

E-news Express Project

Business Statistics

11-11-2011

Contents / Agenda

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

Executive Summary

- The users spend more time on the new landing page than the existing landing page.
- Subscribers are more time spent on the landing page than those not subscribers.
- The new landing page is attracting users to subscribe and spent more time on the page than the existing landing page, so the design team should give more attention to enhance the new landing page to increase users and subscribers.
- The conversion status to subscriber and the time spent on the new landing page is not affected by the preferred languages.

Business Problem Overview and Solution Approach

- **Business Problem overview**
- The E-news Express are of the opinion that there has been a decline in new monthly subscribers compared to the past year because 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.
- **Solution Approach**
- Overview of the data and data sanitary check.
- Review the patterns and trends of the sample data using EDA
- Test hypothesis

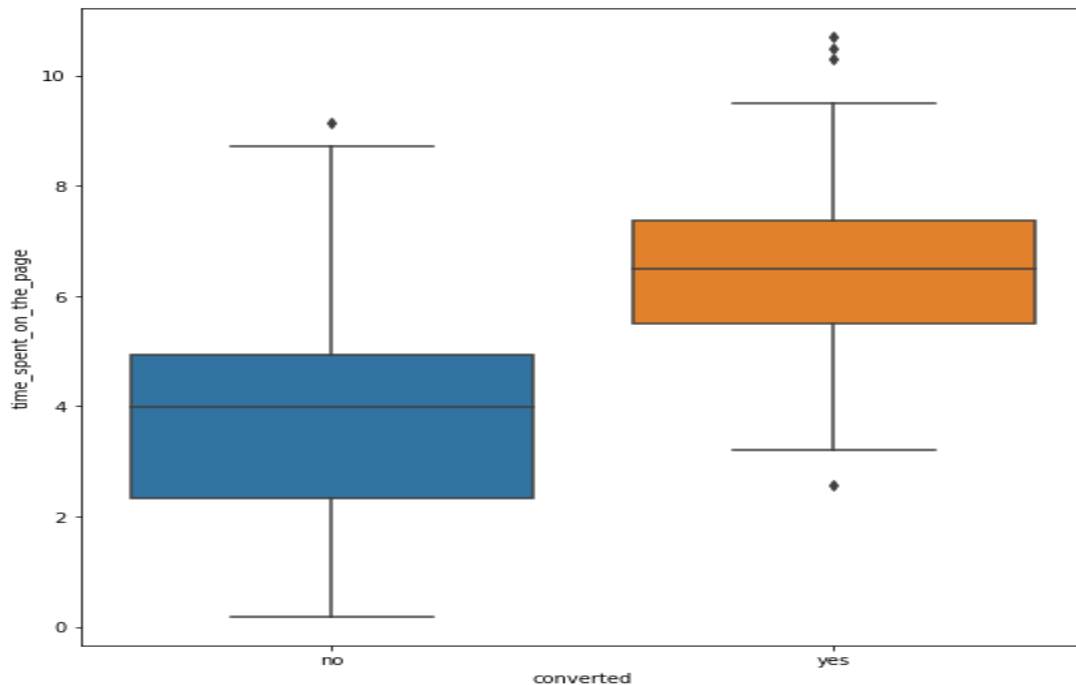
Solution Approach Continue

- Test hypothesis to answer the below questions
 - Do the users spend more time on the new landing page than on the existing 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 time spent on the new page the same for the different language users?
- Conclusions or insights from the test hypothesis
- Recommendation for the findings.

EDA Results

- The mean and median of time spent on the page are almost the same, which means the distribution is symmetric not skewed.
- The number of randomly selected users from the old and new landing page are the same, which is 50 each.
- There is no missing value and duplicate entries.
- The distribution of the time spent on the page is close to normal distribution.
- The average time spent on the new page is higher than the old page.
- Those users who converted to subscribers are more time spent on the page than those not converted to subscriber.
- The average time spent by language preferred is almost the same for all the languages.

EDA Result



- The average time spent on the page for the converted customers is higher than by more than 2Hrs.

