E-News Express Project2

Brandy Murray

Objective

To perform a statistical analysis of the business data. Explore the dataset and extract insights from the data. The idea is for you to get comfortable with doing statistical analysis in Python.

Problems to tackle:

- Explore the dataset and extract insights using EDA.
- Do the users spend more time on the new landing page than the old landing page?
- 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?
- Does the converted status depend on the preferred language?
  [HINT: Create a contingency table using the pandas.crosstab() function]
- Is the mean time spent on the new page same for the different language users?

Data Dictionary

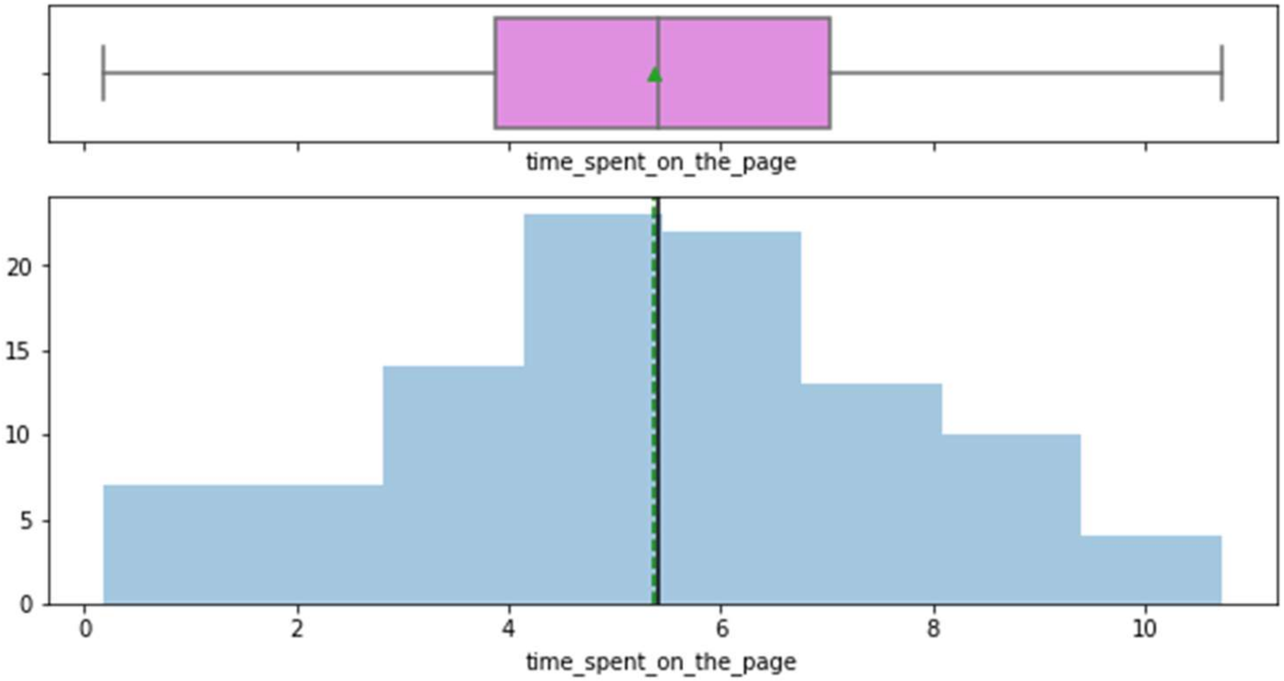
A. user_id : Represents the user ID of the person visiting the website.
B. group - Represents whether the user belongs to the first group (control) or the second group (treatment).
C. landing_page : Represents whether the landing page is new or old.
D. time_spent_on_the_page : Represents the time (IN MINUTES) spent by the user on the landing page.
E. converted : Represents whether the user gets converted to a subscriber of the news portal or not.
F. language_preferred : Represents the language chosen by the user to view the landing page.

EDA

Observations on Time Spent on the Page

Observations

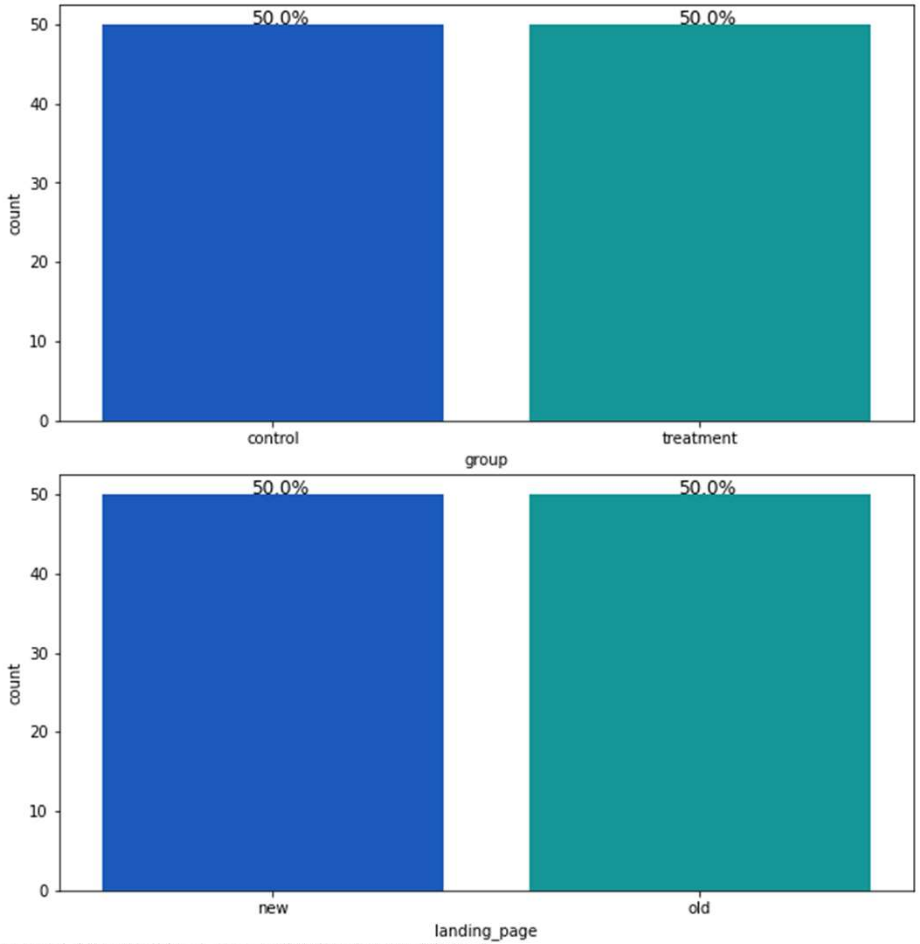
- There are no outliers in with this variable.
- From the boxplot, we can see that the third quartile(Q3) is equal to 7.02 seconds which means that 75% of the customers spent less than 7.02 seconds on the page.



Observations on Landing Page

Observations

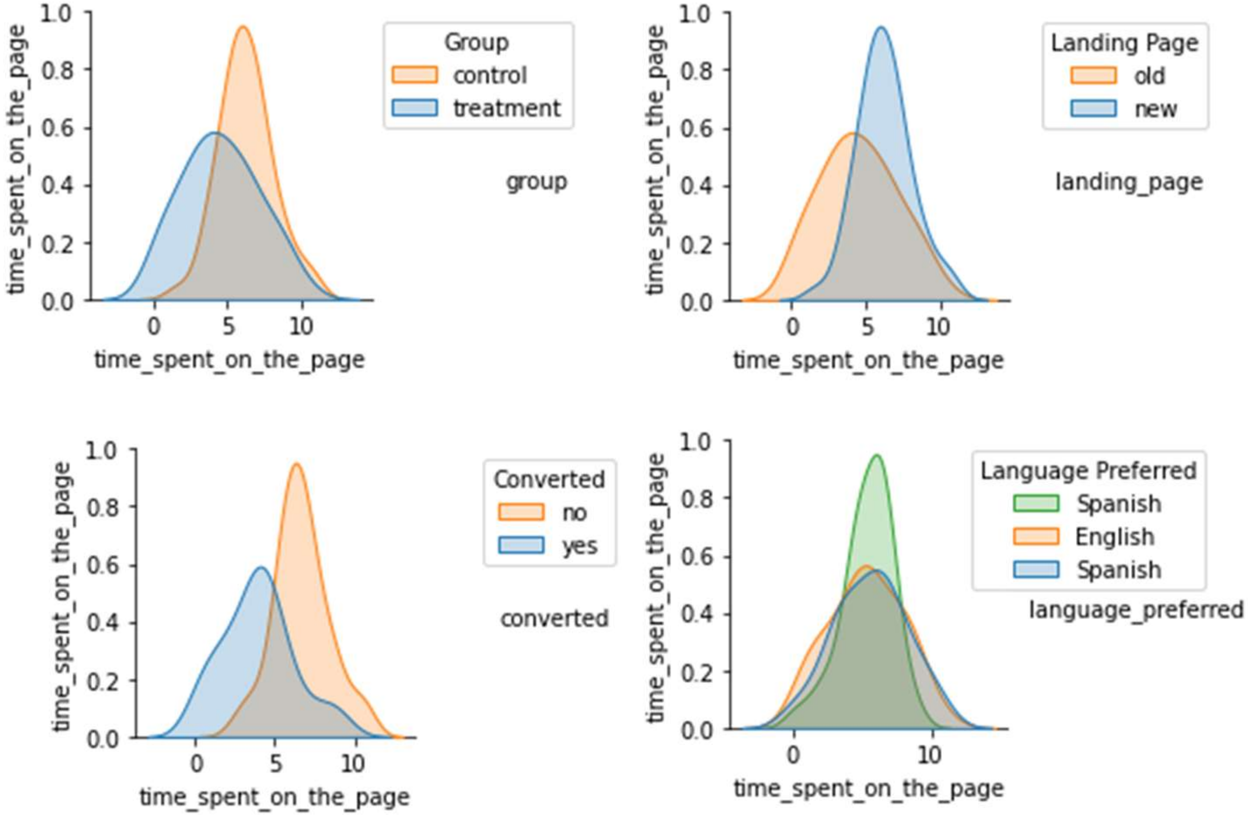
- Here we can prove the sample data was split 50/50 with 50 participants viewing the old landing page and 50 viewing the new landing page.
- As you will see throughout this analysis, since the groups were split between control and treatment and the landing_page was split into old and new, with old being the control and new being the treatment the outcomes of each observation is the same.



