# E-News Express Hypothesis Testing

Michail Mersinias

02/13/2023

# Contents / Agenda

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

# Executive Summary

- The executive summary of the E-News Express Hypothesis Testing presentation is as follows:
    - **New landing page results in higher user engagement time:** Users spend more time on the new landing page than the old landing page.
    - **New landing page results in higher user conversion rate:** The conversion rate (the proportion of users who visit the landing page and get converted) for the new page is greater than the conversion rate for the old page.
    - **Preferred language independent of user conversion rate, but affects user engagement time:** Preferred language correlates with an increase of the time users spend on the page. However, it is not a factor when it comes to the conversion of a user to a subscriber.
    - **Recommendation 1:** The new landing page should be chosen, as it provides higher engagement (more time spent) and higher conversion rate (users becoming subscribers).
    - **Recommendation 2:** While the preferred language does not affect the conversion rate overall, it affects engagement time and is successful in certain cases, i.e. the “English” page has a 66% conversion rate. Thus, improvements should be made on the “French” and “Spanish” versions.

# Business Problem Overview and Solution Approach

- Problem Definition:
  - E-news Express, an online news portal, aims to expand its business by acquiring new subscribers. The design team of the company has researched and created a new landing page that has a new outline and more relevant content shown compared to the old page.
  - The goal is to perform a statistical analysis (at a significance level of 5%) to determine the effectiveness of the new landing page in gathering new subscribers for the news portal.
- Solution Approach and Methodology:
  - This A/B testing experiment is consisted of 100  randomly selected users which are divided equally into two groups. The existing landing page is served to the first group (control group) and the new landing page to the second group (treatment group).
  - For EDA, both univariate and multivariate analysis will be performed to find insights.
  - For Hypothesis Testing, we will answer 4 relevant questions using statistical analysis.
  - Recommendations will be given, based on the results of both EDA and Hypothesis Testing.

# EDA Results: Data Overview

- The dataset is comprised of 100 rows and 6 columns. The columns are as follows:

| | | |
|---|---|---|
| user_id | Unique id of the User | Type: Int64 |
| group | “Control” or “Treatment” | Type: Object (String/Text) |
| landing_page | “Old” or “New” | Type: Object (String/Text) |
| time_spent_on_the_page | Minutes spent on the Page | Type: Float64 |
| converted | “Yes” or “No” | Type: Object (String/Text) |
| language_preferred | “English”, “Spanish, “French” | Type: Object (String/Text) |

# EDA Results: Data Overview

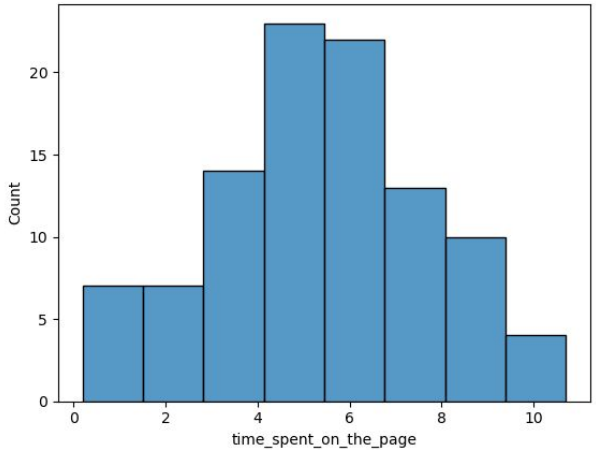
- The dataset has no missing values in any of the columns. Furthermore, there are no duplicate values in any of the columns either. Thus no action to handle missing or duplicate values is necessary.
- Some quick observations and insights from the data:
  - Time Spent on Page:
    - Min = 0.19
    - Mean = 5.38
    - Max = 10.71
  - Converted:
    - Yes = 54 (54%), No = 46 (46%)
  - Language Preferred
    - “Spanish” with 34 (34%), “French” with 34 (34%), “English” with 32 (32%)

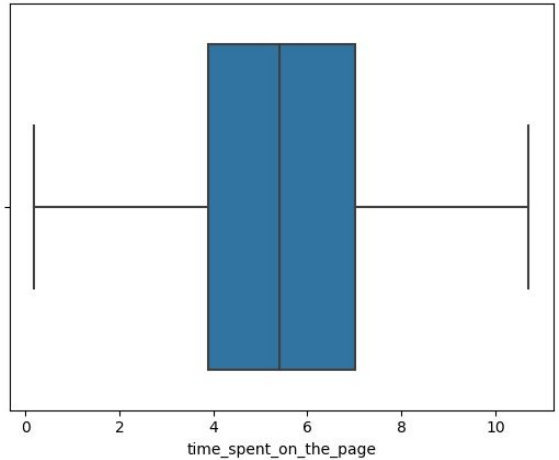
# EDA Results: Univariate Analysis

- In this section, we perform univariate analysis on the data.
  - For each attribute, we perform a descriptive statistical analysis where only that attribute is involved as a variable.
  - We also analyze the corresponding data and present the distribution of the attribute, with confidence intervals to signify variance and statistical significance.
  - Finally, we write the conclusion based on both quantitative and qualitative observations.

# EDA Results: Univariate Analysis

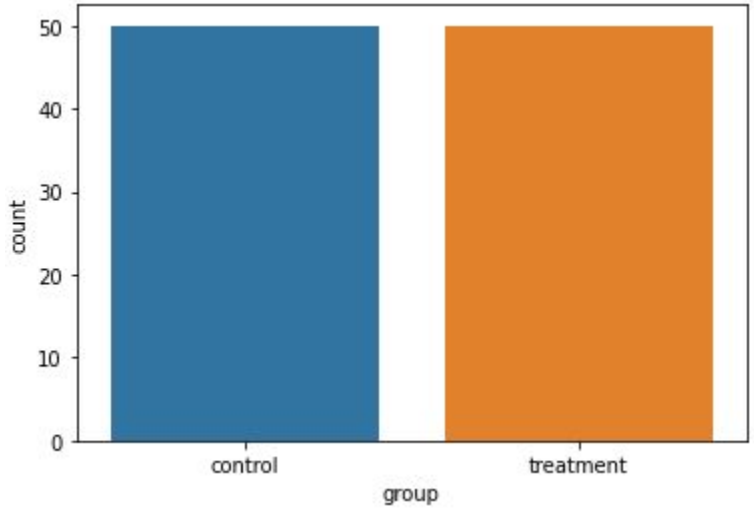
- Analyzing the time_spent_on_the_page attribute, we report the min time (0.19), the median time (5.42), the mean time (5.38) and the max time (10.71), with a standard deviation of 2.38. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



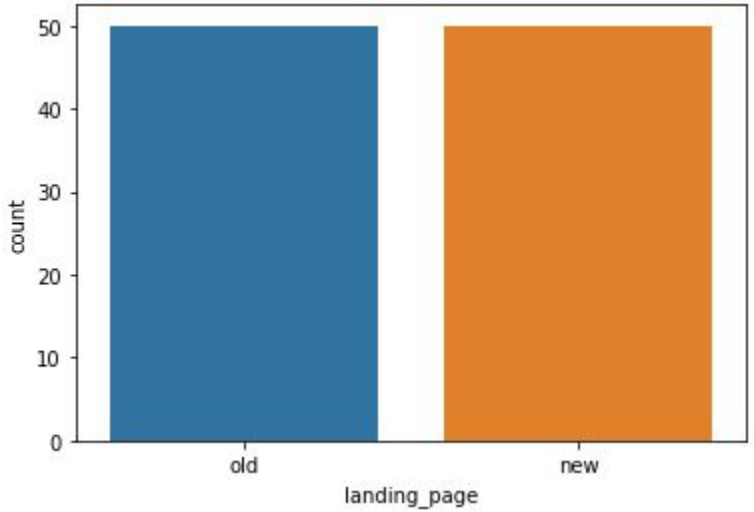
# EDA Results: Univariate Analysis

- Analyzing the group attribute, we observe 50 values of “control” and 50 values of “treatment”, thus confirming our experimental setup for A/B testing. A countplot graph is presented to depict this more clearly.



