INFERENTIAL STATISTICS

GUIDED PROJECT

LIST OF CONTENT

LIST OF TABLES

LIST OF FIGURES

Problem Statement - IS Project - Guided

Problem 1

Q1.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?

Answer 1.1: Following probability is given

(a) The probability of radiation leak by fire (P(L/F)) is 20% or 0.2

(b) The probability of radiation leak by a mechanical failure (P(L/M)) is 50% or 0.5

(c) The probability of radiation leak by human error (P(L/H)) is 10% or 0.1

(d) The probability of radiation leak by fire (P (F ∩ L)) is 0.1% or 0.001

(e) The probability of radiation leak by mechanical failure (P (M ∩ L)) is 0.15% or 0.0015

(f) The probability of radiation leak by human error (P (H ∩ L)) is 0.12% or 0.0012

1.1 **Probability of a fire:**

P (F ∩ L) = P(F)*P(L/F)

0.001 = P(F)* 0.2

P(F) = 0.001/0.2 = **0.005**

P (M ∩ L) = P(M)*P(L/M)

0.0015 = P(M)*0.5

P(M) = 0.0015/0.5 = **0.003**

$P(H \cap L) = P(H)*P(L/H)$

$20.0012 = P(H)*0.1$

$P(H) = 0.0012/0.1 =$ **0.012**

Q1.2) What is the probability of a radiation leak?

Answer1.2: Probability of a radiation leak

$P(L) = P(F \cap L) + P(M \cap L) + P(H \cap L)$

$P(L) = 0.001 + 0.0015 + 0.0012$

= **0.0037**

So, the probability of a radiation leak is 0.0037

Q1.3) Suppose there has been a radiation leak in the reactor for which the definite cause is not known. What is the probability that it has been caused by:

a) a fire?

b) a mechanical failure?

c) a human error?

Answer 1.3) Probability given by radiation leak:

Using Bays's Theorem,

P(A/B) = P(B/A) *P(A)/P(B)

(a) **Probability that radiation leak was caused by fire**

P(F/L) = P(L/F) *P(F)/P(L)

P(F/L) = 0.2*0.005/0.0037

= 0.001/0.0037

= **0.270** (Probability that radiation leak was caused by fire)

(b) **Probability that radiation leak was caused by mechanical failure:**

P(M/L) = P(L/M) *P(M)

P(M/L) = 0.5*0.003/0.0037

= 0.0015/0.0037

= **0.405** (Probability that radiation leak was caused by mechanical failure)

(c) **Probability that radiation leak was caused by human error:**

P(H/L) = P(L/H) *P(H)/P(L)

P(H/L) = 0.1*0.012/0.0037

= 0.0012/0.0037

=**0.324** (Probability that radiation leak was caused by human error)

PROBLEM 2

Q2.1 What is the probability that a randomly chosen student gets a grade below 85 on this exam?

Answer 2.1: Probability of grade below 85

Firstly, we need to calculate z-score for 85

$Z = X-\mu/ \sigma$

X = 85

Z = 85 – 77/8.5 = 8/8.5

= 0.941

So, probability that random chosen student gets grade below 85:

P (X < 85) = **0.8264**

Q2.2 What is the probability that a randomly selected student score between 65 and 87?

Answer 2.2: Probability of grade between 65 & 87:

Here we need to calculate z-score of both 65 & 87

For X = 65

Z = 65-77/8.5 = -12/8.5

= -1.412

For X = 87

Z = 87-77/8.5 = 10/8.5

= 1.176

For Z = -1.412, the C.F is approx. 0.0797

For Z = 1.176, the C.F is approx. 0.8808

The probability that a student scores between 87 & 65 is difference between these 2 probabilities:

P(65<X<87) = P(X<87)-P(X<65)

= 0.8808- 0.0797

= **0.8011**

Q2.3 What should be the passing cut-off so that 75% of the students clear the exam?

Answer 2.3: Passing cut off for 75% of the students

Firstly, we will find grade corresponding to 25^{th} percentile which is approx. -0.674

Z = X-μ/ σ = -0.674

= X-77/8.5

X = 77+ (-0.674*8.5)

X = 77- 5.729

X= **71.27**

So, the passing cut off should be approx. **71.27** for students to clear the exam.

Problem 3

DATA DESCRIPTION & It's DICTIONARY

E- News Express, an online news portal contains 6 attributes regarding interaction of users in both groups with the two versions of the landing page.

DATA OVERVIEW:

1. user_id - Unique user ID of the person visiting the website
2. group - Whether the user belongs to the first group (control) or the second group (treatment)
3. landing_page - Whether the landing page is new or old
4. time_spent_on_the_page - Time (in minutes) spent by the user on the landing page

5. converted - Whether the user gets converted to a subscriber of the news portal or not
6. language_preferred - Language chosen by the user to view the landing page

After importing all the necessary libraries, and loading the data set, we will explore the dataset and extract insights using Exploratory Data Analysis.

Checking the shape of the data set

- There are 100 rows and 6 columns.

Checking the data types of the column of the data set

- There are 2 numerical data types and 4 categorical features.

Checking for null values

- There are no missing values in the data set.

Checking for duplicates

- There is no duplicate row.

Time Spent on the Page:

- Mean: 5.38 minutes
- Standard Deviation: 2.38 minutes
- Minimum: 0.19 minutes
- Maximum: 10.71 minutes
- 25th Percentile: 3.88 minutes
- 50th Percentile (Median): 5.42 minutes
- 75th Percentile: 7.02 minutes

UNIVARIATE ANALYSIS:

- HISTOGRAM: TIME SPENT ON THE PAGE

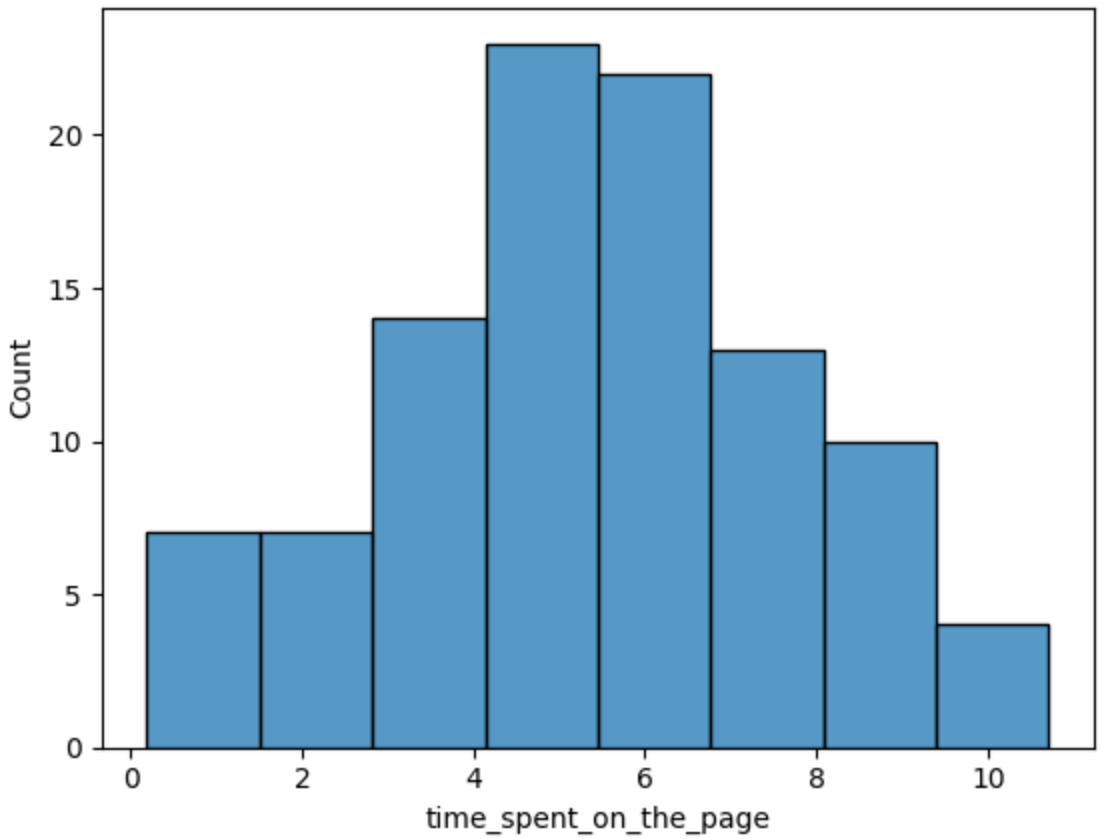

FIGURE 1

- Mostly the time spent on the page is between 4- 6 minutes, the highest being 5 minutes. This shows that most users spend around 4 – 6 minutes on the page.
- The distribution appears to be symmetric, showing normal distribution, suggesting that time spent on the page is distributed evenly around the central value.
- The time spent ranges from 0 – 10 minutes, showing differences in user engagement.
- Most users lie between the middle range of the time spent showing few users who spend very less time (0-2 minutes) and high time (8 – 10 minutes).

- The peak can be observed around 5 minutes, showing that this is the common duration for engagement of users.

➢ BOXPLOT: TIME SPENT ON THE PAGE

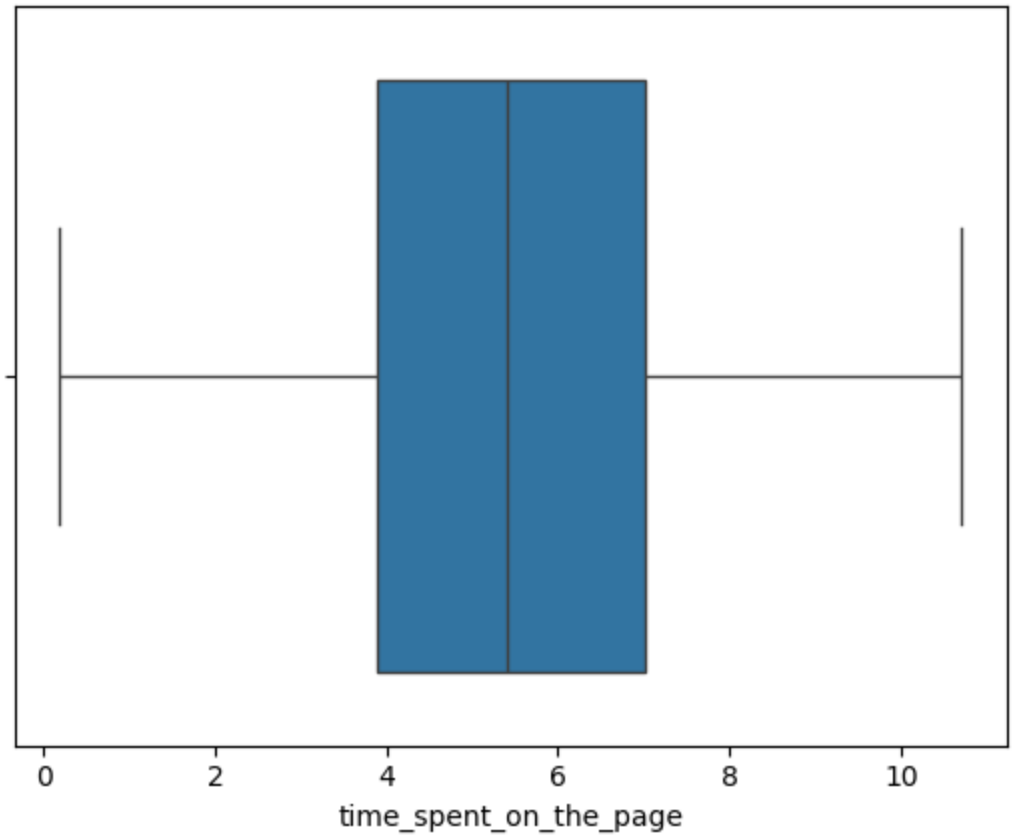

FIGURE 2

- The median time spent on the page is around 5 minutes.
- There is no outlier present.
- The data is centrally distributed in the boxplot.
- The whisker ranges from 1 to 10.

➢ GROUP

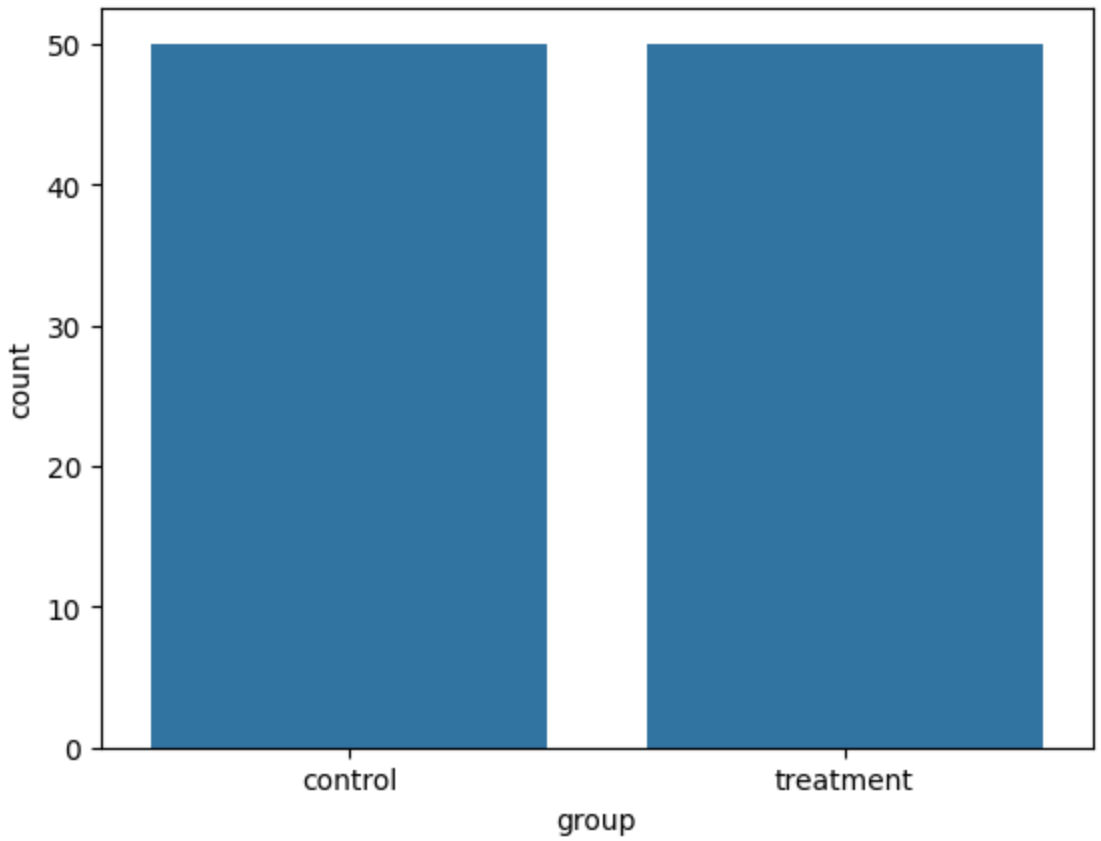

FIGURE 3

- It can be observed from the above histogram that the groups – Control and Treatment have equal counts i.e. 50-50 each respectively.
- The equal number of participants shows that the comparison between the two groups is fair.
- As the group size is equal, any difference observed in the outcome can be confidently attributed to the treatment effect.

