E-news Landing Page Analysis

PGP-DSBA _ E-news Express Project

October 13, 2023

Contents / Agenda

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

Executive Summary: Conclusion (1/2): Exploratory Data Analysis of the Sampled Subjects

- On average, the subjects spent about 5 minutes 23 seconds (5.38 minutes) on the landing pages
- The spread of the time spent on the landing page is characterized by a range from about 11 seconds (0.19 minutes) to 10 minutes 42 seconds (10.71 minutes) and a standard deviation of about 2 minutes 23 seconds (2.38 minutes)
- The randomly selected subjects generated an approximately normal distribution of the time spent on the landing pages
- The number of subjects converted is slightly greater (54%) than those not converted
- Half of the subjects spent between 4 and 7 minutes on the landing pages and half also spent more 5 minutes
- The ratio of subjects converted to those not converted is 27:23
- Both Spanish and French are the favorite languages among the subjects (34 each) while English was the preferred language of only 32 subjects
- Subjects assigned to the new landing page tended to spend close to two more minutes on the landing page than those assigned to the old landing page
- Converted subjects, in general, spent over two more minutes on the landing page than unconverted ones
- Subjects having English as preferred language spent the most time on the landing page, closely followed by those whose preferred language is Spanish; the least time spent on the landing page was registered for subjects whose preferred language is French but the differences among the time spent by subjects of the three languages are a few seconds apart

Executive Summary: Conclusion (2/2): Result of Inferential Evaluation of Key Statistical Hypotheses

- On average, users spend more time on the new landing page than on the existing one
- The rate of conversion for the new page is effectively larger than that for the existing page
- The converted status appears to be independent of the preferred language of the user; specifically at the level of significance of 5%
- The average time spent on the new page is the same across different language users

Executive Summary: Business Recommendations

- We have successfully determined that the new page generates greater engagement by ensuring longer time spent on the page and a greater rate of conversion; so the business might effectively drive increased subscriber acquisition by deploying and continuously ameliorating the new page
- We also uncovered that the language choices of the subjects have little or no bearing on the time spent on the new page or conversion rate of users; so language choice might not be a critical business factor for growth
- By testing and exploring other factors and design considerations, E-news Express may identify other factors that might drive growth and ameliorate customer experience

Business Problem Overview

E-news Express, an online news portal, has been faced with attrition of its monthly subscriber base for the last couple of months compared to last year. The company's management is of the opinion that the trend is a result of inadequate design of the landing page, thus not sufficiently attractive to hold users long enough for them to make a decision that registers as a conversion.

After, getting the design team to create a new page with a revamped outline and renewed content in an effort to expand its business and drive growth by acquiring new subscribers, management reached out to the Data Science Team to evaluate the success of this page redesign and its potential business impact.

Solution Approach

The Data Science Team carried out an A/B testing experiment. The experiment was carried out on a sample of 100 subjects split evenly between two groups: The treatment group was assigned to the new landing page and the control group assigned to the old landing page. The group, time spent on the landing page, the conversion status of the subject, the landing page, and the preferred language were registered for each subject. The data thus collected was then analyzed using the following steps to uncover key business insights:

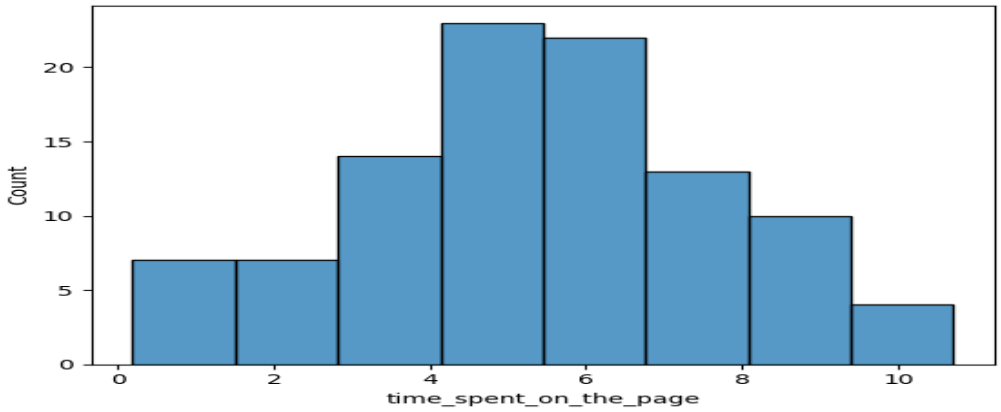
1. Exploratory Data Analysis of the Sample
2. Evaluation of Key Hypotheses using Inferential Statistics

