# Data analysis on E-news Project

Project Express News: Paul-Yvann Djamen

09/13/2022

# Contents / Agenda

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

# Executive Summary and recommendation

- There is enough evidence that users spend more time on the new page and convert into subscribers which means the new page has more traffic than the old page and is effective
- The proportion of users who visit the new page and get converted is greater than the proportion of those who convert accessiing the ol page
- The converted status does not depend on the prferred language
- The time spent on the new page is the same for different language users
- I recommend a larger sample size to divide users into more categories which can be anlaysed to derive more insights and quantify by how much the proportion of subcribers on the new page differs to the existing page
- There were no significant differences on time spent on a page based on different language users. Geographical Location could probably be a better variable used to explain the variability in how much time spent on a page differs
- Overall, users spend 3 to 9 minutes exploring after landing on the new page while users in the control group spend close to 0 minutes to over 10 minutes. There were no missing data or duplicates in the data, but a few outliers for users in the treatment group.
- Executives will be happy to know their concerns were valid although new page users have more outliers than old page users. Outliers could be due to users of different languages since meaning can get lost in translation.
- I recommend conducting testing on specific days and times of the week during A/B testing to have better control on the experiment such that our results can be more interpretable

# Business Problem Overview and Solution Approach

- **Problem Overview**

E-news Express, an online news portal, aims to expand its business by acquiring new subscribers. We are to perform data analysis on actions taken by visitors on the website to understand user interests and determine how to drive better engagement.
Executives at E-news are concerned subscribers have declined compared to last year because of poor website design.

- **Solution approach**

The design team of the company has researched and created a new landing page that has a new outline & more relevant content compared to the old page.
To test its effectiveness, the Data Science team conducted an experiment by randomly selecting 100 users and dividing them equally into two groups. The existing landing page was served to the first group (control group) and the new landing page to the second group.
Statistical analysis will be performed at the 5% level of significance to determine the effectiveness of the new landing page :

For our complete analysis, the following questions will be answered:

# Business Problem Overview and Solution Approach

- **For our complete analysis, the following questions will be answered following four methods**

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 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?

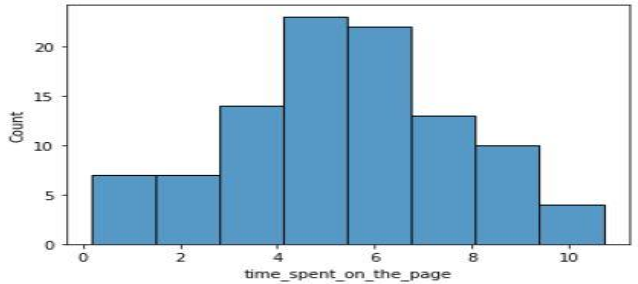
4. Is the time spent on the new page the same for the different language users?

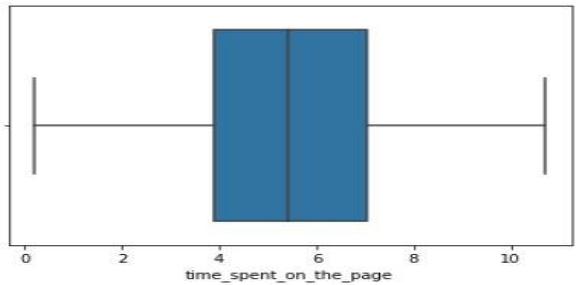
- **The Following tests were used to come up with conclusions an inference:**

2-sample indepent T-test, 2 sample proportions test, a Chi square test of independence and ANOVA F test

# EDA Results

- Univariate Analysis



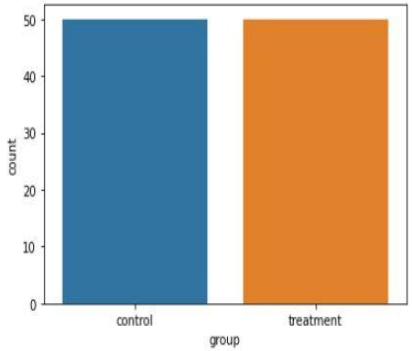

There two numerical variables in our data set which are user id and time spent on a page by visitors.

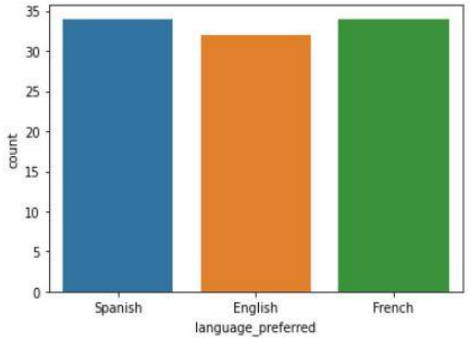
The histogram and boxplot of time_spent_on_the_page has a distribution that appears normal with no outliers

*Link to Appendix slide on data background check*

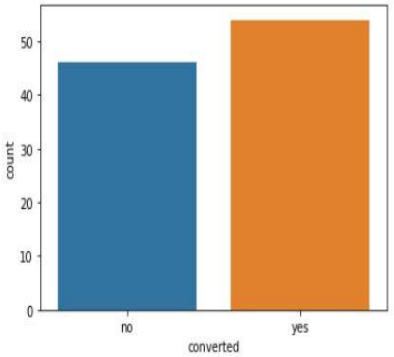
# EDA Results

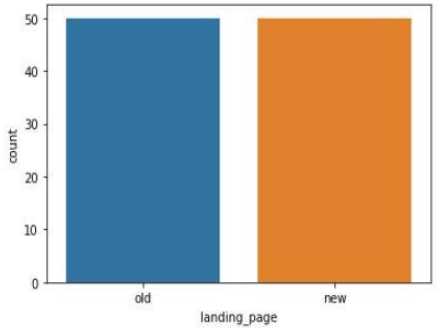
- Univariate Analysis: Categorical variables




- We notice equal count of users in control and treatment group
- French and Spanish have about 3 more users than English
- There is equal count of users landing on the old and new page
- Of all users landing page the count of those who convert and subscribe to the news channel is higher than those who do not subscribe

