E-News Express

Project Business Statistics

Date: November, 2023

Contents / Agenda

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

Executive Summary

Conclusions:

- Users visiting the new landing page convert in statistically more significant numbers to a subscription than visitors to the old landing page.
- Users on the new landing page spend statistically significant more time, as compared to users on the old landing page.
- The preferred language of users also has a statistically significant impact on the conversion rate. There is in particular a positive correlation for English and Spanish.
- French language preferred visitors spent most time on the landing page, yet have the lowest converting rate.

Actionable insights & recommendations:

- Recommendation to permanently switch to the new landing page, as it contributes to an increased conversion rate among visitors, and increased time spent on the site.
- Extra attention should be paid to the content for French language preferred visitors. There is an opportunity to increase the conversion rate among this population.

Business Problem Overview and Solution Approach

The problem: E-news Express aims to expand its business by acquiring new subscribers. There has been a decline in new monthly subscribers compared to the past year because the management team suspects the current webpage is not designed well enough in terms of the outline & recommended content to keep customers engaged long enough to make a decision to subscribe.

The solution approach / methodology: The design team has researched and created a new landing page that has a new outline & more relevant content shown compared to the old page. Our data scientist team has explored the data and performed 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 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 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?

Exploratory Data Analysis - Key Results

The dataset comprised of 100 rows and 6 columns. No data was missing.
Data tracked 100 users with an average visit on the landing page of 5.38 minutes, and a standard deviation of 2.38 minutes. Minimum time spent by a user was 19 seconds, the maximum 10.71 minutes.

Language preferred among 100 users is Spanish and French at 34% each, followed by English at 32%.

100 users were divided in two sample groups of 50. One control group was directed to the existing landing page, of which 21 converted to a subscription. The treatment group was directed to the new landing page, of which 33 converted to a subscription.

The percentage of people opting to convert to a subscription after visiting the landing page in general is 54%, versus 46% of people who do not.

Check for duplicates: There are five duplicates in the column Time_Spent_On_The_Page. It is entirely possible that five users spent an equal amount of time on the landing page. Not cause for concern.

Link to Appendix slide on data background check

EDA key results continued

Univariate Analysis

Based on its histogram and boxplot, the column Time spent on the page, seems to have a normal distribution.

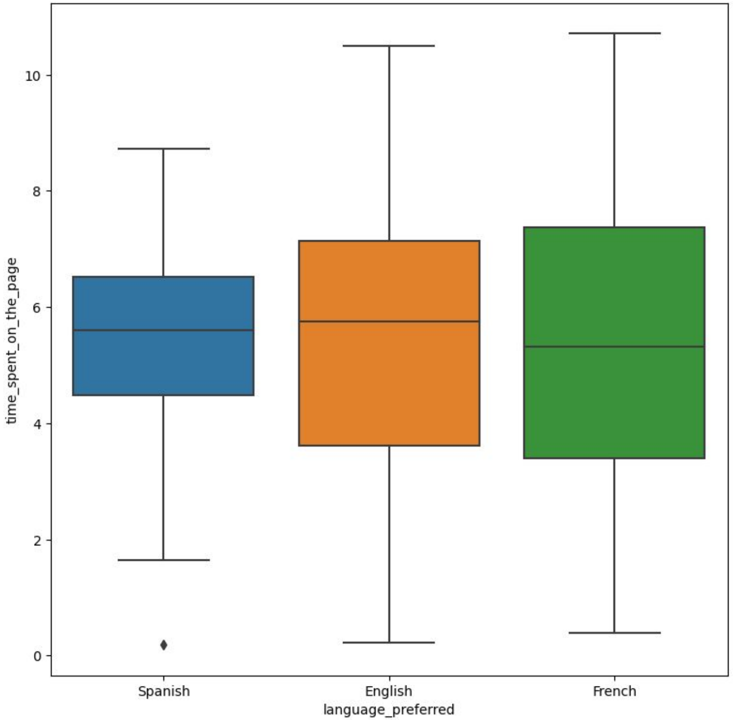
The control and treatment group are equal in size with 50 participants each and have a fixed correlation with the landing page.

The histogram showing the correlation between type of landing page and converted indicates a positive relation for the new landing page.

Histogram for language preferred.
Boxplot for time spent on page and landing page, and conversion status vs time spent on page.