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

- From the sample data, the users spend more time on the new landing page than the existing Landing page.
- As the assumptions of 2 sample t test are satisfied, I can select the 2 independent sample T-test for equality of means- unequal standard deviation for this problem.
- As the p-value (0.00014) is less than the level of significance, we can reject the null hypothesis. Hence, we do have enough evidence to support the claim that the users spend more time on the new landing page than the existing landing page.

Hypothesis Testing

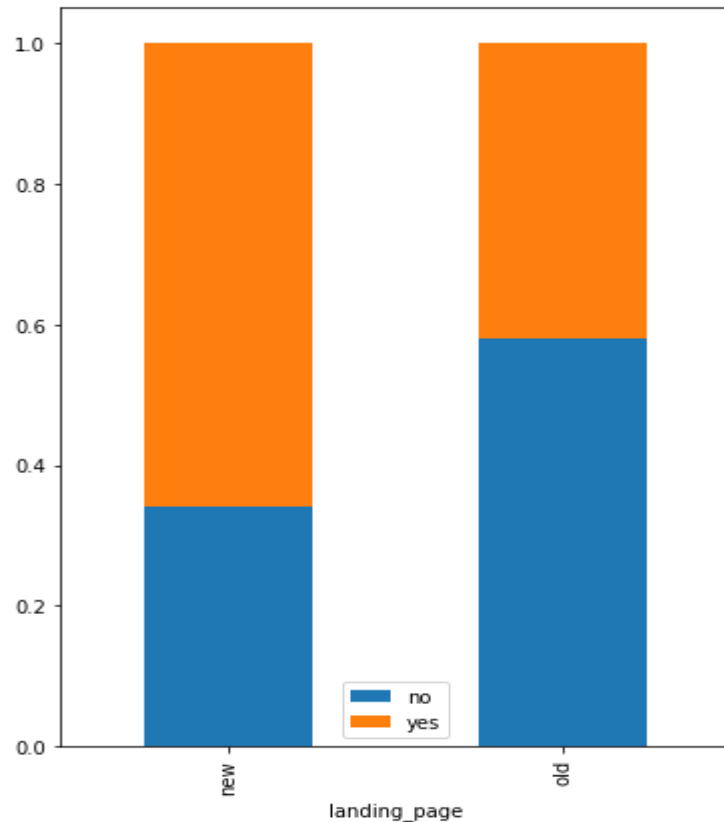
- **Step -1 : Define the null and alternative hypothesis**
- Let μ_1 , μ_2 be the average time spent on the new landing page and existing landing page.
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 > \mu_2$
- **Step -2 : Select Appropriate test**
- This is a one-tailed test concerning two population means from two independent populations
- The time spent on the page is measured on a continuous scale
- The populations are assumed to be normal
- As the sample standard deviations are different, the population standard deviations may be assumed to be different.
- we are informed that the collected sample a simple random sample.

Hypothesis Testing

- Based on the above information, we can select 2 independent sample T-test for equality of means-unequal standard deviation for this problem.
- **Step 3: Decide the significance level**
- As given in the problem statement, we select $\alpha = 0.05$
- **Step 4: Calculate the p-value**
- P-Value = 0.00014
- **Step 5: Compare P-value with α and Draw inference**
- As the p-value (~ 0.00014) is less than the level of significance, we can reject the null hypothesis. Hence, we do have enough evidence to support the claim that the users spend more time on the new landing page than the existing landing page.

Is the conversion rate for the new page greater than the conversion rate for the old page?

- From the graph, we observe that the conversion rate to subscriber on the new page is higher than on the old page.
- As the assumptions of 2 proportion Z test are satisfied, I can select 2 proportion Z test for this problem.
- As the p-value (0.008) is less than the significance level 0.05, we can reject the null hypothesis. Hence, we do have enough evidence to support the claim that 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.