➢ BARPLOT: LANDING PAGE

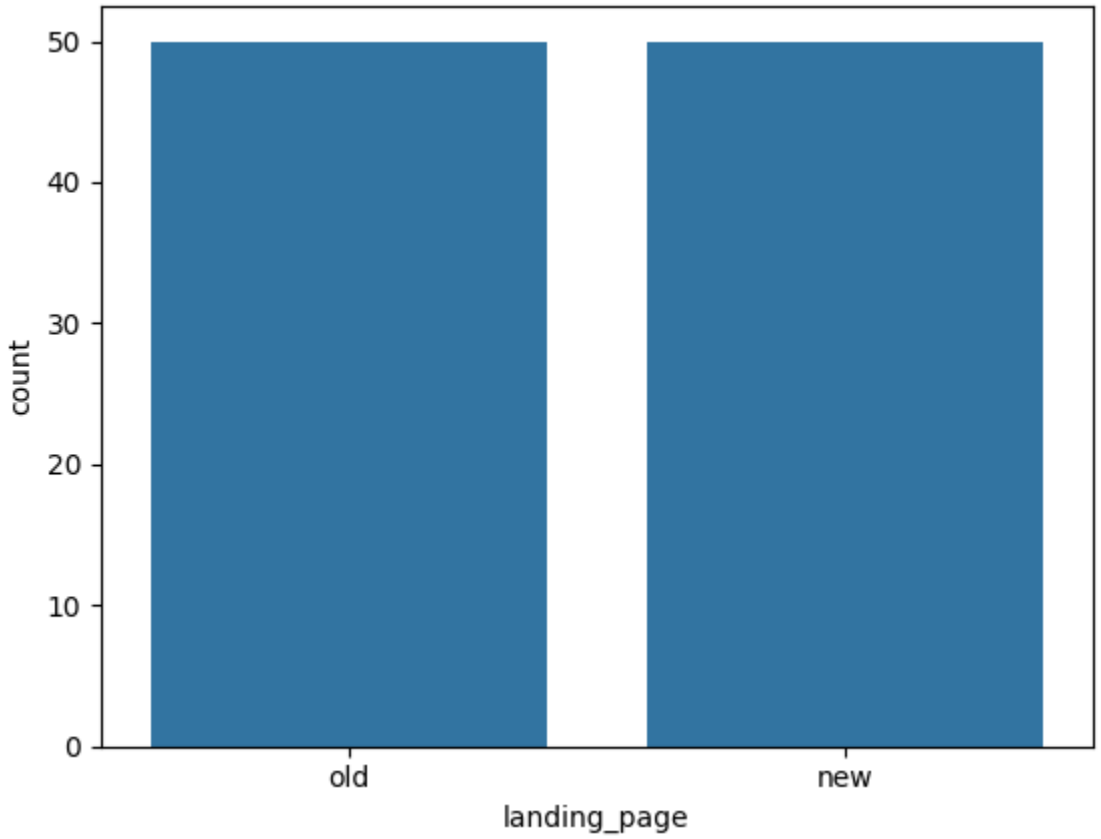

FIGURE 4

- The bar plot shows the distribution of users between the old and new landing pages.
- There are equal number of users showing balanced design.

➢ BAR PLOT: CONVERTED

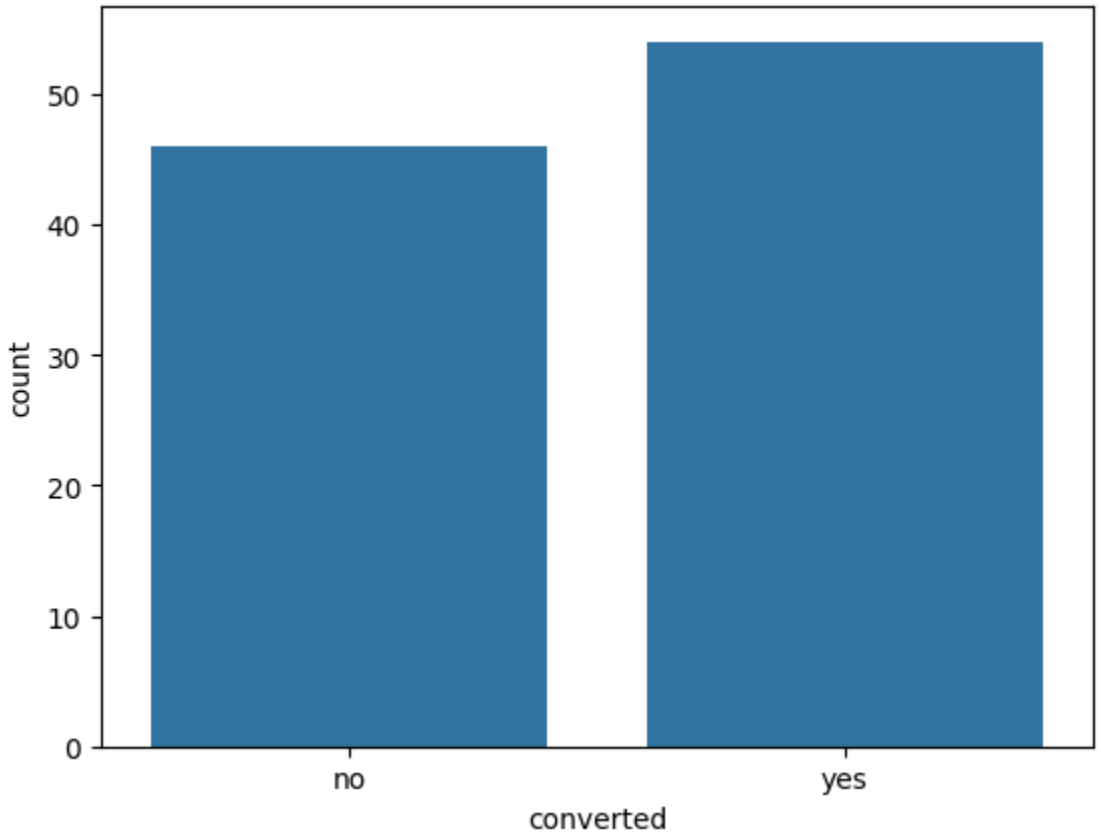

FIGURE 5

- It can be observed from the above figure that the subscribers who got converted are 54. The subscribers who did not turn up are 46 in number.
- There are more users who converted ‘Yes’ as compared to those who did not convert ‘No’.
- This shows that the conversion rate is high.

➢ BAR PLOT: LANGUAGE PREFERRED

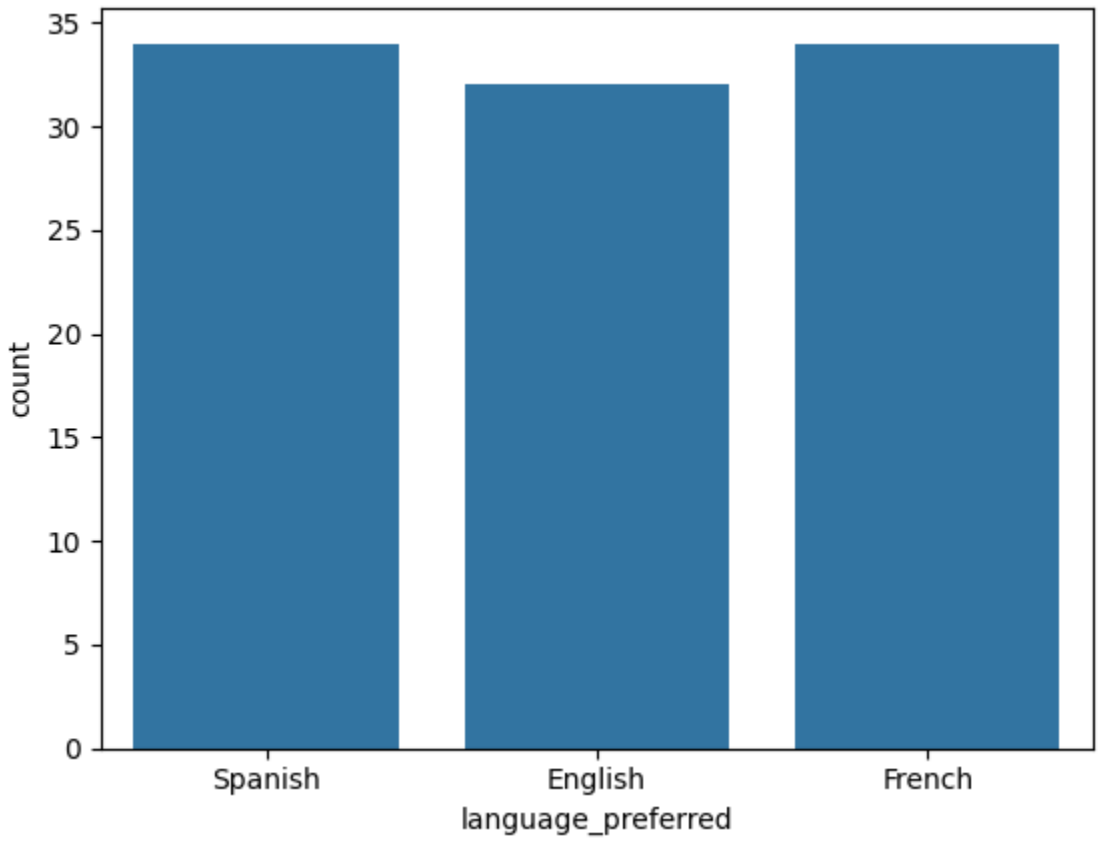

FIGURE 6

- Count of users (Language): **Spanish 34 users, French 31 users & English 33 users.**
- The counts are similar across all three languages showing an even distribution of language preference among users.

BIVARIATE ANALYSIS

➢ LANDING PAGE vs TIME SPENT ON THE PAGE

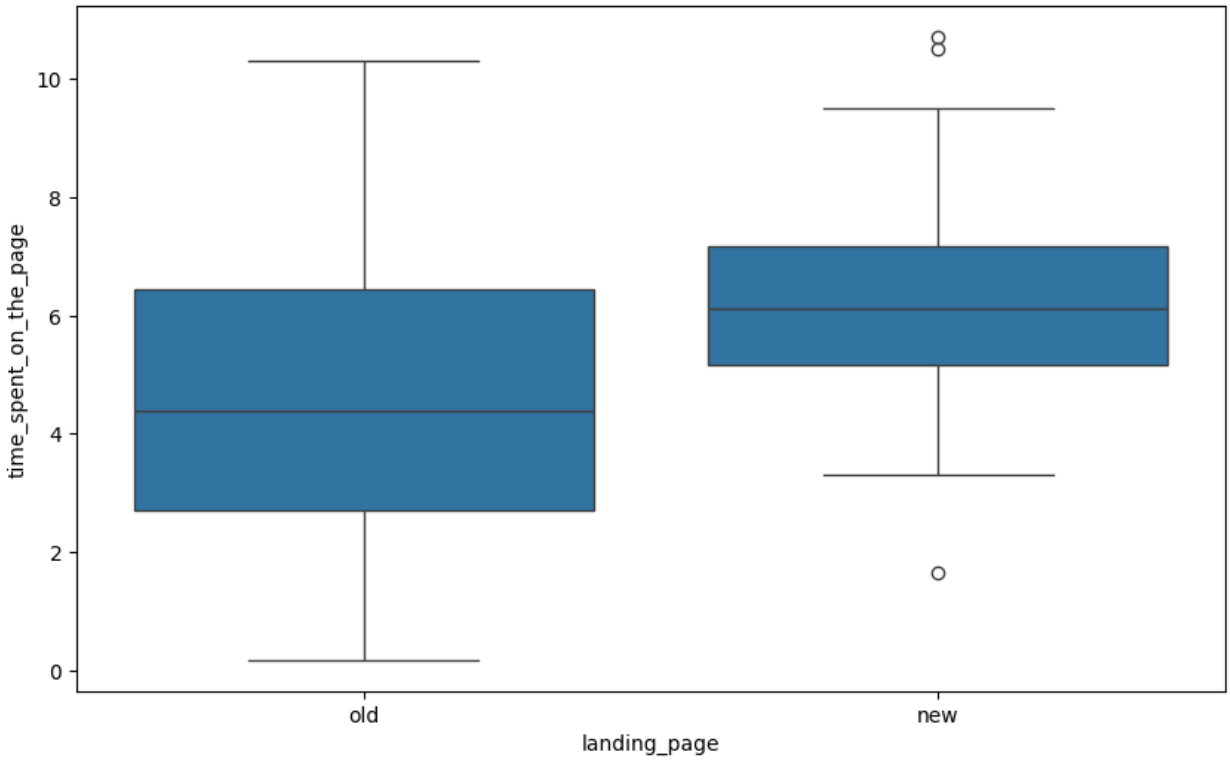

FIGURE 7

- It can be observed that the time spent on the page by subscribers on the ‘old’ landing page is less than compared to the time spent on the new landing page suggesting that user spend more time on the ‘new’ landing page indicating better engagement.
- Time spent on the page by users on old landing page is around **2.5 - 6.5 minutes.**

- There are some outliers present on the landing page ‘new’, which means that while most users spend time, some find the content really engaging.
- The distribution of both the old and new landing page is slightly right skewed.

➢ CONVERSION STATUS vs TIME SPENT ON THE PAGE

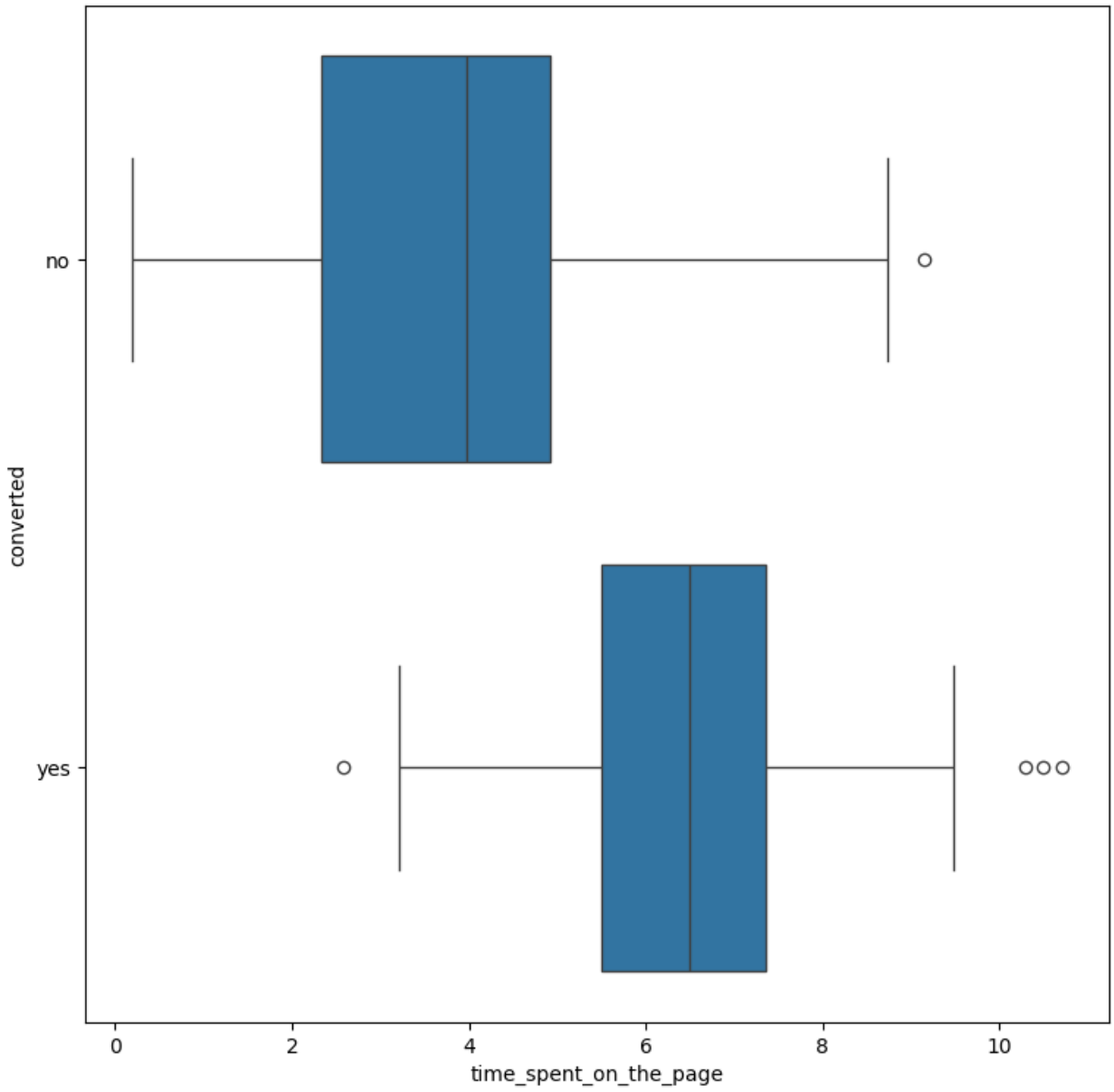

FIGURE 8

- The median time of non-converted users is around 6 whereas median time of converted users is around 5.
- The IQR of non-converted users ranges from approx. 4 – 8.

- The whisker of non-converted user extends from 1 to 10 on the other hand the whisker of converted user extends from 2 to 9.
- There are few outliers on the lower end of the non-converted users whereas few outliers exist on the higher end of the converted users.

➢ LANGUAGE PREFFERED vs TIME SPENT ON THE PAGE

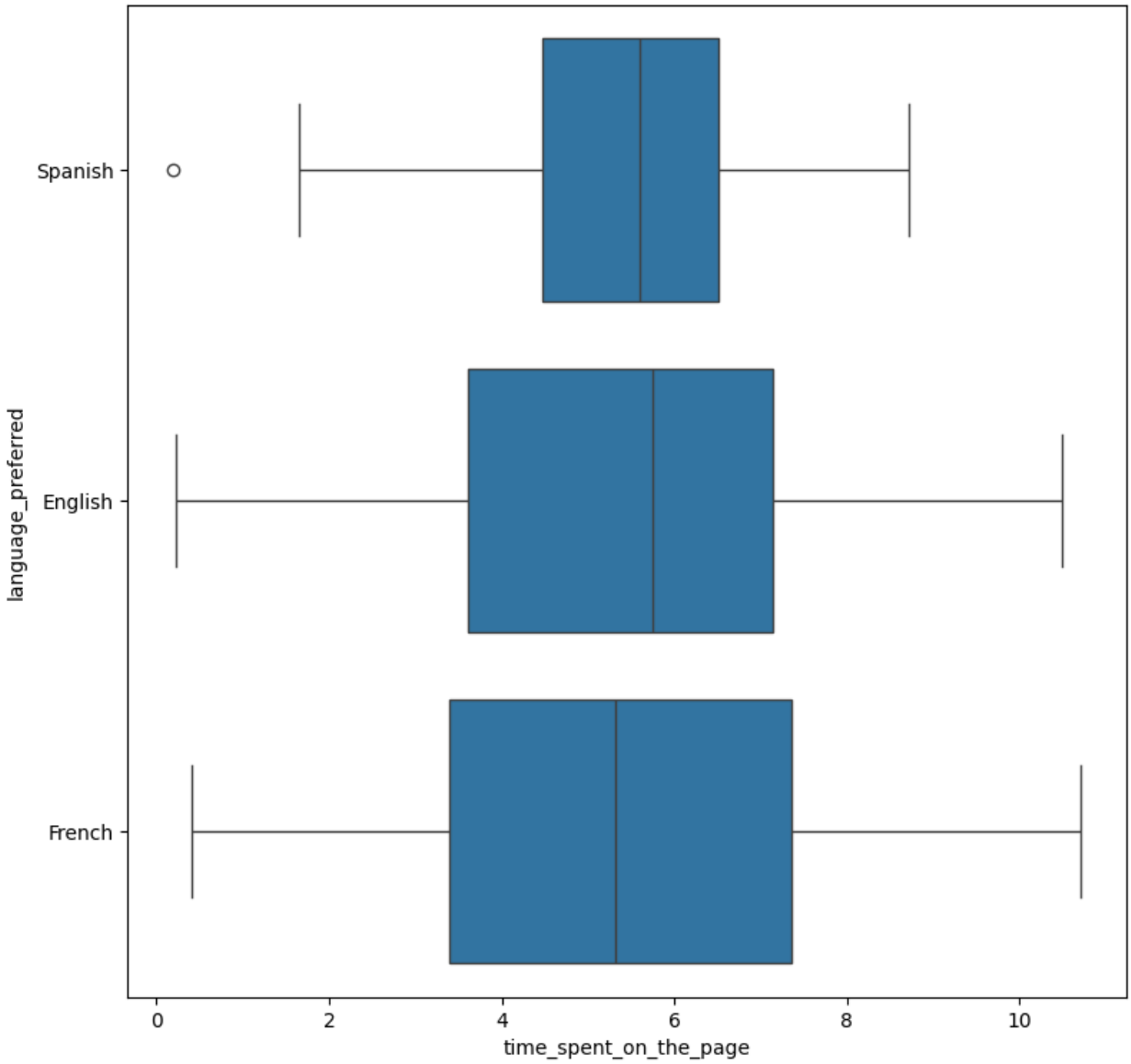

FIGURE 9

- From the above boxplot, we can see that mostly the people who spend time on the page are French than English.
- The least time spent on the page is by the Spanish users.
- The distribution of Spanish users is slightly left skewed whereas English plot column has no skewness.
- The French user's distribution is slightly right side skewed.