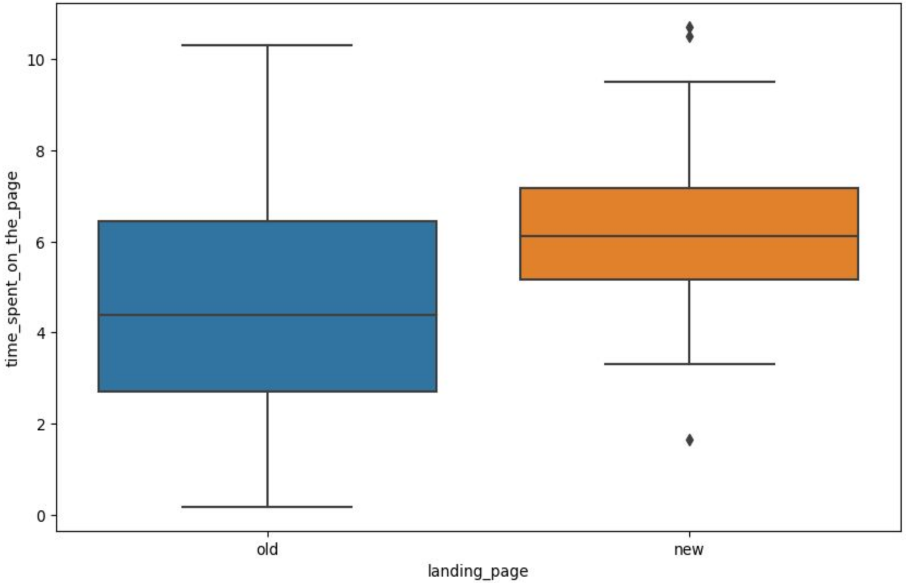
EDA key results continued II

Bivariate Analysis



This boxplot shows the distribution of time spent on the landing page by all 100 users categorized according to their preferred language. French preferred users spent the most time on the page.

1. Do users spend more time on the new landing page or on the old landing page?



Ho:
μ time on new page <= μ time on old page

Ha:
μ time on new page > μ time on old page

Alpha:
0.05

Note*:*

1. *You can use more than one slide if needed*
2. *This template can be followed for all hypotheses tested*

Link to Appendix slide on details of the test performed

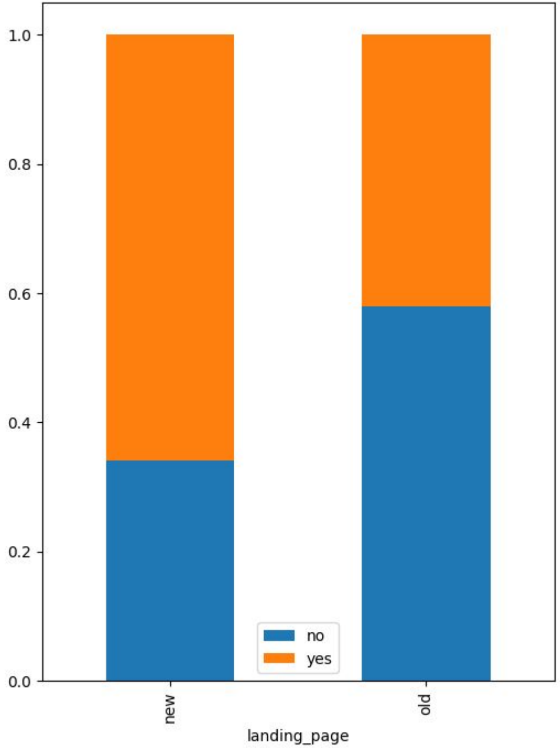
1. Hypothesis tested and result

- This is a one-tailed test concerning two population means from two independent sample populations. The population standard deviations are unknown. **Based on these assumptions, we select the two sample independent t-test** to carry out our statistical analysis.

- Based on this test, the p-value is 0.0001392381225166549, and is less than the pre-set level of significance at 0.05. We therefore reject the null hypothesis in favor of the alternative hypothesis.

- This means users are on average spending more time on the new landing page compared to visitors on the old existing landing page.