Time Spent on the Page by Categorical Variables

Observations

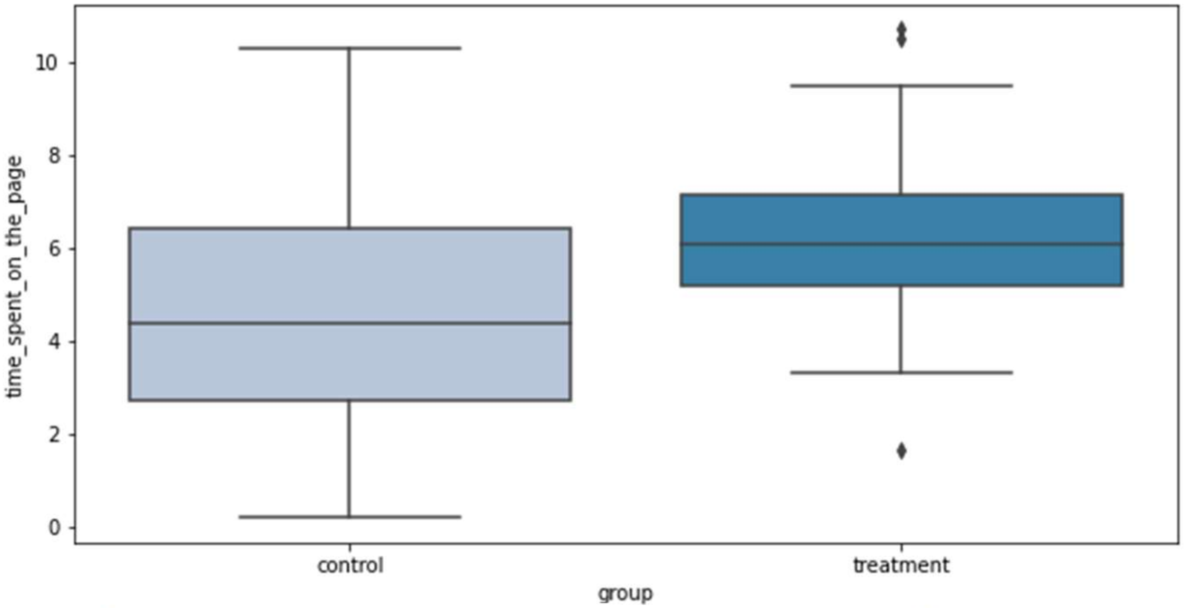
- Here we can see there is some variation between the variables versus the time spent on the landing page. We need to further investigate this.

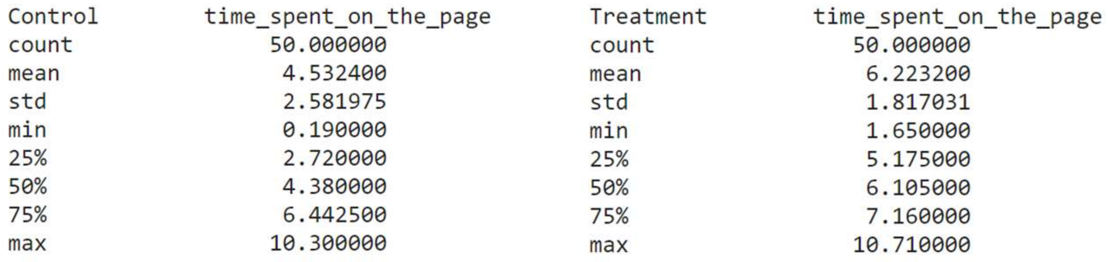


Group vs Time Spent on the Page

Observations

- The control group has a larger variance of 3.72 minutes (Q3-Q1) than the treatment group did with a variance of 1.99 minutes.
- The treatment group had a few outliers but 75% of the people spent at least 7.16 minutes on the new page.



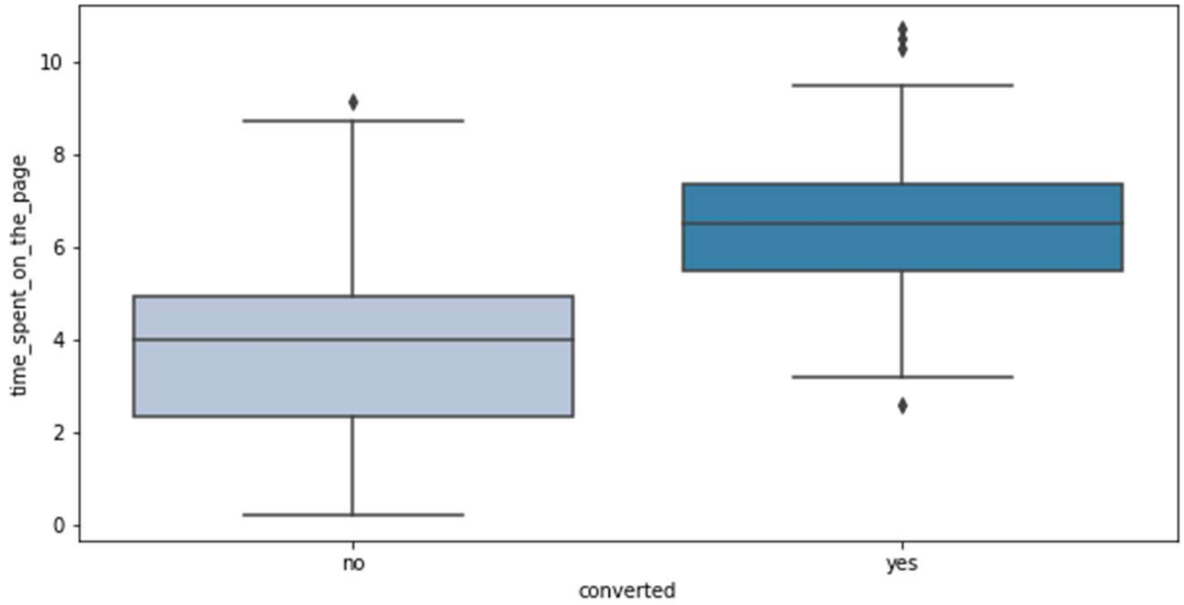


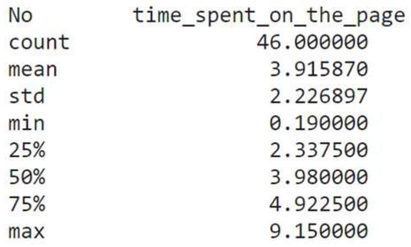
```
Control        time_spent_on_the_page      Treatment       time_spent_on_the_page
count                       50.000000      count                        50.000000
mean                         4.532400      mean                          6.223200
std                          2.581975      std                           1.817031
min                          0.190000      min                           1.650000
25%                          2.720000      25%                           5.175000
50%                          4.380000      50%                           6.105000
75%                          6.442500      75%                           7.160000
max                         10.300000      max                          10.710000
```

Converted vs Time Spent on the Page

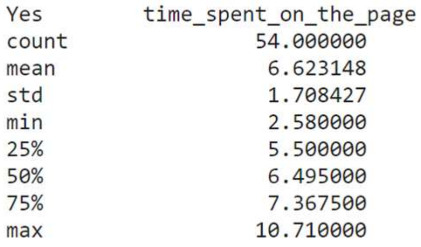
Observations

- This shows us that people who did not convert spent less time on the page than people who did convert.
- 75% of the people who did not convert only spent 4.92 minutes on the page.
- 75% of the people that did convert 7.37 minutes on the page.





```
No       time_spent_on_the_page
count                 46.000000
mean                   3.915870
std                    2.226897
min                    0.190000
25%                    2.337500
50%                    3.980000
75%                    4.922500
max                    9.150000
```

