# E-news Express Project

Post-Grad Program in Data Science and Business Analytics

Date: 8/19/2022

By: Manik Chhabra

# Contents / Agenda

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

# Executive Summary

- Conclusion
  - The users spend more time on the new landing page as compared to the old landing page based on the results from the statistical test – **Two-sample independent t test**
  - The conversion rate (0.66) for the new page is greater than the conversion rate (0.42) for the old page, established from the results in the visual analysis and confirmed by the statistical test – **Two-proportions z test**
  - The converted status does not depend on the preferred language, as supported by the **Chi-squared Test for Independence**
    - If more samples are collected and further statistical analysis is done, there may be a correlation between the converted status and the preferred language
  - Lastly, the time spent on the new page is the same for different language users, as supported by the **One way ANOVA test**
  - All the statistical analysis was performed at a significance level of 5%.

# Executive Summary

- Recommendations
  - Since the users spent more time and were more likely to become subscribers on the new landing page as compared to the old landing page, more digital advertising (increase spend on marketing) needs to be done to attract prospective customers to the newly designed website.
    - Ways to advertise the new landing page are launching email newsletters, targeting audiences on premium publishers sites, utilizing search advertising, and creating compelling referral content
  - The company (including web designers and content creators) needs to maintain the website daily to include the latest and intriguing news content to increase engagement and conversion across the platform
    - The current subscribers should be maintained by giving them easier access to the site by creating an app for the mobile device, providing perks for being a long-term customer, and offering discounts for annual membership of the news website
  - The company should collect more data and perform statistical analysis to discover the relationship between the conversion rate and languages preferred
  - E-News Express should analyze user data in terms of their demographics, including their age, gender, occupation, education, and marriage status to generate cutting edge insights that will continue driving more customers (of a specific type) to the website
    - For example, the business can compare the demographics of non-subscribers vs. subscribers on the new landing page

# Business Problem Overview and Solution Approach

- Business problem
  - To determine the effectiveness of the new landing page in gathering new subscribers (by analyzing the user's behavior) as compared to the effectiveness of the old landing page
- Please mention the solution approach / methodology
  - The solution approach is to explore the data and perform a statistical analysis (at a significance level of 5%) to determine the effectiveness of the new landing page in converting new subscribers by answering the following questions:

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

2. Is the conversion rate (the proportion of users who visit the landing page and become subscribers) for the new page greater than the conversion rate for the old page?

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

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

# EDA Results

- Univariate Analysis
  - Variable - Time spent on the page
    - **Figure 1** and **2** show that the data is evenly distributed. Mean time spent on the page = 5.37 min, max = 10.7 min , and min = 0.19 min.
  - Variable - Converted
    - **Figure 3** shows that there were more subscribers than non-subscribers in the dataset
      - 54 users – subscribers
      - 46 users – non-subscribers

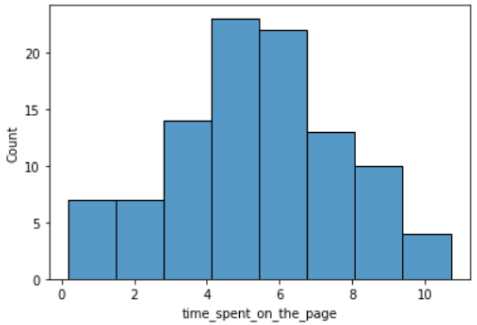


**Figure 1**

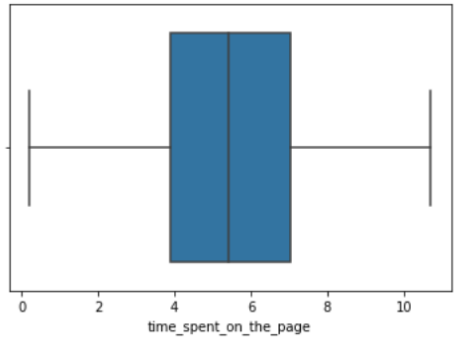


**Figure 2**

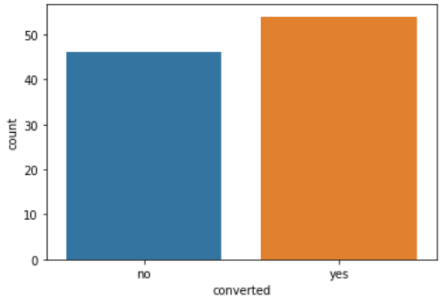


**Figure 3**

*Link to Appendix slide on data background check*

# EDA Results

- Univariate Analysis
  - Variable - Group
    - According to **Figure 4**, 50 users are in the control group and 50 users are in the treatment group
  - Variable – Landing Page
    - According to **Figure 5**, 50 of users land on old page whereas 50 of users land on new page.
  - Variable – Language Preferred
    - **Figure 6** displays most users prefer both Spanish and French as their language
    - 34% of the users prefer Spanish, 34% of the users prefer French and 32% of the users prefer English.

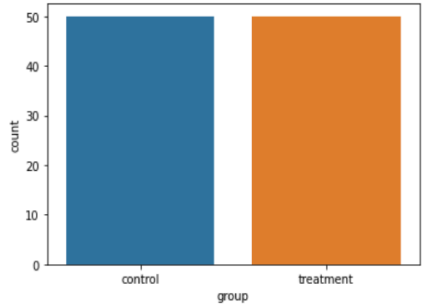


**Figure 4**

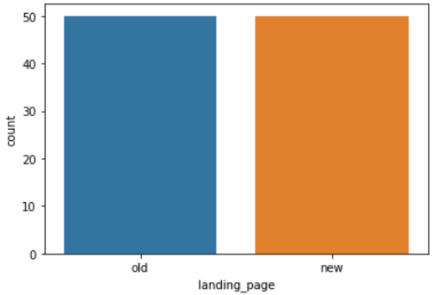


**Figure 5**

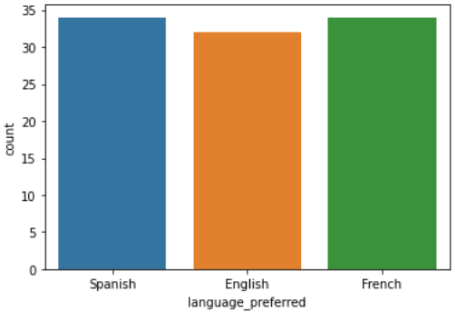


**Figure 6**

*Link to Appendix slide on data background check*

# EDA Results

- Bivariate Analysis
  - Variables – Time spent on the page and Landing Page
    - **Figure 7** - The median time spent on the new landing page (6.2 min) was 1.7 minutes higher than the median time spent on the old landing page (4.5 min)
    - More outliers (in time spent on the page) are present in the new landing page
  - Variables – Time spent on the page and Converted
    - **Figure 8** - Time spent on page was higher (median = 6.5 min) for those who converted; whereas, time spent on page was lower (median = 4.0 min) for those who did not convert.
    - More outliers (in time spent on the page) are present in the converted users vs. the non-converted users

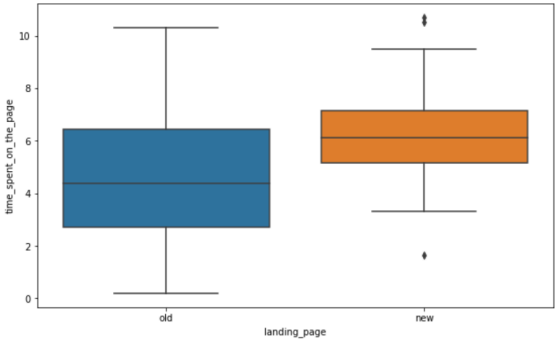


**Figure 7**

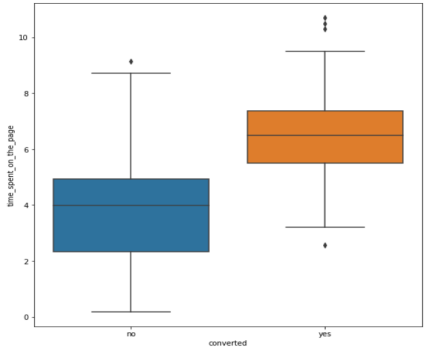


**Figure 8**

*Link to Appendix slide on data background check*

# EDA Results

- Bivariate Analysis
  - Variables – Time spent on the page and Languages Preferred
    - **Figure 9 –** For users who prefer the English language, their median time spent on the page is the highest (5.75 min)
    - Moreover, users who are French spend the least median amount of time on the page (5.32 min); however, 75 % of the French users spend at least 7 min on the page (the highest among all at that data point)
    - Spanish users have a median time spent on the page of (5.61 min)

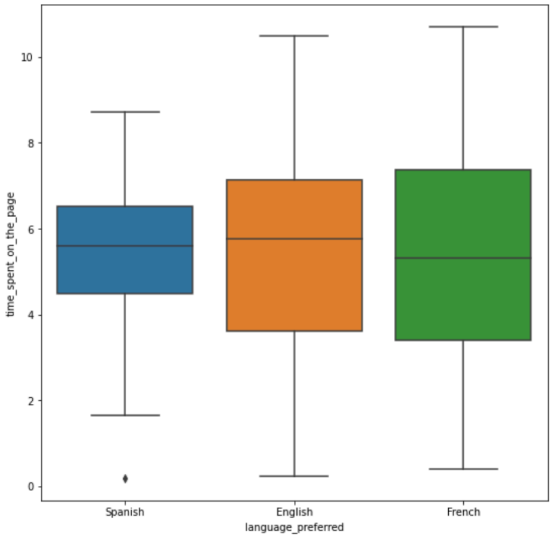


**Figure 9**

*Link to Appendix slide on data background check*

# Hypotheses Tested and Results – Question 1

- Visual analysis of the time spent on the new page as compared to the time spent on the old page

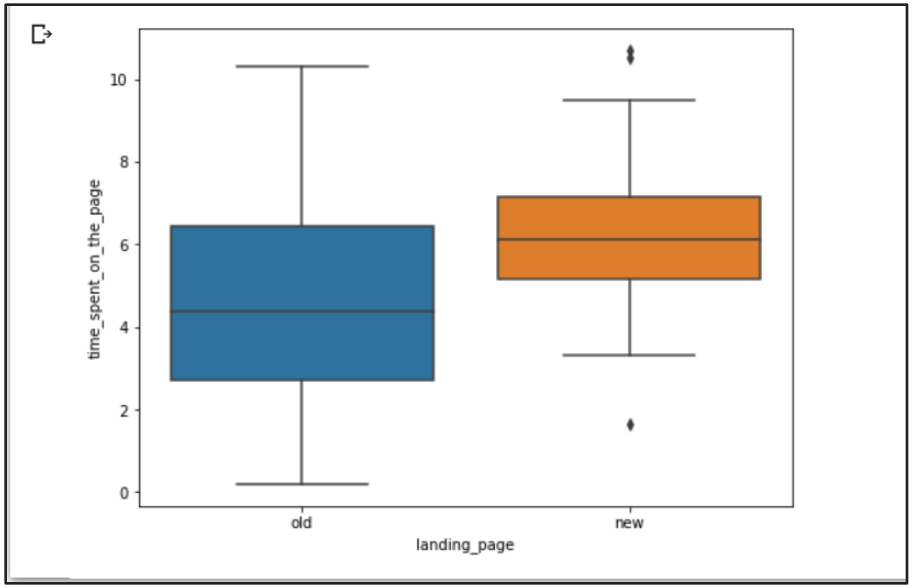


**Figure 10**

- According to the findings from the Boxplot **(Figure 10)**, the median time spent on the "new" page (6.2 min) is 1.7 min greater than the median time spent on the "old" page (4.5 min)