2. Is the conversion rate for the new page greater than that for the old page?



Ho :
P<=0.5

Ha:
P>0.5

Alpha:
0.05

Y axis = Converted

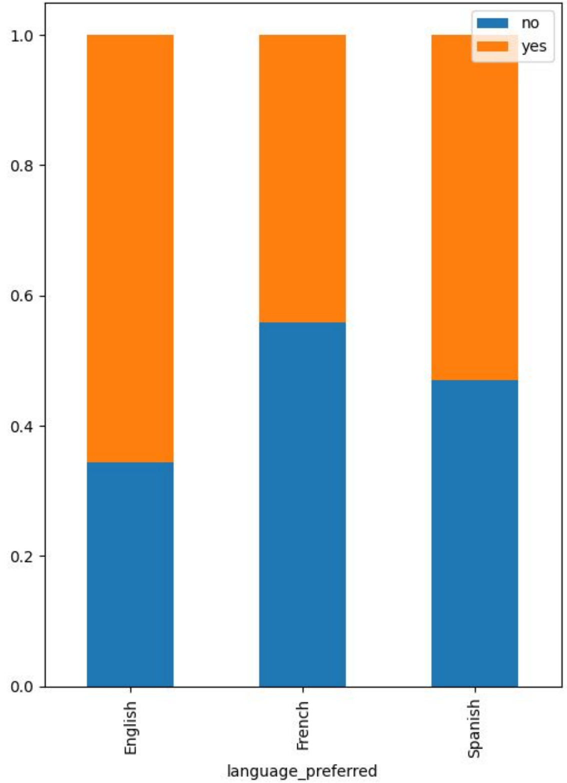
2. Hypothesis tested and results

- This is a one-tailed test concerning two population proportions from two independent populations. **Based on these assumptions, we select the proportions z-test** to carry out our statistical analysis.

- Based on this test, The p-value is 0.008026308204056278. This is lower than the pre-set level of significance of 0.05, and therefore we reject the null hypothesis in favor of the alternative hypothesis.

- The probability rate is greater than half that the conversion rate of users on the new landing page is higher than the conversion rate of users on the old page.

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



Ho:
Conversion rate and language preferred are independent factors.

Ha:
Conversion rate and language preferred are dependent.

Alpha:
0.05

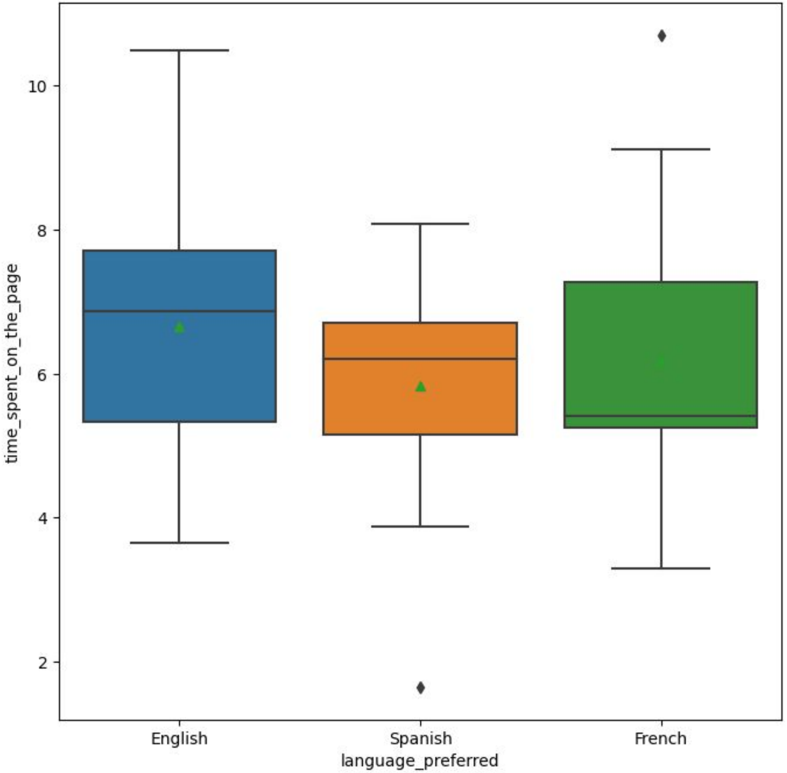
3. Hypothesis tested and results

- This is a problem of the test of independence, concerning two categorical variables - converted status and preferred language. **Therefore we select the chi square test for independence.**

- Based on this test, the p-value is 0.21. This is higher than the pre-set level of level of significance of 0.05, and does not satisfy the predetermined decision rule. Therefore we fail to reject the null hypothesis, in favor of the alternative hypothesis.

- The factors conversion rate and language preferred are statistically dependent of each other.

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



Ho:
μ1 = μ2 = μ3

Ha:
At least one of the means
is not equal to the other means.

Alpha:
0.05

Hypothesis tested and results

- This is a problem, concerning three population means. **Based on this information, we select the analysis of variance (ANOVA) test to compare the three population means.**

- Based on this test, the p-value obtained is 0.43.   This is higher than the preset level of significance of 0.05. Therefore we fail to reject the null hypothesis, in favor of the alternative hypothesis

- At least one of the languages is not similarly preferred on average as the other languages. From looking at the visualized data, we can see that French is on average less preferred than English and Spanish.