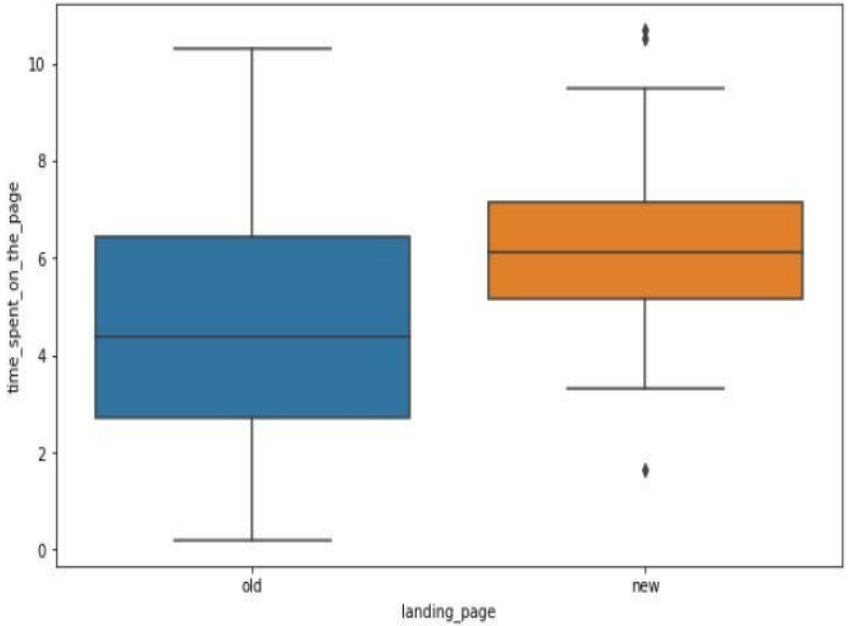



*Link to Appendix slide on data background check*

# EDA Results

- Bivariate Analysis



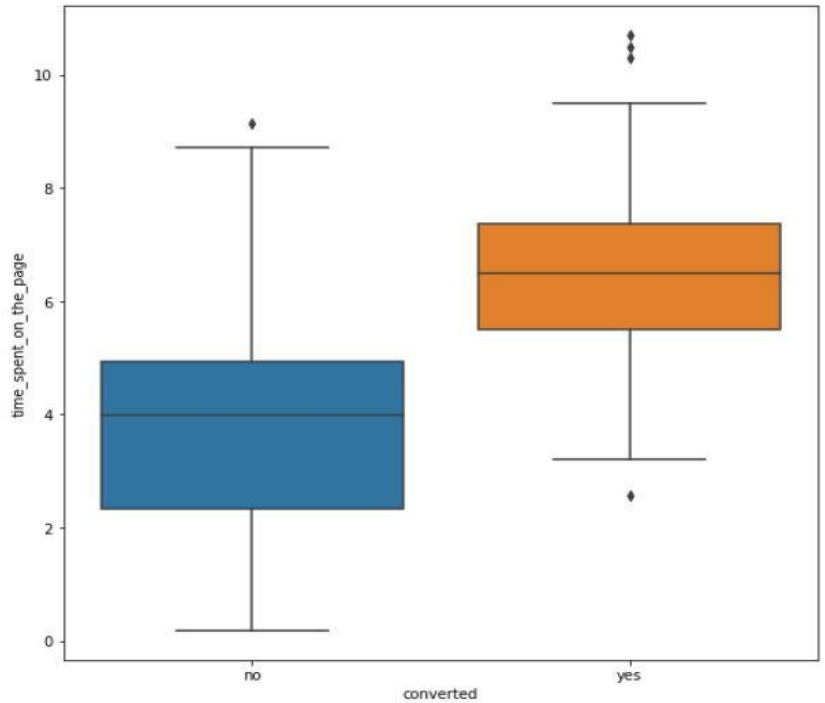

There is lower variability on amount of time spent on a page for users landing on the new page.

users on the old page have a boxplot which is normal with a median of about 5 with no outliers while the boxplot of users on the new page has a few outliers on either sides of the whiskers .
Both plots appear symmetrical and normal

*Link to Appendix slide on data background check*

# EDA Results

- Bivariate Analysis



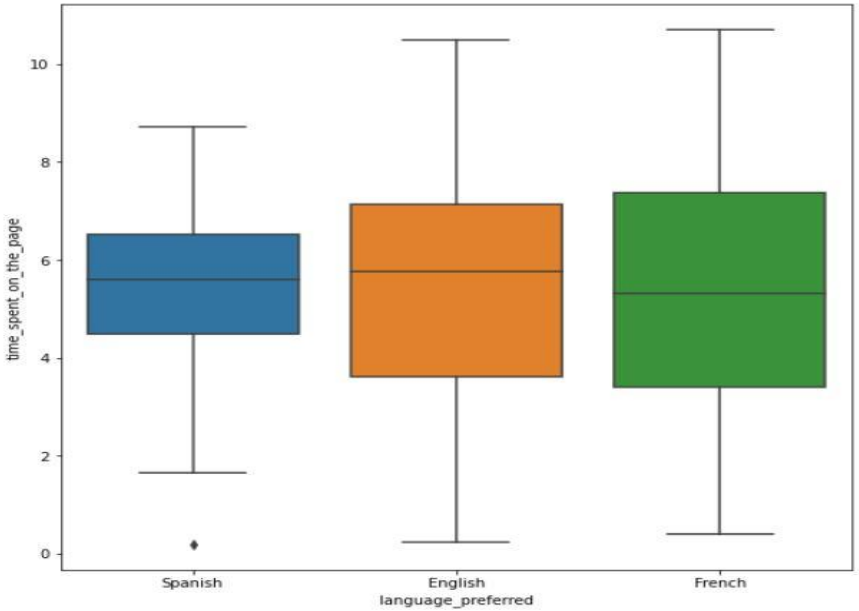

There is lower variability on amount of time spent for users who subscribe to the news channel after landing on a page than those who do not subscribe or convert minimum time for those who convert is about 3 minutes while the maximum is more than 9 minutes.

As expected from the previous plot, users who convert to subscribers after visiting a page have a boxplot with about 4 outliers. Overall, both plots look normal although users who convert generally spend more time on a page they visit

*Link to Appendix slide on data background check*

# EDA Results

- Bivariate Analysis




The boxplot of language-preferred and time spent on a page reveals a similar distribution with a median time slightly below 6minutes.

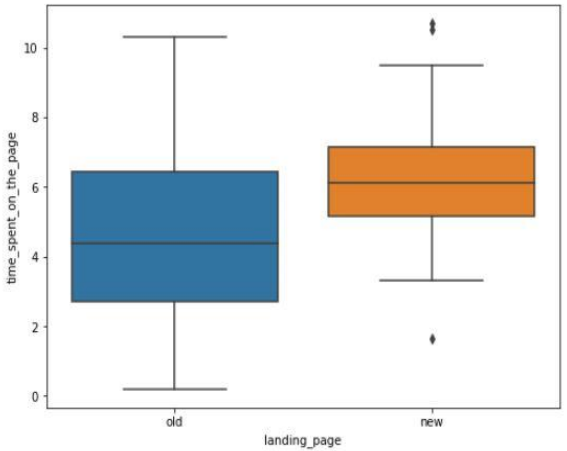
French has the highest variability of users while Spanish has the least.
The median or 50$^{th}$ percentile of all language users are close.

Their distribution are similar, and all appear to be normal with the only outlier present in Spanish users

*Link to Appendix slide on data background check*

# Hypotheses Tested and Results

- *Question 1 answered*




**Step 1: Define the null and alternate hypotheses**

$H_0$: p1 = p2

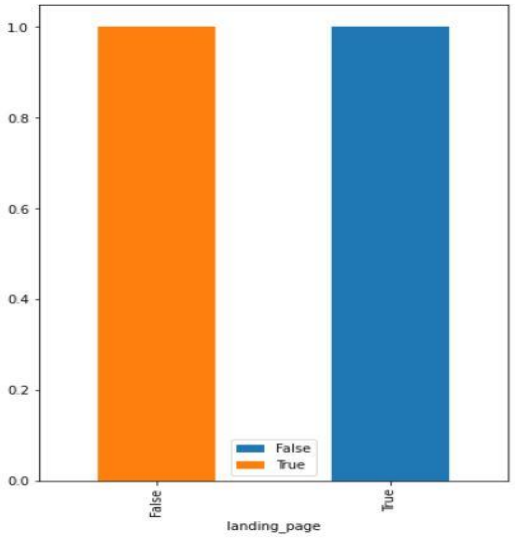
$H_a$: p1 < p2

- Visual analysis of the time spent by users when visiting the new page and old page shows the minimum time spent for those who visit a new page is about 3 minutes. Visitors of new pages have a few outliers
- I tested the hypothesis of whether users spend more time on the new landing page than the existing landing page as follows using a 2 sample independent t-test since we calculated the sample variance.
- H_0 → null hypothesis: mu_1 = mu_2
- H_A → alternative hypothesis: mu_1 < mu_2, where mu_1 and mu_2 represent average time spent my users on the old page and new page respectively
- At the 0.05 significance level, the p-value calculated was `p-value 0.000 to 3 decimal places`
- `Since the p-value is much less than alpha of 0.05, we have enough statistical evidence to reject the null hypothesis that users spend equal amount of time on the existing and new page`

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

# Hypotheses Tested and Results

- Question 2 answered




- The visual compares the conversion rate of users landing on the new page vs those landing on the old page

- I tested the hypothesis of whether the proportion of users who convert to subscribers for the new page is greater than those who subscribe to the old page

- H_0 → null hypothesis: P_1 = P_2
- H_A → alternative hypothesis: P_2 > P_1, where P_1 and P_2 represent proportion of those who convert to subscribers after visiting the old page and new page respectively

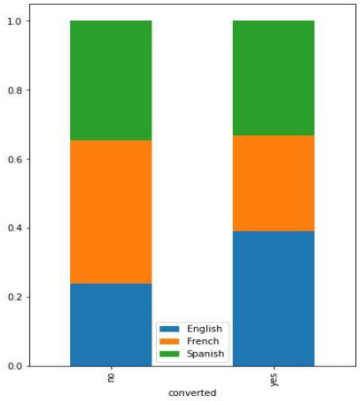
- At the 0.05 significance level, the p-value calculated was `0.008 to 3 decimal places.`
- `Since the p-value is much less than alpha of 0.05, we have enough statistical evidence to reject the null hypothesis that users spend equal amount of time on the existing and new page`

