Inferential Statistics

Project Report

Contents

List of Tables

Problem Statement - IS Project

Problem 1:

An independent research organization is trying to estimate the probability that an accident at a nuclear power plant will result in radiation leakage. The types of accidents possible at the plant are, fire hazards, mechanical failure, or human error. The research organization also knows that two or more types of accidents cannot occur simultaneously.
According to the studies carried out by the organization, the probability of a radiation leak in case of a fire is 20%, the probability of a radiation leak in case of a mechanical 50%, and the probability of a radiation leak in case of a human error is 10%. The studies also showed the following;

- The probability of a radiation leak occurring simultaneously with fire is 0.1%.
- The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%.
- The probability of a radiation leak occurring simultaneously with a human error is 0.12%.

On the basis of the information available, answer the questions below:
1.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?
1.2 What is the probability of a radiation leak?
1.3 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?

Solution:
Given—

- P(Radiation leak/Fire) or P(RL / F) = 0.2
- P(Radiation leak/Mechanical Error) or P(RL / ME) = 0.5
- P(Radiation leak/Human Error) or P(RL / HE) = 0.1
- P(RL ∩ F) = 0.001
- P(RL ∩ ME) = 0.0015
- P(RL ∩ HE) = 0.0012

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

Ans:- P(F) = P(RF ∩ F)/P(RF / F) = 0.001/0.2 = **0.005**
P(ME) = P(RL ∩ ME)/ P(RL / ME) = 0.0015/0.5 = **0.003**
P(HE) = P(RL / HE)/ P(RL ∩ HE) = 0.1/0.0012 = **0.012**
So, the probabilities of a fire, a mechanical error, and a human error are respectively **0.005, 0.003, 0.012**

1.2 What is the probability of a radiation leak?

Ans:- P(No Accident) or P(NA) = 1-(0.005+0.003+0.012) = 0.98
P(RF / NA) = 0
P(RF ∩ NA) = P(RF / NA)/P(NA) = 0/0.98 = 0
With the help of probability theorem,
P(RL) = P(RL ∩ F)+ P(RL ∩ ME)+ P(RL ∩ HE)+ P(RF ∩ NA) = 0.001+0.0015+0.0012+0 = **0.0037**
So, the probability of radiation leak is **0.0037**

1.3. 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?

a) Ans:- Probability of fire radiation leak is-
P(F/RL) = P(RL ∩ F)/P(RL) = 0.001/0.0037 = **0.270**
b) Ans:- Probability of mechanical failure radiation leak is-
P(ME/RL) = P(RL ∩ ME)/P(RL) = 0.0015/0.0037 = **0.405**
c) Ans:- Probability of human error radiation leak is-
P(HE/RL) = P((RL ∩ HE)/P(RL) = 0.0012/0.0037 = **0.324**

Problem 2:

Grades of the final examination in a training course are found to be normally distributed, with a mean of 77 and a standard deviation of 8.5. Based on the given information answer the questions below.

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

2.2 What is the probability that a randomly selected student scores between 65 and 87?

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

Solution:
Given:---
Mean (μ) = 77
Standard Deviation (σ) = 8.5
The formula of Z score is, $z = (n- \mu)/ \sigma$
Where n = 85, 65 and 87 observed values

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

Ans:- $z = (n- \mu)/ \sigma = (85-77)/8.5 = 0.941$

From the cumulative distribution function (cdf) of the 'z' value [1-stats.norm.cdf(z)] we found the probability that a randomly chosen student gets a grade below 85 on this exam is 0.17 or approximately 17%.

2.2 What is the probability that a randomly selected student scores between 65 and 87?

Ans:- $z1 = (n- \mu)/ \sigma = (65-77)/8.5 = -1.41$
$z\ 2 = (n- \mu)/ \sigma = (87-77)/8.5 = 1.17$

From the difference between cdf(z2) and cdf(z1) we found the probability that a randomly selected student scored between 65 and 87 is 0.80 or approximately 80%.

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

Ans:- From the percent point function (ppf) we found that approximately 71% passing cut-off so that 75% of the student clear the exam.

Problem 3:

Business Context

The advent of e-news, or electronic news, portals has offered us a great opportunity to quickly get updates on the day-to-day events occurring globally. The information on these portals is retrieved electronically from online databases, processed using a variety of software, and then transmitted to the users. There are multiple advantages of transmitting news electronically, like faster access to the content and the ability to utilize different technologies such as audio, graphics, video, and other interactive elements that are either not being used or aren't common yet in traditional newspapers.

E-news Express, an online news portal, aims to expand its business by acquiring new subscribers. With every visitor to the website taking certain actions based on their interest, the company plans to analyze these actions to understand user interests and determine how to drive better engagement. The executives at 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 web page is not designed well enough in terms of the outline & recommended content to keep customers engaged long enough to make a decision to subscribe.

[Companies often analyze user responses to two variants of a product to decide which of the two variants is more effective. This experimental technique, known as A/B testing, is used to determine whether a new feature attracts users based on a chosen metric.]

Objective

The design team of the company has researched and created a new landing page that has a new outline & more relevant content shown compared to the old page. In order to test the effectiveness of the new landing page in gathering new

subscribers, the Data Science team conducted an experiment by randomly selecting 100 users and dividing them equally into two groups. The existing landing page was served to the first group (control group) and the new landing page to the second group (treatment group). Data regarding the interaction of users in both groups with the two versions of the landing page was collected. Being a data scientist in E-news Express, you have been asked to explore the data and perform 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:

3.1 Do the users spend more time on the new landing page than on the existing landing page?
3.2 Does the converted status depend on the preferred language?
3.3 Is the mean time spent on the new page the same for the different language users?

Explore the dataset and extract insights using Exploratory Data Analysis

Data Dictionary

The data contains information regarding the interaction of users in both groups with the two versions of the landing page.
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

Data Overview
The initial steps to get an overview of any dataset is to:

- observe the first few rows of the dataset, to check whether the dataset has been loaded properly or not
- get information about the number of rows and columns in the dataset
- find out the data types of the columns to ensure that data is stored in the preferred format and the value of each property is as expected.
- check the statistical summary of the dataset to get an overview of the numerical columns of the data

>Displaying the first few rows of the dataset

	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

Table 1: Top 5 rows of the dataset

>Displaying the last few rows of the dataset

	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

Table 2: Last 5 rows of the dataset

Observation: There are 100 rows and 6 columns in this dataframe.

>Checking the data types of the columns for the dataset

```
<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
```

Table-3 Basic info of the dataset

Observation: There are different data types present in this dadaset

- One is integer type which is 'User Id'
- Four object variables are 'Group', 'Landing Page', 'Converted', 'Language Preferred'.
- There is one float type variable present in the dataset which is 'Time Spent On The Page'.

>Getting the statistical summary for the numerical variables

	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

Table-4 statistical summary of num variables

> Getting the statistical summary for the categorical variables

	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

Table-5 statistical summary of categorical variables

Observation:

- The average time spent on the page is about 5.38 min and the standard deviation is 2.38 minutes
- The 50% of the entries spent about 5.42 minutes on the page
- The minimum time spent on the page was 0.19 minutes and the maximum time spent on the page was 10.71 minutes.

Check for missing values

```
user_id                   0
group                     0
landing_page              0
time_spent_on_the_page    0
converted                 0
language_preferred        0
dtype: int64
```

Observation: There are no missing values present in the dataset.

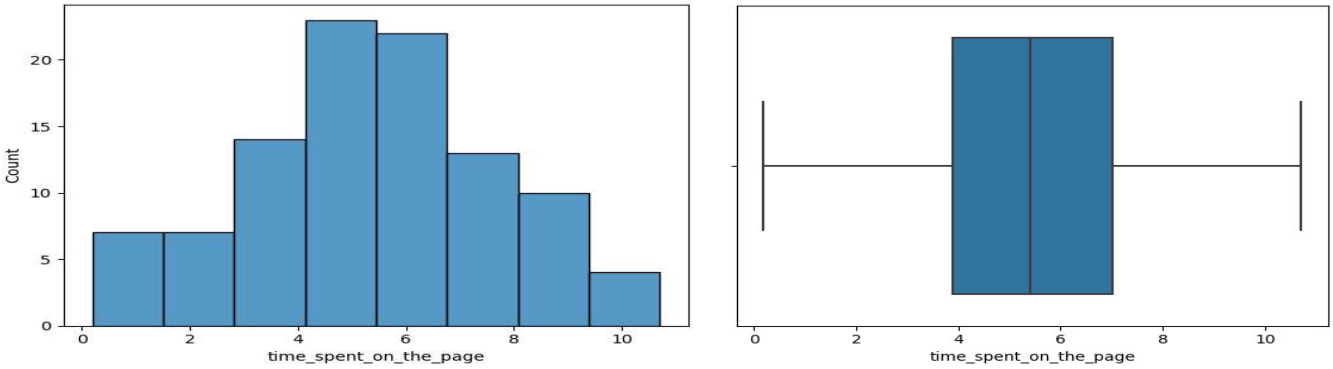
Check for duplicates

```
0     False
1     False
2     False
3     False
4     False
      ...
95    False
96    False
97    False
98    False
99    False
Length: 100, dtype: bool
```