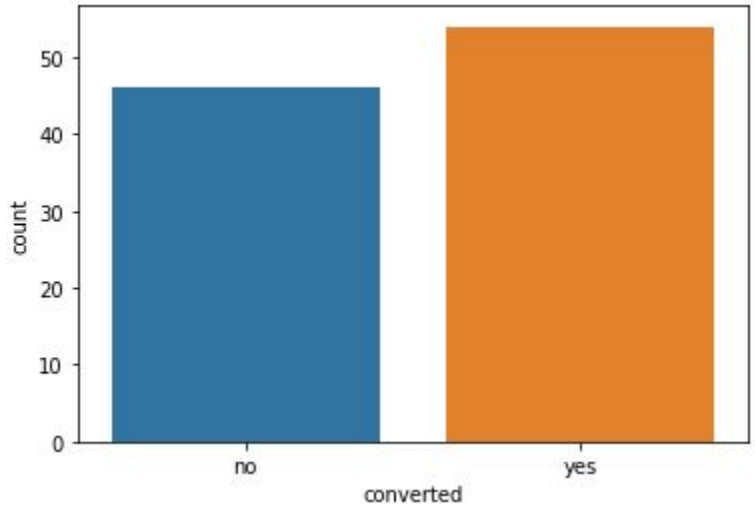
# EDA Results: Univariate Analysis

- Analyzing the landing_page attribute, we observe 50 values of “old” and 50 values of “new”, thus confirming our experimental setup for A/B testing. A countplot graph is presented to depict this more clearly.



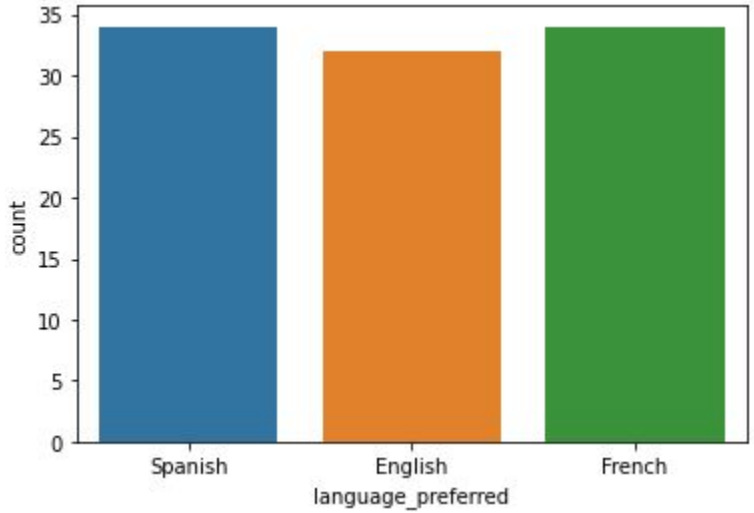
# EDA Results: Univariate Analysis

- Analyzing the converted attribute, we observe 54 values of “yes” and 46 values of “no”. This means that overall, the aggregate conversion rate (regardless of landing page) is 54%. A countplot graph is presented to depict this more clearly.



# EDA Results: Univariate Analysis

- Analyzing the language_preferred attribute, we observe 34 values of “Spanish”, 34 values of “French” and 32 values of “English”. A countplot graph is presented to depict this more clearly.