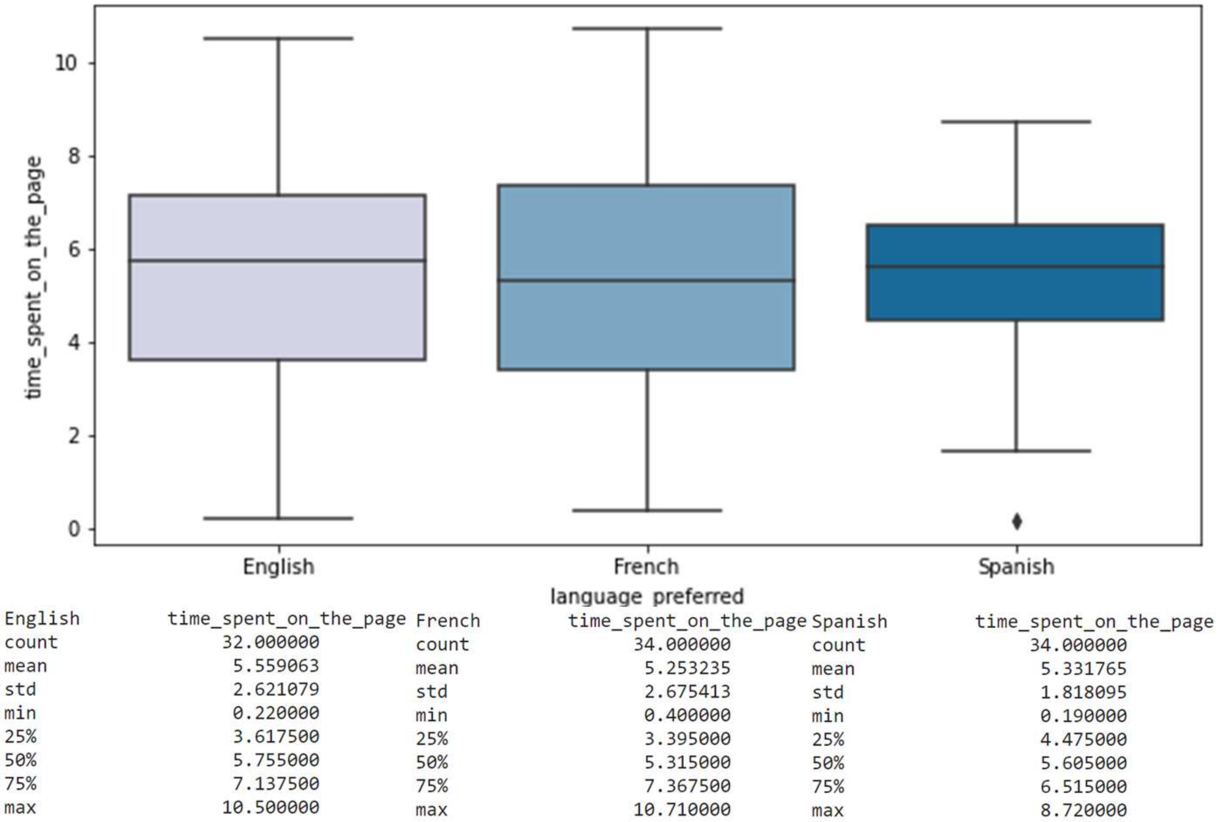


```
Yes      time_spent_on_the_page
count                 54.000000
mean                   6.623148
std                    1.708427
min                    2.580000
25%                    5.500000
50%                    6.495000
75%                    7.367500
max                   10.710000
```

Language Preferred vs Time Spent on the Page

Observations

- The three languages all had very similar medians at approximately 5 minutes.
- Spanish had a much smaller variance though.

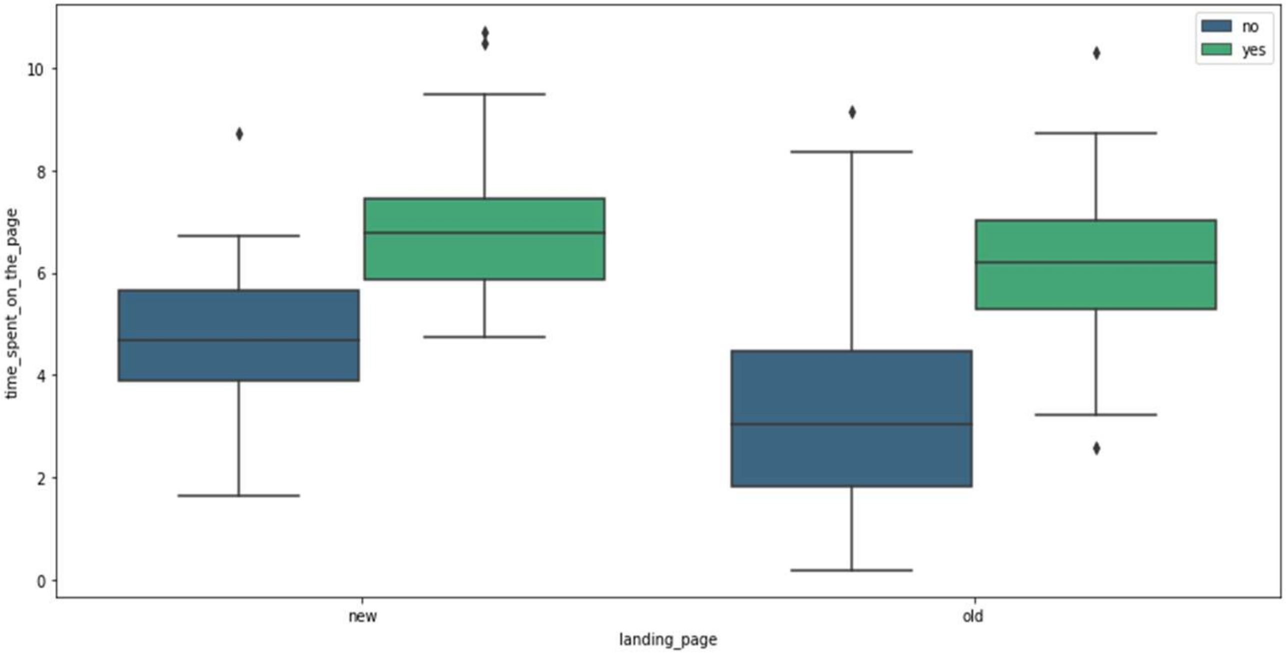


```
English        time_spent_on_the_page French        time_spent_on_the_page Spanish        time_spent_on_the_page
count                      32.000000  count                     34.000000  count                     34.000000
mean                        5.559063  mean                       5.253235  mean                       5.331765
std                         2.621079  std                        2.675413  std                        1.818095
min                         0.220000  min                        0.400000  min                        0.190000
25%                         3.617500  25%                        3.395000  25%                        4.475000
50%                         5.755000  50%                        5.315000  50%                        5.605000
75%                         7.137500  75%                        7.367500  75%                        6.515000
max                        10.500000  max                       10.710000  max                        8.720000
```

Landing Page vs Time Spent on the Page by Converted

Observations

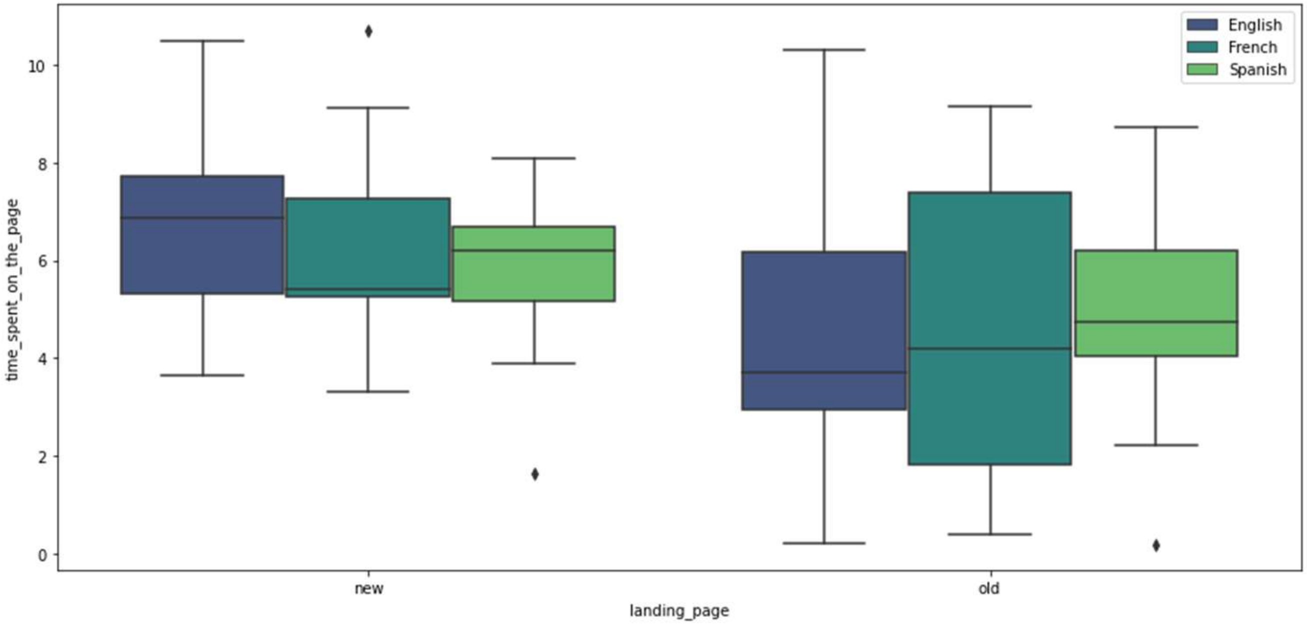
- Here we can see that those spent the most time on either page converted more.
- The old page had a bigger variance, showing that people spent less time on the old page than on the new page.