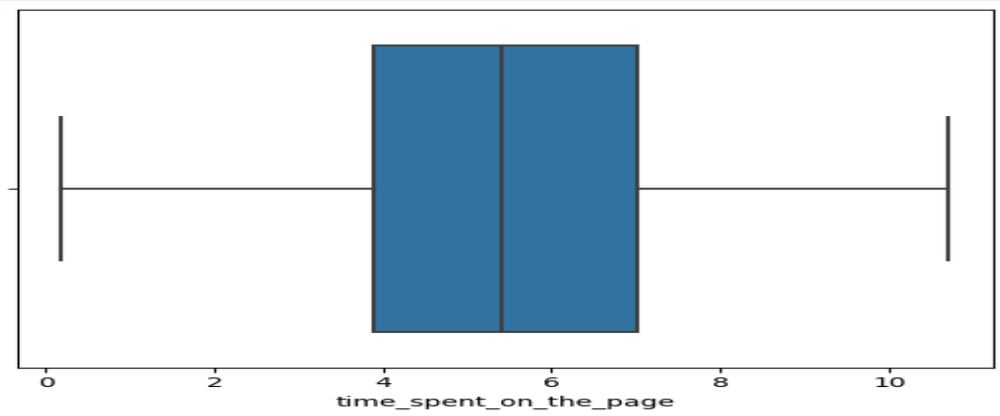
EDA Results: Data Overview

- Each entry contains the user id, the group to which the subject belongs (control or treatment), the landing page to which they were assigned, the time spent on the page, status information on whether they were converted or not, and their preferred language.
- The data appears to consist of a sample of 100 entries
- The data collected contains 4 categorical columns (group, landing_page, converted, and language_preferred) and 2 numerical columns (user_id containing integer values and time_spent_on_the_page containing float values)
- The user id is a unique identifier; thus no interesting insight can be gained from statistical analysis of its values
- The subjects spent about 5 minutes 23 seconds (5.38 minutes) on average on the landing pages
- The spread of the time spent on the landing pages is characterized by a range from about 11 seconds (0.19 minutes) to 10 minutes 42 seconds (10.71 minutes) and a standard deviation of about 2 minutes 23 seconds (2.38 minutes)
- The average and median time spent on the landing pages are close, suggesting that the distribution is at least slightly normal
- The number of subjects converted is slightly greater (54%) than those not converted
- From an initial analysis, the most preferred language among the 3 registered languages appears to be Spanish (34%)

Link to Appendix slide on data background check

EDA Results: Univariate Analysis: Time spent on the page

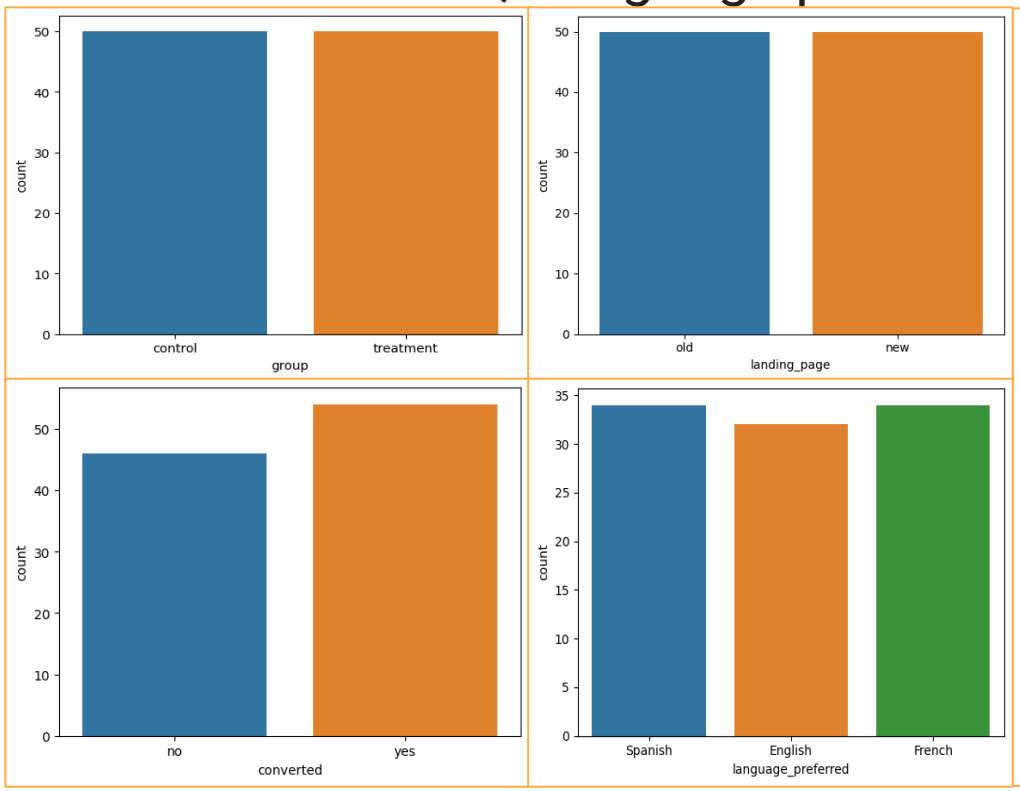




- We have confirmation that the distribution of the time spent on the landing pages is approximately normal
- The time spent on the landing pages has no outliers
- Half of the subjects spent between 4 and 7 minutes on the landing pages and half also spent more 5 minutes

