E-news Express Business Case - Analysis

SARAVANAN RANGARAJAN (SR)

Contents

- Business Problem Overview
- Solution Approach
- Data Overview
- Time Spent by Various language Preference
- Time Spent by Feature & Conversion
- EDA Insights
- Do the users spend more time on the new landing page than the old landing page?
- 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?
- Does the converted status depend on the preferred language?
- Is the mean time spent on the new page same for the different language users?
- Questions

Business Problem Overview

An online news portal aims to expand its business by acquiring new subscribers. Every visitor to the website takes certain actions based on their interest. The company plans to analyze these interests and wants to determine whether a new feature will be effective or not.

Company performed A/B testing with 2 user groups and collected data across various parameters like Language Preference etc.

Analyze conversion rate among users across old and new feature and come-up with statistical justification

Solution Approach

With the Dataset collected, Exploratory Data Analysis to be performed to find some insights and try to prove them with Statistical analysis.

Following are the set of activities performed

- Understand the quality of data
- Try to mark out the data collected is good for analysis by checking the skewness, categorical variable analysis.
- Perform Univariate & Multivariate Analysis to find relation ship between various factors, so that patterns can be identified for efficient Business recommendations
- Answer statistical questions for Business problem stated.

Data Overview

- Data was collected across user group named control and treatment. It contains following fields
  - user_id - This represents the user ID of the person visiting the website.
  - group - This represents whether the user belongs to the first group (control) or the second group (treatment).
  - landing_page - This represents whether the landing page is new or old.
  - time_spent_on_the_page - This represents the time (in minutes) spent by the user on the landing page.
  - converted - This represents whether the user gets converted to a subscriber of the news portal or not.

  - language_preferred - This represents the language chosen by the user to view the landing page.
- It has 2 Numerical values & 4 Categorical values (Group, Landing Page, Converted, Language Preferred)
- Null checks have been done to check the data. No null values exist.
- Duplicate checks have been done. No Duplicate values exist.
- Check the distinct values in the Categorical columns and check the data spread for general information.

Time Spent by Various language Preference
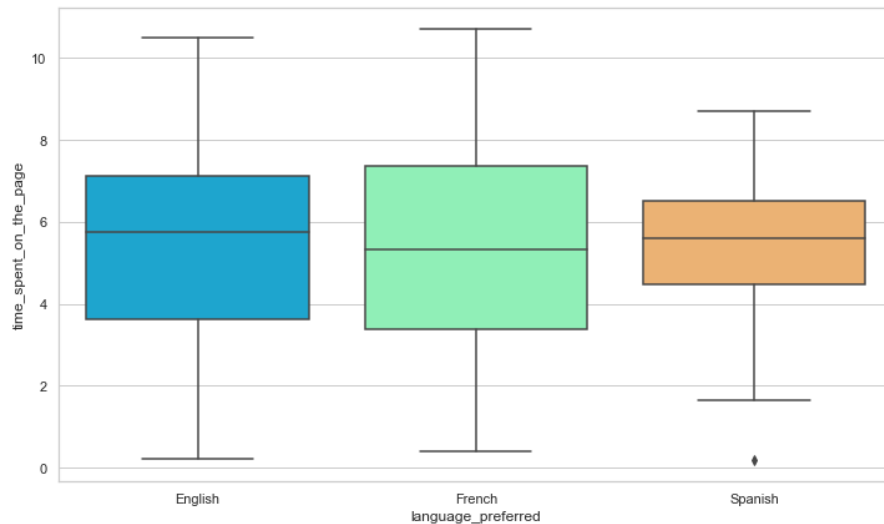
- Above analysis result provides the average time spent by users across various language is almost the same.
- One insight is that Spanish people spent less time in the site.

Time Spent by Feature & Conversion
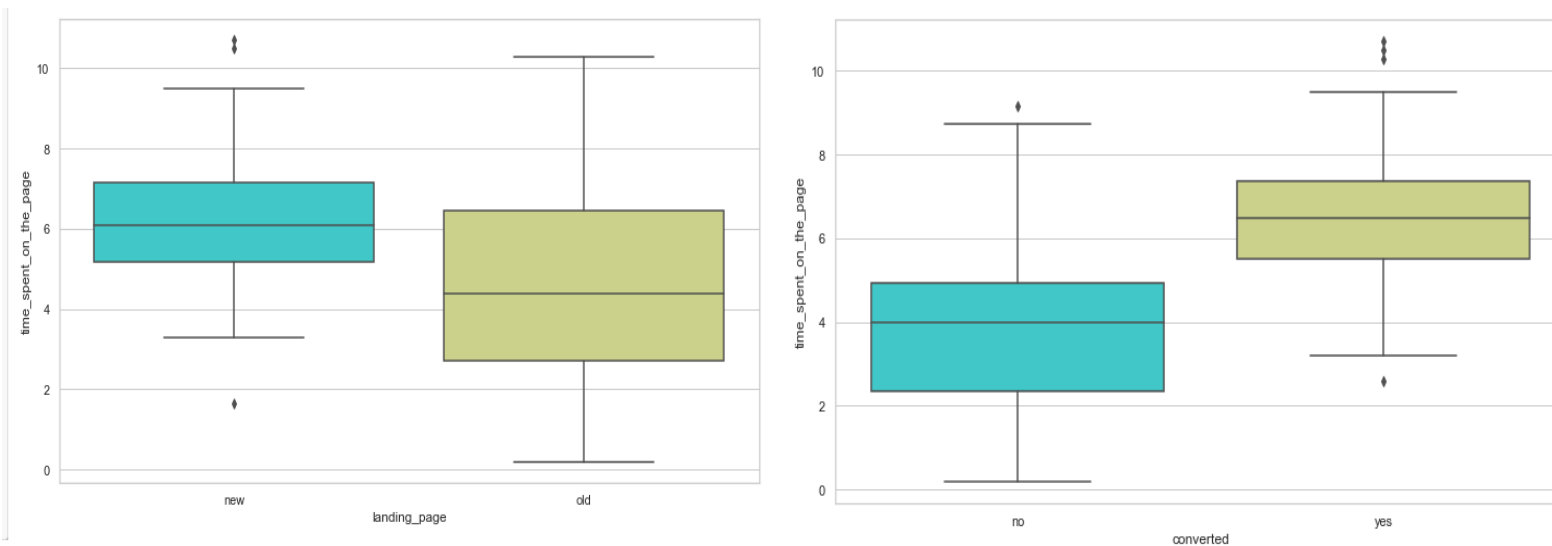
- Above analysis result provides the average time spent by users in new page has increased.
- It can also be noted that user spend more time over-all in the new design
- People who are converted spend more time in the new feature, need to improve design to improve conversion rate.