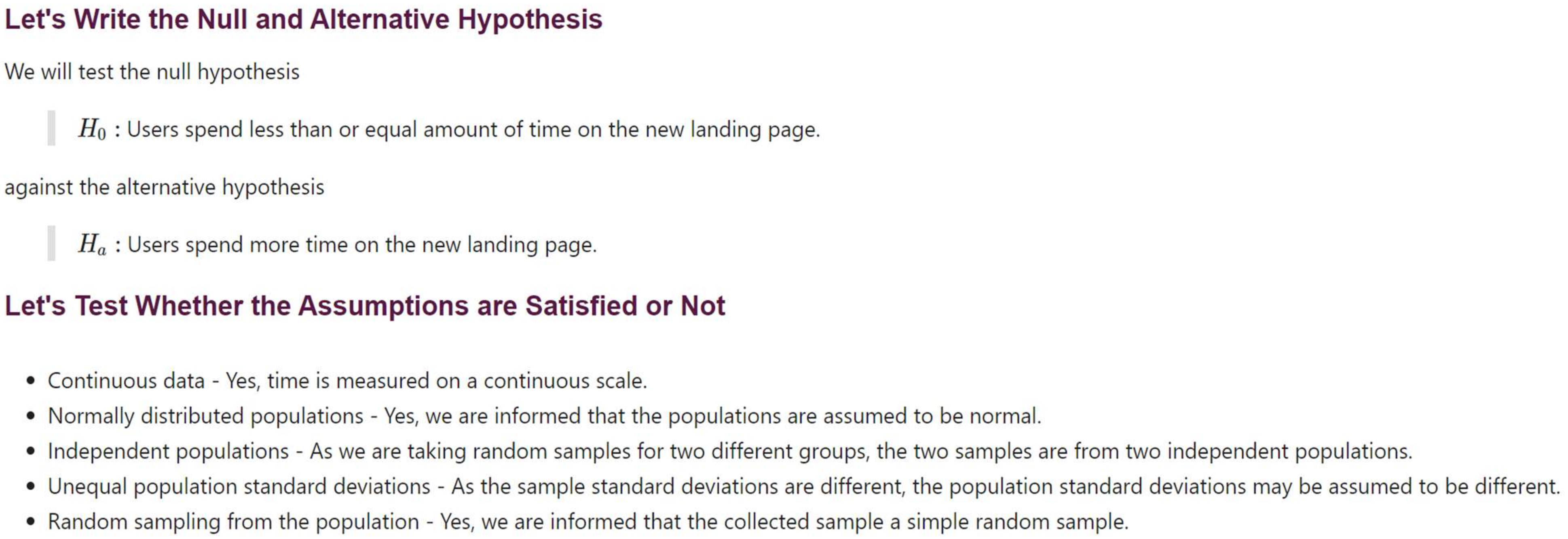
Landing Page vs Time Spent on the Page by Preferred Language

Observations

- Here we can see that there is a degree of variance with each of the page's vs the preferred language.
- The old page still had a bigger variance than the new page.
- One interesting note is how skewed the French is on the new page.



Part Two: Do the users spend more time on the new landing page than the old landing page?



Let's Write the Null and Alternative Hypothesis

We will test the null hypothesis

> H_0 : Users spend less than or equal amount of time on the new landing page.

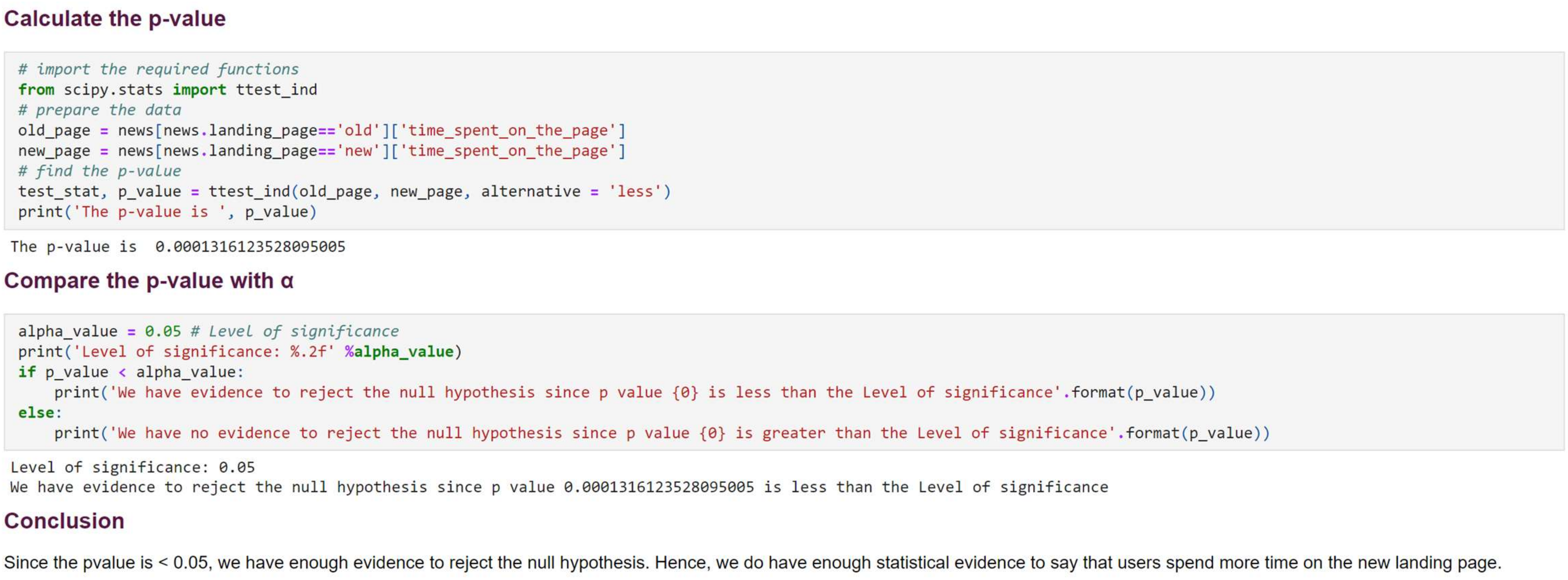
against the alternative hypothesis

> H_a : Users spend more time on the new landing page.

Let's Test Whether the Assumptions are Satisfied or Not

- Continuous data - Yes, time is measured on a continuous scale.
- Normally distributed populations - Yes, we are informed that the populations are assumed to be normal.
- Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.
- Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.
- Random sampling from the population - Yes, we are informed that the collected sample a simple random sample.

Finding the P-Value with the Two Independent Sample T-test



Calculate the p-value

```
# import the required functions
from scipy.stats import ttest_ind
# prepare the data
old_page = news[news.landing_page=='old']['time_spent_on_the_page']
new_page = news[news.landing_page=='new']['time_spent_on_the_page']
# find the p-value
test_stat, p_value = ttest_ind(old_page, new_page, alternative = 'less')
print('The p-value is ', p_value)
```

```
The p-value is  0.0001316123528095005
```