Link to Appendix slide on data background check

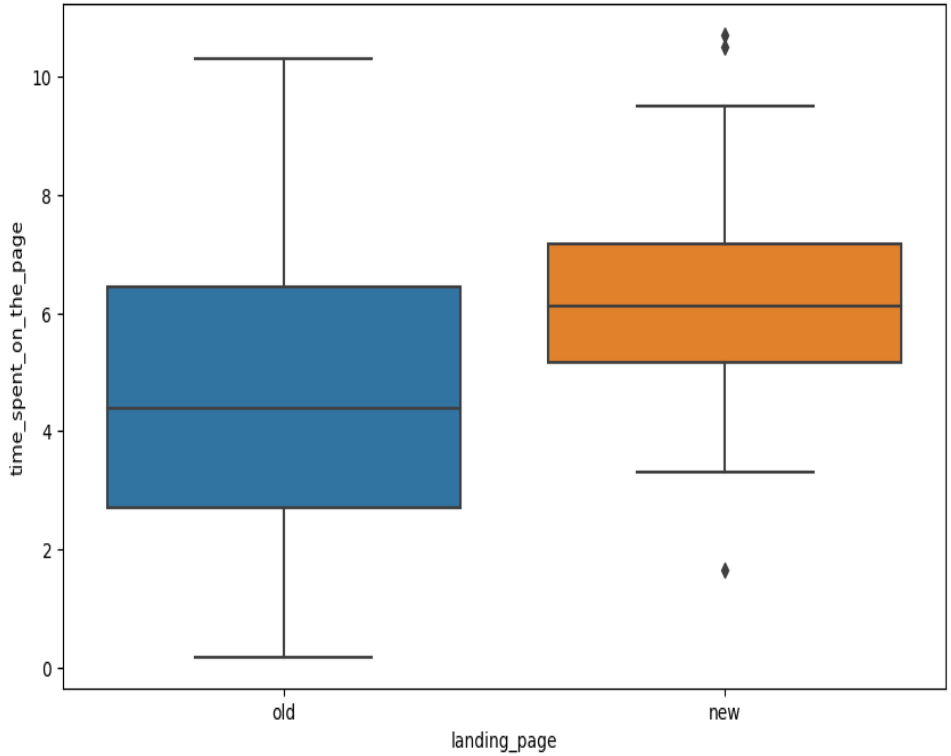
EDA Results: Univariate Analysis: Group, Landing page, Converted Status, Language preferred



- We have confirmation that the subjects are equally distributed between the control and treatment groups
- We also have confirmation that the subjects are equally distributed between the old and new landing pages
- The ratio of subjects converted to those not converted is 27:23
- We now have a clearer perspective on the distribution of preferred languages: Both Spanish and French are the favorite languages among the subjects (34 each) while English was the preferred language of only 32 subjects

Link to Appendix slide on data background check

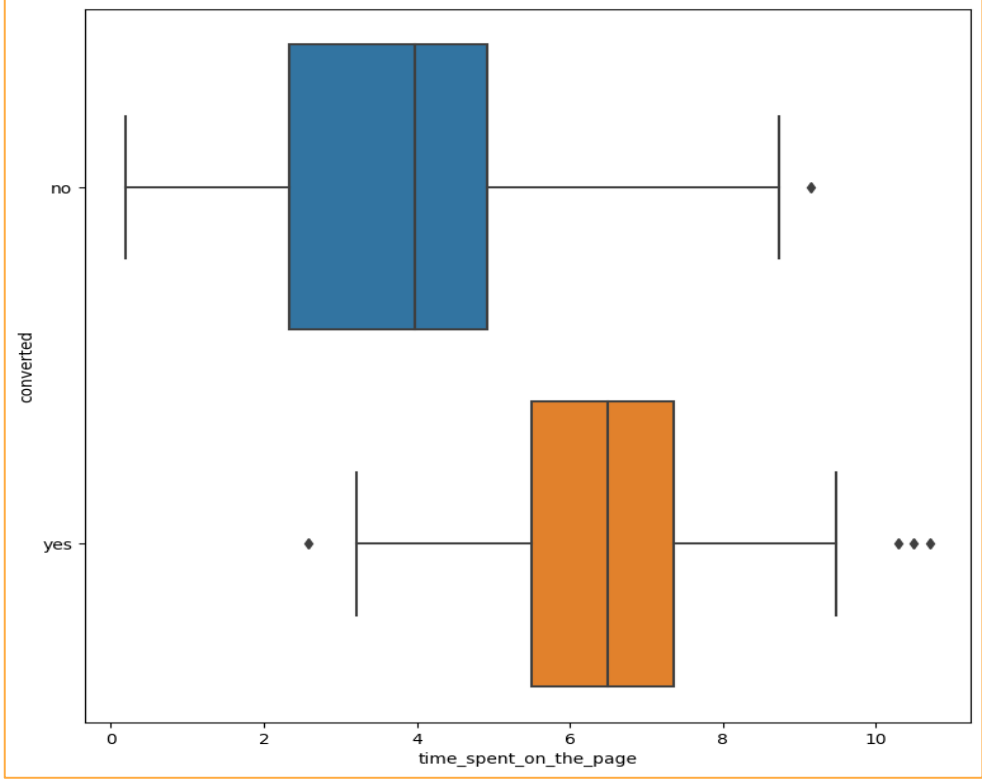
EDA Results: Bivariate Analysis: Landing page vs Time spent on the page



