In [ ]:
```
# import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy
```

In [ ]:
```
# Load dataset
kashti = sns.load_dataset('titanic')
kashti.head()
```

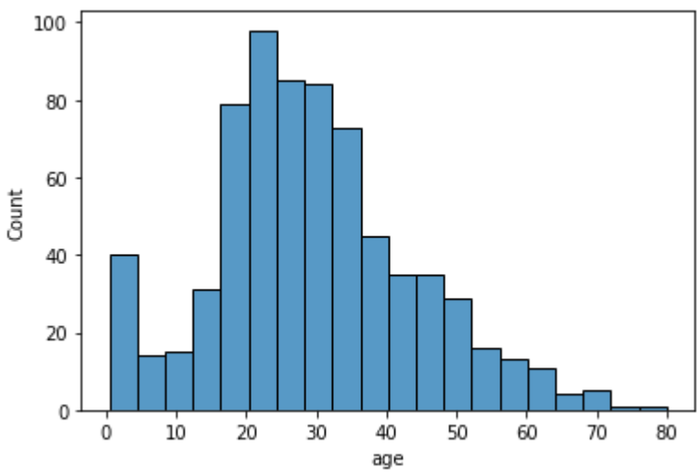
Out[ ]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN

Check Normal Distribution (Guassian)

In [ ]:
```
# Make histogram
sns.histplot(kashti['age'])
```

Out[ ]:
```
<AxesSubplot:xlabel='age', ylabel='Count'>
```



In [ ]:
```
# Shapiro Wilk Test
from scipy.stats import shapiro
shapiro(kashti['age'])

if p > 0.05:
    print('Probably Guassian')
else:
    print('Probably Not Guassian')
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-6-8e2b268d8224> in <module>
      3 shapiro(kashti['age'])
      4
----> 5 if p > 0.05:
      6     print('Probably Guassian')
      7 else:

NameError: name 'p' is not defined
```

In [ ]:
```
kashti.isnull().sum()
```

Out[ ]:
```
survived         0
pclass           0
```

```
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

In [ ]:
```
# drop nan of age
kashti1 = kashti.dropna(subset=['age'], axis=0)
```

In [ ]:
```
kashti1.isnull().sum()
```

Out[ ]:
```
survived         0
pclass           0
sex              0
age              0
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           530
embark_town      2
alive            0
alone            0
dtype: int64
```

In [ ]:
```
# Shapiro Wilk Test
from scipy.stats import shapiro
shapiro(kashti['age'])
p = p_value
if p > 0.05:
    print('Probably Guassian')
else:
    print('Probably Not Guassian')
```

```
Probably Not Guassian
```

In [ ]:
```
# Separate 3 columns age,sex & fare
df = kashti1[['sex','age','fare']]
df.head()
```

Out[ ]:

	sex	age	fare
0	male	22.0	7.2500
1	female	38.0	71.2833
2	female	26.0	7.9250
3	female	35.0	53.1000
4	male	35.0	8.0500

In [ ]:
```
# t. test to compare the age of male vs females

#-1 import libraries
from scipy.stats import ttest_ind
#-2 subsets of male and female
df_male = df[df['sex']=='male']
df_female = df[df['sex']=='female']
#-3 t.test(un-paired, as contain 2 samples or independent t-test)
ttest_ind(df_male['age'], df_female['age'])
stat, p_value = ttest_ind(df_male['age'], df_female['age']) # stored
print('stat=', stat, 'p=', p_value) # to show
#-4 conditional loop, different or not
if p_value > 0.05:
    print("There is no significant difference")
else:
    print("There is a significant difference")
```

```
stat= 2.499206354920835 p= 0.012671296797013709
There is a significant difference
```

In [ ]:
```
# t. test for One  sample Value

#-1 import libraries
from scipy.stats import ttest_1samp
#-2 subsets of male and female
df_male = df[df['sex']=='male']
df_female = df[df['sex']=='female']
#-3 t.test(un-paired, as contain 2 samples or independent t-test)
ttest_1samp(df_male['age'], 36)
stat, p_value = ttest_1samp(df_male['age'], 36) # stored
print('stat=', stat, 'p=', p_value) # to show
#-4 conditional loop, different or not
if p_value > 0.05:
    print("There is no significant difference")
else:
    print("There is a significant difference")
```

```
stat= -7.646511009251602 p= 1.2523613407424712e-13
There is a significant difference
```

Assignment

In [ ]:
```
df1 = kashti1[['sex','class']]
df1.head()
```

Out[ ]:

	sex	class
0	male	Third
1	female	First
2	female	Third
3	female	First
4	male	Third

In [ ]:

```
# t. test to compare the age of male vs class

#-1 import libraries
from scipy.stats import ttest_rel
#-2 subsets of male and female
df1_male = df1[df1['sex']=='male']
df1_1stclass = df1[df1['class']=='First']
#-3 t.test(un-paired, as contain 2 samples or independent t-test)
ttest_rel(df1_1stclass['class'], df_male['age'])
stat, p_value = ttest_rel(df_male['age'], df1_1stclass['class']) # stored
print('stat=', stat, 'p=', p_value) # to show
#-4 conditional loop, different or not
if p_value > 0.05:
    print("There is no significant difference")
else:
    print("There is a significant difference")
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-17-c979f21a7195> in <module>
      7 df1_1stclass = df1[df1['class']=='First']
      8 #-3 t.test(un-paired, as contain 2 samples or independent t-test)
----> 9 ttest_rel(df1_1stclass['class'], df_male['age'])
     10 stat, p_value = ttest_rel(df_male['age'], df1_1stclass['class']) # stored
     11 print('stat=', stat, 'p=', p_value) # to show

c:\Users\kalee\anaconda3\lib\site-packages\scipy\stats\stats.py in ttest_rel(a, b, ax
is, nan_policy, alternative)
   5889     nb = _get_len(b, axis, "second argument")
   5890     if na != nb:
-> 5891         raise ValueError('unequal length arrays')
   5892
   5893     if na == 0:

ValueError: unequal length arrays
```

In [ ]: