Name-: Shubham Neema

CWID-:20007579

Introduction-:

We have a dataset in which we will study whether sex of a person have any impact on the qualifications(UG or PG).We will apply Z proportionity test to check this condition.

In [5]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
#pd.plotting.register_matplotlib_converters() %matplotlib inline
#plt.style.use('seaborn-whitegrid')
#pd.set_option('display.max_columns',
```

In [24]:
```
data = pd.read_csv('Media vs mental health.csv')
data.head(20)
```

Out[24]:

	Age	Age Category	Sex:	Qualifications	Habitat	Media use(hrs/day)	Media category	Well-being score	Well-being category	anxie sco
0	30	Young	Female	PG	Urban Municipal Area	10	High	56	Middle	
1	27	Young	Male	PG	Urban Municipal Area	8	Middle	48	Middle	
2	25	Young	Male	PG	Urban Municipal Area	11	High	45	Low	
3	25	Young	Male	PG	Urban Municipal Area	12	High	50	Middle	
4	22	Young	Male	UG	Urban Municipal Area	8	Middle	41	Low	
5	23	Young	Male	PG	Rural	11	High	26	Low	

	Age	Age Category	Sex:	Qualifications	Habitat	Media use(hrs/day)	Media category	Well-being score	Well-being category	anxie sco
6	26	Young	Male	PG	Rural	7	Low	57	High	
7	25	Young	Male	PG	Rural	7	Low	34	Low	
8	57	Mid	Male	Phd/M.Phil	Urban Municipal Area	6	Low	54	Middle	
9	23	Young	Male	UG	Urban Municipal Area	9	Middle	55	Middle	
10	53	Mid	Female	PG	Metropolitan City	9	Middle	50	Middle	
11	27	Young	Male	PG	Metropolitan City	9	Middle	65	High	
12	27	Young	Male	PG	Rural	8	Middle	56	Middle	
13	24	Young	Female	PG	Metropolitan City	13	High	24	Low	
14	26	Young	Male	UG	Rural	10	High	56	Middle	
15	24	Young	Female	PG	Metropolitan City	9	Middle	57	High	
16	26	Young	Male	PG	Rural	10	High	67	High	
17	25	Young	Male	PG	Urban Municipal Area	9	Middle	47	Middle	
18	35	Young	Female	PG	Metropolitan City	8	Middle	51	Middle	
19	26	Young	Male	PG	Rural	10	High	45	Low	

In [10]:
```
print(f'Shape of the data: {data.shape}')
print(f'There are {data.shape[0]} rows in the dataset.')
```

```
Shape of the data: (426, 11)
There are 426 rows in the dataset.
```

In [11]:
```
data.columns
```

Out[11]:
```
Index(['Age', 'Age Category', 'Sex: ', 'Qualifications', 'Habitat',
       'Media use(hrs/day)', 'Media category', 'Well-being score',
       'Well-being category', 'anxiety score', 'anxiety category'],
      dtype='object')
```

In [26]:
```
data['Qualifications']=data['Qualifications'].replace(['Phd/M.Phil'],'PG')
data.head(4)
```

Out[26]:

	Age	Age Category	Sex:	Qualifications	Habitat	Media use(hrs/day)	Media category	Well-being score	Well-being category	anxiety score
0	30	Young	Female	PG	Urban Municipal Area	10	High	56	Middle	9
1	27	Young	Male	PG	Urban Municipal Area	8	Middle	48	Middle	17
2	25	Young	Male	PG	Urban Municipal Area	11	High	45	Low	18
3	25	Young	Male	PG	Urban Municipal Area	12	High	50	Middle	12

In [61]:
```
data.sex.unique()
```

Out[61]:
```
array(['Female', 'Male   '], dtype=object)
```

In [12]:
```
series = data.columns.to_series().groupby(data.dtypes).groups
series
```

Out[12]:
```
{int64: ['Age', 'Media use(hrs/day)', 'Well-being score', 'anxiety score'], object: ['Ag
e Category', 'Sex: ', 'Qualifications', 'Habitat', 'Media category', 'Well-being categor
y', 'anxiety category']}
```

In [19]:
```
dt = {k.name: v for k, v in series.items()}
attributes_by_datatype = pd.DataFrame(list(dt.values()), index = dt.keys(), columns = [
attributes_by_datatype
```

Out[19]:

	Attribute Set 1	Attribute Set 2	Attribute Set 3	Attribute Set 4	Attribute Set 5	Attribute Set 6	Attribute Set 7
int64	Age	Media use(hrs/day)	Well-being score	anxiety score	None	None	None
object	Age Category	Sex:	Qualifications	Habitat	Media category	Well-being category	anxiety category

In [ ]:
```
sorted(data['sex'].unique())
```

In [28]:
```
n_UG = data['Qualifications'].value_counts()[1]
n_PG = data['Qualifications'].value_counts()[0]
print([n_UG, n_PG])
```

