E-News Express

Business Statistics and PGP-DSBA

07/05/2023

Contents / Agenda

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

Executive Summary

Conclusions:-

- To answer the questions if the users spend more time on the new landing page compared to old landing page a two sample independent t-test is performed. In that the p-value is less than the level of significance it means that null hypothesis is rejected.

- To answer the questions if conversion rate for new page is greater than the conversion rate of old page than a two proportion z-test was performed. In that the p-value is less than the level of significance so the null hypothesis is rejected.

- To answer the third question the conversion status and preferred language is related than a chi-square test for independence was performed. In that the p-value is more than the significance level so the null hypothesis is failed to be rejected.

- In order to answer the fourth question if the time spent on the new landing page based on the preferred language a one way ANNOVA test was performed. In that p-value is more than the significance level i.e. 5%. So the null hypothesis is failed to be rejected.

Recommendations:-

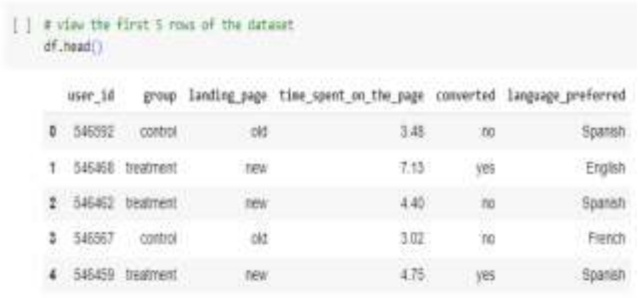
- E-news express should fully implement the new landing page as it gains a lot more attraction than the old landing page. Through this they can be benefited by cut the losses with old landing page as there are minute returns in average time spent and conversion rates.
- The new landing page has increased a lot amongst the users so E-news express have more opportunity to increase memberships.
- Can be reach to wider language preferred users by adding more languages so it gets to the wider audience considering the new landing page.

Business Problem Overview and Solution Approach

- E-news express is an online news portal where hey are planning to expand the business by acquiring the new subscribers.
- So the company wants to analyze the data which is collected from the new subscribers and determines how to drive better engagements.
- As a result the team designed a new landing page and to get the result the data scientist team conducted an experiment and tested an effectiveness of new landing page.
- To get a solution team came up with data and performed a statistical analysis to determine the effectiveness of new landing page in gathering new subscribers for the news portal.

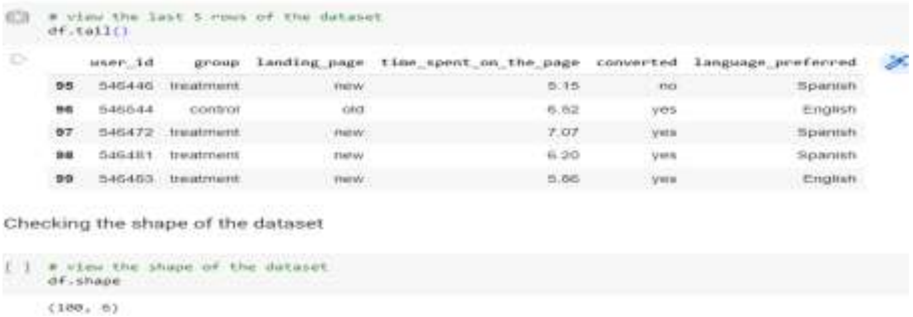
EDA Results

- Displaying the first few rows of the dataset



```
[ ] # view the first 5 rows of the dataset
    df.head()
```

	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



```
# view the last 5 rows of the dataset
df.tail()
```

	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

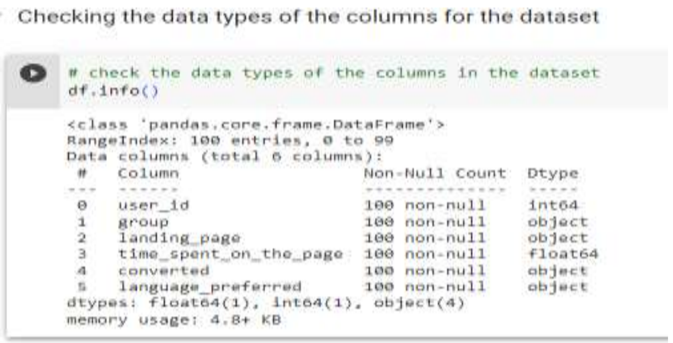
Checking the shape of the dataset

```
[ ] # view the shape of the dataset
    df.shape

    (100, 6)
```

Observations:-

1. We have observed here that from the data it shows the first five as well as last five rows of the dataset.
2. We can see that there are 100 rows and 6 columns in this dataset.