- Though French language is mostly preferred by the users, but it won't be wrong to conclude that **5 – 7 minutes** is an average engagement of all the users on the page.

❑ QUESTIONS AND ANSWERS (As per Rubrics)

Q1: Do the users spend more time on the new page than the existing landing page?

Answer: Yes, the users spend more time on the new page as compared to the existing landing page.

PERFORM VISUAL ANALYSIS

➢ LANDING PAGE vs TIME SPENT ON THE PAGE

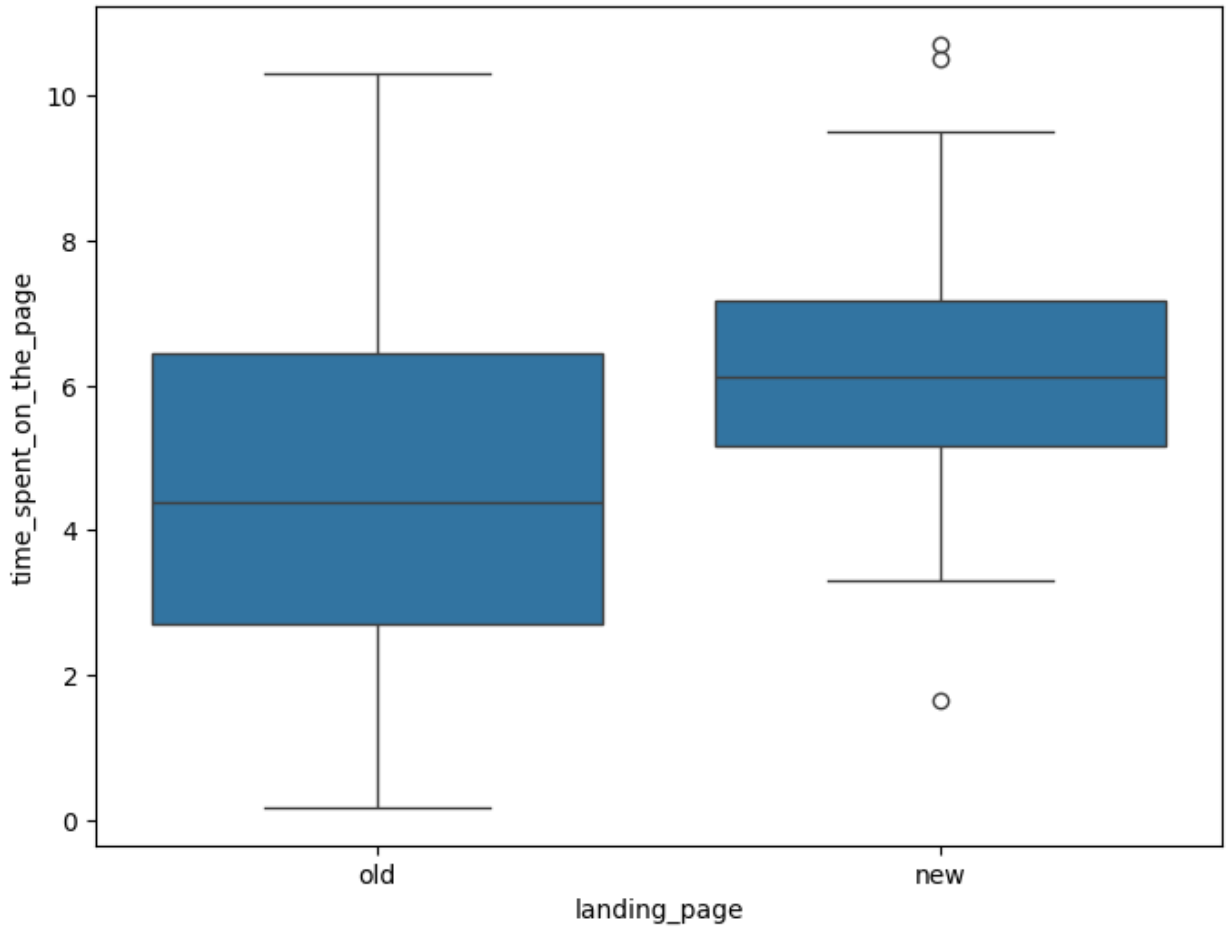

FIGURE 10

- We can observe from the boxplot the users spend more time on the new landing page, i.e. approx. **5-7 minutes,** indicating users engagement for long
- On the other hand, users on the old landing page spend **3-6 minutes.**
- There are some outliers present on the new user landing page distribution, showing that users find new design more engaging.

STEP1: DEFINE NULL AND ALTERNATE HYPOTHESIS

To analyze whether users spend more time on the new landing page compared to the existing landing page, we can formulate the following null and alternative hypotheses:

Null Hypothesis: There is no difference in the mean time spent on the new landing page compared to the existing landing page.

Alternative Hypothesis: The mean time spent on the new landing page is greater than the mean time spent on the existing landing page.

STEP 2: SELECT APPROPRIATE TEST

We'll use a one-tailed independent samples **t-test** to compare the mean time spent on the two landing pages.

STEP 3: DECIDE SIGNIFICANCE LEVEL

Significance level $\alpha = 0.05$.

STEP 4: COLLECT AND PREPARE DATA

- The sample standard deviation of the time spent on the new page is: **1.82 minutes**

- The sample standard deviation of the time spent on the old page is: **2.58 minutes**

Based on the sample standard deviations of the two groups, we can say that the population standard deviations can be assumed to be **unequal.**

STEP 5: CALCULATE THE p-value

- The p- value is: **0.00013923812251665 49**

STEP 6: COMPARE the p- value with α

- **t- statistic: -3.79**
- **P- value: 0.000139**
- The p- value is less than 0.05, so we can reject the null hypothesis.
- This suggests clearly that there is a difference in the time spent on the new landing page as compared to the old landing page.
- The negative t- statistics show that the users spend more time on the new landing page.

STEP 7: DRAW INFERENCE:

- As there is a significant difference in the time spent between the two landing pages, we can reject the null hypothesis.
- Users spend more time on the new landing page as compared to the old landing page.

Q2: 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?

Answer: Yes, there is evidence to suggest that the conversion rate for the new landing page is greater than the conversion rate for the old landing page.