Observation: There are no duplicated values in the dataset.

Univariate Analysis

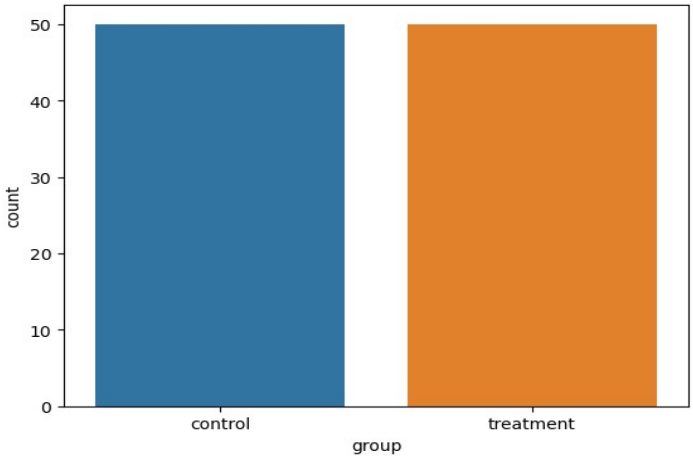
>Time spent on the page



Observations: The time spent on the page appears to be normally distributed and there is no outliers present.

> Group

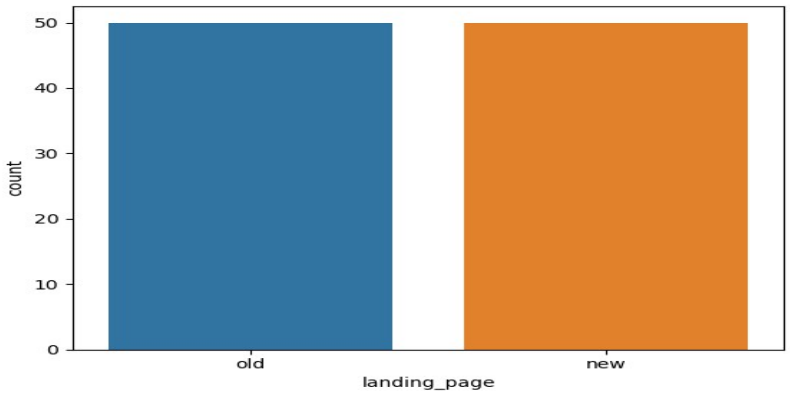
```
group
control      50
treatment    50
Name: count, dtype: int64
```



Observation: The sample is equally distributed amongst the control and the treatment groups.

>Landing Page

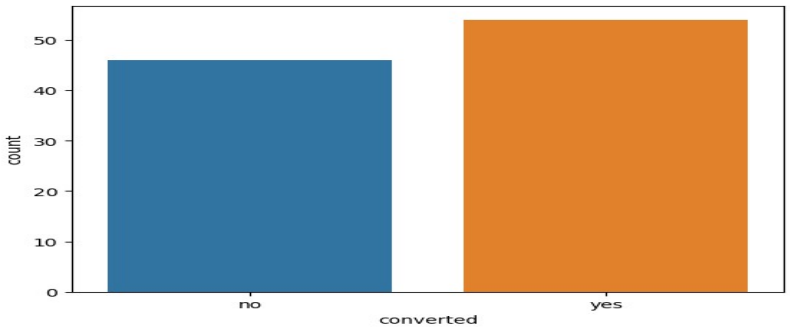
```
landing_page
old    50
new    50
Name: count, dtype: int64
```



Observation: The sample is equally distributed amongst the old and new landing pages.

>Converted

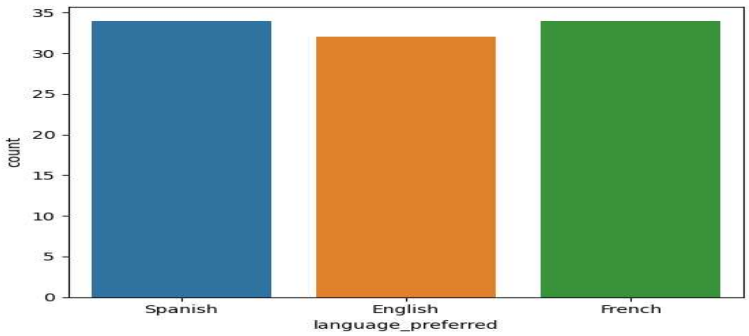
```
converted
yes    54
no     46
Name: count, dtype: int64
```



Observation: From the above plot it is clearly see that there are more people with converted the new landing page opposite to the people who are not converted.

>Language Preferred

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



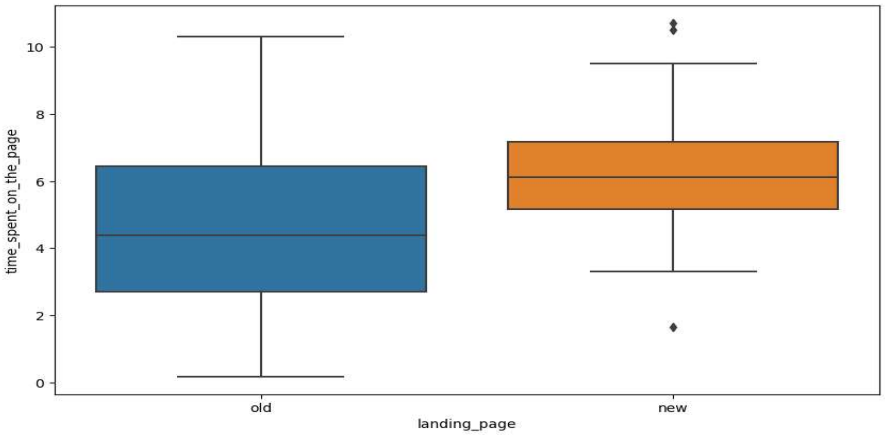
Observation: From the above plot we see that entry o Spanish and French language is 34 and the English language entries is about approximately 32.

Overall Observations On Univariate Analysis:

- From the above analysis group variable has two unique values, control and treatment, with 50 entries each
- From the above analysis landing_page variable has two unique values, old and new, with 50 entries each
- From the above analysis converted variable has two unique values, yes and no, with 54 and 46 entries respectively
- From the above analysis language_preferred variable has three unique values: Spanish, French, English.They have 34, 34, and 32 entries respectively.