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 1

- Hypothesis tested
  - The mean time spent on the new landing page is greater than the mean time spent on the old landing page (Alternative hypothesis)
- Test result and inference
  - The p-value (~0.000131) is less than .05 significance, thus rejecting the null hypothesis. Since the null hypothesis was rejected, the mean time spent on the new landing page is greater than the mean time spent on the old landing page.

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 2

- Visual analysis performed to compare the conversion rate for the new page and the conversion rate for the old page

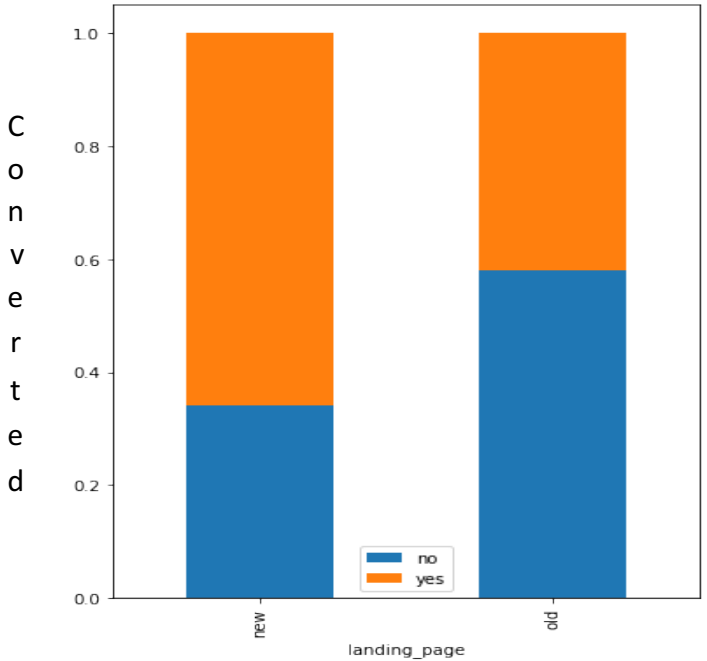


**Figure 11**

- Based on the results from the Stacked Bar chart **(Figure 11),** the conversion rate on the new page was at least 20% greater than the conversion rate on the old page

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 2

- Hypothesis tested
  - The conversion rate for the new page is greater than conversion rate for the old page (Alternative hypothesis)
- Test result and inference
  - Since the p-value (~0.00803) is less than the level of significance (.05), we reject the null hypothesis. Therefore, the conversion rate for the new page (0.66) is greater than the conversion rate for the old page (0.42).

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 3

- Visual analysis performed to view the dependency between conversion status and preferred language

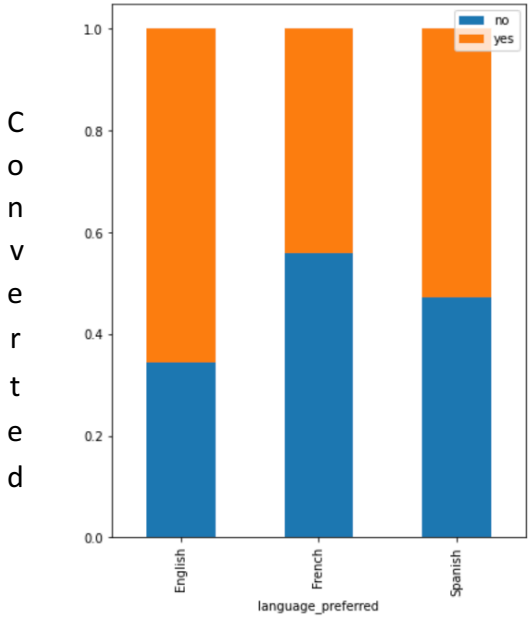


**Figure 12**

- Based on the results from the Stacked Bar chart **(Figure 12),** the conversion rate is the highest for users who prefer English language, while the conversion rate is the lowest for users who prefer French language.