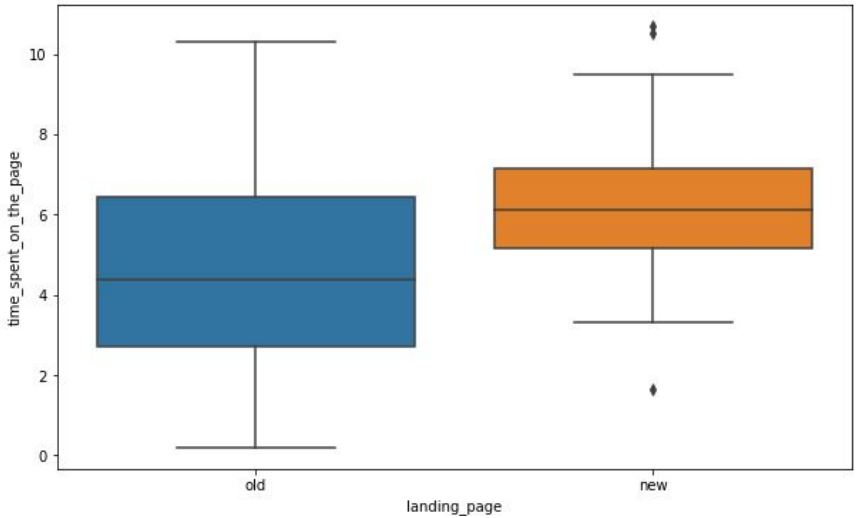


# Multivariate Analysis

- In this section, we perform multivariate analysis on the data.
    - For selected attributes, we perform a descriptive statistical analysis where each attribute is jointly examined along with other attributes in order to evaluate their relationship and correlation.
    - We also analyze the corresponding data and present the joint distribution of the attributes, with confidence intervals to signify variance and statistical significance.
    - Finally, we write the conclusion based on both quantitative and qualitative observations.

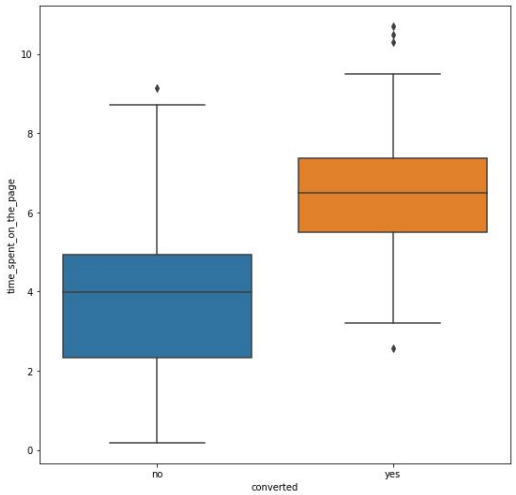
# Multivariate Analysis

- Analyzing the relationship between the landing_page and time_spent_on_the_page attributes, we observe that the new landing page overall displays higher time spent on the page, as it has a higher mean value and a lower variance. However, while the new landing page seems to be particularly better in boosting the time users spend on the page, the result is not statistically significant.



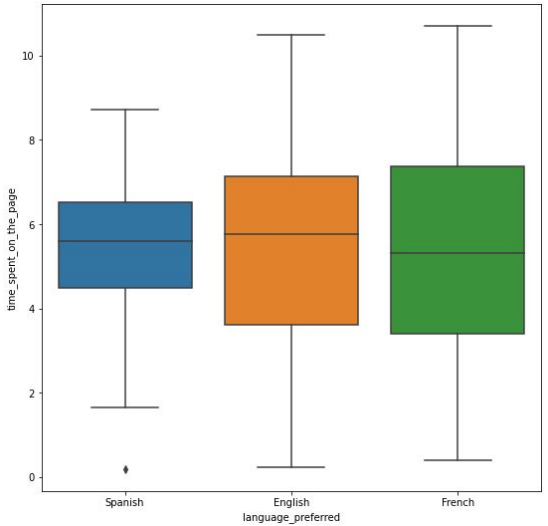
# Multivariate Analysis

- Analyzing the relationship between the converted and time_spent_on_the_page attributes, we observe that the more time users spend on the page, the more likely they are to be converted to subscribers. The result is statistically significant.