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

The p-value is

# Hypotheses Tested and Results

- Question 3




Out[65]:

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

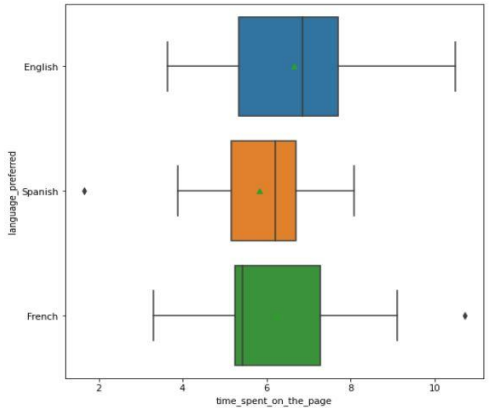
- The visual compares dependency between conversion status and the preferred language together with a contingency table. The bar plots vary differently with French having the largest frequency on non converted while it has the least frequency on the converted users
- I tested the hypothesis of conversion status and preferred language are independent using the chi square test of independence since all the assumptions of normal distribution, simple random sample and constant variance are met
- H_0 → null hypothesis: conversion status and preferred language are independent
- H_A → alternative hypothesis: conversion status and preferred language are dependent.

- At the 0.05 significance level, the p-value calculated was `0.213 to 3 decimal places`
- `Since the p-value is greater than alpha of 0.05, we fail to reject the null hypothesis since we have enough statistical evidence to say that conversion status and language preferred are independent`

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

# Hypotheses Tested and Results

- Question 4 answered




```
Out[83]: language_preferred
English    6.663750
French     6.196471
Spanish    5.835294
Name: time_spent_on_the_page, dtype: float64
```

- The visual plot compares time spent on the new page for different language users. The box plot shows English users spend the highest amount of time on a new page with a mean of about 6.6 minutes while Spanish users have the lowest mean. This is supported by the table with Out[83]. French and Spanish users have one outlier while English users do not. Spanish users have the lowest variability
- I tested the hypothesis of all means being the same using the one- way anova f-test

- H_0 → null hypothesis: all means are the same; mu_1 = mu_2 =mu_3
- H_A → alternative hypothesis: at least one of the means are different.

- At the 0.05 significance level, the p-value calculated was `0.432 to 3 decimal places`
- `Since the p-value is greater than alpha of 0.05, we fail to reject the null hypothesis since we have enough evidence to conclude the means are not statistically different.`

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

# Hypotheses Tested and Results

- Q4 Confirming the assumptions using anova

```
from scipy.stats import levene
statistic, p_value = levene(df_new['time_spent_on_the_page'][df_new['language_preferred'] == 'English'],
                            df_new['time_spent_on_the_page'][df_new['language_preferred'] == 'French'],
                           df_new['time_spent_on_the_page'][df_new['language_preferred' ] == 'Spanish'])
print('The p_value is', p_value)
```

```
The p_value is 0.46711357711340173
```

```
P greater than alpha fail the reject the null of homogeneity of contant variance
```

```
from scipy import stats

w,p_value = stats.shapiro(df_new['time_spent_on_the_page'])
print('The p_value is', p_value)
```

```
The p_value is 0.8040016293525696
```

```
P greater than alpha fail to reject the null of constant variance assumptions
```

The results from the Shapiro test and Levene test are in accordance with the assumptions needed to perform anova test on our dataset with time_spent_on_the_page being the numerical variable and response while language preferred is the categorical variable

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

# Appendix

- Number 1 with test performed: from scipy.stats import ttest_ind

```
test_stat, p_value =  ttest_ind(time_spent_new, time_spent_old, equal_var = False, alternative = 'greater')

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

- Number 2 with test performed:  from statsmodels.stats.proportion import proportions_ztest

```
test_stat, p_value = proportions_ztest([new_converted, old_converted] , [n_treatment, n_control], alternative = 'larger)

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

- Number 3 with test performed:from scipy.stats import chi2_contingency

```
chi2, p_value, dof, exp_freq = chi2_contingency(contingency_table)

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

- Number 4 with test performed :from scipy.stats import f_oneway

```
test_stat, p_value = f_oneway(time_spent_English, time_spent_French, time_spent_Spanish)

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

Great Learning

# Happy Learning !