- Subjects assigned to the new landing page tended to spend close to two more minutes on the landing page than those assigned to the old landing page
- Much larger variability was observed on the time spent on the landing page by the subjects assigned to the old landing page
- Some outliers were registered on both sides of the distribution of the time spent on the landing page by subjects assigned to the new landing page

Link to Appendix slide on data background check

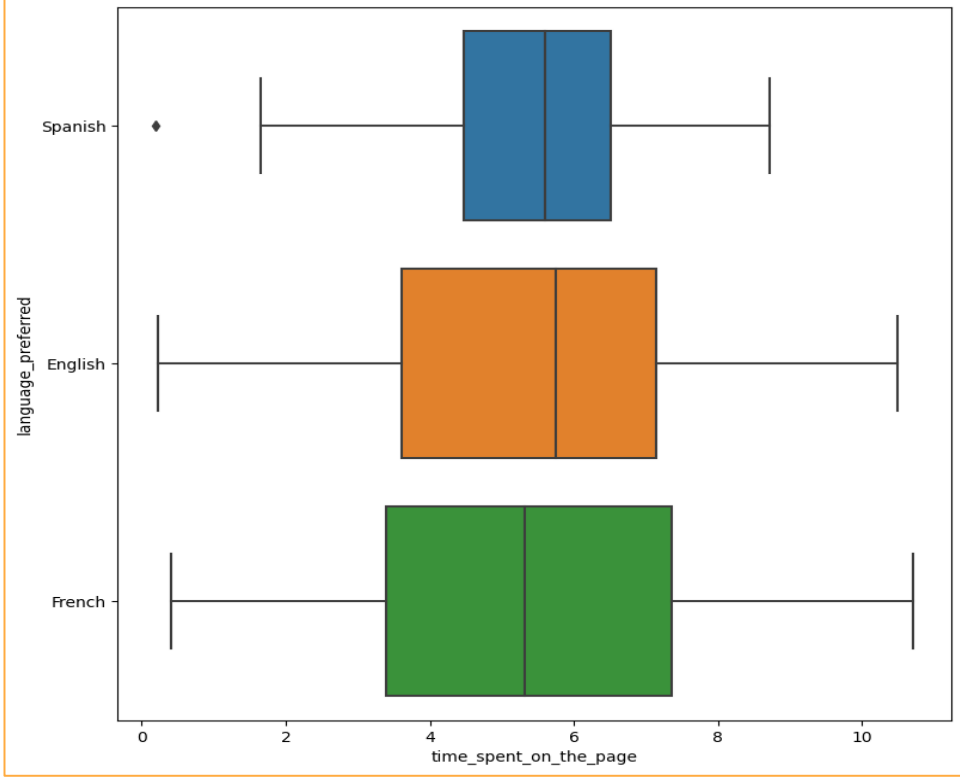
EDA Results: Bivariate Analysis: Conversion status vs Time spent on the page



- Converted subjects, in general, spent over two more minutes on the landing page than unconverted ones
- A much larger variability of the time spent on the landing page was registered for unconverted subjects
- Outliers of time spent on the landing page were registered for both converted and unconverted subjects; converted subjects appear to have many more outliers

Link to Appendix slide on data background check

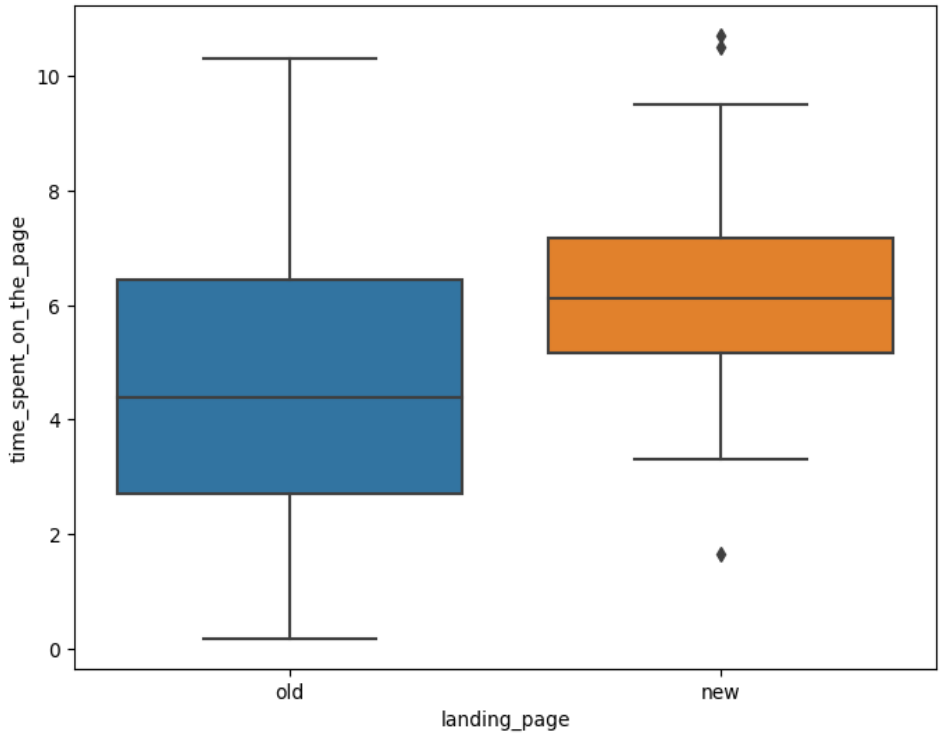
EDA Results: Bivariate Analysis: Language preferred vs Time spent on the page



- Subjects having English as preferred language spent the most time on the landing page, closely followed by those whose preferred language is Spanish; the least time spent on the landing page was registered for subjects whose preferred language is French but the differences among the time spent by subjects of the three languages are a few seconds apart
- The greatest variability of time spent on the landing page was registered for subjects whose preferred language is French while the least variability was observed for Spanish subjects
- At least an outlier was registered to the left of the distribution of time spent on the landing page by subjects whose preferred language is Spanish

Link to Appendix slide on data background check

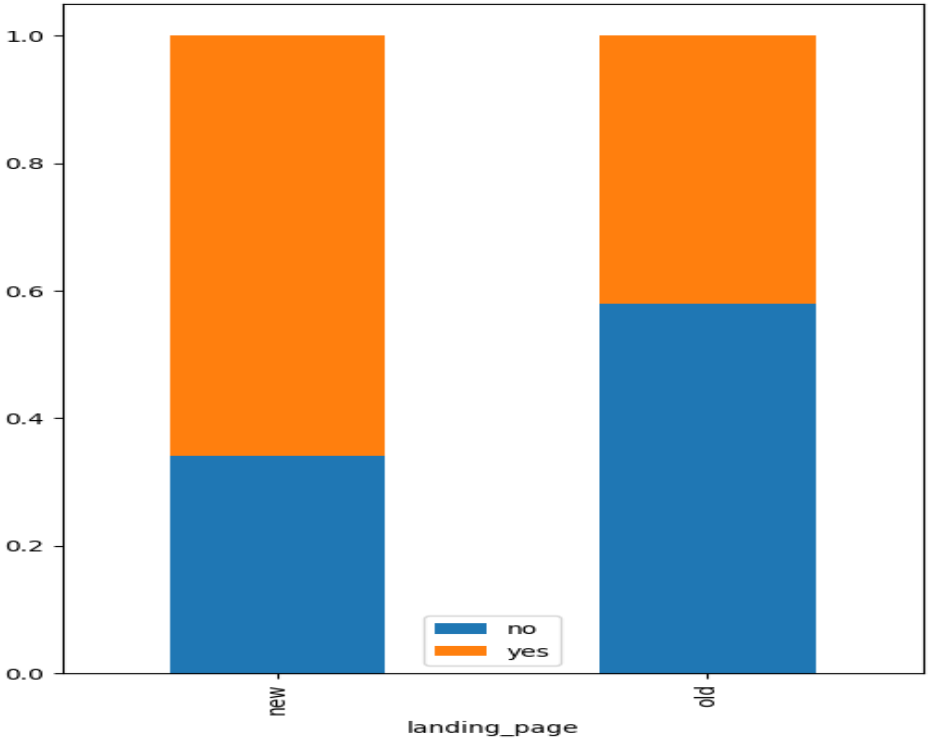
Hypotheses Tested and Results: Verify whether users spend more time on the new landing page than the existing landing page



- Subjects assigned to the new landing page tended to spend close to two more minutes on the landing page than those assigned to the old landing page
- Much larger variability was observed on the time spent on the landing page by the subjects assigned to the old landing page
- Some outliers were registered on both sides of the distribution of the time spent on the landing page by subjects assigned to the new landing page
- Since the p-value (about 0.01%) is less than the level of significance (α = 5%), we reject the null hypothesis; thus we have sufficient statistical evidence to conclude that users spend more time on the new landing page than on the existing one