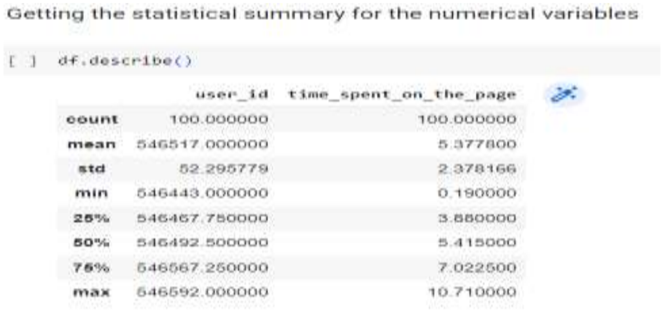
EDA results



Checking the data types of the columns for the dataset

```
# check the data types of the columns in the dataset
df.info()
```

```
<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
```



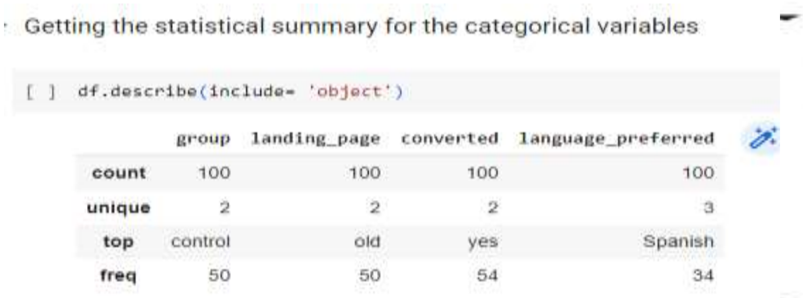
Getting the statistical summary for the numerical variables

```
[ ]  df.describe()
```

	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

Observations:-

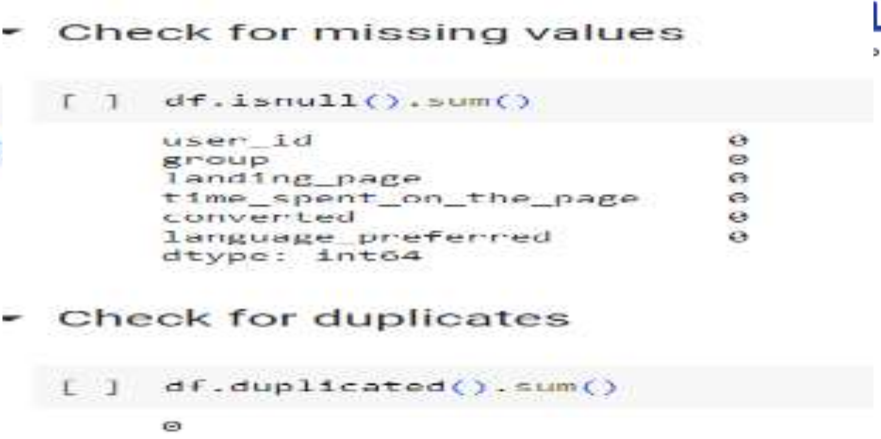
1. There are different data types which includes user_id as a integer type variable
2. The object type variable is group, landing_page, converted, and language preferred.
3. There is one float type which includes time_spent_on_the_page.
4. The average time spent on the page is 5.38 minutes with a std of 2.38minutes.
5. The minimum time spent on the page was 0.19 minutes and maximum was 10.71 minutes.
6. 50% of the time spent about 5.42 minutes.



Getting the statistical summary for the categorical variables

```
[ ]  df.describe(include= 'object')
```

	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



Check for missing values

```
[ ]  df.isnull().sum()

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

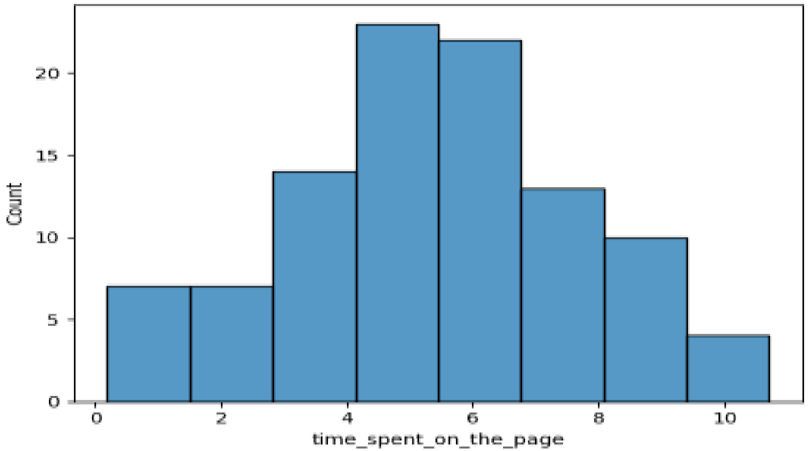
Check for duplicates

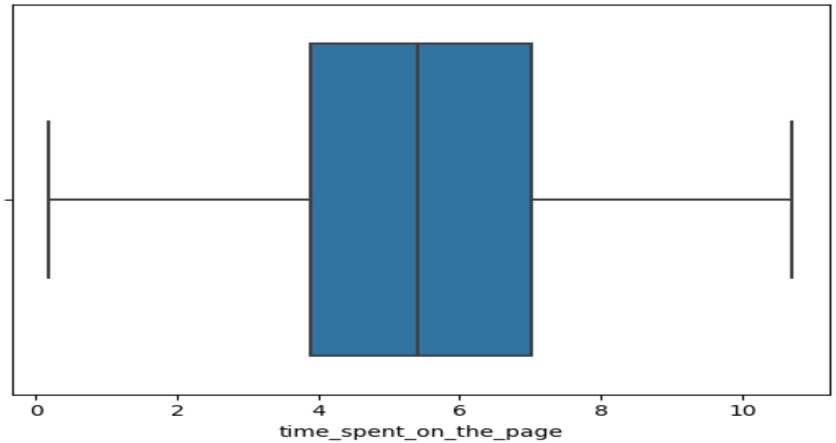
```
[ ]  df.duplicated().sum()

     0
```

Observations:-

1. All the categorical variables have 100 entries each and each has its unique values included in it.
2. The group has 2 unique value which includes control and treatment, landing page has 2 unique value which includes old an new, converted has 2 unique values which includes yes and no, an the language has 3 unique values which includes Spanish, French , and English.
3. There are no missing values as well as no duplicates for this dataset.

Univariate Analysis for time spent on page

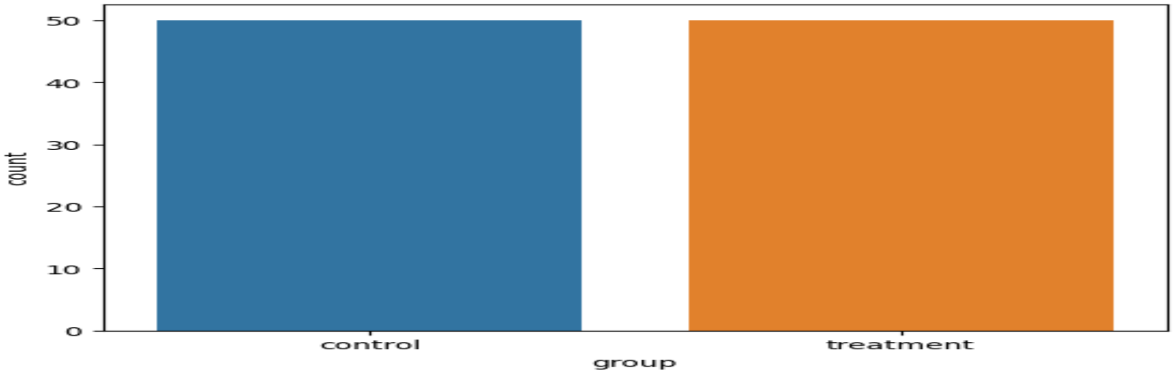


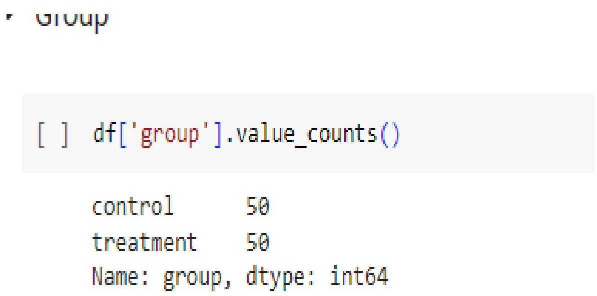


Observations:-

1. We have observed here that through histplot and box we can conclude that the time spent on the page is normally distributed and have no outliners.

Univariate analysis for group





Group

```
[ ]  df['group'].value_counts()
```

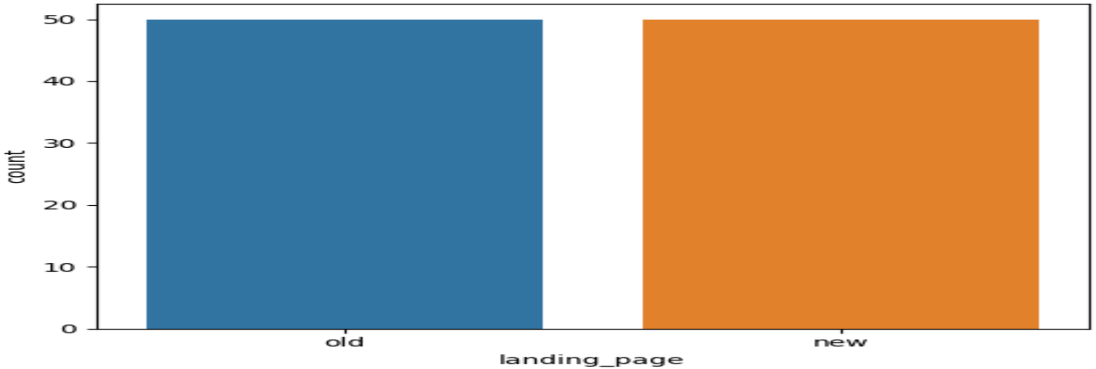
```
control      50
treatment    50
Name: group, dtype: int64
```

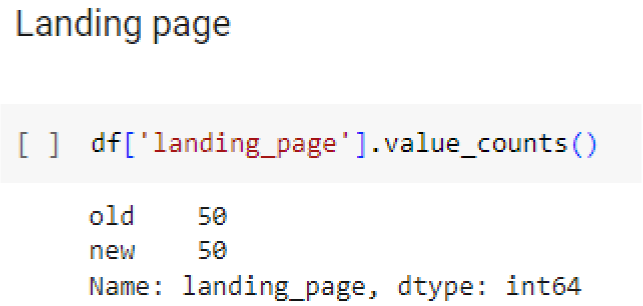
Observations:-

1. We have observed here that from the countplot and value counts we got control and treatment which has its value 50-50.

2. From the graph it appears that sample is equally distributed between control and treatment groups.

Univariate analysis for landing page.





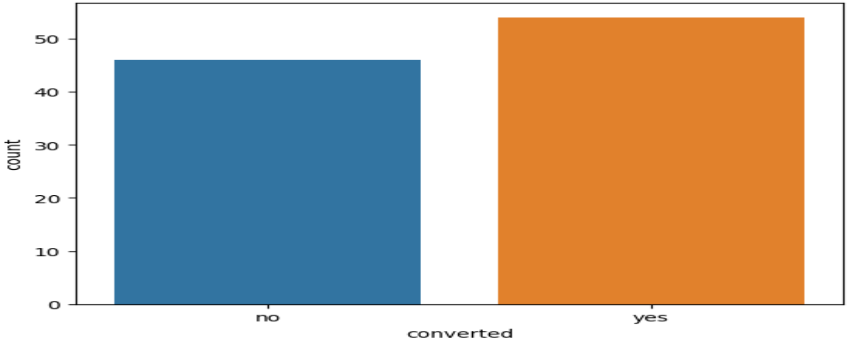
Landing page

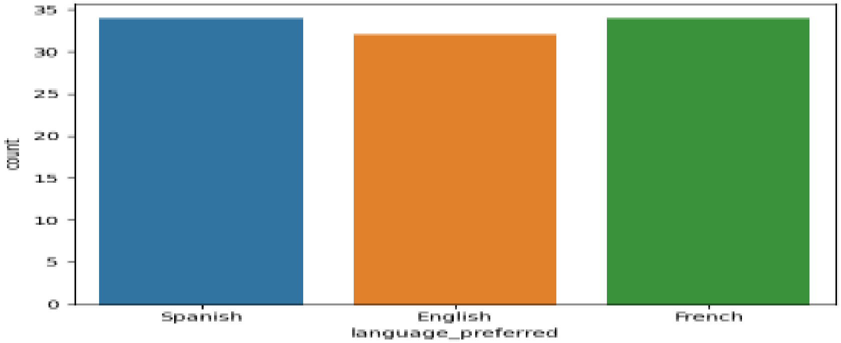
```
[ ]  df['landing_page'].value_counts()

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

Observations:-

1. We have observed that landing page has its two counts that is old and new and it’s value is 50-50.
2. From the graph we can conclude that this sample is equally split between the old and new landing pages.

Univariate analysis for converted and language preferred.

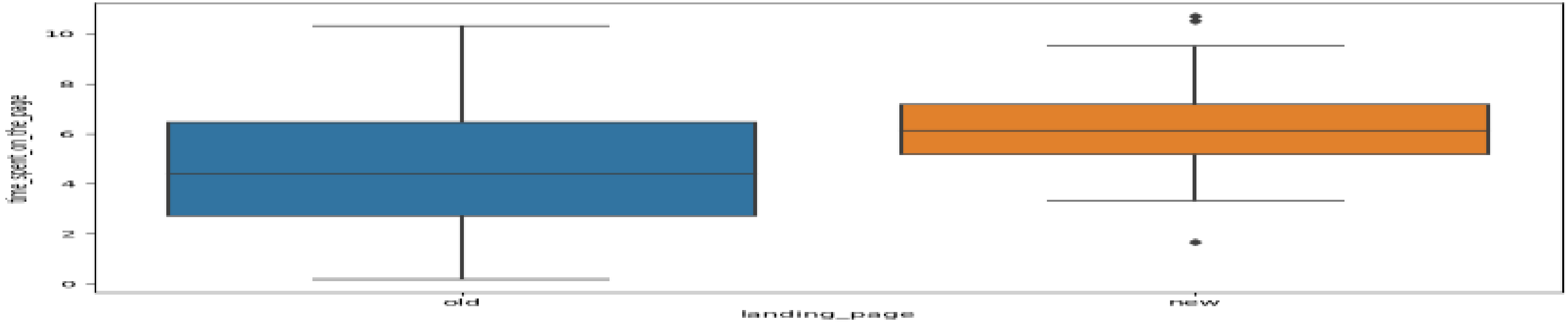




Observations:-

1. We have observed that the value counts for converted is 54 for yes and 46 for no so we can conclude here that there are more people who converted to the new landing page than the old landing page.
2. We have even observed here that the values for the Spanish, French ,and English are 34,34, and 32. So to a conclusion we can conclude that the maximum entries were Spanish and French while the minimum entries were English language preferred.

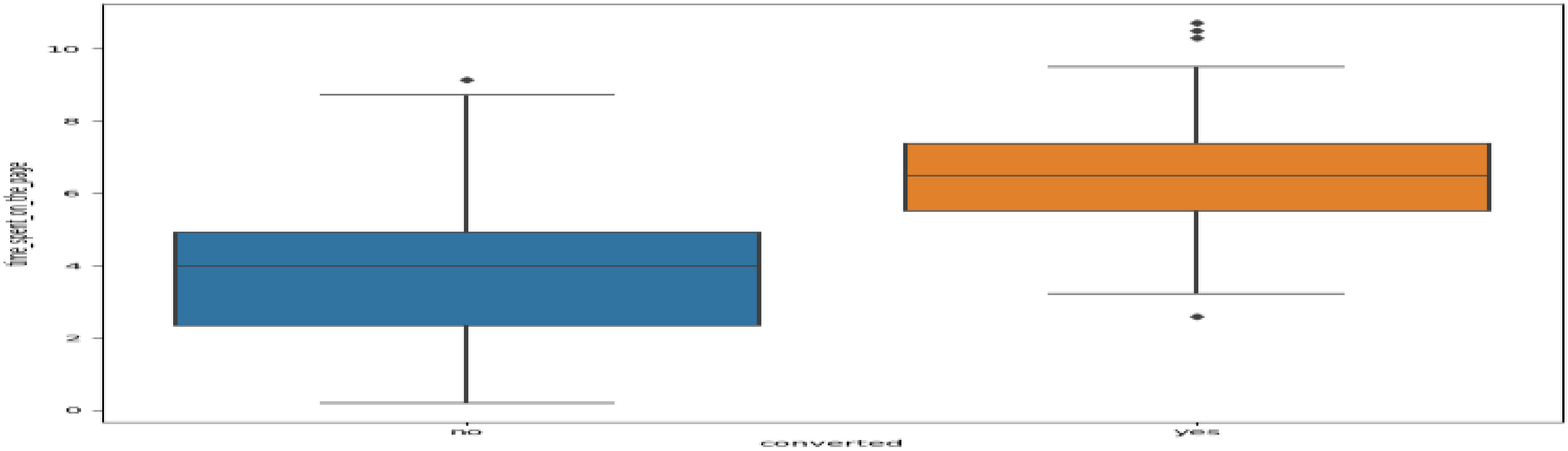
Bivariate Analysis for Landing time v/s Time spent on page



Observation:-

- We have observed here from the box plot that user spends more time on the new landing page compared to old landing page.

