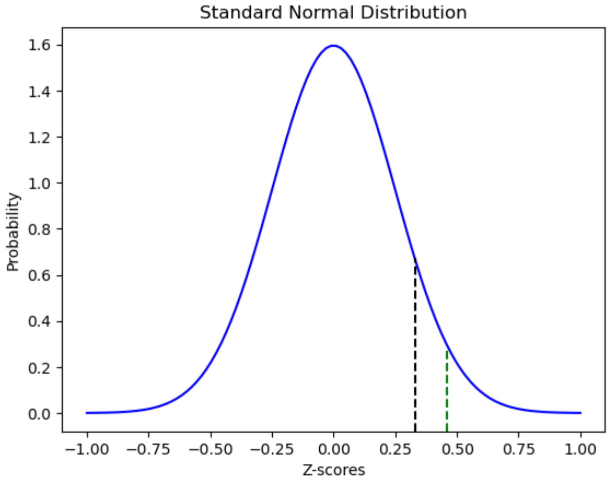
Hypothesis Testing Details

- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - **Based on the sample standard deviations of the two groups, decide whether the population standard deviations can be assumed to be equal or unequal**

```
from scipy.stats import norm

# plot the standard normal distribution and visualize the standardized scores
# We are plotting the distributions here to better visualize the calculations.

fig, ax = plt.subplots()
x = np.linspace(-1,1,100)
ax.plot(x, norm.pdf(x, loc = 0, scale = 0.25), color = 'b')
ax.set_title('Standard Normal Distribution')
ax.set_xlabel('Z-scores')
ax.set_ylabel('Probability')
ax.axvline(new_page, ymax = 0.2, linestyle = '--', color = 'green')
ax.axvline(old_page, ymax = 0.43, linestyle = '--', color = 'black')
plt.show()
```

Hypothesis Testing Details

- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - **Based on the sample standard deviations of the two groups, decide whether the population standard deviations can be assumed to be equal or unequal**
    - Based on z-scores and plotting standard normal distribution, both of them can be assume unequal. However, we need to re-confirm using P-value calculation.



Hypothesis Testing Details

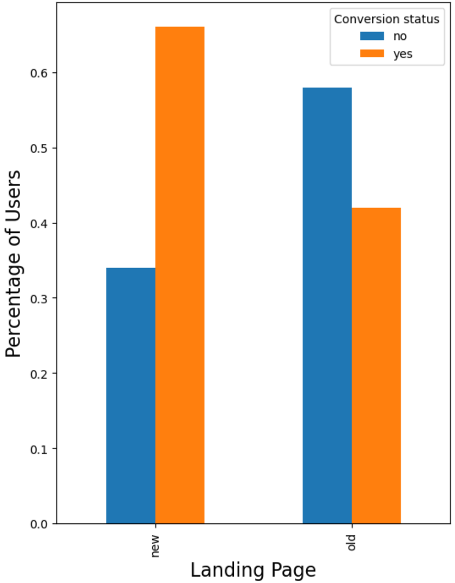
- Q1: **Do the users spend more time on the new landing page than the existing landing page?**
  - P-value

```
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 = True,
alternative = 'less')  #complete the code by filling appropriate parameters in the
blanks

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

  - Based on the result, P-value is 0.9998683876471904 > α = 0.05, therefore we fail to reject the null hypothesis.

Hypothesis Testing Details

- 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?**
  - Visual analysis



Hypothesis Testing Details

- 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?**
  - Null and alternative hypotheses
    - H_0: Conversion rate for new page is equal and greater than old page, $\mu1 \geq \mu2$

      H_a: Conversion rate for new page is lower than old page, $\mu1 < \mu2$
  - Hypothesis test selected
    - This is a one-tailed test concerning two population proportions from two independent populations. It is random sampling from the population and the appropriate test would be 2-proportion z-test
    - Significance level, $\alpha = 0.05$

Hypothesis Testing Details

- 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?**
    - P-value

```
# 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']

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 ))
```

Hypothesis Testing Details

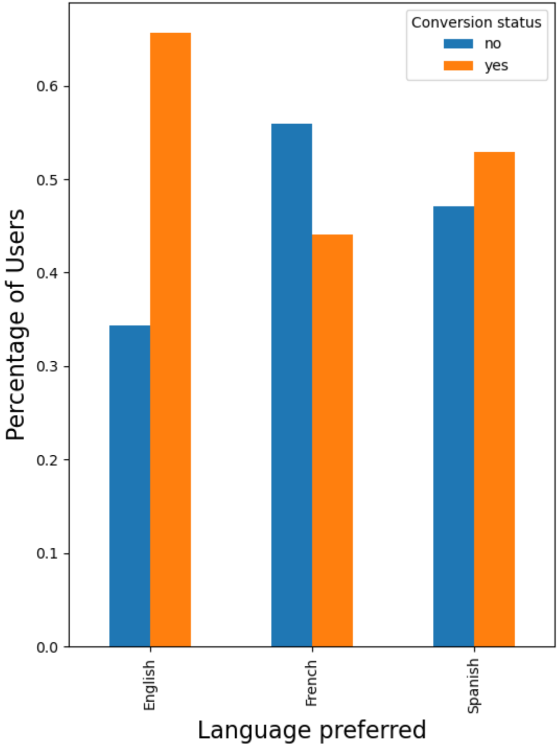
- Q2: **Is the conversion rate (the proportion of users who visit the landing page and get converted) for the new page greater than the conversion rate for the old page?**
    - P-value

      ```
      from statsmodels.stats.proportion import proportions_ztest

      test_stat, p_value = proportions_ztest([new_converted, old_converted] , [n_treatment,
      n_control], alternative ='smaller')   #complete the code by filling appropriate parameters in the
      blanks

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

        - Based on the result, P-value is 0.9919736917959437 > α = 0.05, therefore we fail to reject the null hypothesis.

Hypothesis Testing Details

- Q3: **Does the converted status depend on the preferred language?**
    - Visual analysis



Hypothesis Testing Details

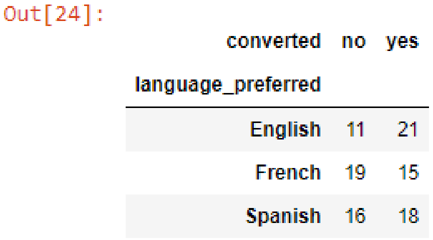
- Q3: **Does the converted status depend on the preferred language?**
  - Null and alternative hypotheses
    - H_0: Conversion status is not depend on language preferred

      H_a: Conversion status depends on language preferred
  - Hypothesis test selected
    - This is a problem of the test of independence, concerning two categorical variables - converted status and preferred language. We can try to use chi-square test.
    - Significance level, $\alpha = 0.05$

Hypothesis Testing Details

- Q3: **Does the converted status depend on the preferred language?**
  - P-value

```
contingency_table = pd.crosstab(df['language_preferred'], df['converted'])

contingency_table
```



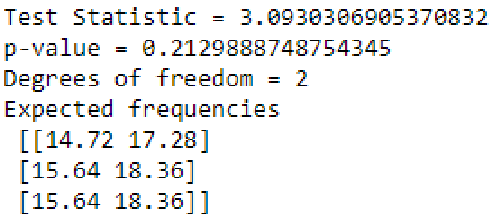
```
Out[24]:
```

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

Hypothesis Testing Details

- Q3: **Does the converted status depend on the preferred language?**
  - P-value

```
from scipy.stats import chi2_contingency

chi, p_value, dof, exp_freq = chi2_contingency(contingency_table)

print("Test Statistic =",chi)
print("p-value =",p_value)
print("Degrees of freedom =",dof)
print("Expected frequencies \n", exp_freq)
```

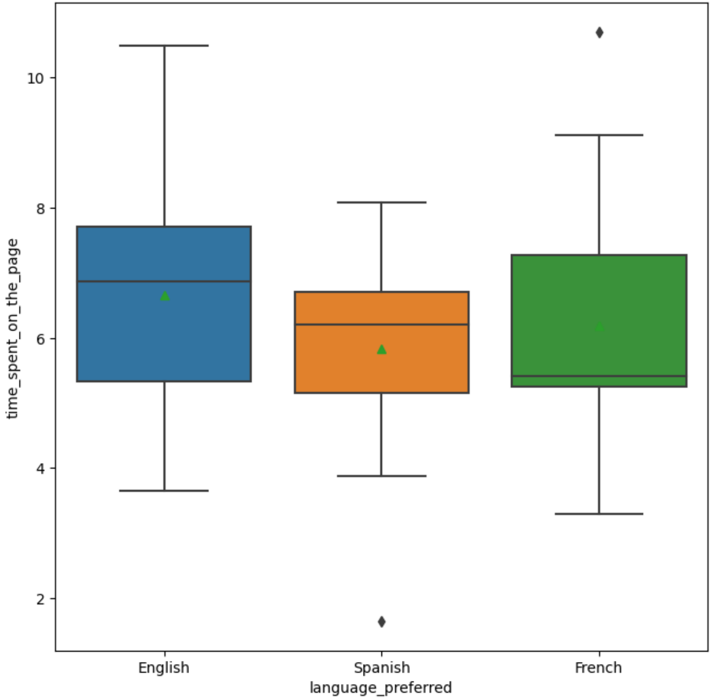


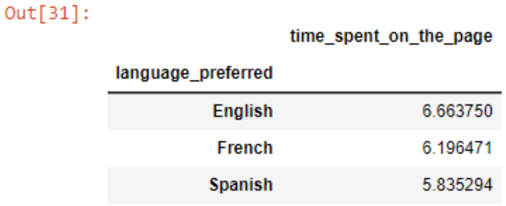
```
Test Statistic = 3.0930306905370832
p-value = 0.2129888748754345
Degrees of freedom = 2
Expected frequencies
 [[14.72 17.28]
 [15.64 18.36]
 [15.64 18.36]]
```

- Based on the result above, P-value > α = 0.05, therefore we fail to reject the null hypothesis.

Hypothesis Testing Details

- Q4: **Is the time spent on the new page same for the different language users?**
    - ○ Visual analysis





```
Out[31]:
```

	time_spent_on_the_page
language_preferred	
English	6.663750
French	6.196471
Spanish	5.835294

Hypothesis Testing Details

- Q4: **Is the time spent on the new page same for the different language users?**
  - Null and alternative hypotheses
    - H_0: All time spent on the new page is same for all language users, μ1 = μ2 = μ3

      H_a: At least one group of language users is different
  - Hypothesis test selected
    - This is a problem, concerning three population means. Therefore, ANOVA test is more appropriate.
    - Significance level, α = 0.05
    - We need to test for normality testing, Shapiro-Wilk’s test and equality of variance test, Levene’s test

Hypothesis Testing Details

- Q4: **Is the time spent on the new page same for the different language users?**
    - Shapiro_Wilk's test
        - H_0: The time spent on the new page is follows normal distribution

          H_a: The time spent on the new page is does not follows normal distribution
    - P-value

      ```
      from scipy import stats

      w, p_value = stats.shapiro(df['time_spent_on_the_page'])
      print('The p-value is', p_value)
      ```

        - P-value is 0.5643193125724792 > α = 0.05, therefore it follows normal distribution

Hypothesis Testing Details

- Q4: **Is the time spent on the new page same for the different language users?**
  - Levene’s test
    - H_0: All population variance is equal

      H_a: At least one variance is different
  - P-value

    ```
    from scipy.stats import levene

    statistic, p_value = levene(time_spent_English,time_spent_French,time_spent_Spanish)
    print('The p-value is', p_value)
    ```

    - P-value is 0.46711357711340173 > $\alpha = 0.05$, therefore all population variance is equal

Hypothesis Testing Details

- Q4: **Is the time spent on the new page same for the different language users?**
    - P-value

      ```
      from scipy.stats import f_oneway

      # perform one-way anova test
      test_stat, p_value = f_oneway(time_spent_English,time_spent_French,time_spent_Spanish)
      print('The p-value is ', p_value)
      ```

        - P-value is 0.43204138694325955 > α = 0.05, therefore we fail to reject the null hypothesis. Therefore, all time spent on the new page is same for all language users

Great Learning

Happy Learning !