PERFORM VISUAL ANALYSIS

- CONVERSION RATE OLD PAGE vs NEW PAGE

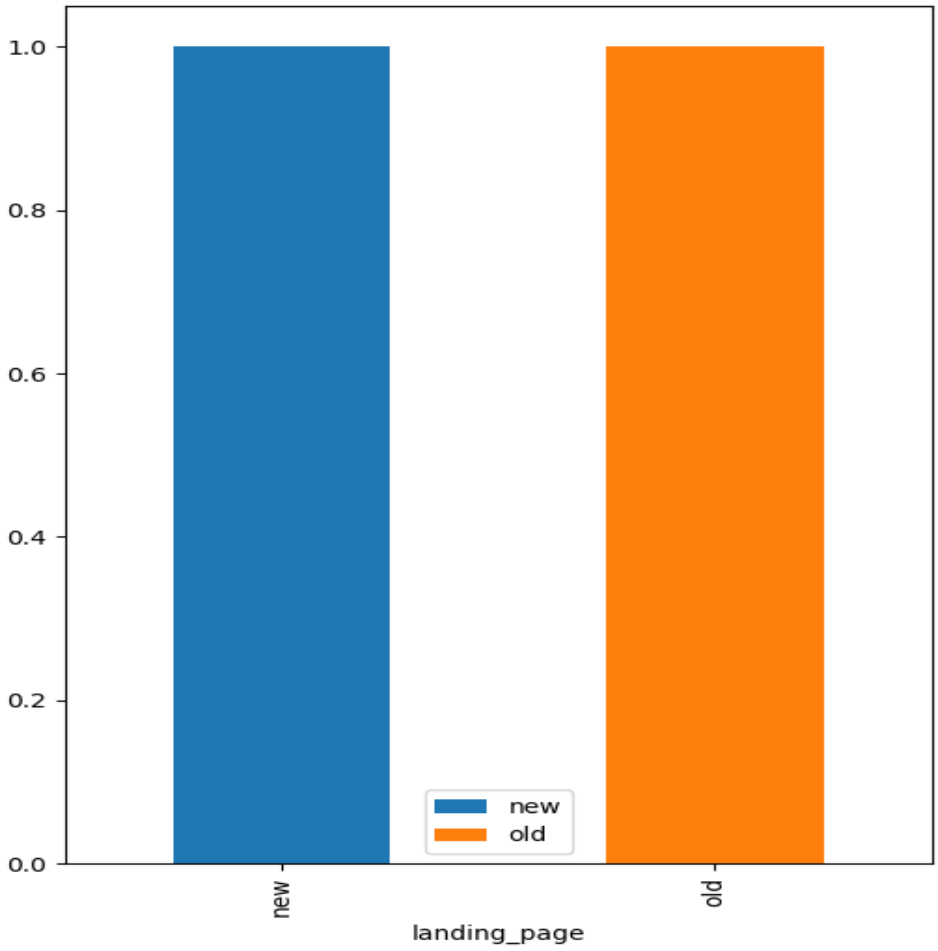

FIGURE 11

- From the above histogram we can observe that the conversion rate of both the old and new landing page is equal.

STEP 1: DEFINE NULL AND ALTERNATE HYPOTHESIS

Null Hypothesis H_0:
The conversion rate for the new landing page is equal to or less than the conversion rate for the old landing page.

Alternative Hypothesis H_a:
The conversion rate for the new landing page is greater than the conversion rate for the old landing page.

STEP 2: SELECT APPROPRIATE TEST

- We'll use a one-tailed hypothesis test to compare the proportion of users converted between the old and new landing pages.
- We'll perform a **one-tailed z-test** for two independent proportions.

STEP 3: DECIDE SIGNIFICANCE LEVEL

Set the significance level α to 0.5

STEP 4: COLLECT AND PREPARE DATA

- The number of users served the new and old pages are 50-50 respectively.

STEP 5: CALCULATE THE p- value

- The p-value is (0.0, 1.0)

STEP 6: COMPARE THE p- value with α

- The p-value is 1.0
- As the p-value 1.0 is greater than the level of significance, we fail to reject the null hypothesis.

STEP 7: DRAW INFERENCE:

- In this test, the p-value is greater than the significance level (0.05), we fail to reject the null hypothesis and conclude that there is evidence to suggest that the conversion rate for the new landing page is greater than the conversion rate for the old landing page.

Q3: Does the converted status depend on the preferred language?

PERFORM VISUAL ANALYSIS

- DEPENDANCY BETWEEN CONVERSION STATUS & PREFERRED LANGUAGE

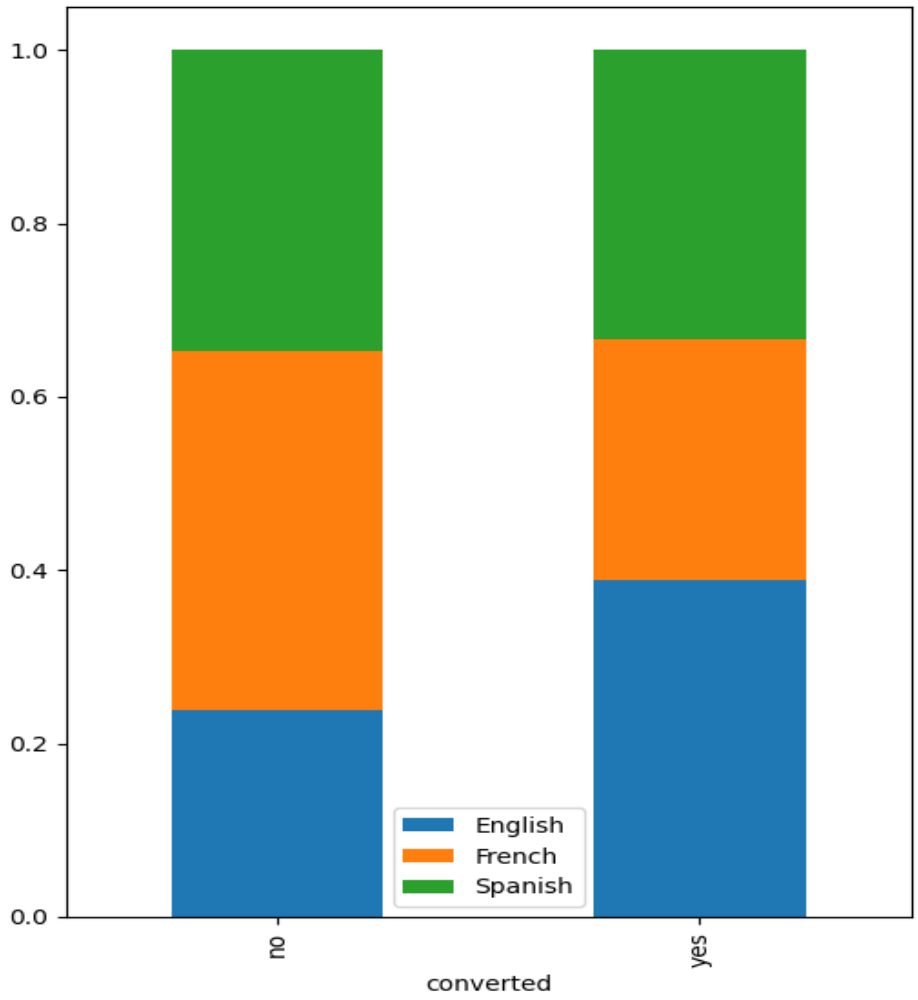

FIGURE 12

- Among users who converted English was mostly preferred as those who did not convert.
- Mostly the users preferred Spanish who did not get converted.
- Spanish speakers have the lowest conversion rate.
- English speakers have the highest conversion rate.

STEP 1: DEFINE NULL AND ALTERNATE HYPOTHESIS

- Null Hypothesis: There is no association between conversion status and preferred language.
- Alternative Hypothesis: There is an association between conversion status and preferred language.

STEP 2: SELECT APPROPRIATE TEST

- We'll use the **Chi-Square test of independence** because it's suitable for analyzing the relationship between two categorical variables.

STEP 3: DECIDE SIGNIFICANCE LEVEL

- As given in the problem statement, we select $\alpha = 0.05$

STEP 4: COLLECT AND PREPARE DATA

Language preferred	**English**	**French**	**Spanish**
Converted			
No	11	19	16
Yes	21	15	18

TABLE 1

STEP 5: CALCULATE THE p- value

- The p- value is 0.21298887487543447

STEP 6: COMPARE THE p- value with α

- As the p- value 0.21298887487543447 is greater than the level of significance, we fail to reject the null hypothesis.

STEP 7: DRAW INFERENCE:

- As the p-value is greater than the significance level, we fail to reject the null hypothesis, indicating there is no evidence of an association between conversion status and preferred language.

Q4: Is the time spent on the new page the same for the different language users?

PERFORM VISUAL ANALYSIS

➢ TIME SPENT ON THE PAGE vs LANGUAGE PREFERRED

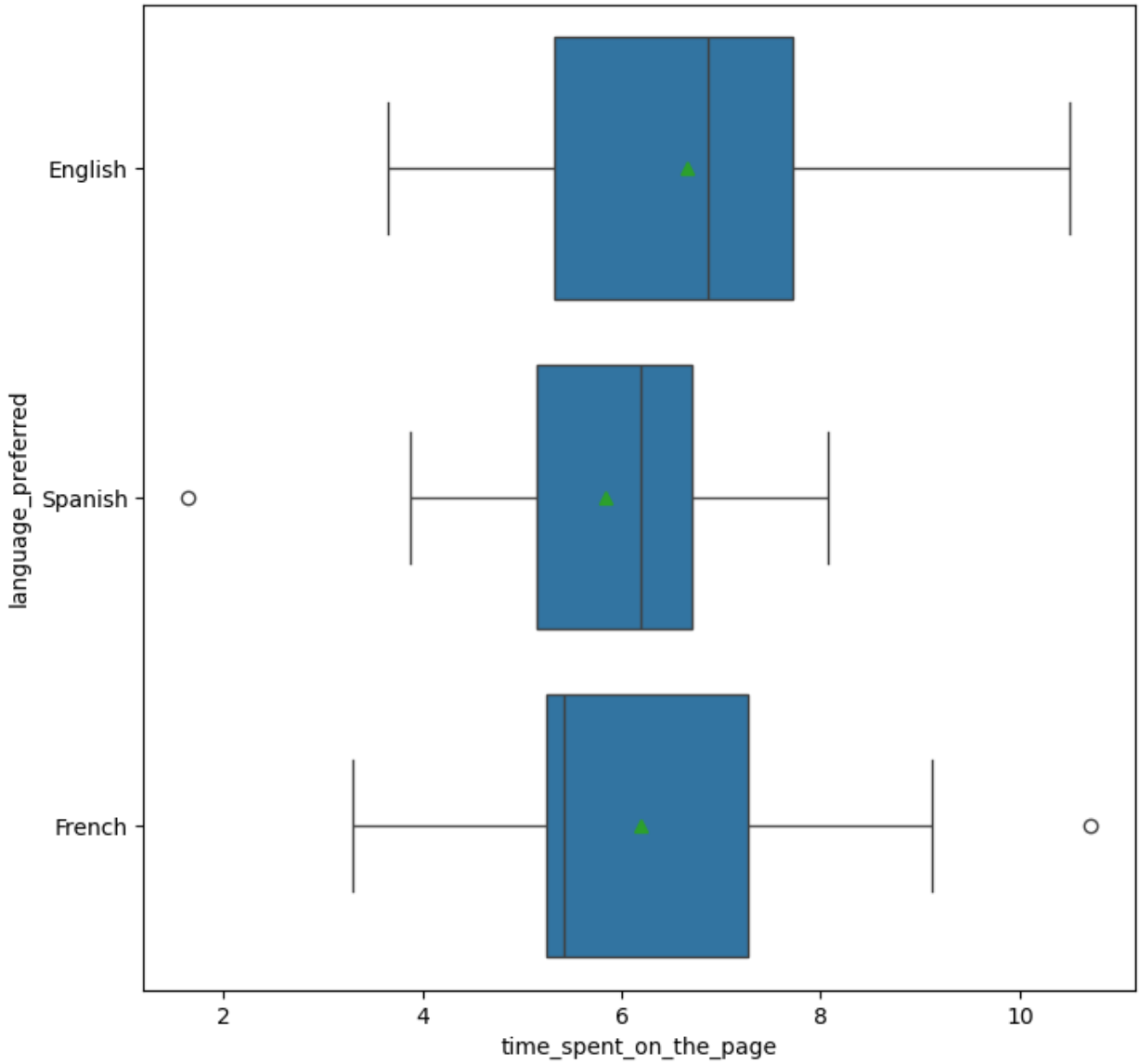

FIGURE 13

- English speakers tend to spend the most time on the page with median time around 6.5 minutes.
- Spanish speakers have slight lower median time of 6 minutes with less difference in the time spent.
- French speakers have the least median time spent on the page approx. 5.5 minutes with an outlier spending more time.

- The outlier present in both Spanish & French groups show small number of users spend more time on the page
- English speakers who spend the highest time shows a possible correlation between the time spent and likely to get converted.

Language preferred	Time spent on the page
English	6.663750
French	6.196471
Spanish	5.835294

TABLE 2

STEP 1: DEFINE NULL AND ALTERNATE HYPOTHESIS

- Null Hypothesis: There is no difference in the mean time spent on the page across different preferred languages.
- Alternative Hypothesis: There is a difference in the mean time spent on the page across different preferred languages.

STEP 2: SELECT APPROPRIATE TEST

- We'll use a one-way analysis of variance **(ANOVA)** test to compare the means of time spent on the page for different preferred languages.
- ANOVA is appropriate when comparing means across more than two groups.

STEP 3: DECIDE SIGNIFICANCE LEVEL

- Significance level α = 0.5

STEP 4: COLLECT AND PREPARE DATA

- We'll conduct the ANOVA test to determine if there are significant differences in mean time spent on the page across different language groups.

STEP 5: CALCULATE THE p- value

- The p- value is 0.43204138694325955

STEP 6: COMPARE THE p- value with α

- As the p- value 0.43204138694325955 is greater than the level of significance, we fail to reject the null hypothesis, indicating no significant difference.

STEP 7: DRAW INFERENCE:

- As the p-value is greater than the significance level, we fail to reject the null hypothesis, indicating no evidence of an

association of time spent on the page and language preferred.

CONCLUSION & BUSINESS RECOMMENDATIONS:

Based on the analysis following insights have been observed:

Conclusion:

- **Time Spent**: As the p-value for the t-test is less than 0.05, users spend significantly more time on the new landing page.

- **Conversion Rate by Language**: As the p-value for the chi-square test is less than 0.05, the conversion status depends on the preferred language.

- **Time Spent by Language**: As the p-value for the ANOVA is less than 0.05, the mean time spent on the new page differs across languages.

- Users on the new landing page spend more time on the new landing page than those on the old landing page. This

suggests that the new landing page is more engaging and holds user attention better.

- The new landing page shows a high conversion rate as compared to the old landing page suggesting new design is more successful in conversion.

- Users preferring English have the highest conversion rates, followed by those who prefer French and least by the Spanish users.

- French and Spanish speakers spend less time on the page indicating some serious issues with the engagement and content.

Business Recommendations:

- **Implement the New Landing Page**: As users spend more time on the new landing page, it indicates better engagement thus beneficial.

- **Personalize Content**: As the conversion depends on language, tailor the content to different language preferences to improve conversion rates on the landing page

- **Optimize Engagement by Language**: As the time spent on the page varies by language, focus on optimizing content for the languages with lower engagement.

- Observing new landing page performance in terms of conversion rates, adopt it as the default landing page for all the users.

- Consider conducting user surveys with non-English speakers to gain deeper insights.