# Multivariate Analysis

- Analyzing the relationship between the language_preffered and time_spent_on_the_page attributes, we do not conclude in any statistically significant observation. However, it is worth noting that users who have a preference for the Spanish language appear to have lower variance than the rest.

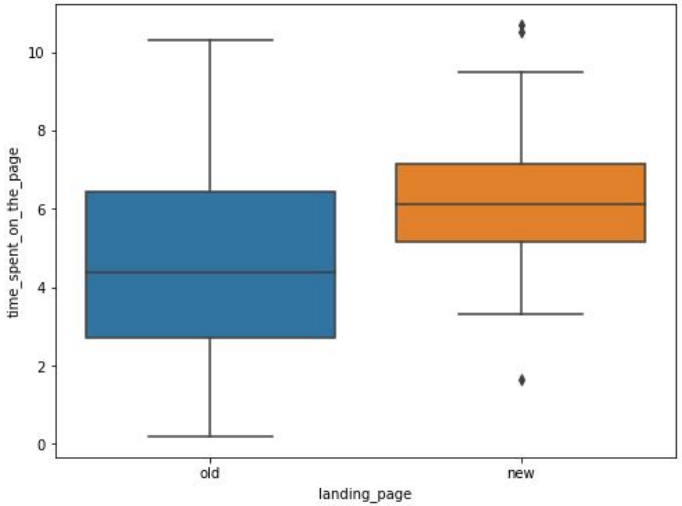


# Hypotheses Testing: Questions

1. Do the users spend more time on the new landing page than the existing landing page?
2. 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?
3. Does the converted status depend on the preferred language?
4. Is the time spent on the new page same for the different language users?

# Hypothesis Testing: Question 1

- We first perform a visual analysis, which shows a boxplot of landing_page and time_spent_on_the_page, as follows:



# Hypothesis Testing: Question 1

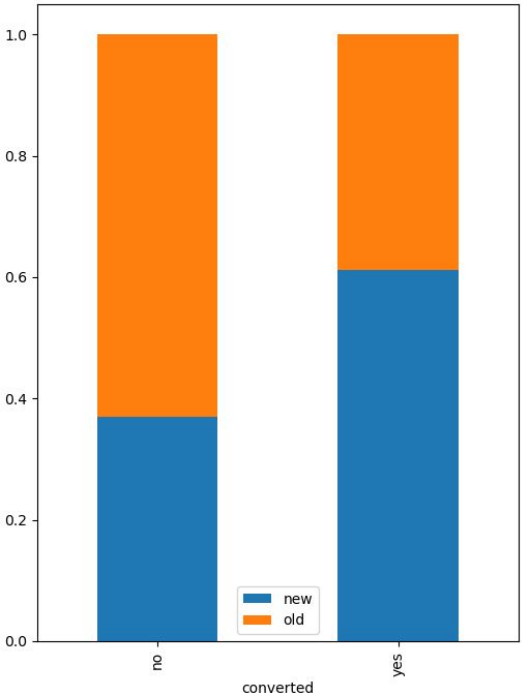
- Define the null hypothesis and the alternative hypothesis:
  - H0: Users spend equal or less time on the new landing page than the existing landing page.
  - Ha: Users spend more time on the new landing page than the existing landing page.
- Select an appropriate statistical test:
  - Properties: One-tailed test for two population means from two independent populations whose standard deviation values are unknown.
  - Thus, we will perform an one-tailed 2-sample independent t-test.
- Select an appropriate significance threshold:
  - We select a = 0.05 as the significance threshold.