```
[113, 313]
```

In [66]:

```
n_females = data['sex'].value_counts()[1] # number of females in the data
n_males = data['sex'].value_counts()[0] # number of females in the data
print([n_females, n_males])
```

```
[190, 236]
```

In [56]:
```
import math
female_postgraduation = len(data[(data.sex =='Female') & (data.Qualifications=='PG')])
```

In [62]:
```
male_postgraduation = len(data[(data.sex=='Male   ') & (data.Qualifications=='PG')])
```

In [63]:
```
print([female_postgraduation,male_postgraduation])
```

```
[155, 158]
```

In [65]:
```
print([female_postgraduation,male_postgraduation],[n_females,n_males])
```

```
[155, 158] [190, 236]
```

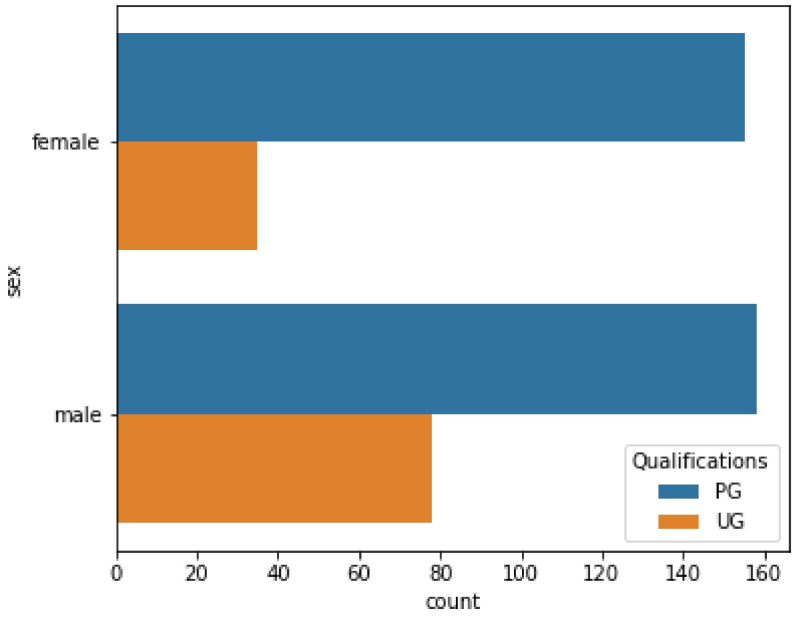
In [69]:
```
print('Proportion of female post graduates are', (female_postgraduation/n_females)*100,
```

```
Proportion of female post graduates are 81.57894736842105 % and that of male graduates a
re  66.94915254237289
```

Number of female post graduates are significantly higher than male post graduates.According to this data more females opt for PG than Males

In [70]:
```
plt.figure(figsize=(6,5))
chart = sns.countplot(y='sex', hue = 'Qualifications', data = data)
chart.set_yticklabels(['female','male'])
```

Out[70]:
```
[Text(0, 0, 'female'), Text(0, 1, 'male')]
```



In [71]:
```
from statsmodels.stats.proportion import proportions_ztest
stat, pval = proportions_ztest([female_postgraduation , male_postgraduation], [n_female
print(f'Statistic: {stat}\np_value: {pval}')
```

```
Statistic: 3.3998867146741216
p_value: 0.000674137769826977
```

In [74]:
```
if pval < 0.05:
  print(f'With a p-value of {pval} the difference is significant. Thus, We reject the N
else:
  print(f'With a p-value of {pval} the difference is not significant. Thus, We fail to
```

```
With a p-value of 0.000674137769826977 the difference is significant. Thus, We reject th
e Null Hypothesis.
```

In [ ]:
```

```

Analytical Observations after performing Hypothesis Testing

1. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.
2. Since the pval<0.05 we reject the null hypothesis and state that at 5% significance level, the proportion of PG in males and females are equal.
3. Hence, proportion of PG in males and females are significantly different

In [ ]:
```

```

ANOVA

To see, if the well being score across person with PG or UG degree are same or not. Analysis of variance (ANOVA)

ANOVA is a hypothesis testing technique tests the equality of two or more population means by examining the variances of samples that are taken. ANOVA tests the general rather than specific differences among means.

Assumptions of ANOVA

All populations involved follow a normal distribution

All populations have the same variance

The samples are randomly selected and independent of one another

Step1: Formulate the null hypothesis H0 and the alternate hypothesis HA

H0 : μ0 = μ1 = μ2

HA : Atleast one of the means are different.

Where:

μi = Population mean of well being score of people having UG and PG degree.

Step 2: Select appropriate statistical test and the corresponding test statistic

We select one way ANOVA as our test and mean well being score of the different groups as our test statistic.

Step 3: Choose the level of significance α

#we Select α = 0.005

Step 4: Collect data and calculate value of test statistic

Here we have 3 groups. Analysis of variance can determine whether the means of three or more groups are different.

ANOVA uses F-tests to statistically test the equality of means.

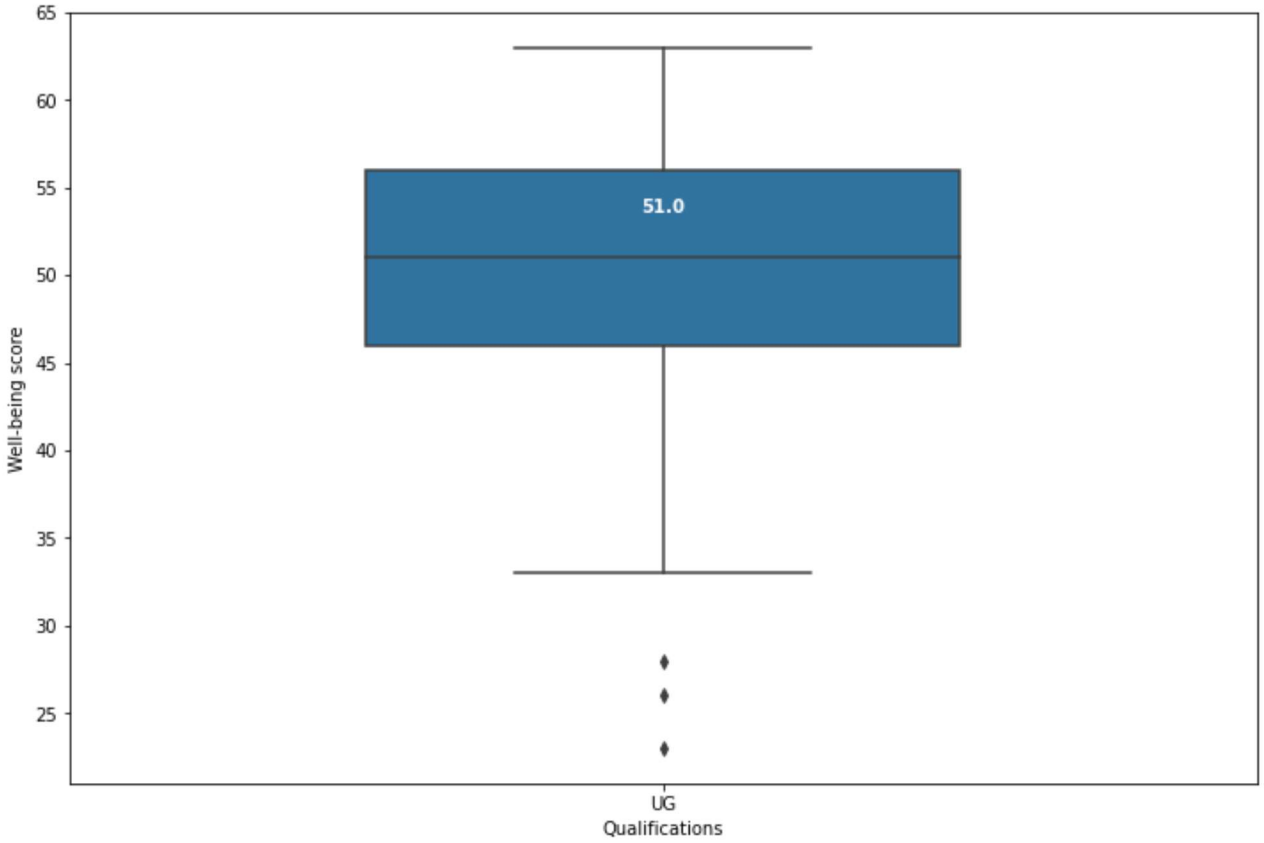
data.rename(columns = {'Well-being score':'score'}, inplace = True)

In [109…
```
data.head()
```

Out[109…

	Age	Age Category	sex	Qualifications	Habitat	media use	Media category	score	Well-being category	anxiety score	anxie catego