EDA Insights

- As overall, new_page design has positive impact as time spent by users have increased from initial analysis
- Language does seem to influence conversion rate
- Users who spent more time seems to get converted, that is good sign for business
- Average time spent by users across language seems to be same
- All above insights has been validated by adopting Statistical tests at significance level of 5%

Do the users spend more time on the new landing page than the old landing page?

- Performed T-test to validate the Business question
- Following Qualification about data is validated
  - Continuous data - Yes, the time spent is measured on a continuous scale.
  - Normally distributed populations - Yes, we are informed that the populations are assumed to be normal.
  - Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.
  - Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.
  - Random sampling from the population - Yes, we are informed that the collected sample a simple random sample.
- Some Statistical Figures

  Average time spent on new page :6.2232
  Average time spent on old page :4.532400000000001
  SD of new page :1.7987689568146321
  SD of old page :2.556024694716387
  Mean time spent diff is :1.6907999999999994

Formulated the Hypothesis and found P value much less that rejecting null Hypothesis of Users spend same time in Old and new thus validating the findings

# Conversion rate for new page greater than conversion rate of old page

- Performed 2 tailed Z – Test to validate the Business question
- Following Qualification about data is validated
  - Binomially distributed population - Yes, a user is either converted or not
  - Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.
  - Can the binomial distribution have approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10. Here, n and p refer to the sample size and sample proportion respectively.
  - All values are above >= 10
  - Some Statistical Figures

    Conversion rate for new users is 66%

    Conversion rate of Old users is 42%

Formulated the Hypothesis and found P value much less that rejecting null Hypothesis of Conversion rate across old and new features are same.

Does Converted status depend on Pref Language

- Performed Chi Square Independence test to check relationship between the categorical values.
- Following Qualification about data is validated
  - ➢ Categorical variables - yes
  - ➢ Expected value of the number of sample observations in each level of the variable is at least 5 - Yes
  - ➢ Random Sampling from the population - Yes.
  - ➢ Some Statistical Figures


```
language_preferred  English  French  Spanish
converted
no                       11      19       16
yes                      21      15       18
```
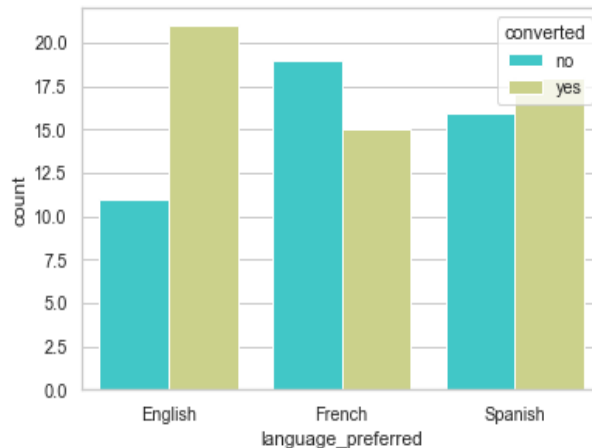
__Formulated the Hypothesis and found P value much greater that fail to reject null Hypothesis, Hence we do have enough statistical evidence to conclude that Converted status depend on Preferred language at 5% significance level__

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

- Performed one way ANOVA f-test to check on the equality of means.
- Following Qualification about data is validated
  - ➢ Perform Levene test for variance equality – Population variance are equal.
  - ➢ Perform Shapiro-Wilk's test for normality – Follows normal distribution
  - ➢ Random Sampling from the population - Yes.
  - ➢ Some Statistical Figures

    English   6.663750

    French   6.196471

    Spanish   5.835294
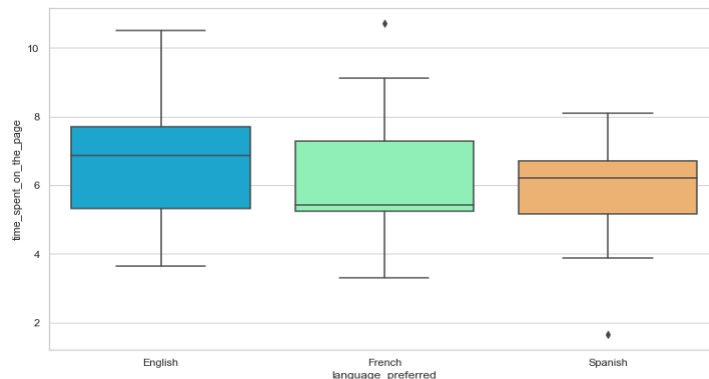
Formulated the Hypothesis and found P value much greater that fail to reject null Hypothesis, Hence we do have enough statistical evidence to conclude that the mean time spent by users on new landing page is almost same at 5% significance level.

Questions

- Questions give different Perspective, most of the time result in idea.