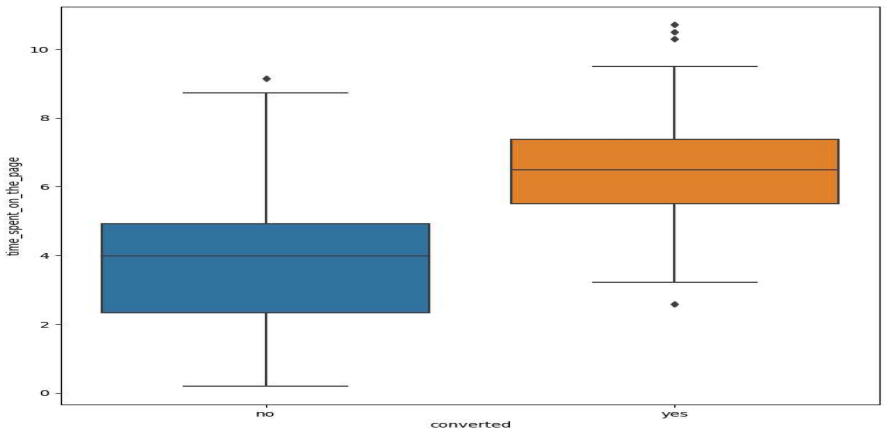
Bivariate Analysis

>Landing page vs Time spent on the page



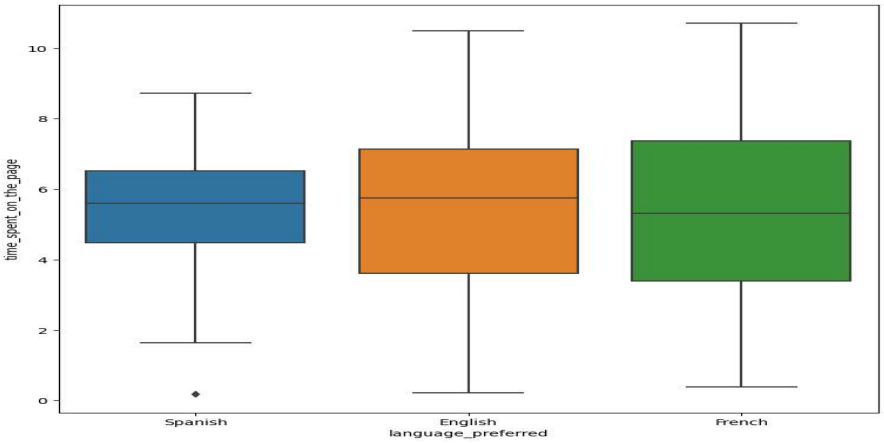
Observation: From the above plot we see that users spent more time on the new landing page compare to the old landing page.

>Conversion status vs Time spent on the page



Observation: From the above plot we see that which users converted a subscriber spent more time on the page.

>Language preferred vs Time spent on the page

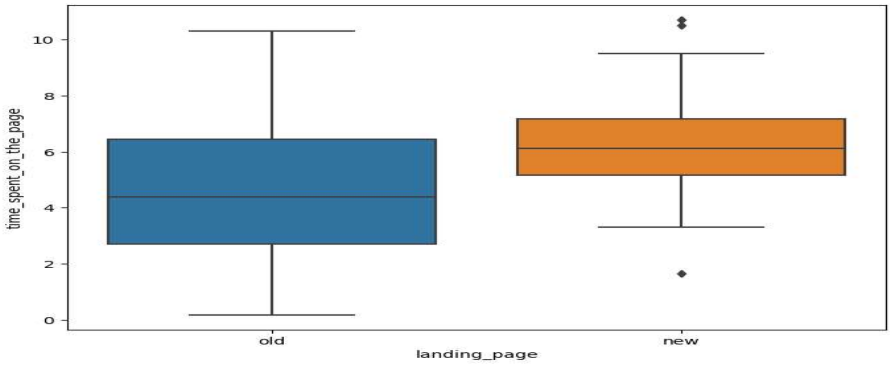


Observation: From the above plot, it is clear that preferred language are appear similar time spent on the page. Also preferred language Spanish is smallest spread in the time spent on the page.

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

- State the null and alternate hypotheses - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results

Ans:- Perform Visual Analysis



Step 1: Define the null and alternate hypotheses

H_0: The mean time spent by the users on the new page is equal to the mean time spent by the users on the old page.

H_a: The mean time spent by the users on the new page is greater than the mean time spent by the users on the old page.

Step 2: Select Appropriate test

This is a one-tailed test concerning two population means from two independent populations. The population standard deviations are unknown. Based **on this information, a two-sample independent t-test would be the most appropriate.**

Step 3: Decide the significance level

As given in the problem statement, we select α=0.05.

Step 4: Collect and prepare data

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

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

Based on the sample standard deviations of the two groups, decide whether the population standard deviations can be assumed to be equal or unequal

=Two-sample independent t-test assumptions:

- The time spent on the pages is measured on a continuous scale so it is continuous data.
- It is a normally distributed populations.
- From the population we are taking random samples for two different groups, the two samples are from two independent populations so it is independent populations.
- Standard deviation of the population are different or unequal.
- We are informed that the collected sample from the population is a simple random sample.

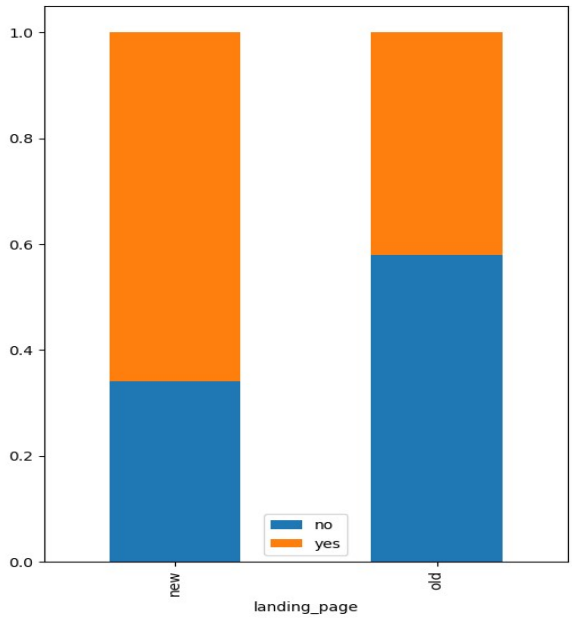
Step 5: Calculate the p-value
The p-value is 0.0001392381225166549

Step 6: Compare the p-value with α
As the p-value 0.0001392381225166549 is less than the level of significance, we reject the null hypothesis.

Step 7: Draw inference
Here the p-value is less than the level of significance of 0.05 or 5%, so we can easily reject the null hypothesis. So it means that there is significant evidence that the mean time spent by the users on the new page is greater than the mean time spent by the users on the old page.

3.2 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?

Ans:- Perform Visual Analysis



Step 1: Define the null and alternate hypotheses
H_0: The conversion rate of the new page is equal to the conversion rate of the old page.
H_a: The conversion rate of the new page is greater than the conversion rate of the old page.

Step 2: Select Appropriate test
This is a one-tailed test concerning two population proportions from two independent populations. **Based on this information, a two proportion z-test would be the most appropriate**.

Step 3: Decide the significance level
As given in the problem statement, we select α = 0.05.

Step 4: Collect and prepare data
The numbers of users served the new and old pages are 50 and 50 respectively.

=Two-proportion z-test assumptions:

- We see that a user is either converted or not converted so the population is binomially distributed.
- We are informed that the collected sample from the population is a simple random sample.
- The binomial distribution approximated to normal distribution.

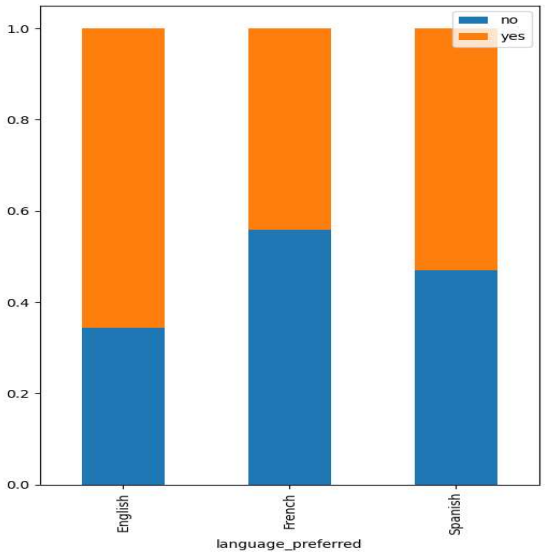
Step 5: Calculate the p-value
The p-value is 0.008026308204056278

Step 6: Compare the p-value with α
As the p-value 0.008026308204056278 is less than the level of significance, we reject the null hypothesis.

Step 7: Draw inference
Here the p-value is less than the level of significance at 0.05 or 5%, so we can easily reject the null hypothesis. This means that, there is significant evidence that the conversion rate of the new page is greater than the conversion rate of the old page.

3.3 Does the converted status depend on the preferred language?
- State the null and alternate hypotheses - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results

Ans:- Perform Visual Analysis



Step 1: Define the null and alternate hypotheses
$H0$: The converted status is independent of the preferred language.
Ha: The converted status is dependent of 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. **Based on this information, a chi-square test for independence would be the most appropriate.**

Step 3: Decide the significance level
As given in the problem statement, we select α = 0.05

Step 4: Collect and prepare data

converted	no	yes
language_preferred		
English	11	21
French	19	15
Spanish	16	18

=Chi-Squared test for independence assumptions:¶

- Categorical variables.
- The number of observations in each level is greater than 5.
- We are informed that the collected sample from the population is a simple random sample

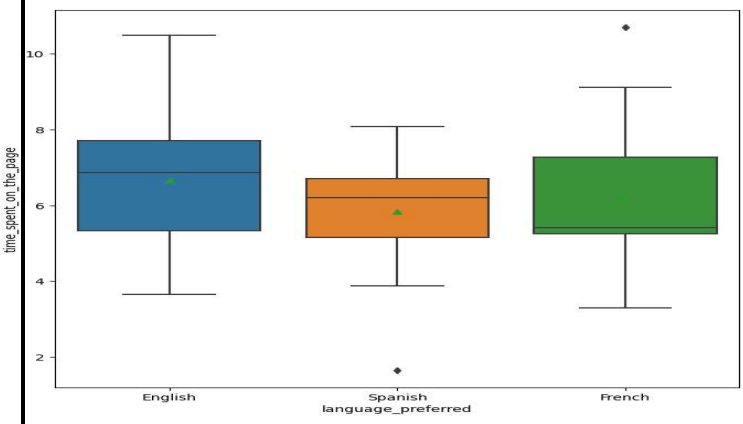
Step 5: Calculate the p-value
The p-value is 0.2129888748754345

Step 6: Compare the p-value with α
As the p-value 0.2129888748754345 is greater than the level of significance, we fail to reject the null hypothesis.

Step 7: Draw inference
Here the p-value is greater than the level of significance of 0.05 or 5%, so we fail to reject the null hypothesis. This means that that the converted status is independent of the preferred language.

3.4 Is the mean time spent on the new page same for the different language users?
- State the null and alternate hypotheses - Check the assumptions of the hypothesis test. - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results

Ans:- Perform Visual Analysis




```
language_preferred
English    6.663750
French     6.196471
Spanish    5.835294
Name: time_spent_on_the_page, dtype: float64
```

Step 1: Define the null and alternate hypotheses
$H0$: The mean time spent on the new lading page is the same across all preferred languages.

Ha: At least one of the mean times spent on the new landing page is different amongst the preferred languages.

Step 2: Select Appropriate test
This is a problem, concerning three population means. **Based on this information, a one-way ANOVA test would be the most appropriate.**

Step 3: Decide the significance level
As given in the problem statement, we select α = 0.05.

Step 4: Collect and prepare data
From the Shapiro-Wilk's test we found the p-value is 0.8040016293525696 which is very large so, we to reject the null hypothesis and it is a normal distribution.
From the Levenes test we found the p-value is 0.46711357711340173 it is also very large so we fail to reject the null hypothesis and the variances are equal.

= One-way ANOVA test assumptions:

- The populations are normally distributed .
- Population variances are equal.

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.

Step 7: Draw inference
Here the p-value is greater than the level of significance at 0.05 or 5%, so we fail to reject the null hypothesis. This means that the mean time spent on the new landing page is relatively similar regardless of the preferred language.

Conclusion and Business Recommendations

Conclusion:

- From the above analysis of the question if users spend more time on the new landing page than the old landing page, a two-sample independent t-test was performed. A p-value of 0.0001 was resulted from the test, which is very small than the level of significance of 0.05 or 5%. Therefore, we rejected the null hypothesis. What this means in context is that there is significant evidence that the mean time spent by the users on the new page is greater than the mean time spent by the users on the old page.
- From the above analysis of the question if the conversion rate for the new page is greater than the conversion rate of the old page, a two-proportion z-test was performed. A p-value of 0.008 was resulted from the test, which is less than the level of significance of0.05 or 5%. Therefore we rejected the null hypothesis. What this means in context is that there is significant evidence that the conversion rate of the new landing page was greater than the conversion rate of the old landing page.
- From the above analysis of the question if the conversion status and preferred language are related, a chi-square test for independence was performed. A p-value of 0.213 was resulted from the test, which is more than the level of significance of 0.05 or 5%. Therefore we failed to reject the null hypothesis. What this means in context is that conversion status and the preferred language of the landing page are independent of each other.
- From the above analysis of the question if the time spent on the new landing page differed based on preferred language, a one-way ANOVA test was performed. A p-value of 0.432 was resulted from the test, which is more than the level of significance of 0.05 or 5%. Therefore we failed to reject the null hypothesis. What this means in context is that the mean time spent on the new landing page was relatively similar across all the preferred languages.

Recommendations:

- E-News Express should fully implement the new landing page as it appears to gain a lot more traction than the old landing page. The time spent on the new landing page is greater than the time spent on the old landing page is evidence that users prefer it.
- It might be beneficial to cut the losses with the old landing page as there are diminutive returns in average time spent and conversion rate. The new landing page has an increased conversion rate, therefore, more resources should be directed towards it as it has more opportunity to increase membership.

- Deploy the new landing page incorporating all the exiting preferred language. As there is no signficant difference between the average time spent on the new page across the preferred languages, the conversion rate to subscribers will be the similar throughout. Perhaps consider adding more languages to the portal to reach a wider audience.

.