Hypothesis Testing

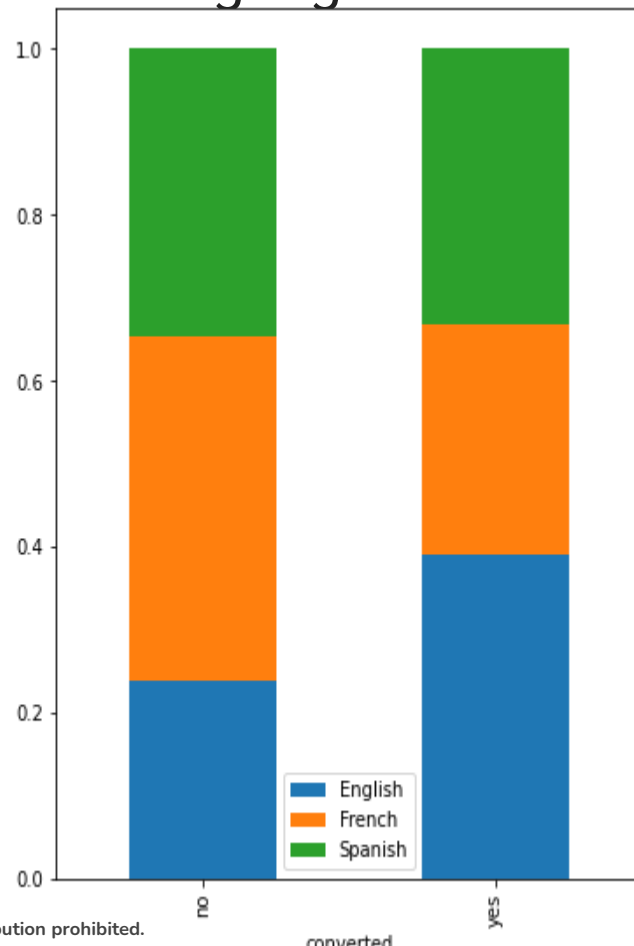
- **Step -1 : Define the null and alternative hypothesis**
- Let p_1 , p_2 be the proportion of users who visit the landing page and get converted to subscriber for the new page and old page.
- $H_0 : p_1 = p_2$
- $H_1 : p_1 > p_2$
- **Step -2 : Select Appropriate test**
- This is a one-tailed test concerning two population proportions from two independent populations
- As the user is either converted to subscriber or not, it's a binomially distributed population.
- we are informed that the collected sample a simple random sample.
- As both the mean (np) and $n(1-p)$ are greater than or equal to 10, the binomial distribution can be approximated by a normal distribution.

Hypothesis Testing

- Based on the above information, we can select 2 proportion Z test for this problem.
- **Step 3: Decide the significance level**
- As given in the problem statement, we select $\alpha = 0.05$
- **Step 4: Calculate the p-value**
- P-Value = 0.008
- **Step 5: Compare P-value with α and Draw inference**
- As the p-value (0.008) is less than the significance level 0.05, we can reject the null hypothesis. Hence, we do have enough evidence to support the claim that 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?

- From the graph, we observe that most of English language users are converted to subscriber.
- As the assumptions of Chi-Square test of independence are satisfied, I can select Chi-Square test of independence for this problem.
- As the p-value (0.213) is greater than the 5% level of significance, we fail to reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that the converted status depend on the preferred language.



Hypothesis Testing

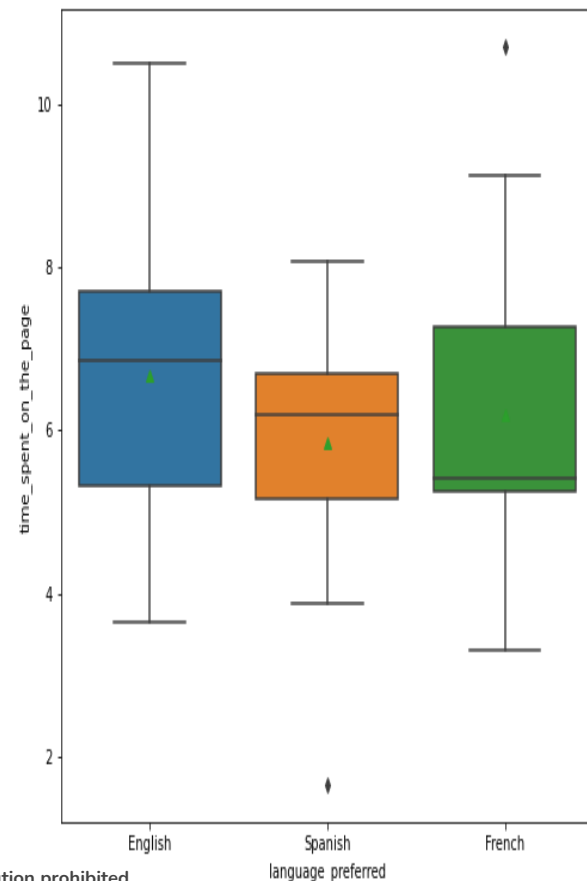
- **Step -1 : write the null and alternative hypothesis**
- H_0 : The Conversion status is independent of preferred language
- H_1 : The Conversion status depends on the preferred language
- **Step -2 : Select Appropriate test**
- This is a problem of the test of independence, concerning two categorical variables - converted status and preferred language.
- The number of observations in each level is greater than 5.
- we are informed that the collected sample a simple random sample.