0	30	Young	Female	PG	Urban Municipal Area	10	High	56	Middle	9	Lc
1	27	Young	Male	PG	Urban Municipal Area	8	Middle	48	Middle	17	Lc
2	25	Young	Male	PG	Urban Municipal Area	11	High	45	Low	18	Lc
3	25	Young	Male	PG	Urban Municipal Area	12	High	50	Middle	12	Lc
4	22	Young	Male	UG	Urban Municipal Area	8	Middle	41	Low	10	Lc

In [102…
```
fig = plt.figure(figsize=(12, 8))
box_plot = sns.boxplot(x = "Qualifications", y = "Well-being score", data = df, width =
medians = df.groupby(['Qualifications'])['Well-being score'].median().round(2)
vertical_offset = df['Well-being score'].median() * 0.05 # offset from median for displ
medians
for xtick in box_plot.get_xticks():
  box_plot.text(xtick, medians[xtick] + vertical_offset,medians[xtick],
    horizontalalignment='center',color='w',weight='semibold')
```



In [110…

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
mod = ols('score ~ Qualifications', data = data).fit()
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)
```

```
                    sum_sq     df         F    PR(>F)
Qualifications    170.783527    1.0  2.161401  0.142257
Residual        33502.448868  424.0       NaN       NaN
```

Analytical approach after perfroming ANOVA We Fail to Reject the null hypothesis that for well being score for 2 groups of Qualifications having UG and PG. mean Score of both groups are equal. Hence, the distribution of well being score across under graduates and post graduates are same.

Determine which mean(s) is / are different An ANOVA test will test that at least one mean is different. We have falied to reject the null hypothesis but do not know which mean(s) is / are different. We use Tukey-krammer HSD test to detect which mean(s) is / are different.

In [111…

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
print(pairwise_tukeyhsd(data['score'], data['Qualifications']))
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj   lower  upper  reject
---------------------------------------------------
    PG     UG  -1.4342 0.1423 -3.3517 0.4833  False
---------------------------------------------------
```

Here reject is false, means they are not significantly different because they are equal.

In [ ]: