Contents

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%.**

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

Ans : The probabilities of a fire, a mechanical failure, and a human error are 0.20, 0.50, and 0.10, respectively

1.2 What is the probability of a radiation leak?

Ans : The probability of a radiation leak is 1.0 or 100%.

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 **= 0.0002**

b) a mechanical failure**= 0.00075**

c) a human error **= 0.00012**

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 = **82.64%**

2.2 What is the probability that a randomly selected student scores between 65 and 87 = **79.97%**

2.3 What should be the passing cut-off so that 75% of the students clear the exam = **82.73425%**

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:

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

Does the converted status depend on the preferred language?

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

Data Dictionary

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

user_id - Unique user ID of the person visiting the website

group - Whether the user belongs to the first group (control) or the second group (treatment)

landing_page - Whether the landing page is new or old

time_spent_on_the_page - Time (in minutes) spent by the user on the landing page

converted - Whether the user gets converted to a subscriber of the news portal or not

language_preferred - language chosen by the user to view the landing page

Q1. "Define the problem and perform an Exploratory Data Analysis"

- Problem definition, questions to be answered - Data background and contents - Uni-variate analysis - Bi-variate analysis

The first few rows of the data-set

	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

The last few rows of the data-set

	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

Shape : Rows - 100,Column = 6

Data types of the columns for the data-set

```
<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
```

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

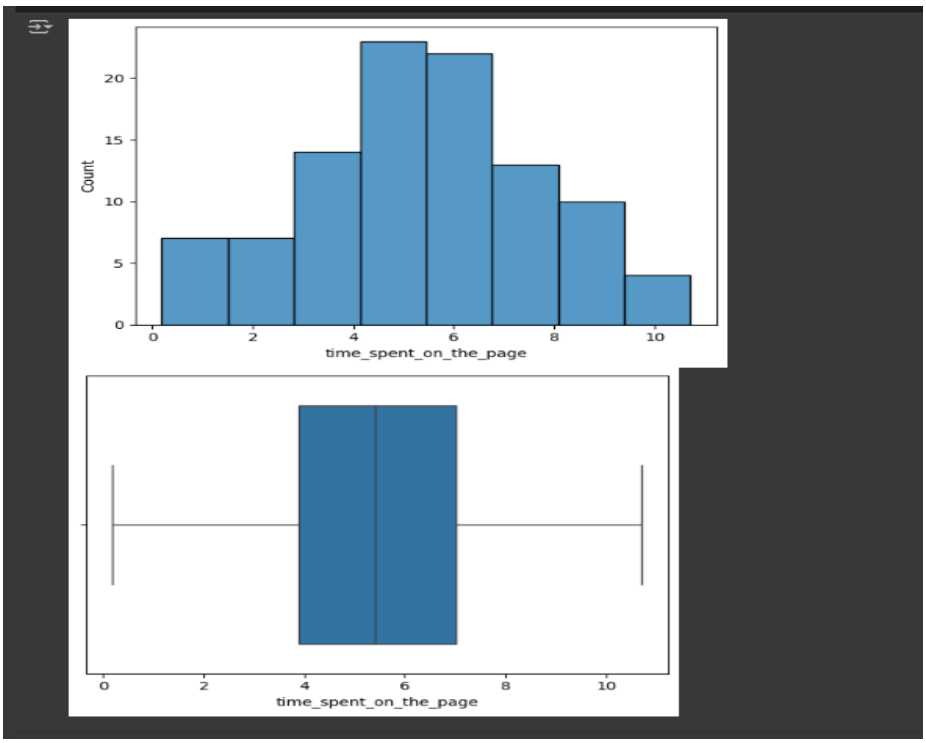
Missing Value :

	0
user_id	0
group	0
landing_page	0
time_spent_on_the_page	0
converted	0
language_preferred	0

dtype: int64

No of Duplicate Rows is Zero

Univariate Analysis

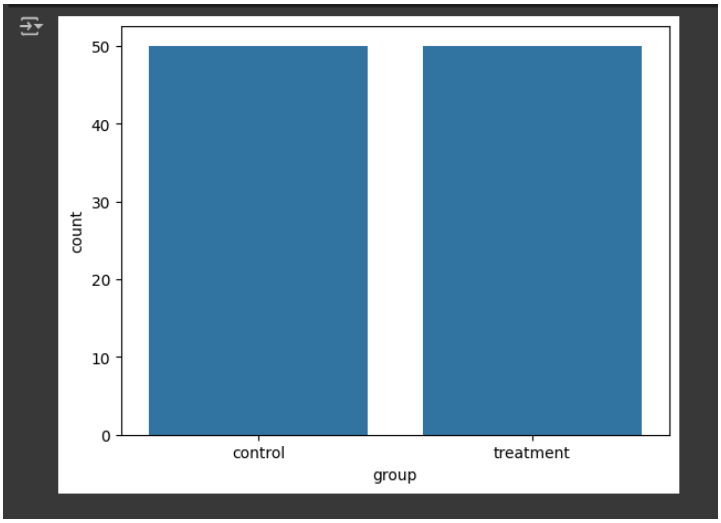


No Outliers in Box plot and and the shape of the distribution is biomodel

Univariate analysis of 'group' variable in the dataset

group	count
control	50
treatment	50

dtype: int64



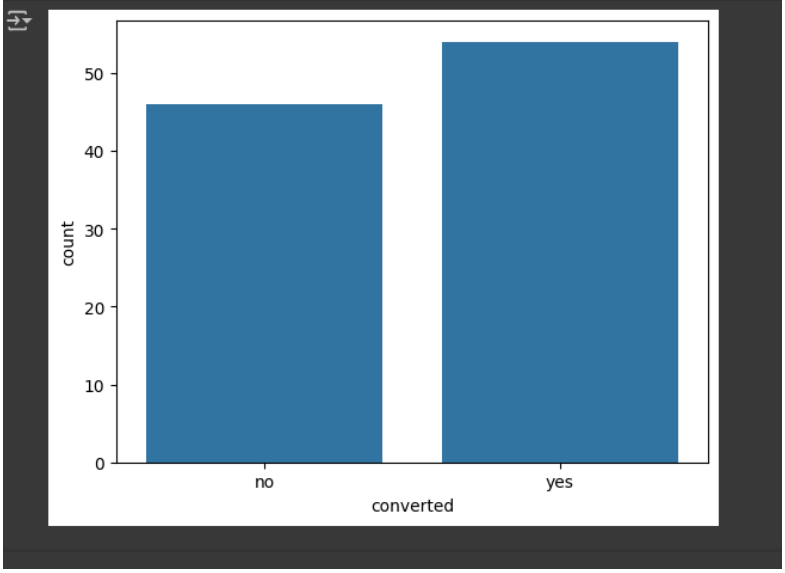
Here the counts for both the unique values of the ‘group’ column are approximately same.

Landing page

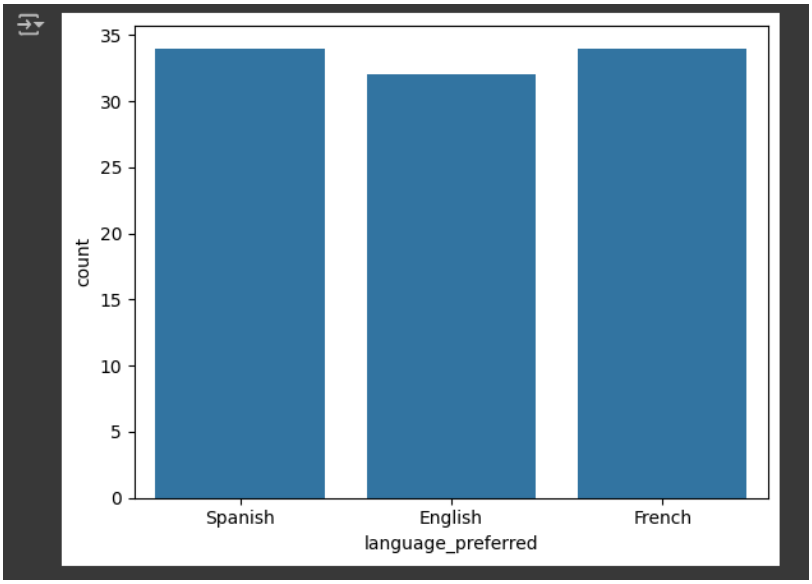
landing_page	count
old	50
new	50

dtype: int64

Here the counts for both the unique values of the landing_page(old,new) column are same.

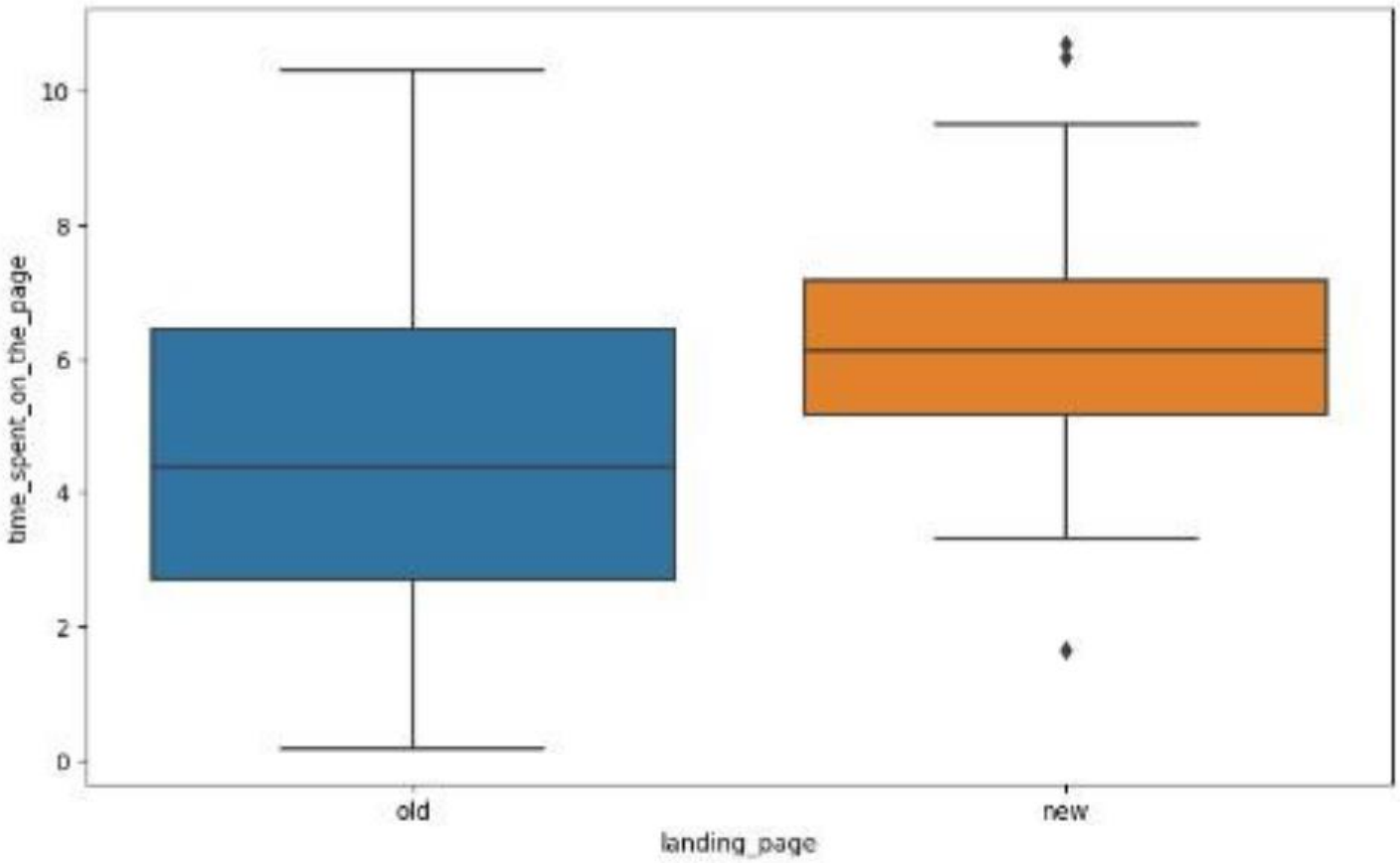


Here the count of ‘yes’ is more than the count of ‘no’.



Here count of the preferred language Spanish and French are equal values, but count of English as preferred language is little less.

Bi-variate Analysis :



New landing page has outliers.

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

Results of the Hypothesis Test for Time Spent:

Normality (Shapiro-Wilk Test):

Control group: $p=0.456p = 0.456p=0.456$ (data is normally distributed).

Treatment group: $p=0.804p = 0.804p=0.804$ (data is normally distributed).

Homogeneity of Variances (Levene's Test):

$p=0.009p = 0.009p=0.009$: Variances are not equal, so we should consider a Welch's t-test for unequal variances.

T-Test (Welch's adjustment for unequal variances):

$t=−3.79t = -3.79t=−3.79$, $p=0.00013p = 0.00013p=0.00013$ (one-tailed test).

Since $p<0.05p < 0.05p<0.05$, we reject the null hypothesis.

Conclusion:

Users spend significantly more time on the new landing page compared to the old one. This supports the idea that the new page engages users better.

Actionable Insights

Based on the exploratory data analysis and hypothesis testing, here are the key insights:

1.

New Landing Page Performance:

2.
1. Users spend significantly more time on the new landing page than the old one.
2. The conversion rate for the treatment group (new landing page) is 66%, compared to 42% for the control group. This indicates a substantial improvement in engaging users and converting them into subscribers.
3.

Language Preferences:

4.

1. Language preference is likely to have an influence on user engagement and conversion. Further testing would clarify if specific languages perform better with the new landing page.
5.

Time on Page Across Languages:

6.

1. If differences in time spent on the new page by language are found (requires completing ANOVA), optimization could target underperforming language groups to further improve engagement.

Business Recommendations

Deploy the New Landing Page:

1. Given its significantly higher engagement and conversion rates, the new landing page should replace the old one across all users.

Optimize Content for Different Languages:

1. Use the insights from language-specific analysis to tailor content or layout for each language group.
2. Identify underperforming language segments and conduct targeted experiments to improve engagement in those groups.

Focus on Time Spent Metrics:

Encourage actions that increase time spent on the page, as this correlates positively with higher conversion rates. Consider adding interactive features like quizzes, videos, or personalized call-to-actions to further enhance engagement

Monitor and Analyze User Feedback:

Gather qualitative feedback from users who converted and those who didn’t, to understand their preferences and address barriers to subscription.