Compare the p-value with α

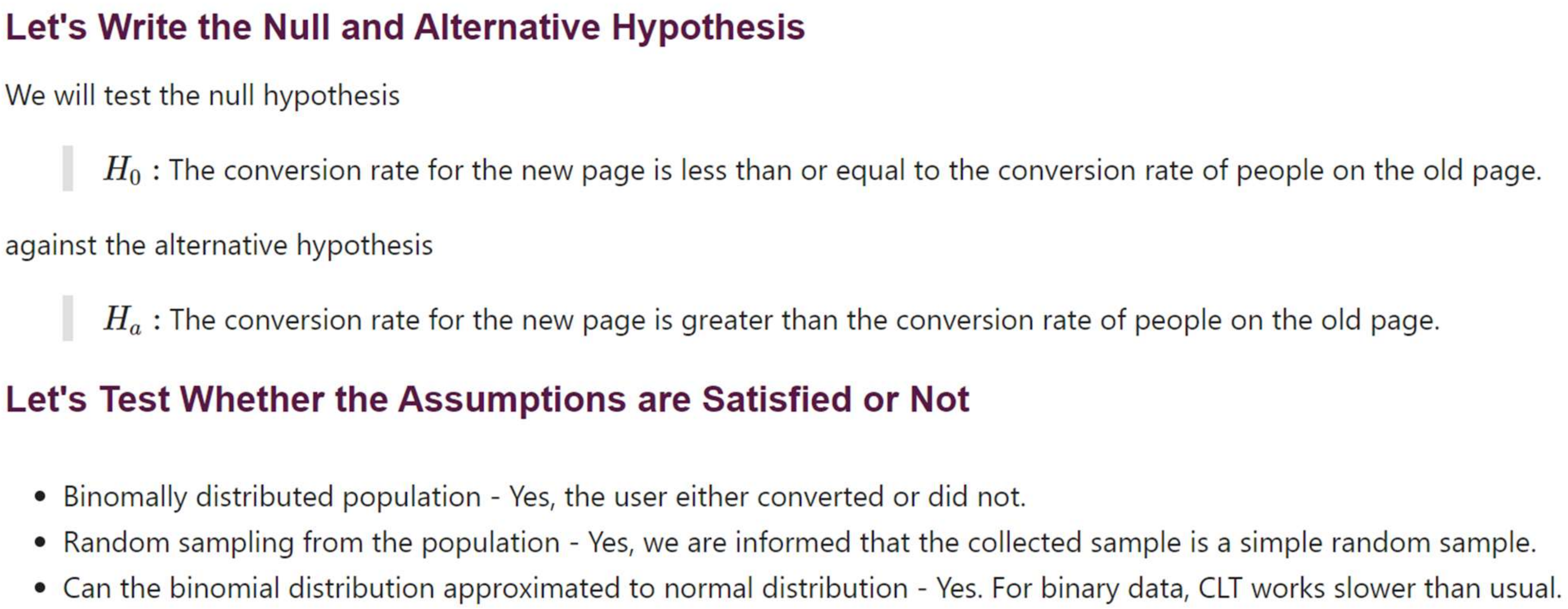
```
alpha_value = 0.05 # Level of significance
print('Level of significance: %.2f' %alpha_value)
if p_value < alpha_value:
    print('We have evidence to reject the null hypothesis since p value {0} is less than the Level of significance'.format(p_value))
else:
    print('We have no evidence to reject the null hypothesis since p value {0} is greater than the Level of significance'.format(p_value))
```

```
Level of significance: 0.05
We have evidence to reject the null hypothesis since p value 0.0001316123528095005 is less than the Level of significance
```

Conclusion

Since the pvalue is < 0.05, we have enough evidence to reject the null hypothesis. Hence, we do have enough statistical evidence to say that users spend more time on the new landing page.

Part Three: 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's Write the Null and Alternative Hypothesis

We will test the null hypothesis

> H_0 : The conversion rate for the new page is less than or equal to the conversion rate of people on the old page.

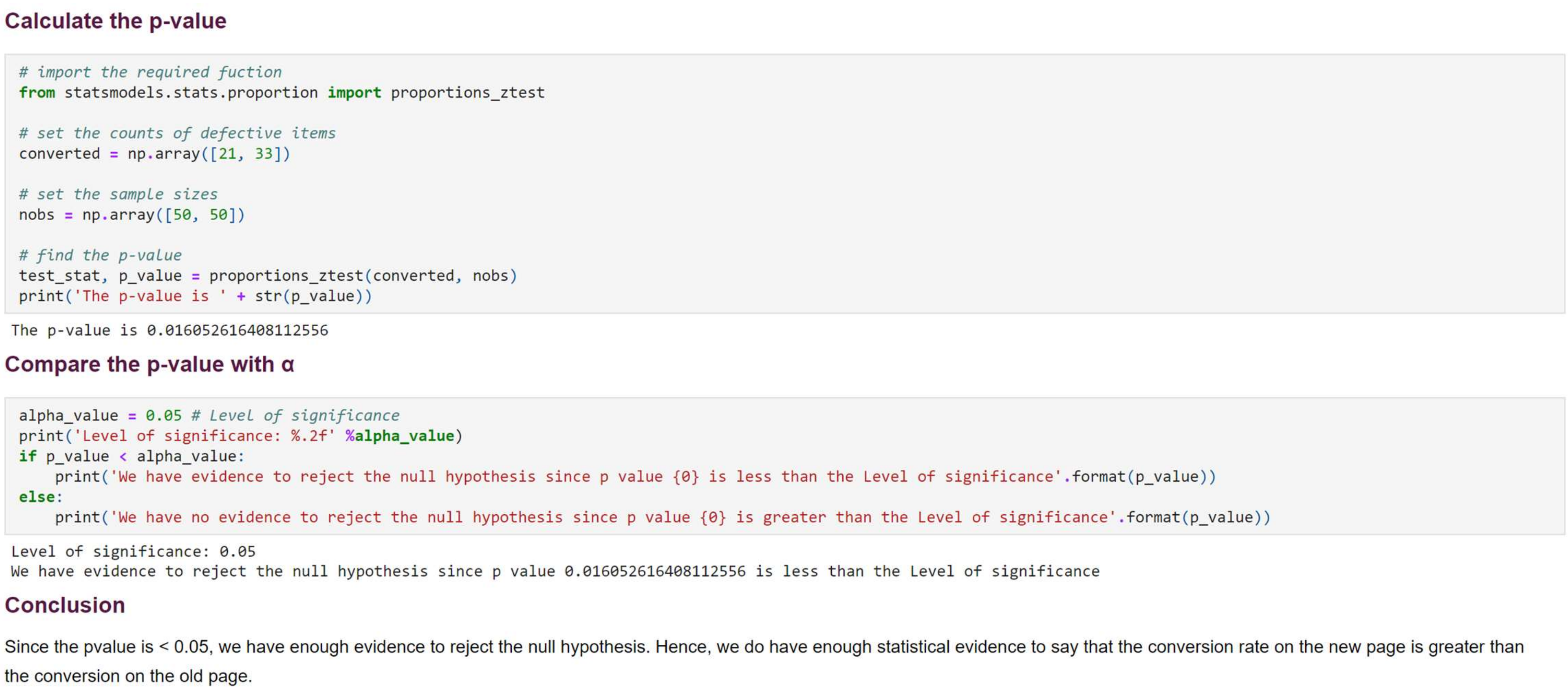
against the alternative hypothesis

> H_a : The conversion rate for the new page is greater than the conversion rate of people on the old page.

Let's Test Whether the Assumptions are Satisfied or Not

- Binomally distributed population - Yes, the user either converted or did not.
- Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual.

Finding the P-Value with the Two Proportion Z-Test



Calculate the p-value

```
# import the required fuction
from statsmodels.stats.proportion import proportions_ztest

# set the counts of defective items
converted = np.array([21, 33])

# set the sample sizes
nobs = np.array([50, 50])

# find the p-value
test_stat, p_value = proportions_ztest(converted, nobs)
print('The p-value is ' + str(p_value))
```

```
The p-value is 0.016052616408112556
```

Compare the p-value with α

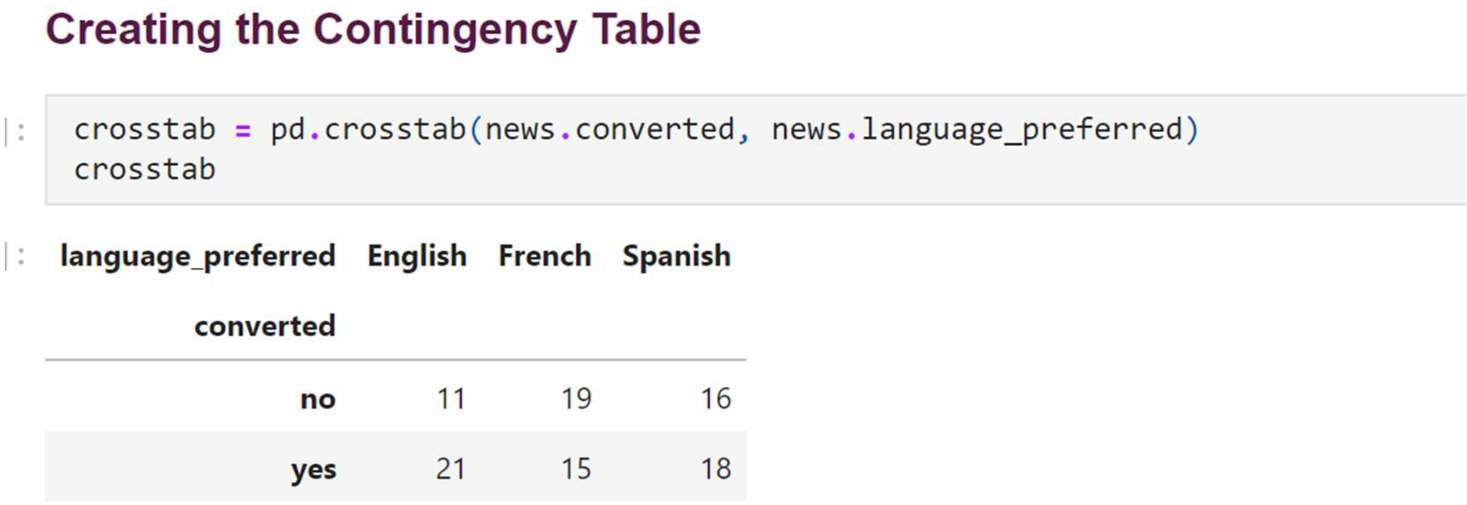
```
alpha_value = 0.05 # Level of significance
print('Level of significance: %.2f' %alpha_value)
if p_value < alpha_value:
    print('We have evidence to reject the null hypothesis since p value {0} is less than the Level of significance'.format(p_value))
else:
    print('We have no evidence to reject the null hypothesis since p value {0} is greater than the Level of significance'.format(p_value))
```

```
Level of significance: 0.05
We have evidence to reject the null hypothesis since p value 0.016052616408112556 is less than the Level of significance
```

Conclusion

Since the pvalue is < 0.05, we have enough evidence to reject the null hypothesis. Hence, we do have enough statistical evidence to say that the conversion rate on the new page is greater than the conversion on the old page.

Part Four: Create a contingency table using the pandas.crosstab() function]

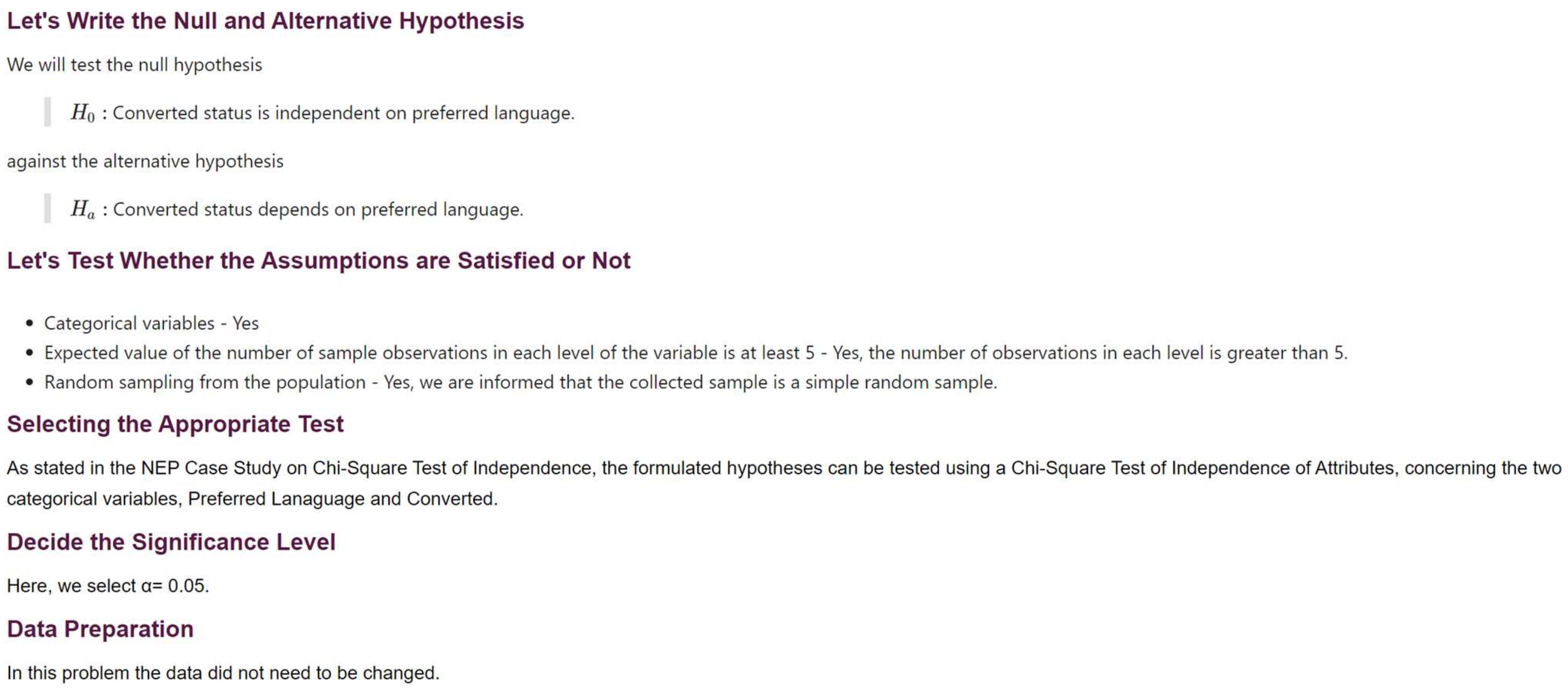


Creating the Contingency Table

```
crosstab = pd.crosstab(news.converted, news.language_preferred)
crosstab
```

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

Does the converted status depend on the preferred language?



Let's Write the Null and Alternative Hypothesis

We will test the null hypothesis

> H_0 : Converted status is independent on preferred language.

against the alternative hypothesis

> H_a : Converted status depends on preferred language.

Let's Test Whether the Assumptions are Satisfied or Not

- Categorical variables - Yes
- Expected value of the number of sample observations in each level of the variable is at least 5 - Yes, the number of observations in each level is greater than 5.
- Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.

Selecting the Appropriate Test

As stated in the NEP Case Study on Chi-Square Test of Independence, the formulated hypotheses can be tested using a Chi-Square Test of Independence of Attributes, concerning the two categorical variables, Preferred Lanaguage and Converted.

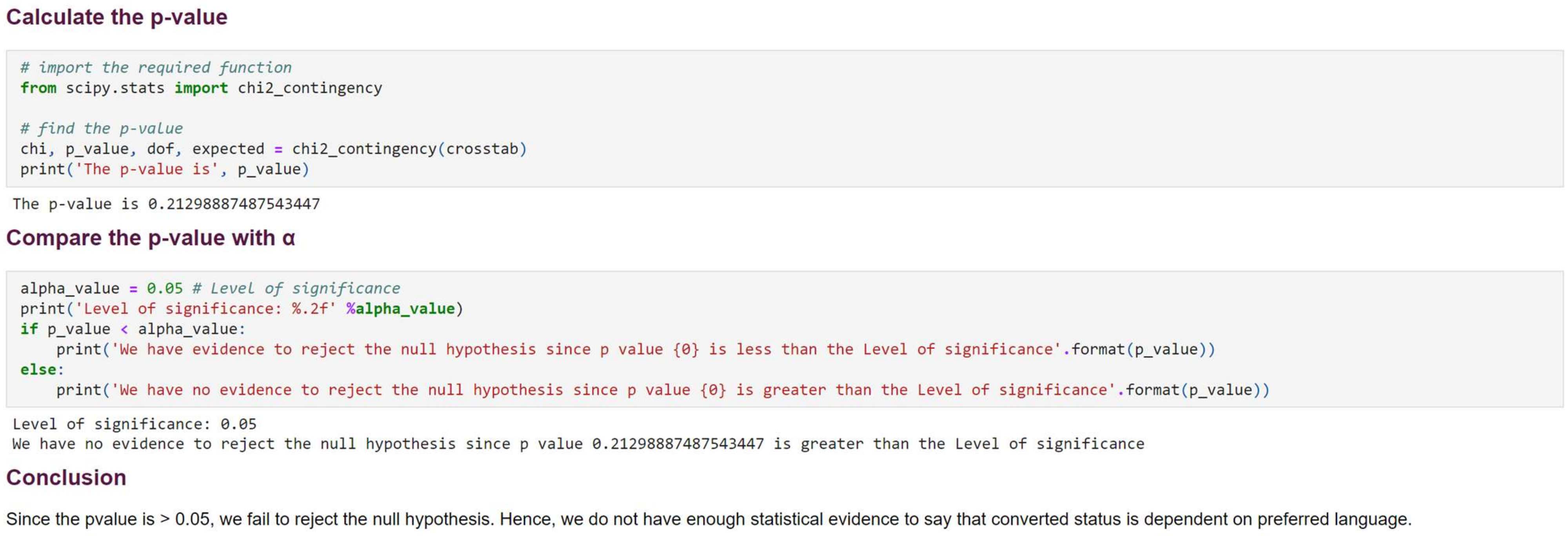
Decide the Significance Level

Here, we select α= 0.05.

Data Preparation

In this problem the data did not need to be changed.

Calculate the P-Value with the Chi-Square Test of Independence



Calculate the p-value

```
# import the required function
from scipy.stats import chi2_contingency

# find the p-value
chi, p_value, dof, expected = chi2_contingency(crosstab)
print('The p-value is', p_value)
```

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

Compare the p-value with α

```
alpha_value = 0.05 # Level of significance
print('Level of significance: %.2f' %alpha_value)
if p_value < alpha_value:
    print('We have evidence to reject the null hypothesis since p value {0} is less than the Level of significance'.format(p_value))
else:
    print('We have no evidence to reject the null hypothesis since p value {0} is greater than the Level of significance'.format(p_value))
```

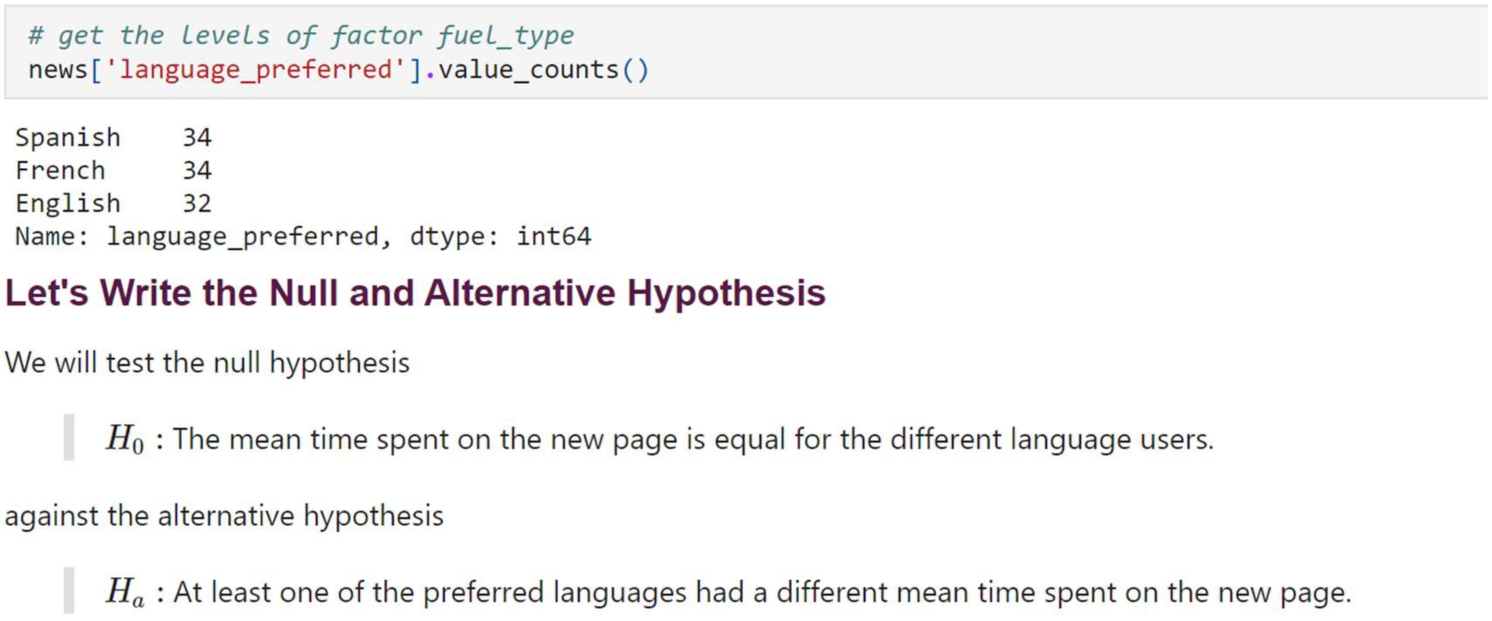
```
Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value 0.21298887487543447 is greater than the Level of significance
```

Conclusion

Since the pvalue is > 0.05, we fail to reject the null hypothesis. Hence, we do not have enough statistical evidence to say that converted status is dependent on preferred language.

Part Five: Is the mean time spent on the new page same for the different language users?

Here, time_spent_on_the_page is the response and language_preferred is the factor.



```
# get the levels of factor fuel_type
news['language_preferred'].value_counts()
```

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

Let's Write the Null and Alternative Hypothesis

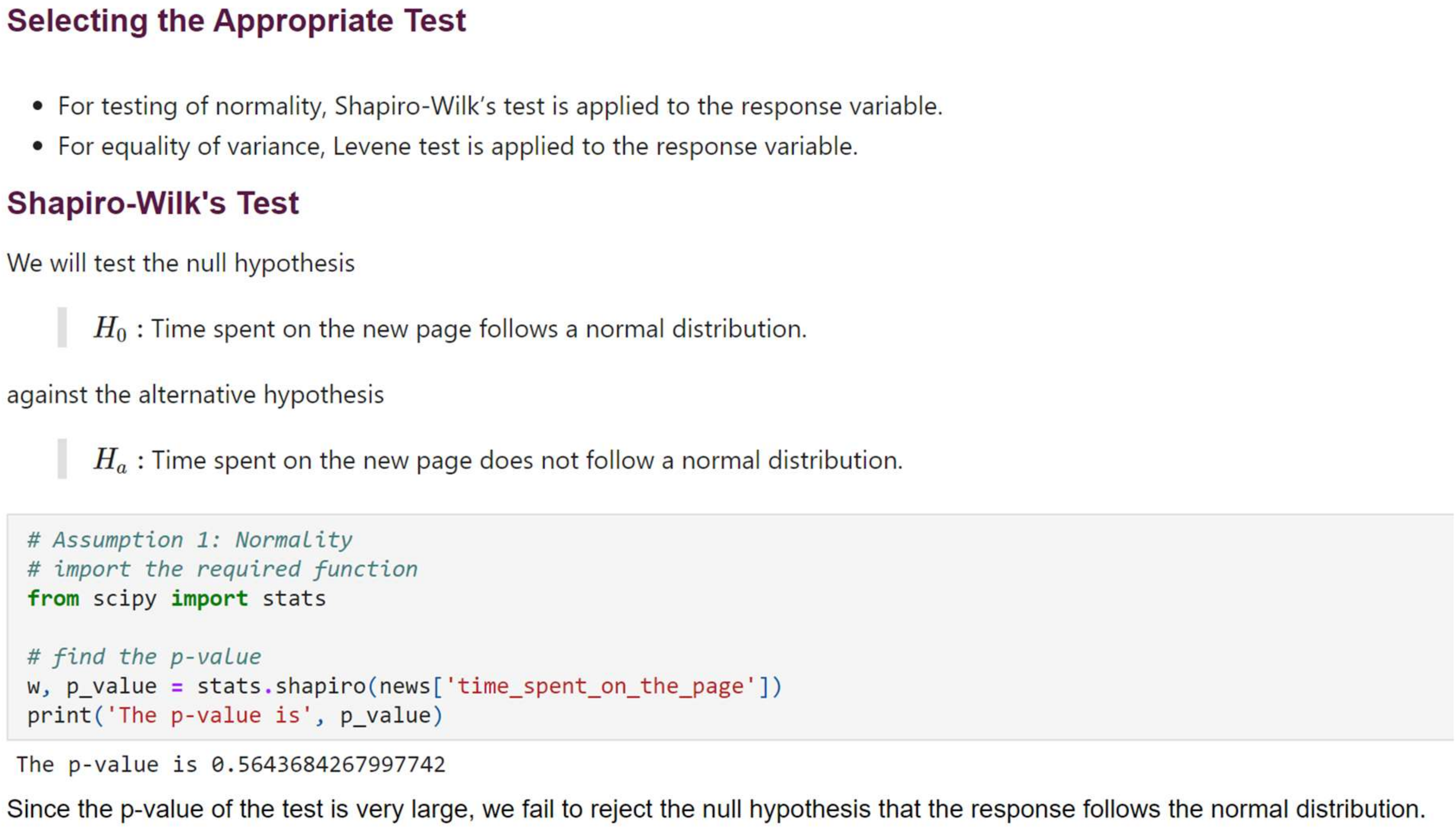
We will test the null hypothesis

> H_0 : The mean time spent on the new page is equal for the different language users.

against the alternative hypothesis

> H_a : At least one of the preferred languages had a different mean time spent on the new page.

Shapiro-Wilk’s Test



Selecting the Appropriate Test

- 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

> H_0 : Time spent on the new page follows a normal distribution.

against the alternative hypothesis

> H_a : Time spent on the new page does not follow a normal distribution.

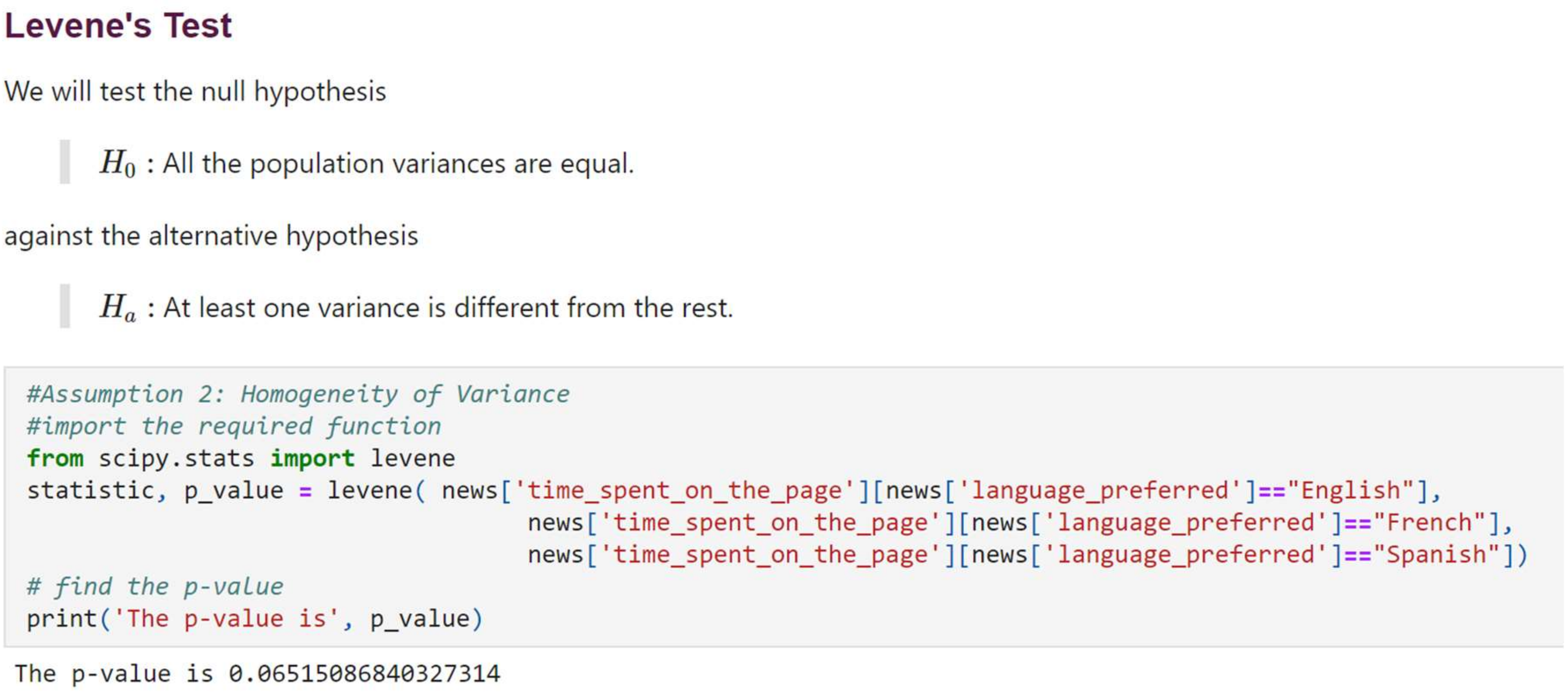
```python
# Assumption 1: Normality
# import the required function
from scipy import stats

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

```
The p-value is 0.5643684267997742
```

Since the p-value of the test is very large, we fail to reject the null hypothesis that the response follows the normal distribution.

Levene’s Test



Levene's Test

We will test the null hypothesis

> H_0 : All the population variances are equal.

against the alternative hypothesis

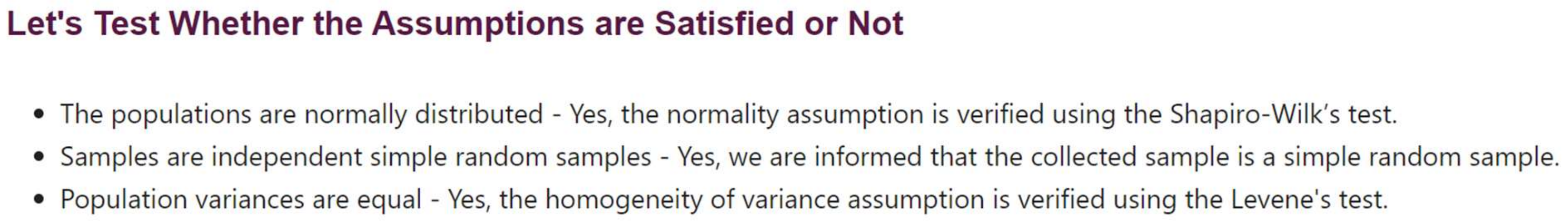
> H_a : At least one variance is different from the rest.

```
#Assumption 2: Homogeneity of Variance
#import the required function
from scipy.stats import levene
statistic, p_value = levene( news['time_spent_on_the_page'][news['language_preferred']=="English"],
                                   news['time_spent_on_the_page'][news['language_preferred']=="French"],
                                   news['time_spent_on_the_page'][news['language_preferred']=="Spanish"])
# find the p-value
print('The p-value is', p_value)
```

```
The p-value is 0.06515086840327314
```

Since the p-value is bigger than 0.05, we fail to reject the null hypothesis that the response follows the normal distribution.

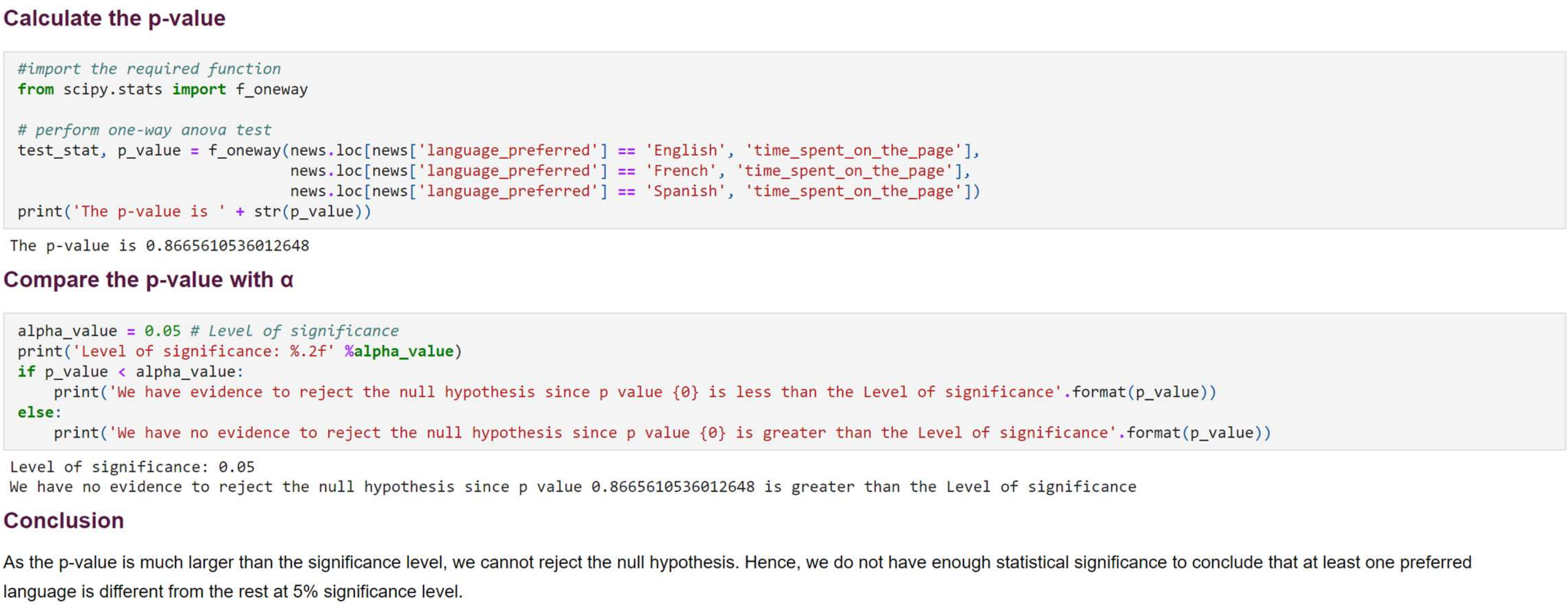
Continued



Let's Test Whether the Assumptions are Satisfied or Not

- The populations are normally distributed - Yes, the normality assumption is verified using the Shapiro-Wilk’s test.
- Samples are independent simple random samples - Yes, we are informed that the collected sample is a simple random sample.
- Population variances are equal - Yes, the homogeneity of variance assumption is verified using the Levene's test.

Calculate the P-Value with the One-Way Anova



Calculate the p-value

```
#import the required function
from scipy.stats import f_oneway

# perform one-way anova test
test_stat, p_value = f_oneway(news.loc[news['language_preferred'] == 'English', 'time_spent_on_the_page'],
                              news.loc[news['language_preferred'] == 'French', 'time_spent_on_the_page'],
                              news.loc[news['language_preferred'] == 'Spanish', 'time_spent_on_the_page'])
print('The p-value is ' + str(p_value))
```

```
The p-value is 0.8665610536012648
```

Compare the p-value with α

```
alpha_value = 0.05 # Level of significance
print('Level of significance: %.2f' %alpha_value)
if p_value < alpha_value:
    print('We have evidence to reject the null hypothesis since p value {0} is less than the Level of significance'.format(p_value))
else:
    print('We have no evidence to reject the null hypothesis since p value {0} is greater than the Level of significance'.format(p_value))
```

```
Level of significance: 0.05
We have no evidence to reject the null hypothesis since p value 0.8665610536012648 is greater than the Level of significance
```

Conclusion

As the p-value is much larger than the significance level, we cannot reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that at least one preferred language is different from the rest at 5% significance level.

greatlearning
Power Ahead

Happy Learning !