Hypothesis Testing

- Based on the above information, we can select Chi-Square test of independence for this problem.
- **Step 3: Decide the significance level**
- As given in the problem statement, we select $\alpha = 0.05$
- **Step 4: Calculate the p-value**
- P-Value = 0.213
- **Step 5: Compare P-value with α and Draw inference**
- As the p-value (0.213) is greater than the 5% level of significance, we fail to reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that the converted status depend on the preferred language.

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

- From the graph, we observe that the average time spent on the new page is different for the different language users.
- This is a problem, concerning three population means and as the assumptions of normal distribution and the homogeneity of variance are satisfied, I can select one-way ANOVA test for this problem.
- As the p-value (0.432) is greater than the significance level, we fail to reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that the average time spent on the new page for the different language users are different.



Hypothesis Testing

- **Step -1 : write the null and alternative hypothesis**
- Let μ_e , μ_f , μ_s be the means of the time spent on the new page for English, French and Spanish language users respectively.
- $H_0 : \mu_e = \mu_f = \mu_s$
- H_1 : At least one of the mean time spent on the new page with respect to the three language user is different.
- **Step -2 : Select Appropriate test**
- This is a problem, concerning three population means.
- As the p-value of Shapiro-wiki's test of normality is 0.804, the time spent on the new page follows normal distribution.
- As the p-value of Levene's test of equality of variance is 0.467, the population variances are equal.
- Based on the above information, we can select one-way ANOVA test for this problem.

Hypothesis Testing

- **Step 3: Decide the significance level**
- As given in the problem statement, we select $\alpha = 0.05$
- **Step 4: Calculate the p-value**
- P-Value = 0.432
- **Step 5: Compare P-value with α and Draw inference**
- As the p-value (0.432) is greater than the significance level, we fail to reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that the average time spent on the new page for the different language users are different.

Conclusions

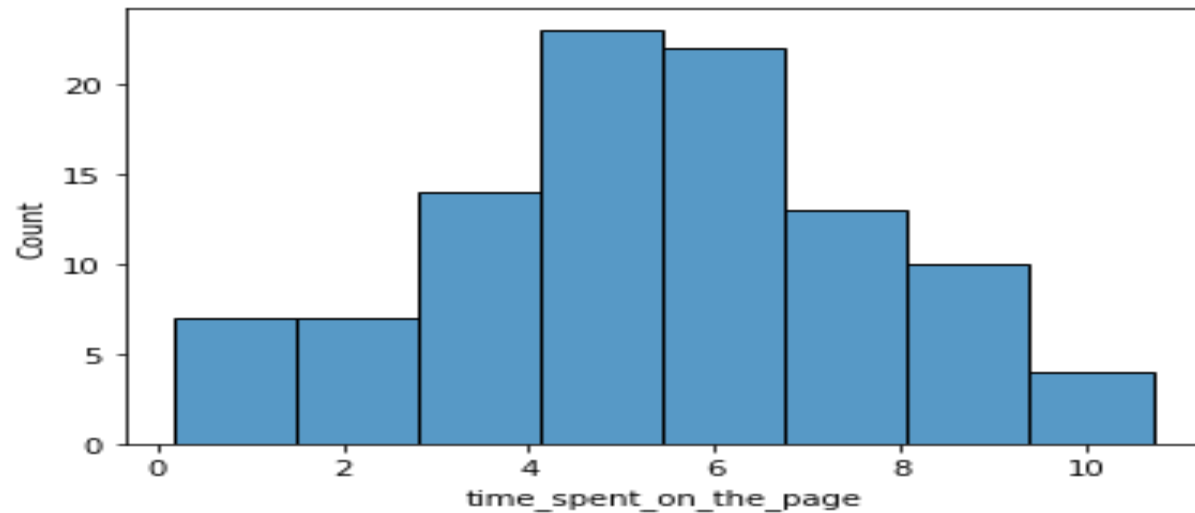
- The users spend more time on the new landing page than the existing landing page.
- Subscribers are more time spent on the landing page than those not subscribers.
- The new landing page is attracting users to subscribe and spent more time on the page than the existing landing page.
- The conversion status to subscriber is not depend on the preferred languages.
- We do not have enough statistical significance to conclude that the average time spent on the new page for the different language users are different.

Recommendations

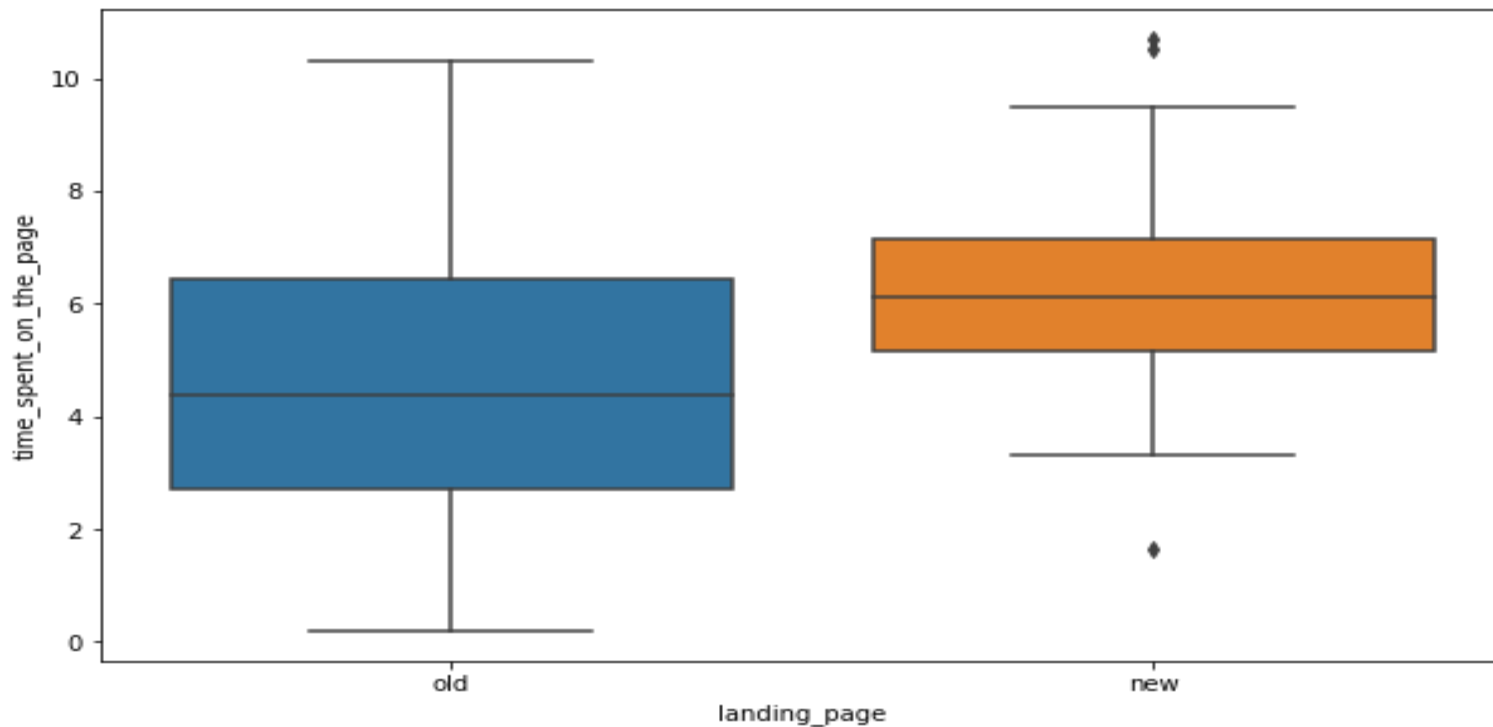
- As the new landing page is attracting users to subscribe it and spent more time, the design team should give more attention and enhance the new landing page to increase the number of users and to spent more time on the page.