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 3

- Hypothesis tested
  - The conversion rate is dependent on the preferred language (Alternative hypothesis)
- Test result and inference
  - As the p-value (~0.213) is greater than the level of significance (.05), we fail to reject the null hypothesis. Therefore, the conversion status is independent of the preferred language.
  - Even though the visual analysis shows that the conversion rate is the highest for users who prefer the English language, there is not enough statistical evidence (based on the p-value) to support this claim

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 4

- Visual analysis performed to compare the time spent on the new page for different language users

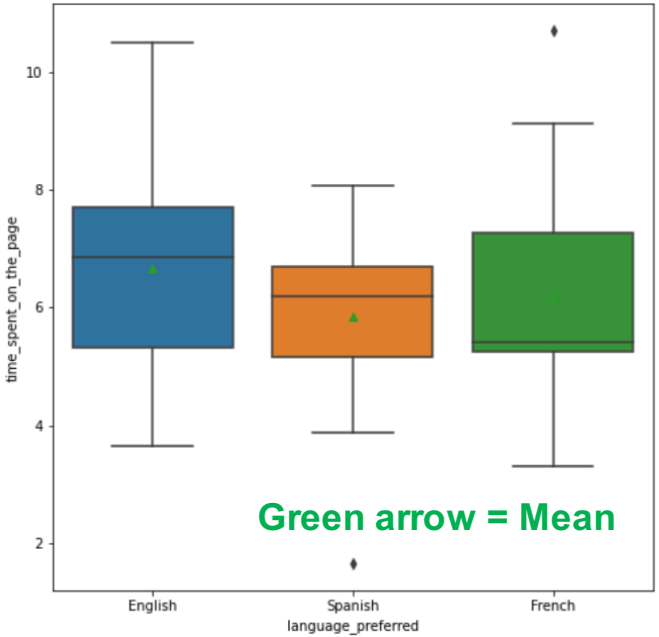


**Figure 13**

- According to the diagram **(Figure 13),** the mean time spent on the new page is the highest (6.66 min) for users who prefer the English language
- Mean time spent on the new page is the lowest (5.83 min) for Spanish users.
- Mean time spent on the new page is in the middle (6.19 min) for French users.

*Link to Appendix slide on details of the test performed*

# Hypotheses Tested and Results – Question 4

- Hypothesis tested
  - The mean time spent on the new page for at least one language user is different from the other two (Alternative hypothesis)
- Test result and inference
  - As the p-value (~0.432) is greater than the level of significance, we fail to reject the null hypothesis. Therefore, the mean time spent on the page for each different language user is the same.
  - Even though the visual analysis shows that the mean time spent on the new page is the highest (6.66 min) for the English language user, there is not enough statistical evidence (based on the p-value) to support this claim

*Link to Appendix slide on details of the test performed*

# APPENDIX

# Data Background and Contents

- The data set consists of information about the interaction of users in both (treatment and control group) groups with two landing pages (old and new)
  - There are 100 unique users (rows) and 6 columns
    - The columns consist of *user_id*
      - *user_id*
      - *group*
      - *landing page*
      - *time_spent_on_the_page*
      - *converted*
      - *language_preferred*
  - No null or duplicate values exist in the data set

**Displays first 5 rows of the data set**

| | 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 |

# Data Background and Contents

- In regards to the types of variables in the data set, there is 1 integer type (unique User ID), 1 float type (time spent on the page), and 4 string types
- In terms of the statistical summary of the numerical variables, the mean *time_spent_on_the_page* is 5.37 minutes, minimum time spent is 0.19 minutes, and max time spent is 10.71 minutes.
  - The statistical summary of the user IDs is not applicable for analysis, since the IDs are unique
- Contents of each categorical variable
  - Group
    - 50 users are in the control group while 50 users are in the treatment group
  - Converted
    - 54 users are in the converted group (subscribers) while there are 46 users in the non-converted group (non-subscribers)
  - Landing Page
    - 50 users spend time on the old landing page while 50 users spend time on the new landing page
  - Language Preferred
    - There are 34 Spanish speakers, 34 French speakers, and 32 English speakers

# Hypothesis Testing Details – Question 1

- Null and alternative hypotheses
  - Null Hypothesis – $H_0$: u1 = u2, The mean time spent on the new landing page is equal to the mean time spent on the old landing page.
  - Alternative Hypothesis – $Ha$: u1 > u2, The mean time spent on the new landing page is greater than the mean time spent on the old landing page.
    - (where u1 and u2 are the mean time spent on the new landing page and old landing page, respectively)
- Hypothesis Test selected
  - 2 sample Independent T-test
    - The population standard deviations can be assumed to be unequal since the two sample standard deviations are unequal.
    - Considered a one-tailed t test for comparing the means between two independent populations
- p-value obtained
  - As the p-value = ~( 0.00013) is less than the level of significance (.05), we reject the null hypothesis.
- Any other computational/mathematical details
  - Computational method of calculating the p-value, using 'ttest_ind' function

```
# 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')  #complete the code by filling appropriate parameters in the blanks

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

```
The p-value is 0.00013161235280950005
```

# Hypothesis Testing Details – Question 2

- Null and alternative hypotheses
  - Null Hypothesis – $H_0$: p1=p2 The conversion rate for the new page is equal to the conversion rate for the old page.
  - Alternative Hypothesis – $Ha$: p1>p2 The conversion rate for the new page is greater than the conversion rate for the old page.
    - p1 and p2 are the proportions of the users that get converted on the new and old landing page, respectively
- Hypothesis Test selected
  - Two-proportions z-test
- p-value obtained
  - As the p-value = (~ 0.00803) is less than the level of significance (.05), we reject the null hypothesis. Therefore, the conversion rate for the new page is greater than the conversion rate for the old page.
- Any other computational/mathematical details
  - Computational method of calculating the p-value, using the 'proportions_ztest' function

```
# 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')   #complete the code by filling appropriate parameters in the blanks


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

# Hypothesis Testing Details – Question 3

- Null and alternative hypotheses
  - Null Hypothesis – $H_0$: The conversion status is independent of the preferred language.
  - Alternative Hypothesis – $Ha$: The conversion status is dependent on the preferred language.
- Hypothesis Test selected
  - Chi-squared Test for Independence
- p-value obtained
  - As the p-value (~ 0.212) is greater than the level of significance (.05), we fail to reject the null hypothesis. Therefore, the conversion status is independent of the preferred language.
- Any other computational/mathematical details
  - Computational method of calculating the p-value, using 'chi2_contingency' function

```
# 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)   # #complete the code by filling appropriate parameters in the blanks

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

# Hypothesis Testing Details – Question 4

- Null and alternative hypotheses
  - Null Hypothesis – $H_0$: u1=u2=u3, The mean time spent on the new page for different language users is the same.
  - Alternative Hypothesis – $Ha$: The mean time spent on the new page for at least one language user is different from the other two.
- Hypothesis Test selected
  - One-way ANOVA test.
- p-value obtained
  - As the (p-value = ~0.432)is greater than the level of significance, we fail to reject the null hypothesis. Therefore, there is not enough evidence to conclude that the mean time spent on the new page for at least one language user is different from the other two users.
- Any other computational/mathematical details
  - Computational method of calculating the p-value, using the 'f_oneway' function

```
# 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)   #complete the code by filling appropriate parameters in the blanks

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

Great Learning

# Happy Learning !