Link to Appendix slide on data background check

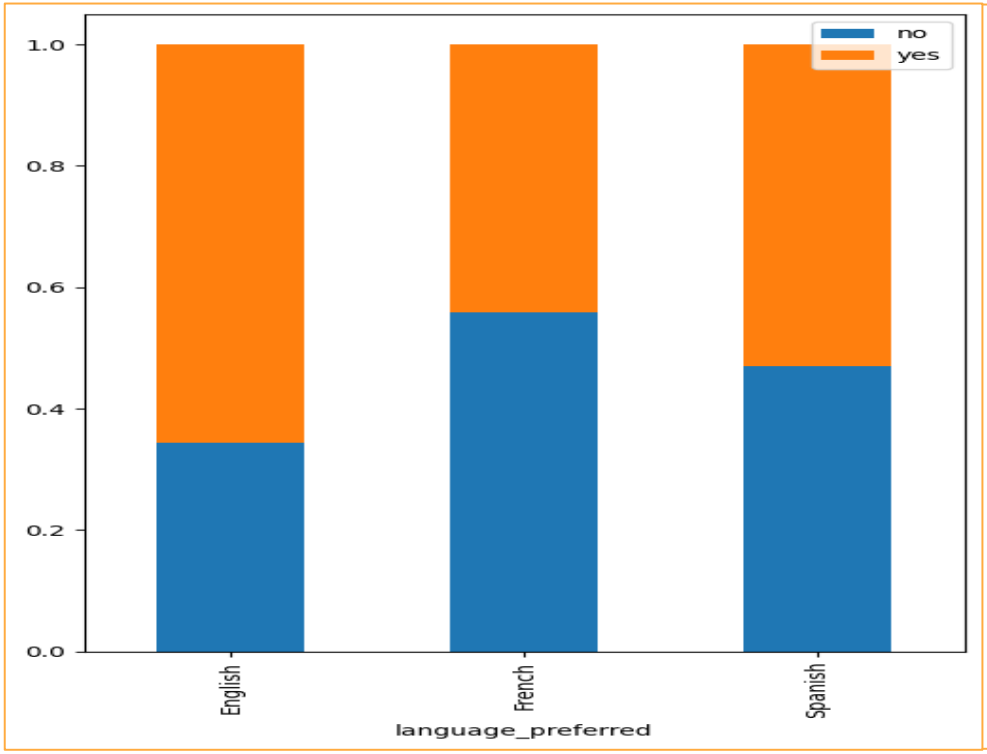
Hypotheses Tested and Results: Verify whether the conversion rate (the proportion of users who visit the landing page and get converted) for the new page is greater than the conversion rate for the old page



- From the samples, it appears that the new landing page drove conversion, a hypothesis we will be testing shortly
- Since the p-value (about 0.8%) is less than the level of significance (5%), we reject the null hypothesis; we can thus conclude that the rate of conversion for the new page is effectively larger than that for the existing page

Link to Appendix slide on data background check

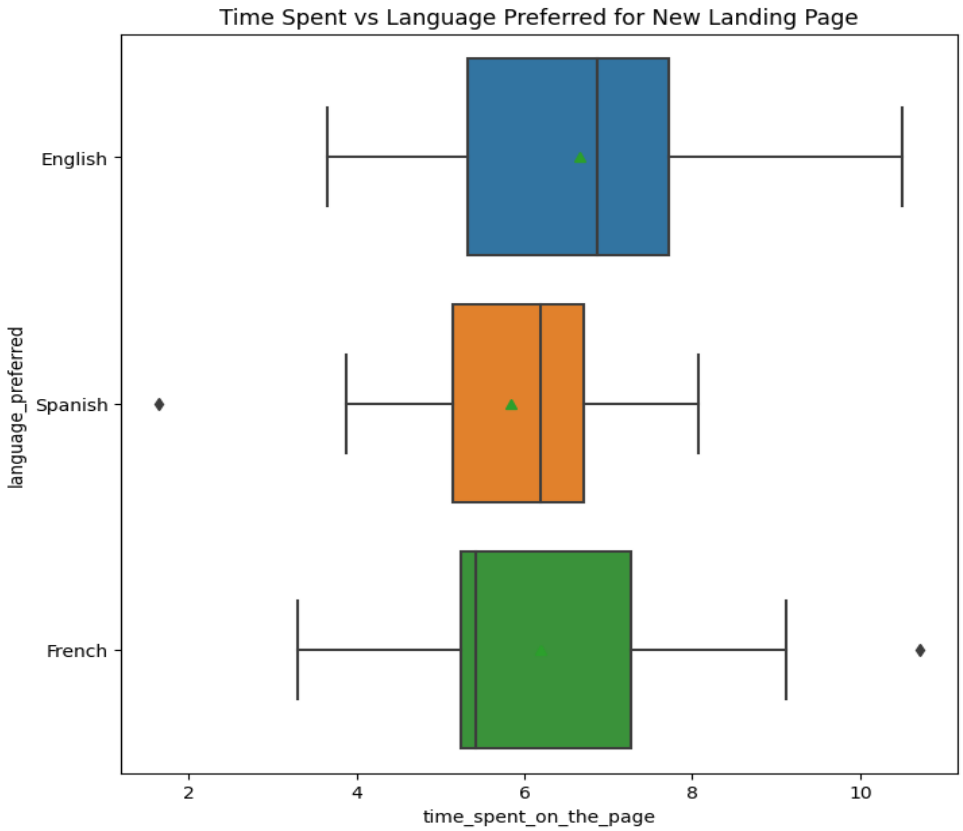
Hypotheses Tested and Results: Verify whether the converted status depend on the preferred language



- Subjects whose preferred language is English appear to register the greatest conversion rate whereas French-speaking subjects had the lowest conversion rate
- Since the p-value (about 21.3%) is greater than the level of significance (α = 5%), we fail to reject the null hypothesis. Thus, contrary to our expectation drawn from the descriptive analysis of the sample, the converted status appears to be independent of the preferred language of the user; specifically at the level of significance of 5%

Link to Appendix slide on data background check

Hypotheses Tested and Results: Verify whether the time spent on the new page same for the different language users



- Among subjects assigned to the new landing page, those whose preferred language is English spent the longest time on the landing page whereas French-speaking subjects spent the shortest time on the page
- The time spent by English and Spanish subjects is left-skewed and at least a left outlier is registered for Spanish subjects; French subjects have a right-skewed time spent distribution and at least a right outlier
- The average time time spent increases from Spanish subjects through French to English subjects; however, the values are pretty close to each other
- The time spent by English subjects has the greatest variability whereas that of Spanish subjects is the least variable
- Since the p-value (about 43.2%) is greater than the level of significance (α = 5%), we fail to reject the null hypothesis. So, in line with the observation obtained from the descriptive analysis of the sample (specifically that the means are close to each other), we can conclude that the average time spent on the new page is the same across different language users

Link to Appendix slide on data background check

APPENDIX

Data Background and Contents: Data Overview

	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

	user_id	group	landing_page	time_spent_on_the_page	converted	language_preferred
95	546446	treatment	new	5.15	no	Spanish
96	546544	control	old	6.52	yes	English
97	546472	treatment	new	7.07	yes	Spanish
98	546481	treatment	new	6.20	yes	Spanish
99	546483	treatment	new	5.86	yes	English

	group	landing_page	converted	language_preferred
count	100	100	100	100
unique	2	2	2	3
top	control	old	yes	Spanish
freq	50	50	54	34

	user_id	time_spent_on_the_page
count	100.000000	100.000000
mean	546517.000000	5.377800
std	52.295779	2.378166
min	546443.000000	0.190000
25%	546467.750000	3.880000
50%	546492.500000	5.415000
75%	546567.250000	7.022500
max	546592.000000	10.710000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   user_id                 100 non-null    int64
 1   group                   100 non-null    object
 2   landing_page            100 non-null    object
 3   time_spent_on_the_page  100 non-null    float64
 4   converted               100 non-null    object
 5   language_preferred      100 non-null    object
dtypes: float64(1), int64(1), object(4)
memory usage: 4.8+ KB
```

Data Background and Contents: Univariate and Bivariate EDA Analysis

```
control      50
treatment    50
Name: group, dtype: int64
```

```
old    50
new    50
Name: landing_page, dtype: int64
```

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

```
yes    54
no     46
Name: converted, dtype: int64
```

- Code snippets:

```
# Visual plot of time spent distribution
sns.histplot(data=df,x='time_spent_on_the_page')
plt.show()
sns.boxplot(data=df,x='time_spent_on_the_page')
plt.show()

# number of subjects in each group
df['group'].value_counts()

# visual display of groups
sns.countplot(data=df,x='group')
plt.show()

# number of subjects in assigned to each landing page
df['landing_page'].value_counts()

# visual display of landing page distribution
sns.countplot(data=df,x='landing_page')
plt.show()

# number of converted and unconverted subjects
df['converted'].value_counts()

# visual display of conversion distribution
sns.countplot(data=df,x='converted')
plt.show()

# number of subjects per preferred language
df['language_preferred'].value_counts()

# visual display of preferred language distribution
sns.countplot(data=df,x='language_preferred')
plt.show()

#Landing page vs Time spent on the page
plt.figure(figsize=(10,6))
sns.boxplot(data=df,x='landing_page',y='time_spent_on
_the_page')
plt.show()

# Conversion status vs Time spent on the page
plt.figure(figsize=(9, 9))
sns.boxplot(data = df, x = 'time_spent_on_the_page', y
= 'converted')
plt.show()

# Language preferred vs Time spent on the page
plt.figure(figsize=(9, 9))
sns.boxplot(data = df, x = 'time_spent_on_the_page', y
= 'language_preferred')
plt.show()
```

Hypothesis Testing Details: Average time spent on the new landing page greater than the existing landing page?

- Hypotheses:
  - $H0$: On average, users spend the same or less time on the new landing page than on the existing landing page
  - Ha: On average, users spend more time on the new landing page than on the existing landing page
- Hypothesis Test selected: Since we are analyzing two independent samples and the population standard deviations are unknown, it would be advisable to use the independent two-sample t-test and since the sample standard deviation of the time spent on the old page is almost 1.5 times larger than that of the time spent on the new page, it can be assumed that the corresponding population standard deviations are unequal. So we will be performing **Welch's t-test** which does not assume equal population variance
- P-value obtained: 0.01%
- Code snippets:

```
# visual analysis of the time spent on the new page and the time spent on the old page
plt.figure(figsize=(8,6))
sns.boxplot(x = 'landing_page', y = 'time_spent_on_the_page', data = df)
plt.show()

# subsetted data frame for new landing page users
time_spent_new = df[df['landing_page'] == 'new']['time_spent_on_the_page']

# subsetted data frame for old landing page users
time_spent_old = df[df['landing_page'] == 'old']['time_spent_on_the_page']

# import required function
from scipy.stats import ttest_ind

# calculation of p-value
test_stat, p_value = ttest_ind(time_spent_new, time_spent_old, equal_var = False, alternative = 'greater')
```

Hypothesis Testing Details: 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?

- Hypotheses:
  - $H0$: The rate of conversion for the new page is less than or equal to the conversion rate for the old page
  - Ha: The rate of conversion for the new page is greater than the conversion rate for the old page
- Hypothesis Test selected: Since we are analyzing two population proportions from independent populations, we will use **proportions z-test**
- P-value obtained: 0.8%
- Code snippets:

```
# visual comparison of the conversion rate for the new page and the conversion rate for the old page
pd.crosstab(df['landing_page'],df['converted'],normalize='index').plot(kind="bar", figsize=(6,8),stacked=True)
plt.legend()
plt.show()

# computation of the number of converted users in the treatment group
new_converted = df[df['group'] == 'treatment']['converted'].value_counts()['yes']

# computation of the number of converted users in the control group
old_converted = df[df['group'] == 'control']['converted'].value_counts()['yes']

n_control = df.group.value_counts()['control'] # total number of users in the control group
n_treatment = df.group.value_counts()['treatment'] # total number of users in the treatment group

# import required function
from statsmodels.stats.proportion import proportions_ztest

# computation of the p-value
test_stat, p_value = proportions_ztest ([new_converted, old_converted] , [n_treatment, n_control], alternative ='larger')
```

Hypothesis Testing Details: Converted status depends on the preferred language?

- Hypotheses:
  - $H0$: The converted status of a user is independent of their preferred language
  - Ha: The converted status of users is dependent on their preferred language
- Hypothesis Test selected: Since we are dealing with a problem of analyzing the independence of two categorical variables and the number in each cell is greater than 5, we will be using **chi-square test of independence**
- P-value obtained: 21.3%
- Code snippets:

```
# Visual plot of the dependency between conversion status and preferred langauge
pd.crosstab(df['language_preferred'],df['converted'],normalize='index').plot(kind="bar", figsize=(6,8), stacked=True)
plt.legend()
plt.show()

# creation of contingency table showing the distribution of the two categorical variables
contingency_table = pd.crosstab(df['language_preferred'], df['converted'])

# import required function
from scipy.stats import chi2_contingency

# computation of the p-value
chi2, p_value, dof, exp_freq = chi2_contingency(contingency_table)
```

Hypothesis Testing Details: Time spent on the new page same for the different language users?

- Hypotheses:
  - $H0$: The time spent on the new page is the same for the different language users?
  - Ha: The time spent on the new page is different for at least one group of language users?
- Hypothesis Test selected: Assuming that the time spent on the new page has a normal distribution and that the variances for the three language groups are the same (of course, these are not strict assumptions as observed from the samples; we are considering approximations for practical purposes here; alternatively, we could have verified these assumptions using the **Shapiro-Wilk** and **Levene's tests** respectively), we will be using the **one-way ANOVA** test to evaluate the hypothesis that the time spent on the new page is the same across different language users
- P-value obtained: 43.2%
- Code snippets:

```
# new DataFrame for users who got served the new page
df_new = df[df['landing_page'] == 'new']

# visual plot of the time spent on the new page for different language users
plt.figure(figsize=(8,8))
sns.boxplot(x = 'time_spent_on_the_page', y = 'language_preferred', showmeans = True, data = df_new)
plt.title("Time Spent vs Language Preferred for New Landing Page")
plt.show()

# computation of the mean time spent on the new page for different language users
df_new.groupby(['language_preferred'])['time_spent_on_the_page'].mean()

# subsetted data frame of the time spent on the new page by English language users
time_spent_English = df_new[df_new['language_preferred']=="English"]['time_spent_on_the_page']

# subsetted data frames of the time spent on the new page by French and Spanish language users
time_spent_French = df_new[df_new['language_preferred']== "French"]['time_spent_on_the_page']
time_spent_Spanish = df_new[df_new['language_preferred']== "Spanish"]['time_spent_on_the_page']

# import the required function
from scipy.stats import f_oneway

# computation of the p-value
test_stat, p_value = f_oneway (time_spent_English, time_spent_French, time_spent_Spanish)
```

Great Learning

Happy Learning !