APPENDIX

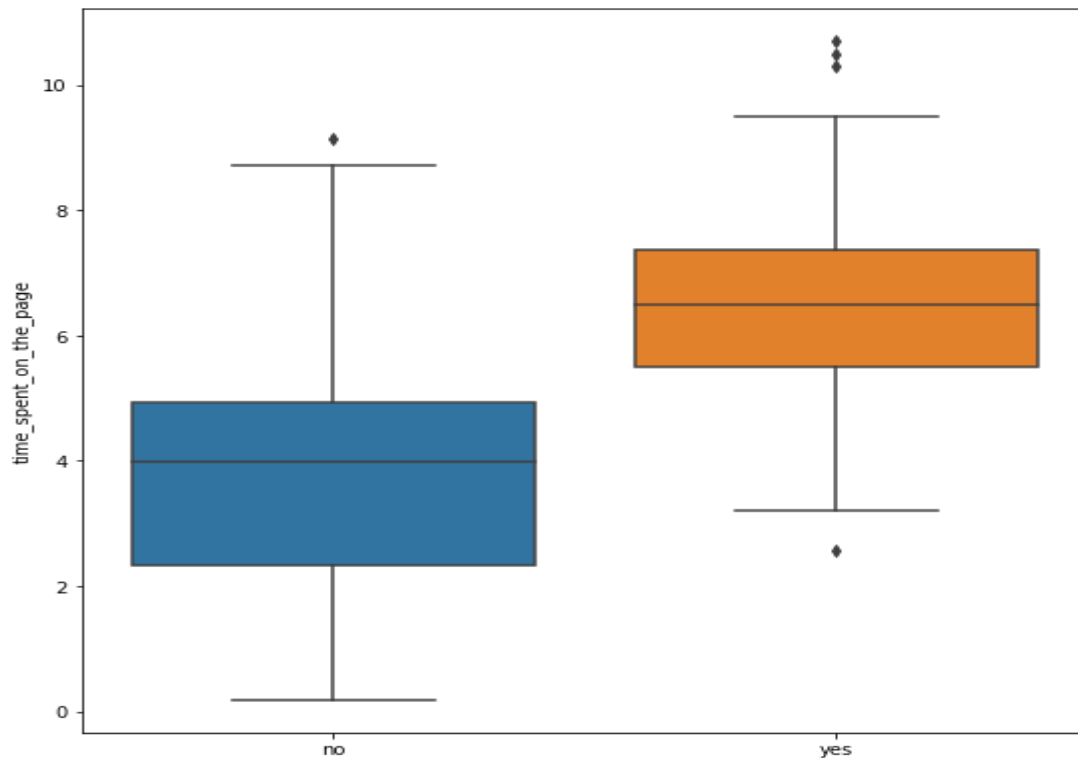
Histogram of Time Spent on the page



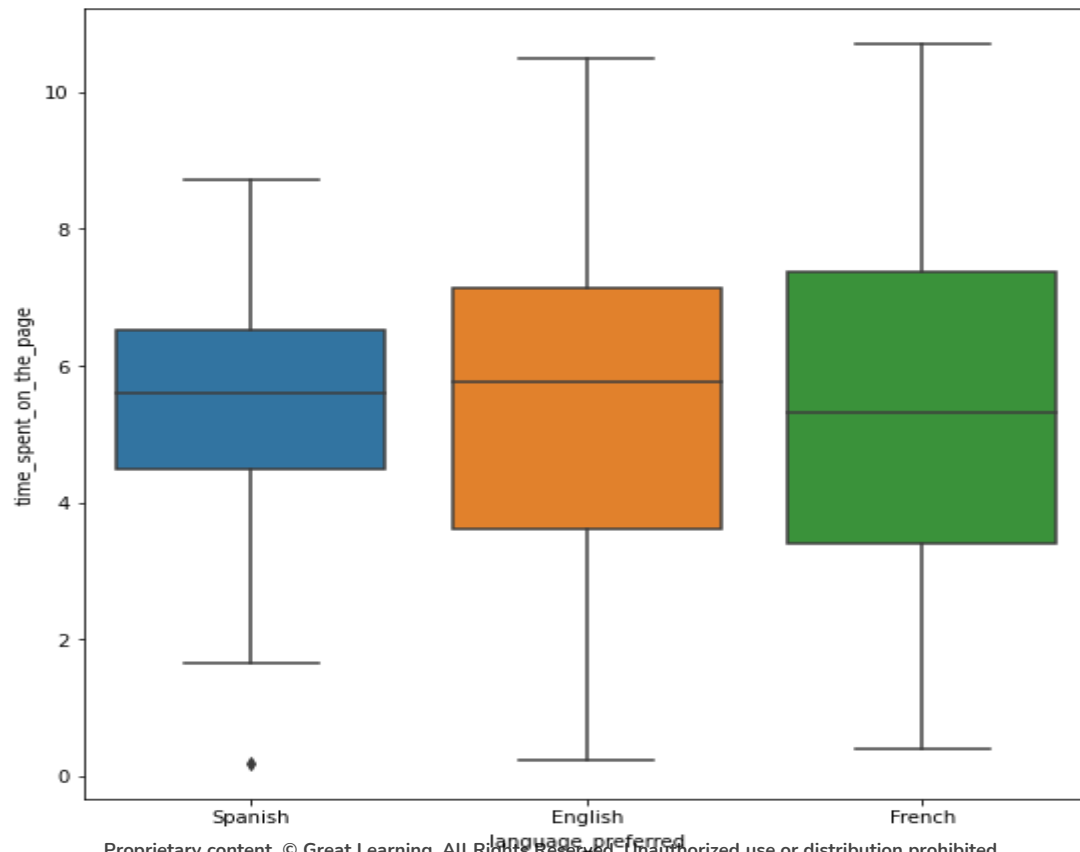
Box plot of Landing page Vs Time spent on the page



Box plot of Conversion status Vs Time spent on the page



Box plot of Language preferred Vs Time spent on the page



Checking the assumption of 2 proportion Z test

- $np_1 = 50 * (33/50) = 33 \geq 10$
- $n(1-p_1) = 50 * (17/50) = 17 \geq 10$
- $np_2 = 50 * (21/50) = 21 \geq 10$
- $n(1-p_2) = 50 * (29/50) = 29 \geq 10$
- As both the mean (np) and $n(1-p)$ are greater than or equal to 10, the binomial distribution can be approximated by a normal distribution.

Checking Assumptions of One-Way ANOVA test

- **Shapiro-Wiki's test of normality**
- H_0 : The time spent on the new page follows a normal distribution against
- H_a : The time spent on the new page does not follow a normal distribution
- P-value = 0.804
- Since p-value of the test is very large, we fail to reject the null hypothesis that the response follows the normal distribution.
- **Levene's test of equality of variance**
- H_0 : All the population variances are equal
- H_a : At least one variance is different from the rest
- P-value = 0.467
- Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances.

- Both assumptions of one-way ANOVA test are satisfied.

Data Background and Contents

- The data has 100 rows and 6 columns. Which means we have randomly selected 100 users and we have 6 variables.
- There is 2 numeric (User-id & time spent on the page) and 4 object type variables on our data.
- There E-news is published by 3 languages (English, French and Spanish)
- We have equal number of users selected from both control and treatment groups.



Happy Learning !