APPENDIX

Data Background and Contents

- To check shape dataset: `df.shape`
- To check for missing data: `df.isnull().sum()`
- To check numerical summary statistics: `df.describe()`
- To check categorical summary statistics: `df['landing_page'].value_counts(normalize=True)`
- To check language preferred: `df['language_preferred'].value_counts(normalize=True)`
- To create arrays for the treatment and control group respectively.

```
control_group = df.loc[(df['group'] == 'control')]
```

```
treatment_group = df.loc[(df["group"]== "treatment")]
```

- To check for each group language preferred:

```
treatment_group['language_preferred'].value_counts()
```

```
control_group['language_preferred'].value_counts()
```

Data Background and Contents continued

- Check for each group statistical summary :

```
control_group.describe()
treatment_group.describe()
```

- Check for each group the correlation with conversion:

```
treatment_group['converted'].value_counts()
control_group['converted'].value_counts()
```

- Check for duplicates:

```
df.nunique()
```

Univariate analysis

- Check how many users per type of landing page converted:

```
df.groupby('landing_page')['converted'].value_counts()
```

Data Background and Contents -continued II

- Check how many users converted per type of landing page continued:

```
# complete the code to plot the countplot
sns.countplot(data=df,x='landing_page', hue='converted')
plt.show()
```

- Check how many users converted: `df['converted'].value_counts()`

```
# complete the code to plot the countplot
sns.countplot(data=df,x='converted')
plt.show()
```

- Check how many users preferred which language: `df['language_preferred'].value_counts()`

```
# complete the code to plot the countplot
sns.countplot(data=df,x='language_preferred')
plt.show()
```

Data background and contents -continued III

Bivariate Analysis

```
# complete the code to plot a suitable graph to understand the relationship between
'time_spent_on_the_page' and 'converted' columns
plt.figure(figsize=(9, 9))
sns.boxplot(data = df, x = 'converted', y = 'time_spent_on_the_page')
plt.show()
```

```
# write the code to plot a suitable graph to understand the distribution of 'time_spent_on_the_page'
among the 'language_preferred'
plt.figure(figsize=(9, 9))
sns.boxplot(data = df, x = 'language_preferred', y = 'time_spent_on_the_page')
plt.show()
```

1. Hypothesis Testing Details

***H*0:** μ new page <= μ old page

***Ha*:** μ new page > μ old page

Alpha: 0.05

Hypothesis: 2 sample independent t-test

```
from scipy.stats import norm
from scipy.stats import ttest_ind

test_stat, p_value = ttest_ind(time_spent_new, time_spent_old, equal_var = False, alternative
='greater')
print('The p-value is', p_value)
```

```
The p-value is 0.0001392381225166549
```

Boxplot:

```
plt.figure(figsize=(8,6))
sns.boxplot(x = 'landing_page', y = 'time_spent_on_the_page', data = df)
plt.show()
```

2. Hypothesis Testing Details

Ho: Conversion rate for the new page is equal or lower than for the old page. P<=0.5

Ha: Conversion rate for the new page is higher than for the old page. P>0.5

Alpha: 0.05

Hypothesis selected: proportions z-test

```
from statsmodels.stats.proportion import proportions_ztest
test_stat, p_value = proportions_ztest([new_converted, old_converted] , [n_treatment, n_control],
alternative ='larger')
print('The p-value is', p_value)
```

The p-value : `0.008026308204056278`

Histoplot:

```
pd.crosstab(df['landing_page'],df['converted'],normalize='index').plot(kind="bar",
figsize=(6,8),stacked=True)
plt.legend()
plt.show()
```

3. Hypothesis Testing Details

Ho: Conversion rate and language preferred are independent factors

.

Ha: Conversion rate and language preferred are dependent.

Alpha: 0.05

Hypothesis Test: chi square test for independence

```
contingency_table = pd.crosstab(df['language_preferred'], df['converted'])
contingency_table_alternative = pd.crosstab(df['converted'], df['language_preferred'])
chi2, p_value, dof, exp_freq = chi2_contingency(contingency_table)
print('The p-value is', p_value)
```

P-value: 0.2129888748754345

Histoplot:

```
pd.crosstab(df['language_preferred'],df['converted'],normalize='index').plot(kind="bar", figsize=(6,8),
stacked=True)
plt.legend()
plt.show()
```

4. Hypothesis testing details

Ho: μ1 = μ2 = μ3

Ha: At least one of the means is not equal to the other means.

Alpha: 0.05

Hypothesis selected:

```
 # one way ANOVA F-test
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)
```

The p-value:

```
0.43204138694325955
```

Histoplot:

```
plt.figure(figsize=(8,8))
sns.boxplot(x = 'language_preferred', y = 'time_spent_on_the_page', showmeans = True, data = df_new)
plt.show()
```

Slide Header

- Please add any other pointers (if needed)

Great Learning

Happy Learning !