# Hypothesis Testing: Question 1

- For the one-tailed 2-sample independent t-test, we need to examine if the standard deviation values of the two populations are equal.
    - Standard Deviation for “Old Page” Sample: 2.58
    - Standard Deviation for “New Page” Sample: 1.82
    - Thus, the standard deviation values of the two populations can be assumed to be not equal.
- Using all the above information, we perform “ttest_ind” from the “scipy.stats” Python library.
    - Resulting p-value: 0.000139
    - Thus, as the p-value (0.000139) is less than the significance threshold alpha (0.05), we reject the null hypothesis H0.
- Inference and Recommendation:
    - Indeed, users spend more time on the new landing page than the existing landing page. This is a positive indication to choose the new landing page over the old one.

# Hypothesis Testing: Question 2

- We first perform a visual analysis, which shows a barplot of converted and landing_page , as follows:



# Hypothesis Testing: Question 2

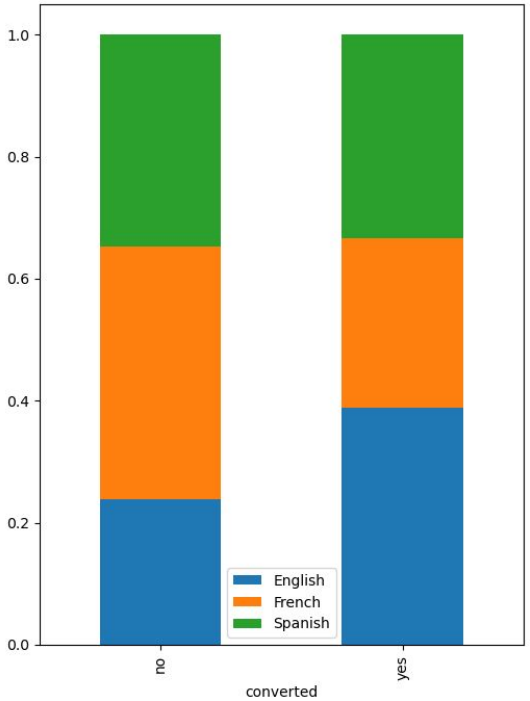
- Define the null hypothesis and the alternative hypothesis:
  - H0: The conversion rate (the proportion of users who visit the landing page and get converted) for the new page is less than or equal to the conversion rate for the old page.
  - Ha: The conversion rate (the proportion of users who visit the landing page and get converted) for the new page is greater than the conversion rate for the old page.
- Select an appropriate statistical test:
  - Properties: One-tailed test for two population proportions from two independent populations.
  - Thus, we will perform an one-tailed 2-sample z-test.
- Select an appropriate significance threshold:
  - We select a = 0.05 as the significance threshold.

# Hypothesis Testing: Question 2

- Using all the above information, we use Python in order to perform “proportions_ztest” from the “statsmodels.stats.proportion” library.
    - Resulting p-value: 0.008
    - Thus, as the p-value (0.008) is less than the significance threshold alpha (0.05), we reject the null hypothesis H0.
- Inference and Recommendation:
    - Indeed, the conversion rate (the proportion of users who visit the landing page and get converted) for the new page is greater than the conversion rate for the old page. This is another positive indication to choose the new landing page over the old one.

# Hypothesis Testing: Question 3

- We first perform a visual analysis, which shows a barplot of converted and language_preffered, as follows:



# Hypothesis Testing: Question 3

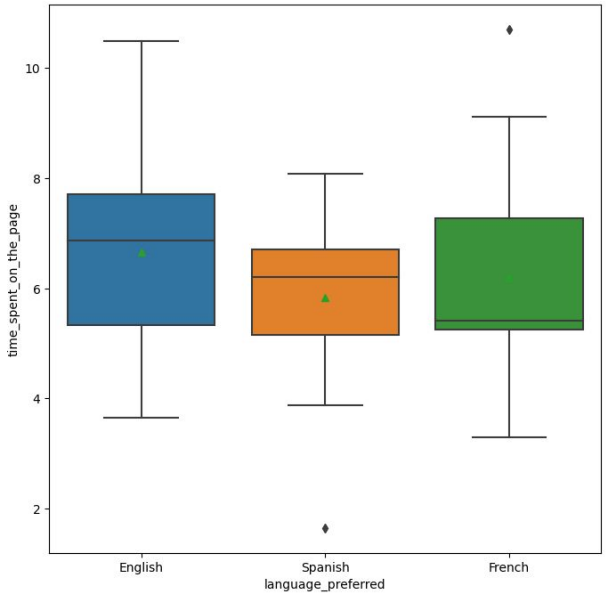
- Define the null hypothesis and the alternative hypothesis:
  - H0: The converted status is independent of the preferred language.
  - Ha: The converted status depends on the preferred language.
- Select an appropriate statistical test:
  - Properties: Test of independence for two categorical variables: converted and language_preffered.
  - Thus, we will perform a chi-square test of independence.
- Select an appropriate significance threshold:
  - We select a = 0.05 as the significance threshold.

# Hypothesis Testing: Question 3

- The Contingency Table is created for the two categorical variables: converted and language_preferred.
    - Converted: “No” <=> Language_Preffered: “English” (11), “French” (19), “Spanish” (16).
    - Converted: “Yes” <=> Language_Preffered: “English” (21), “French” (15), “Spanish” (18).
- Using all the above information, we perform “chi2_contingency” from the “scipy.stats” Python library.
    - Resulting p-value: 0.213
    - Thus, as the p-value (0.213) is greater than the significance threshold alpha (0.05), we fail to reject the null hypothesis H0.
- Inference and Recommendation:
    - No, the converted status does NOT depend on the preferred language. Therefore, preferred language should not be a factor in the choice of landing page regarding conversion.

# Hypothesis Testing: Question 4

- We first perform a visual analysis, which shows a boxplot of language_preffered and time_spent_on_the_page, as follows:



# Hypothesis Testing: Question 4

- Define the null hypothesis and the alternative hypothesis:
  - H0: The time spent on the new page is different for the different language users.
  - Ha: The time spent on the new page is the same for the different language users.
- Select an appropriate statistical test:
  - Properties: Test for three population means.
  - Thus, we will perform an ANOVA test.
- Select an appropriate significance threshold:
  - We select a = 0.05 as the significance threshold.

# Hypothesis Testing: Question 4

- Using all the above information, we perform a “f_oneway” from the “scipy.stats” Python library.
    - Resulting p-value: 0.432
    - Thus, as the p-value (0.432) is greater than the significance threshold alpha (0.05), we fail to reject the null hypothesis H0.
- Inference and Recommendation:
    - No, the time spent on the new page is NOT the same for the different language users. Therefore, preferred language should be a factor in the choice of landing page regarding time spent on the page.

# Executive Summary

- The executive summary of the E-News Express Hypothesis Testing presentation is as follows:
    - **New landing page results in higher user engagement time:** Users spend more time on the new landing page than the old landing page.
    - **New landing page results in higher user conversion rate:** The conversion rate (the proportion of users who visit the landing page and get converted) for the new page is greater than the conversion rate for the old page.
    - **Preferred language independent of user conversion rate, but affects user engagement time:** Preferred language correlates with an increase of the time users spend on the page. However, it is not a factor when it comes to the conversion of a user to a subscriber.
    - **Recommendation 1:** The new landing page should be chosen, as it provides higher engagement (more time spent) and higher conversion rate (users becoming subscribers).
    - **Recommendation 2:** While the preferred language does not affect the conversion rate overall, it affects engagement time and is successful in certain cases, i.e. the “English” page has a 66% conversion rate. Thus, improvements should be made on the “French” and “Spanish” versions.

# Thank you for your time!

Michail Mersinias

02/13/2023