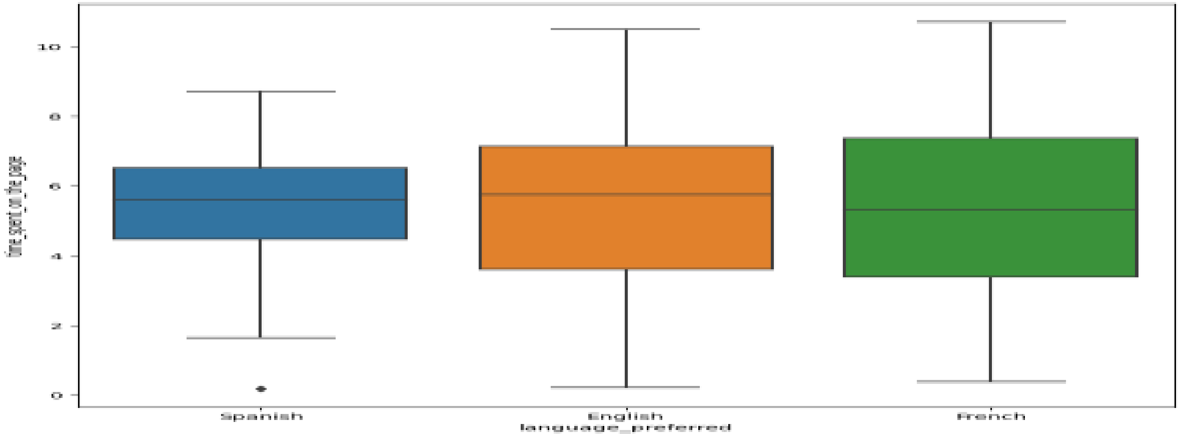
Bivariate Analysis for Conversion status v/s Time spent on page



Observation:-

- We have observed here from boxplot that the conversion status of yes and no category concludes that those who converted to a subscriber agent spend more time on the landing page compared to old one.

Bivariate Analysis for Language Preferred v/s Time spent on page



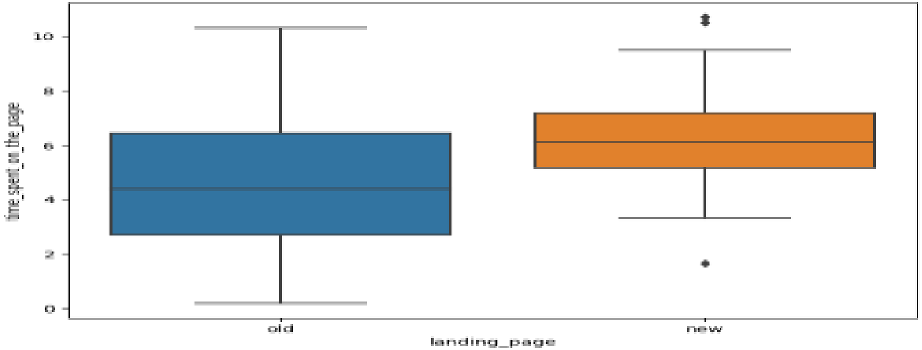
Observation:-

- We have observed here that the average time spent on the page is mostly similar for all languages but the people who preferred Spanish language have the smallest spread on the time spent.

Hypotheses Tested and Results

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

Visual Analysis:-



Observation:-

- The average time spent on the new page is greater than old landing page.

Hypothesis Testing Details

Step 1: Define the null and alternate hypotheses:-

- H0: The mean time spent on the new page is equal to the mean time spent on the old page by users
- Ha: The mean time spent on the new page is greater than mean time spent on the old page by users

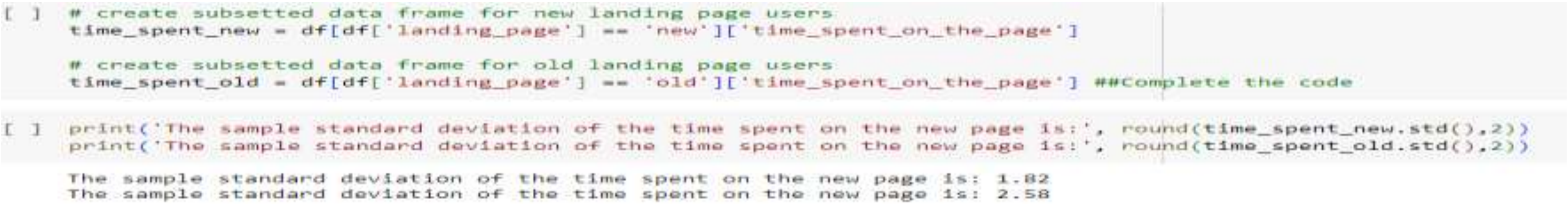
Step 2: Select Appropriate test

- 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



```
[ ]  # create subsetted data frame for new landing page users
     time_spent_new = df[df['landing_page'] == 'new']['time_spent_on_the_page']

     # create subsetted data frame for old landing page users
     time_spent_old = df[df['landing_page'] == 'old']['time_spent_on_the_page'] ##Complete the code
```

```
[ ]  print('The sample standard deviation of the time spent on the new page is:', round(time_spent_new.std(),2))
     print('The sample standard deviation of the time spent on the new page is:', round(time_spent_old.std(),2))

     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 new page is: 2.58
```

- Based on the assumptions of the t-test it has following datapoints to be considered:-

1. The time spent on the pages is measured on a continuous scale.
2. The populations are assumed to be a normal distributed populations.
3. As the standard deviation are different the population standard deviation assumed to be different.

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.

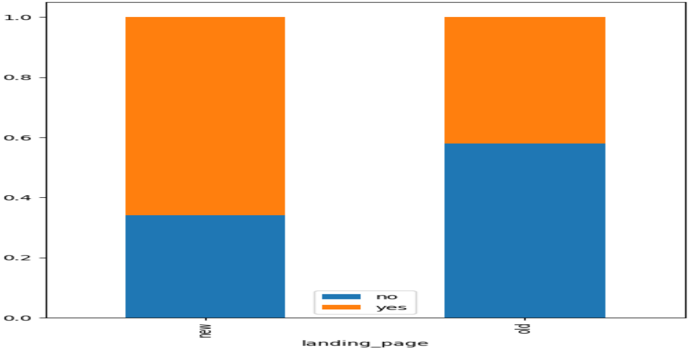
Step 7: Draw inference

- We can see that p-value is below its significance level of 5% he null hypothesis is rejected. This concludes that there is significance evidence that the mean time spent on the new landing page is greater than mean time spent on the old page by users.

Hypotheses Tested and Results

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?

Visual Analysis:-



Observation:-

- We have observed here that users on the new landing page show that they are more likely to convert the subscribers than users on the old landing page.

Hypothesis Testing Details

Step 1: Define the null and alternate hypotheses

- H0:The conversion rate of new page is equal to the conversion of the old page
- Ha: The conversion rate of new page is greater than the conversion of the old page

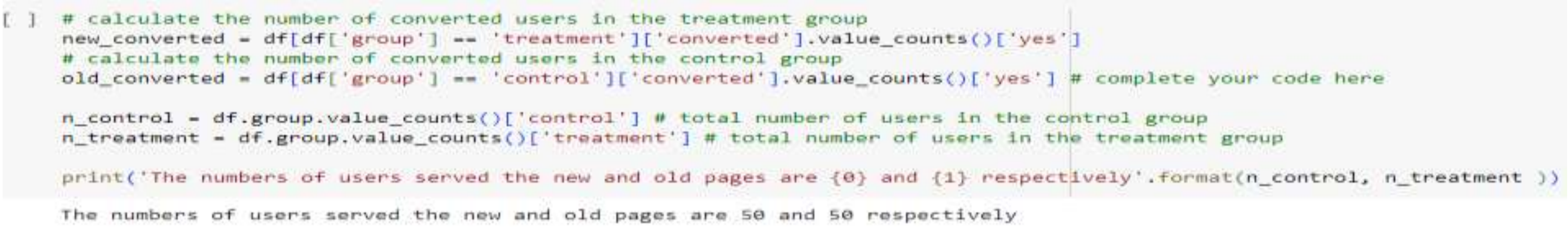
Step 2: Select Appropriate test

- 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 appropriate test**.

Step 3: Decide the significance level

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

Step 4: Collect and prepare data



```
[ ]  # calculate the number of converted users in the treatment group
     new_converted = df[df['group'] == 'treatment']['converted'].value_counts()['yes']
     # calculate the number of converted users in the control group
     old_converted = df[df['group'] == 'control']['converted'].value_counts()['yes'] # complete your code here

     n_control = df.group.value_counts()['control'] # total number of users in the control group
     n_treatment = df.group.value_counts()['treatment'] # total number of users in the treatment group

     print('The numbers of users served the new and old pages are {0} and {1} respectively'.format(n_control, n_treatment ))

     The numbers of users served the new and old pages are 50 and 50 respectively
```

Two proportion z-test assumptions:-

- It is either converted or not converted so it proves that it is binomially distributed population.
- It is also accepted that it is a simple random sampling from the population.

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.

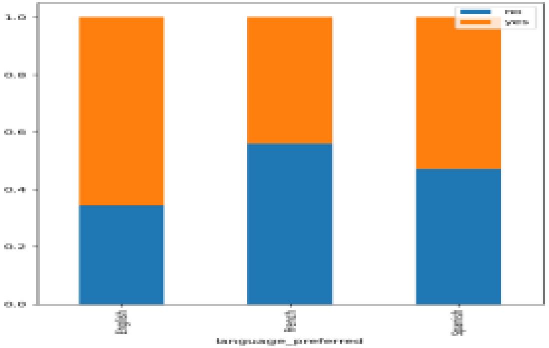
Step 7: Draw inference

- We can see that p-value is less than the significance level of 5% so the null hypothesis is rejected. This means that there is significant evidence that the conversion rate of new page is greater than the old page by users.

Hypotheses Tested and Results

3. Does the converted status depend on the preferred language?

Visual Analysis:-



Observation:-

- As it shows in graph, it has two categorical values as in yes or no so it means that as per the language users are converted to new landing page or not ready to convert.

Hypothesis Testing Details

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.

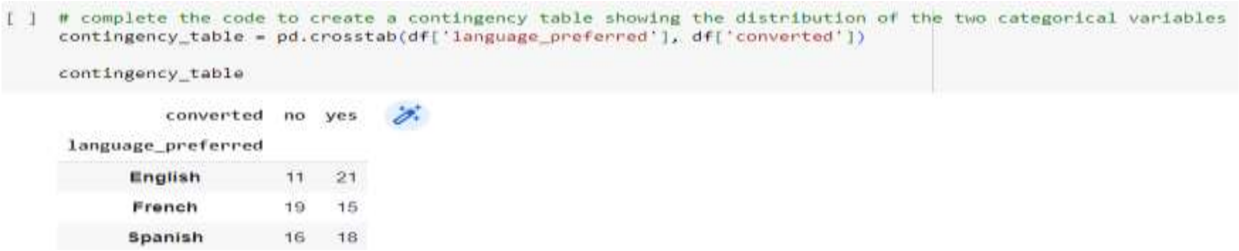
Step 2: Select Appropriate test

- 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 is most appropriate test.**

Step 3: Decide the significance level

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

Step 4: Collect and prepare data



```
# complete the code to create a contingency table showing the distribution of the two categorical variables
contingency_table = pd.crosstab(df['language_preferred'], df['converted'])

contingency_table
```

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

Chi-Square test for independence assumptions:-

- The expected value of sample observations in each level is greater than 5 so the assumptions of each level of the variable is atleast 5 is hence proved.
- We were informed that the collected sample is a simple random sampling.

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.

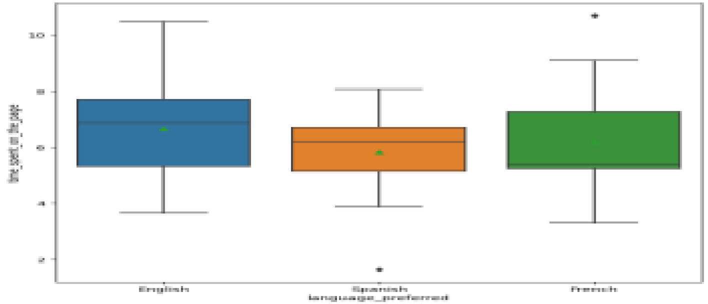
Step 7: Draw inference

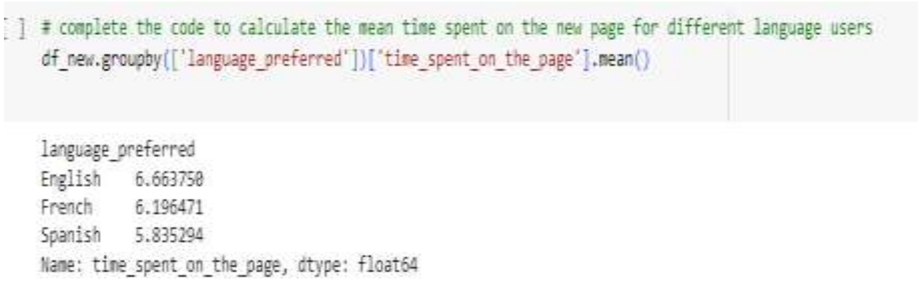
- We have observed that p-value is greater than the level of significance 5% the null hypothesis fails to be rejected. This means that the converted status is independent of the preferred language.

Hypothesis Tested and Results

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

Visual Analysis:-





```
[ ] # complete the code to calculate the mean time spent on the new page for different language users
    df_new.groupby(['language_preferred'])['time_spent_on_the_page'].mean()
```

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

Observation:-

- The average time spent on the page by language shows the maximum language by English and minimum time spent by Spanish language.

Hypothesis Testing Details

Step 1: Define the null and alternate hypotheses

- H0: The mean time spent on the new landing page is same across all preferred languages.
- Ha: Atleast one of the mean time spent on new landing page is different across the preferred language

Step 2: Select Appropriate test

- This is a problem, concerning three population means. **Based on this information, one way ANNOVA would be appropriate test to compare the three population means.**

Step 3: Decide the significance level

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

Step 5: Calculate the p-value

- The p-value is 0.46711357711340173

Step 6: Compare the p-value with α

- As the p-value 0.46711357711340173 is greater than the level of significance, we fail to reject the null hypothesis.

One-way ANNOVA test assumptions:-

- Populations variances are equal using Leavens test.
- Samples are informed as simple random sampling.

Step 7: Draw inference

- Since p-value is greater than the level o f significance i.e. 5% we fail to reject the null hypothesis so that means that variances are equal.

Happy Learning !