Seaborn EDA on Titanic Dataset

July 16, 2022

[1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

[2]:
```
df = pd.read_csv('titanic.csv', index_col = 0)
df
```

[2]:
```
             Survived  Pclass  \
PassengerId
1                   0       3
2                   1       1
3                   1       3
4                   1       1
5                   0       3
…               …     …
887                 0       2
888                 1       1
889                 0       3
890                 1       1
891                 0       3

                                                          Name     Sex   Age  \
PassengerId
1                                      Braund, Mr. Owen Harris    male  22.0
2            Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0
3                                       Heikkinen, Miss. Laina  female  26.0
4                 Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
5                                     Allen, Mr. William Henry    male  35.0
…                                                        …     …   …
887                                      Montvila, Rev. Juozas    male  27.0
888                               Graham, Miss. Margaret Edith  female  19.0
889                   Johnston, Miss. Catherine Helen "Carrie"  female   NaN
890                                      Behr, Mr. Karl Howell    male  26.0
891                                        Dooley, Mr. Patrick    male  32.0

             SibSp  Parch            Ticket     Fare Cabin Embarked
```

```
PassengerId
1                1      0         A/5 21171   7.2500   NaN        S
2                1      0          PC 17599  71.2833   C85        C
3                0      0  STON/O2. 3101282   7.9250   NaN        S
4                1      0            113803  53.1000  C123        S
5                0      0            373450   8.0500   NaN        S
…            …    …               …      …   …      …
887              0      0            211536  13.0000   NaN        S
888              0      0            112053  30.0000   B42        S
889              1      2       W./C. 6607  23.4500   NaN        S
890              0      0            111369  30.0000  C148        C
891              0      0            370376   7.7500   NaN        Q

[891 rows x 11 columns]
```

[3]:
```
df.dtypes
```

[3]:
```
Survived      int64
Pclass        int64
Name         object
Sex          object
Age         float64
SibSp         int64
Parch         int64
Ticket       object
Fare        float64
Cabin        object
Embarked     object
dtype: object
```

[20]:
```
r = df.select_dtypes(exclude = object)
r
```

[20]:
```
             Survived  Pclass   Age  SibSp  Parch     Fare
PassengerId
1                   0       3  22.0      1      0   7.2500
2                   1       1  38.0      1      0  71.2833
3                   1       3  26.0      0      0   7.9250
4                   1       1  35.0      1      0  53.1000
5                   0       3  35.0      0      0   8.0500
…               …     …   …    …    …      …
887                 0       2  27.0      0      0  13.0000
888                 1       1  19.0      0      0  30.0000
889                 0       3   NaN      1      2  23.4500
890                 1       1  26.0      0      0  30.0000
891                 0       3  32.0      0      0   7.7500
```

```
[891 rows x 6 columns]
```

[90]:

```python
df.set_index('Pclass', inplace = True, append = True, drop = False)
df
```

[90]:

```
                    Survived  Pclass  \
PassengerId Pclass
1           3              0       3
2           1              1       1
3           3              1       3
4           1              1       1
5           3              0       3
…                       …     …
887         2              0       2
888         1              1       1
889         3              0       3
890         1              1       1
891         3              0       3

                                                                  Name     Sex  \
PassengerId Pclass
1           3                                  Braund, Mr. Owen Harris    male
2           1       Cumings, Mrs. John Bradley (Florence Briggs Th…   female
3           3                                   Heikkinen, Miss. Laina  female
4           1             Futrelle, Mrs. Jacques Heath (Lily May Peel)  female
5           3                                 Allen, Mr. William Henry    male
…                                                                …     …
887         2                                    Montvila, Rev. Juozas    male
888         1                             Graham, Miss. Margaret Edith  female
889         3                 Johnston, Miss. Catherine Helen "Carrie"  female
890         1                                    Behr, Mr. Karl Howell    male
891         3                                      Dooley, Mr. Patrick    male

                     Age  SibSp  Parch            Ticket     Fare Cabin  \
PassengerId Pclass
1           3       22.0      1      0         A/5 21171   7.2500   NaN
2           1       38.0      1      0          PC 17599  71.2833   C85
3           3       26.0      0      0  STON/O2. 3101282   7.9250   NaN
4           1       35.0      1      0            113803  53.1000  C123
5           3       35.0      0      0            373450   8.0500   NaN
…                  …    …    …                …      …   …
887         2       27.0      0      0            211536  13.0000   NaN
888         1       19.0      0      0            112053  30.0000   B42
889         3        NaN      1      2        W./C. 6607  23.4500   NaN
890         1       26.0      0      0            111369  30.0000  C148
891         3       32.0      0      0            370376   7.7500   NaN
```

```
                   Embarked
PassengerId Pclass
1           3             S
2           1             C
3           3             S
4           1             S
5           3             S
…                      …
887         2             S
888         1             S
889         3             S
890         1             C
891         3             Q

[891 rows x 11 columns]
```

[4]:

```
c = df.corr()     # Correlation Matrix
c
```

[4]:

```
          Survived    Pclass       Age     SibSp     Parch      Fare
Survived  1.000000 -0.338481 -0.077221 -0.035322  0.081629  0.257307
Pclass   -0.338481  1.000000 -0.369226  0.083081  0.018443 -0.549500
Age      -0.077221 -0.369226  1.000000 -0.308247 -0.189119  0.096067
SibSp    -0.035322  0.083081 -0.308247  1.000000  0.414838  0.159651
Parch     0.081629  0.018443 -0.189119  0.414838  1.000000  0.216225
Fare      0.257307 -0.549500  0.096067  0.159651  0.216225  1.000000
```

[5]:

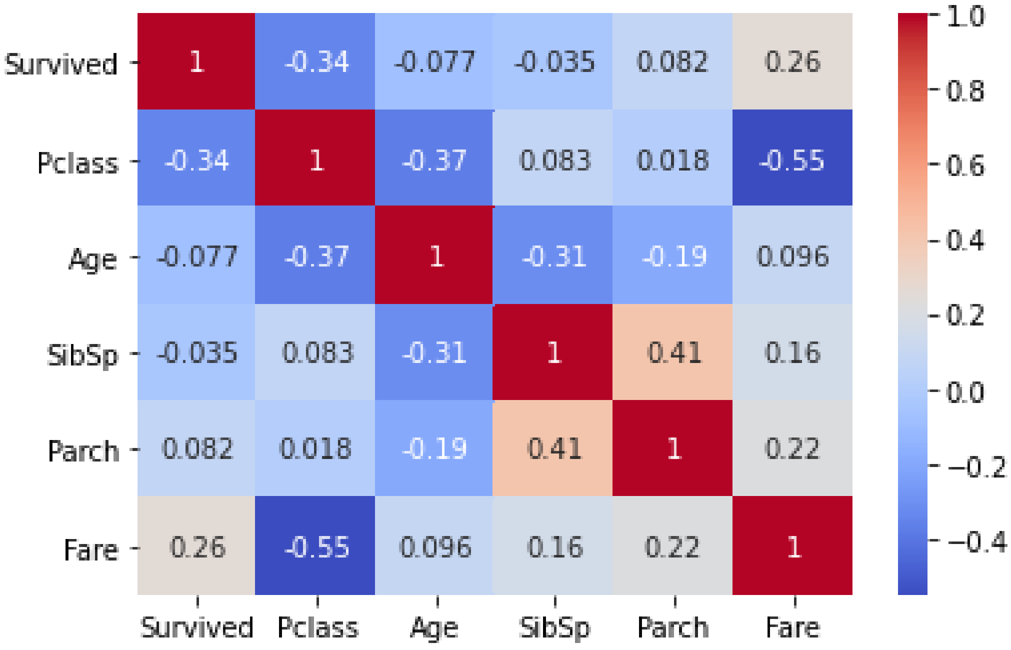
```
df.corr?
```

[12]:

```
sns.heatmap(c, annot = True, cmap = 'coolwarm')        # linewidth = 1
```

[12]:

```
<AxesSubplot:>
```



[22]:

```
sns.heatmap?
```

[25]:

```
df['Fare'].max()
```

```
[25]: 512.3292
```

[33]:

```
df['Fare'].mean()
```

```
[33]: 32.204207968574636
```

[36]:

```
df[df['Fare'] <= df['Fare'].mean()]
```

```
[36]:             Survived  Pclass                                    Name  \
PassengerId
1                   0       3                 Braund, Mr. Owen Harris
3                   1       3                  Heikkinen, Miss. Laina
5                   0       3                Allen, Mr. William Henry
6                   0       3                        Moran, Mr. James
8                   0       3          Palsson, Master. Gosta Leonard
…                …      …                                          …
887                 0       2                   Montvila, Rev. Juozas
888                 1       1            Graham, Miss. Margaret Edith
889                 0       3  Johnston, Miss. Catherine Helen "Carrie"
890                 1       1                   Behr, Mr. Karl Howell
891                 0       3                     Dooley, Mr. Patrick
```

```
                Sex   Age  SibSp  Parch            Ticket     Fare Cabin  \
PassengerId
1              male  22.0      1      0         A/5 21171   7.2500   NaN
3            female  26.0      0      0  STON/O2. 3101282   7.9250   NaN
5              male  35.0      0      0            373450   8.0500   NaN
6              male   NaN      0      0            330877   8.4583   NaN
8              male   2.0      3      1            349909  21.0750   NaN
…             …   …     …     …                 …       …   …
887            male  27.0      0      0            211536  13.0000   NaN
888          female  19.0      0      0            112053  30.0000   B42
889          female   NaN      1      2        W./C. 6607  23.4500   NaN
890            male  26.0      0      0            111369  30.0000  C148
891            male  32.0      0      0            370376   7.7500   NaN

            Embarked
PassengerId
1                  S
3                  S
5                  S
6                  Q
8                  S
…                …
887                S
888                S
889                S
890                C
891                Q

[680 rows x 11 columns]
```

[34]:
```
df[df['Fare'] >= df['Fare'].mean()]   # Query
```

[34]:
```
             Survived  Pclass  \
PassengerId
2                   1       1
4                   1       1
7                   0       1
24                  1       1
28                  0       1
…                 …      …
857                 1       1
864                 0       3
868                 0       1
872                 1       1
880                 1       1
```

```
                                                          Name     Sex   Age  \
PassengerId
2            Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0
4                 Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
7                                      McCarthy, Mr. Timothy J    male  54.0
24                                Sloper, Mr. William Thompson    male  28.0
28                              Fortune, Mr. Charles Alexander    male  19.0
…                                                        …     …   …
857                 Wick, Mrs. George Dennick (Mary Hitchcock)  female  45.0
864                          Sage, Miss. Dorothy Edith "Dolly"  female   NaN
868                       Roebling, Mr. Washington Augustus II    male  31.0
872           Beckwith, Mrs. Richard Leonard (Sallie Monypeny)  female  47.0
880              Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)  female  56.0

             SibSp  Parch    Ticket       Fare        Cabin Embarked
PassengerId
2                1      0  PC 17599   71.2833          C85        C
4                1      0    113803   53.1000         C123        S
7                0      0     17463   51.8625          E46        S
24               0      0    113788   35.5000           A6        S
28               3      2     19950  263.0000  C23 C25 C27        S
…            …    …       …       …          …      …
857              1      1     36928  164.8667          NaN        S
864              8      2  CA. 2343   69.5500          NaN        S
868              0      0  PC 17590   50.4958          A24        S
872              1      1     11751   52.5542          D35        S
880              0      1     11767   83.1583          C50        C

[211 rows x 11 columns]
```

```
[29]: df[df['Fare'] == 0]    # Query
```

```
[29]:             Survived  Pclass                              Name   Sex   Age  \
PassengerId
180                 0       3               Leonard, Mr. Lionel  male  36.0
264                 0       1             Harrison, Mr. William  male  40.0
272                 1       3      Tornquist, Mr. William Henry  male  25.0
278                 0       2       Parkes, Mr. Francis "Frank"  male   NaN
303                 0       3   Johnson, Mr. William Cahoone Jr  male  19.0
414                 0       2    Cunningham, Mr. Alfred Fleming  male   NaN
467                 0       2             Campbell, Mr. William  male   NaN
482                 0       2  Frost, Mr. Anthony Wood "Archie"  male   NaN
598                 0       3               Johnson, Mr. Alfred  male  49.0
634                 0       1     Parr, Mr. William Henry Marsh  male   NaN
675                 0       2        Watson, Mr. Ennis Hastings  male   NaN
733                 0       2              Knight, Mr. Robert J  male   NaN
807                 0       1            Andrews, Mr. Thomas Jr  male  39.0
```

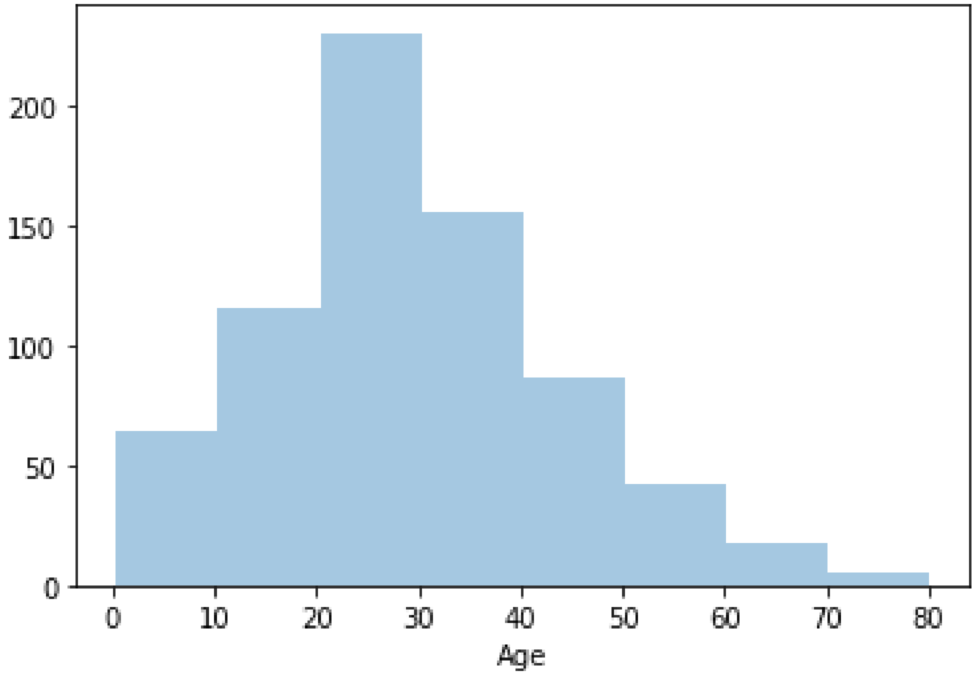
```
816                 0       1                  Fry, Mr. Richard  male   NaN
823                 0       1   Reuchlin, Jonkheer. John George  male  38.0

             SibSp  Parch  Ticket  Fare Cabin Embarked
PassengerId
180              0      0    LINE   0.0   NaN        S
264              0      0  112059   0.0   B94        S
272              0      0    LINE   0.0   NaN        S
278              0      0  239853   0.0   NaN        S
303              0      0    LINE   0.0   NaN        S
414              0      0  239853   0.0   NaN        S
467              0      0  239853   0.0   NaN        S
482              0      0  239854   0.0   NaN        S
598              0      0    LINE   0.0   NaN        S
634              0      0  112052   0.0   NaN        S
675              0      0  239856   0.0   NaN        S
733              0      0  239855   0.0   NaN        S
807              0      0  112050   0.0   A36        S
816              0      0  112058   0.0  B102        S
823              0      0   19972   0.0   NaN        S
```

[20]:

```python
sns.distplot(df['Age'], bins = 8, kde = False)    # Distribution Plot
```

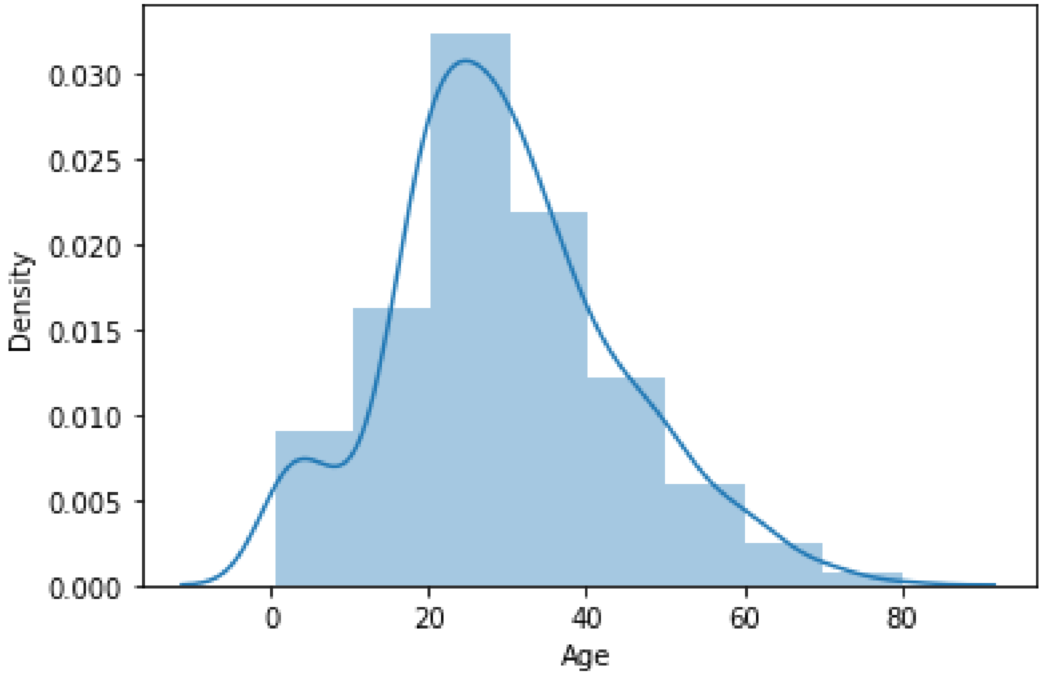
[20]:

```
<AxesSubplot:xlabel='Age'>
```



```
[21]: sns.distplot(df['Age'], bins = 8, kde = True)     # Distribution Plot
```

```
[21]: <AxesSubplot:xlabel='Age', ylabel='Density'>
```



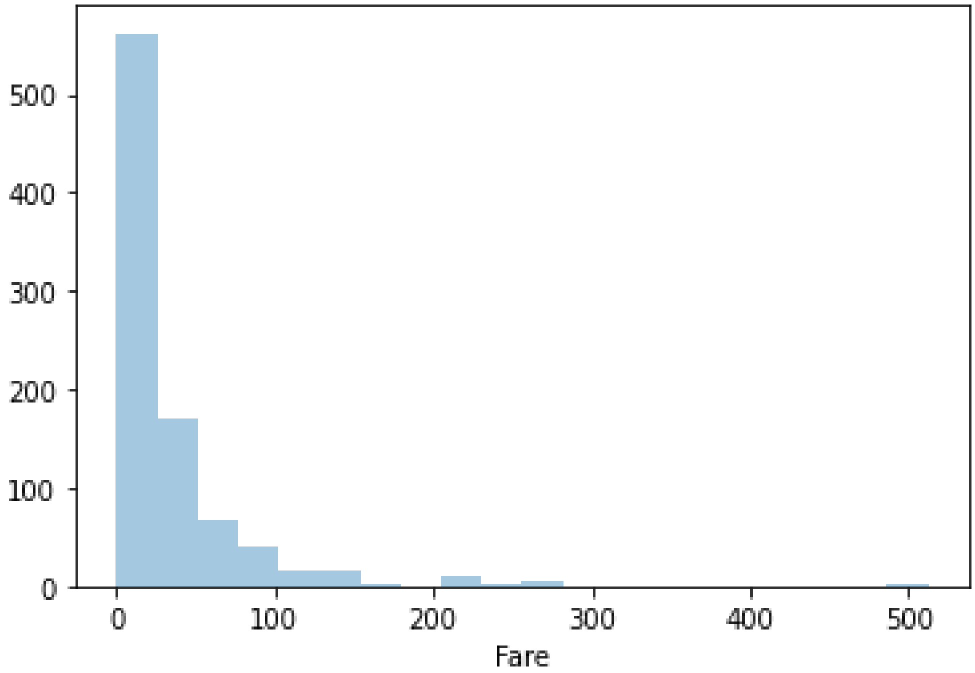
```
[1]: sns.displot(df['Fare'], bins = 20, kde = False)     # Distribution Plot
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_13876/2126389053.py in <module>
----> 1 sns.displot(df['Fare'], bins = 20, kde = False)     # Distribution Plot

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

```
[53]: sns.displot?
```

```
[44]: sns.distplot(df['Fare'], bins = 20, kde = False, vertical = False)     #␣
       ↪Distribution Plot
```

```
[44]: <AxesSubplot:xlabel='Fare'>
```



[43]:

```
sns.distplot?
```

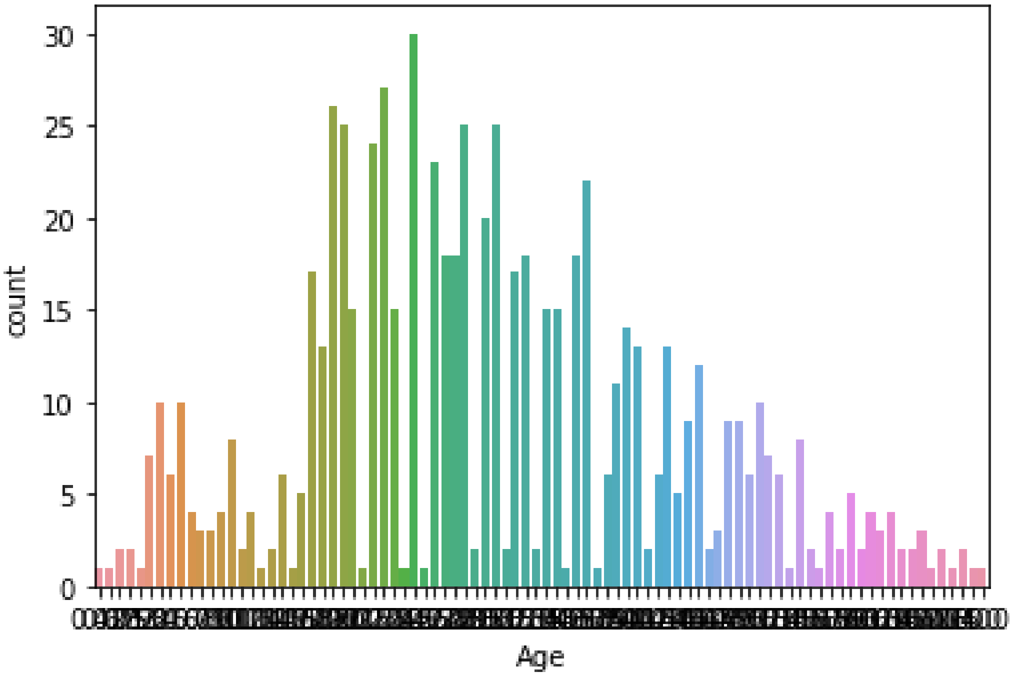
[58]:

```
sns.countplot(df['Age'], hue = df['Embarked'])
```

```
d:\installed softwares\python\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```

[58]:

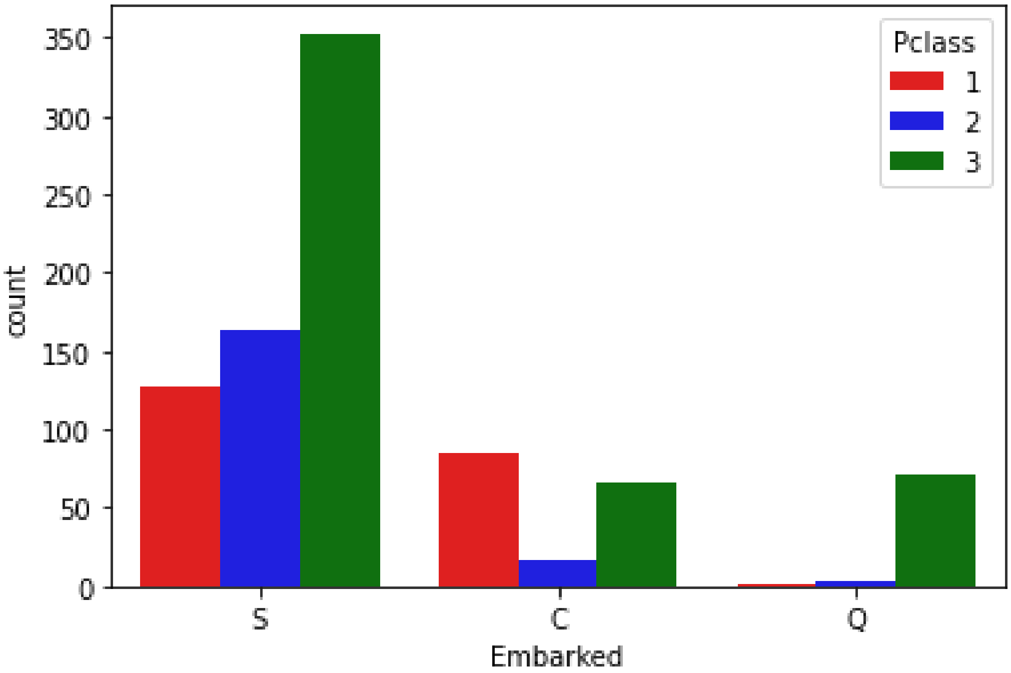
```
<AxesSubplot:xlabel='Age', ylabel='count'>
```



```
[34]: sns.countplot(df['Embarked'], hue = df['Pclass'], dodge = True, palette =␣
       ↪['red', 'blue', 'green'])   #color = 'red'
```

```
d:\installed softwares\python\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```

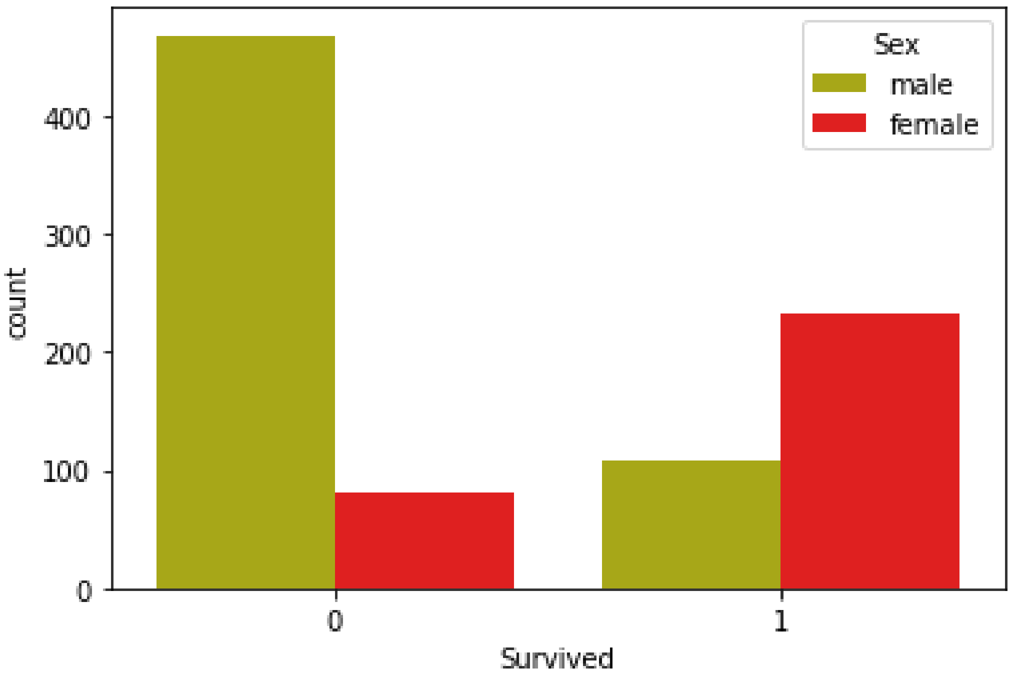
```
[34]: <AxesSubplot:xlabel='Embarked', ylabel='count'>
```



[4]:
```
sns.countplot(df['Survived'], hue = df['Sex'], dodge = True, palette = ['y', 'r'])    #Combine survived column also...
```

```
d:\installed softwares\python\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```

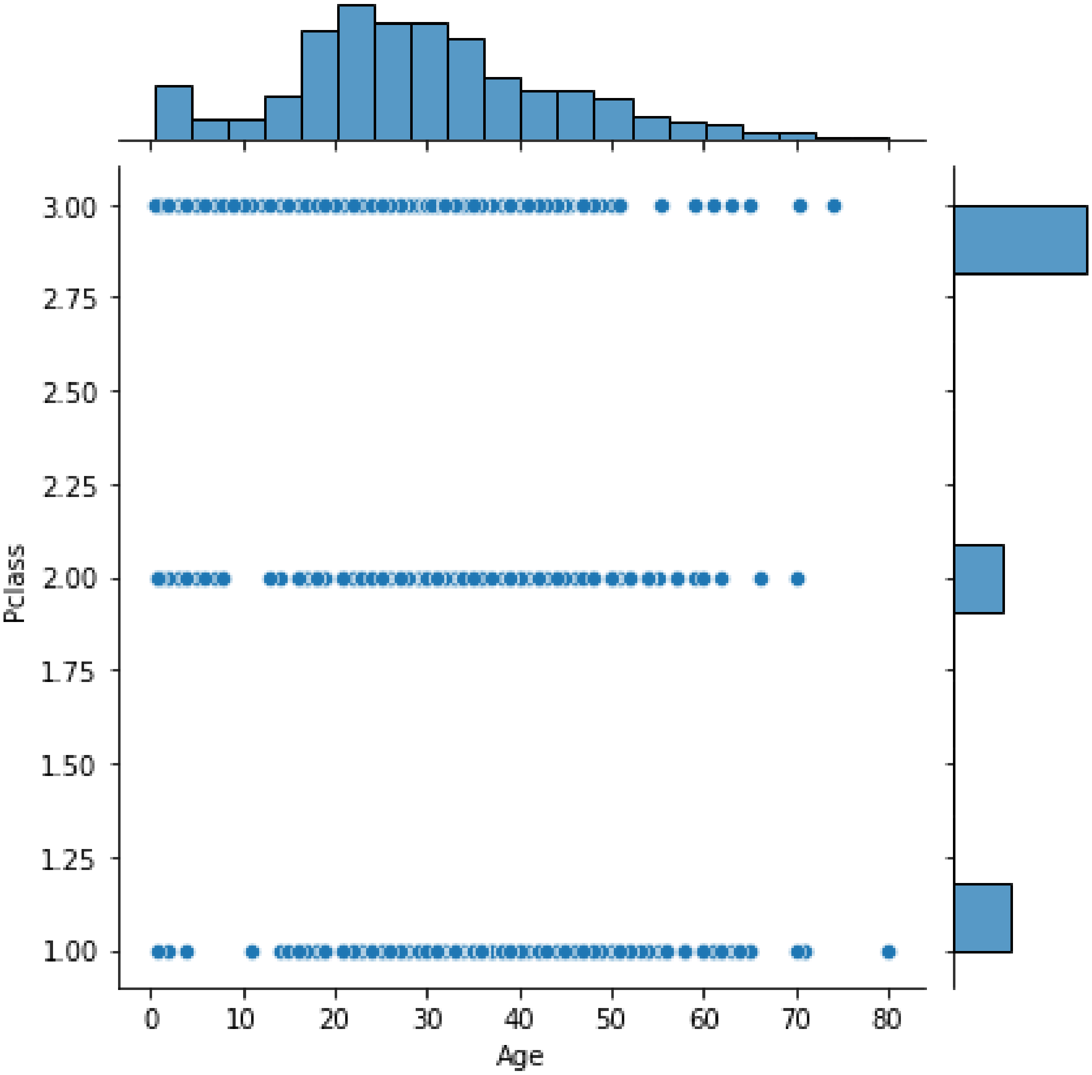
[4]:
```
<AxesSubplot:xlabel='Survived', ylabel='count'>
```



[38]:
```
sns.jointplot?
```

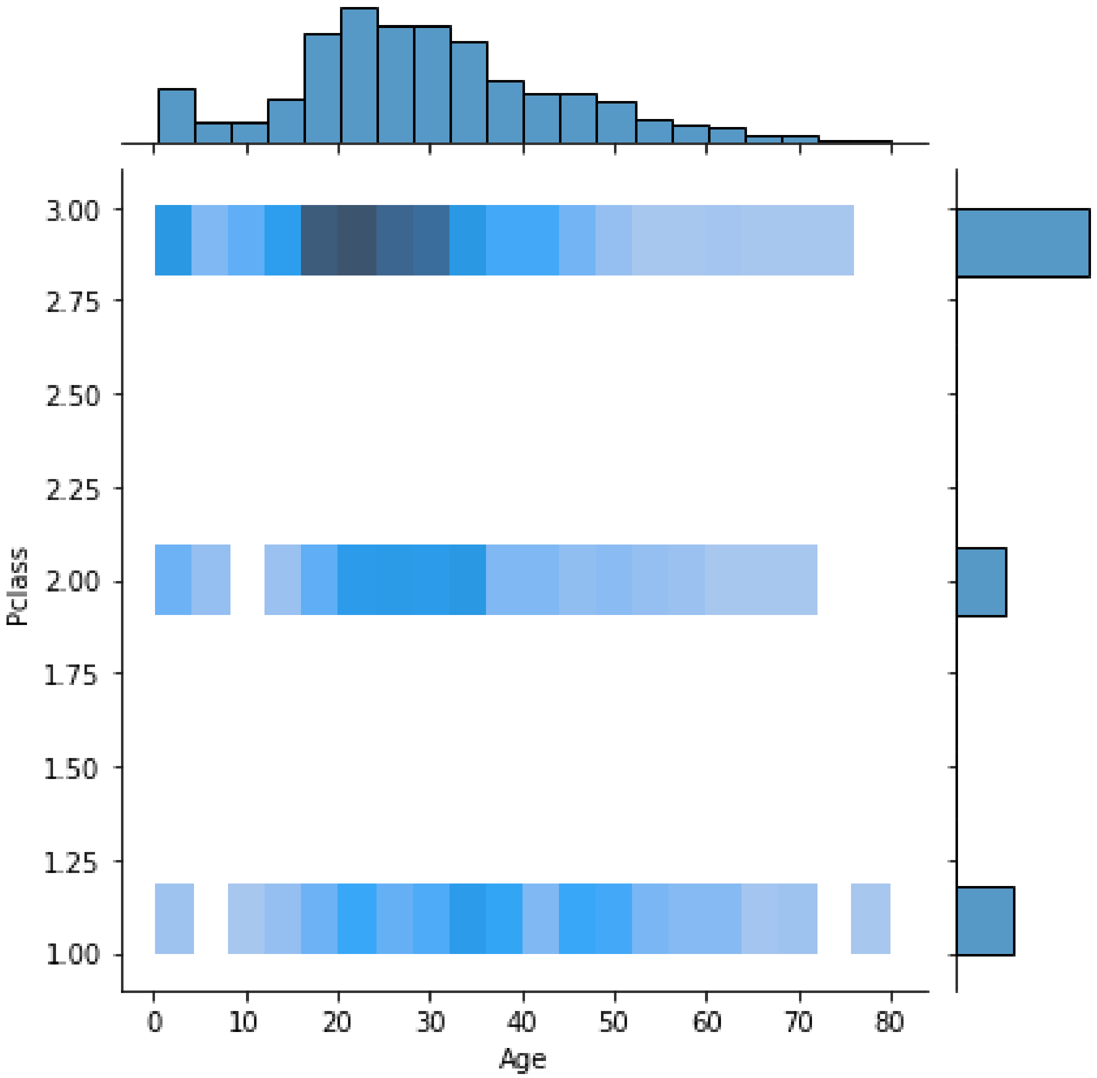
[36]:
```
sns.jointplot(x = df['Age'], y = df['Pclass'])
```

[36]:
```
<seaborn.axisgrid.JointGrid at 0x2354192ca60>
```



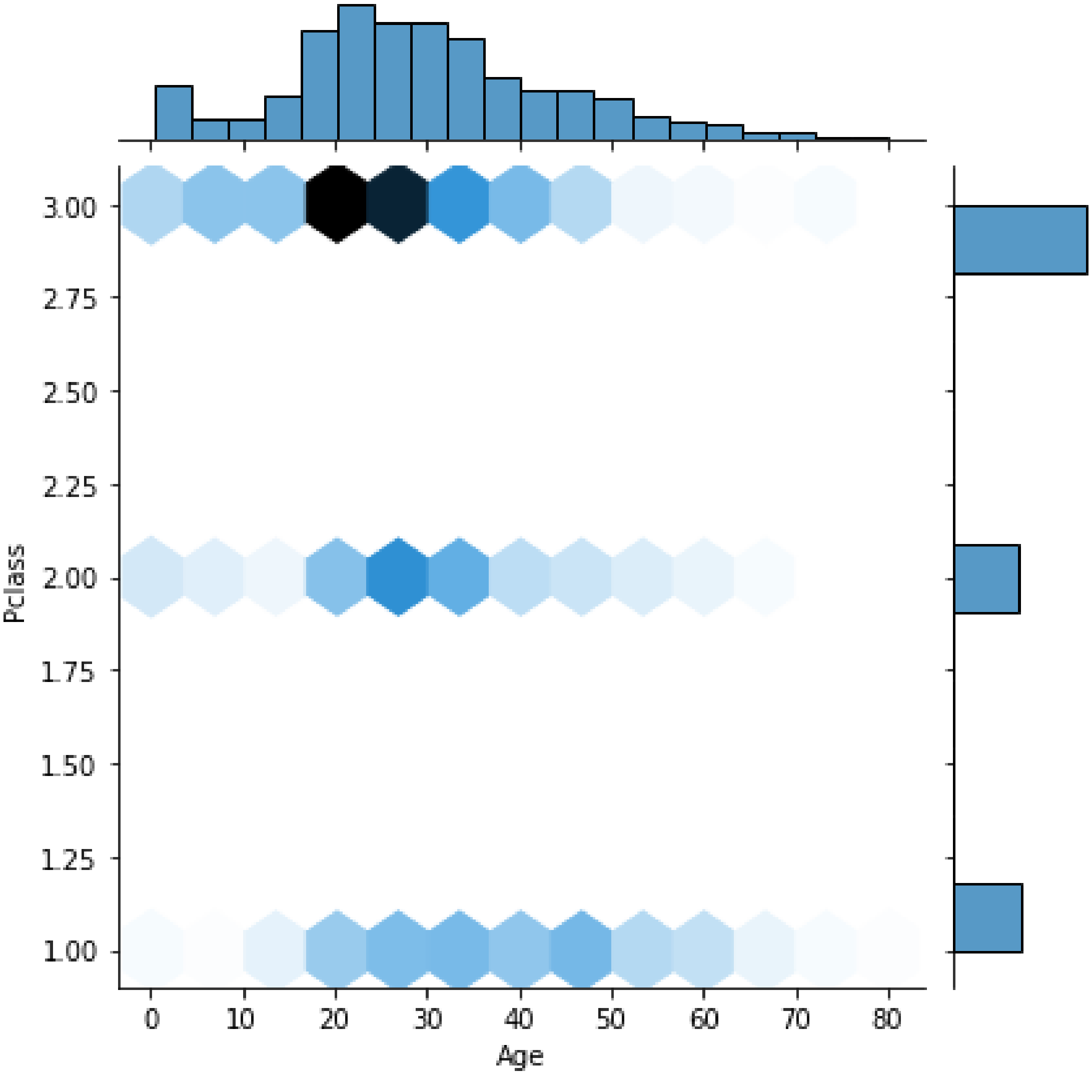
```
[37]: sns.jointplot(x = df['Age'], y = df['Pclass'], kind = 'hist')
```

```
[37]: <seaborn.axisgrid.JointGrid at 0x235420d8b80>
```



```
[41]: sns.jointplot(x = df['Age'], y = df['Pclass'], kind = 'hex')
```

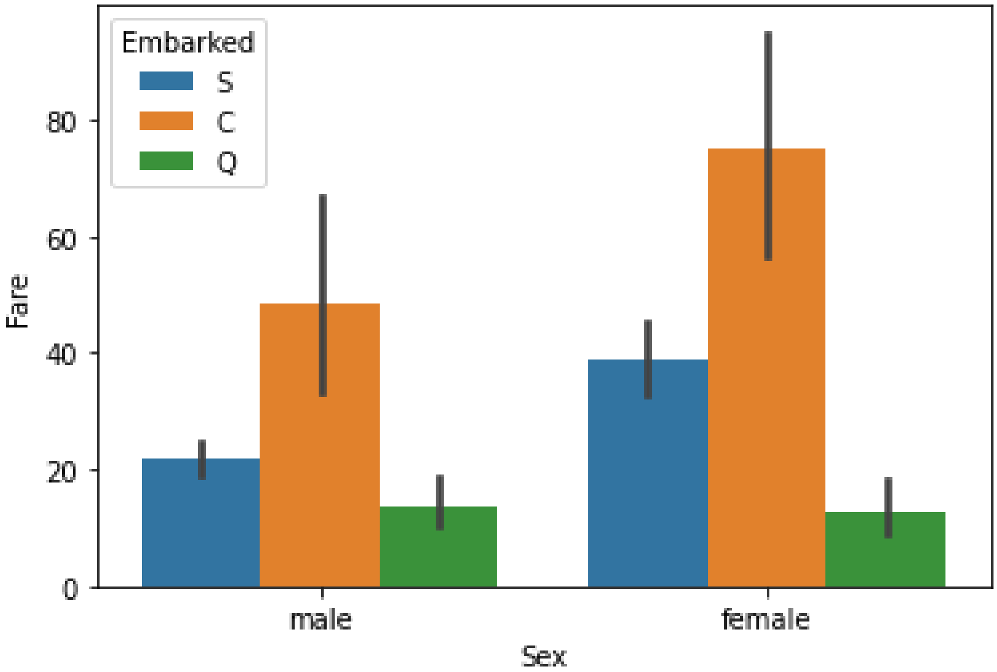
```
[41]: <seaborn.axisgrid.JointGrid at 0x235423d5be0>
```



```
[21]: sns.jointplot?
```

```
[45]: sns.barplot(y = df['Fare'], x = df['Sex'], hue = df['Embarked'])
      # black lines are showing uncertainity in the variable data of the barplot␣
       ↪which might there because of missing data,
      # maybe getting effected from there columns
```

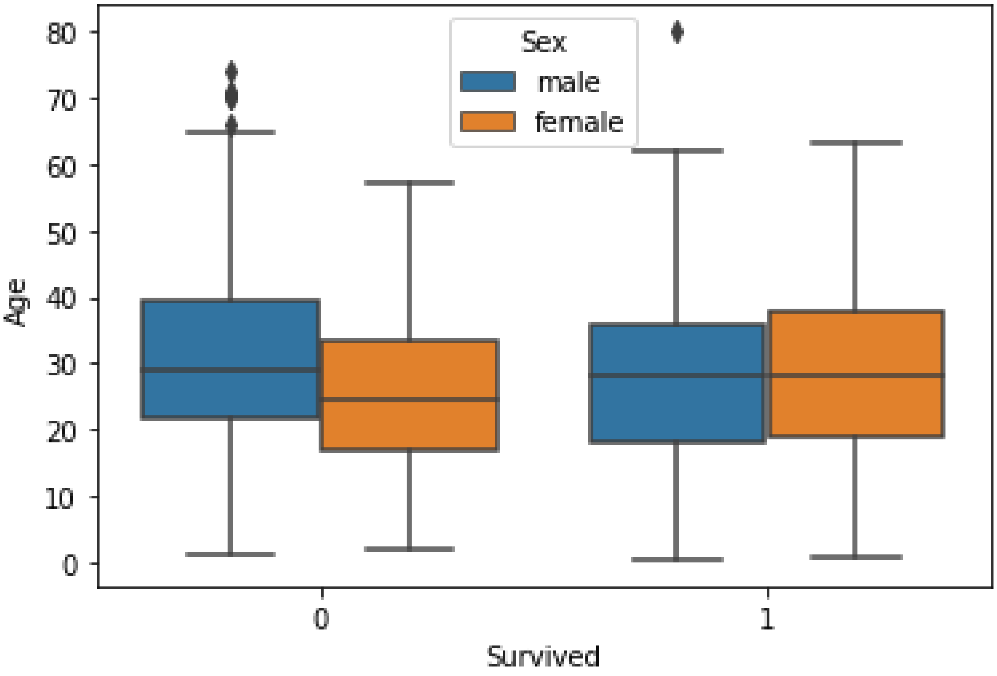
```
[45]: <AxesSubplot:xlabel='Sex', ylabel='Fare'>
```



```
[46]: sns.barplot?
```

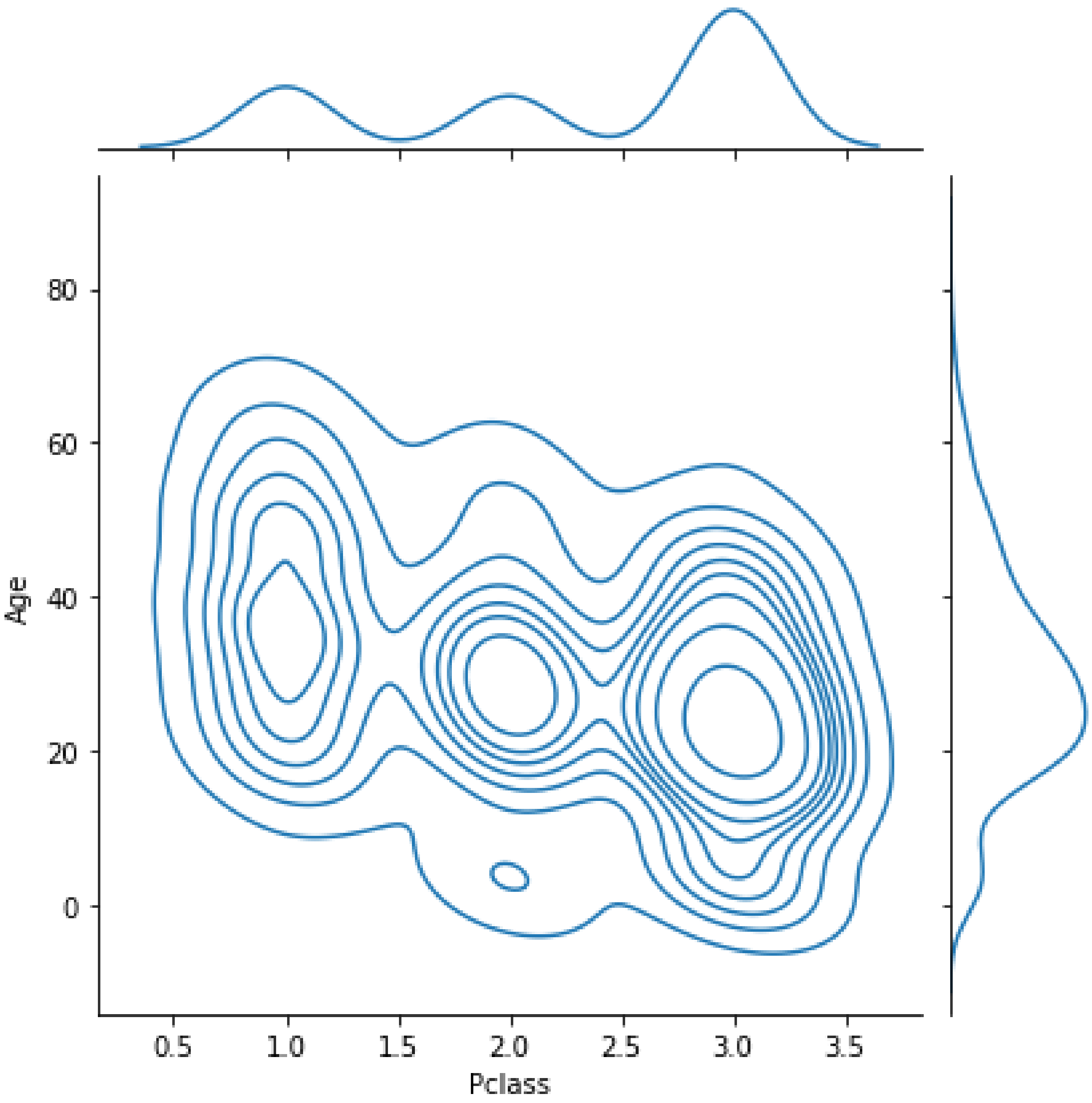
```
[56]: sns.boxplot(y = df['Age'], x = df['Survived'], hue = df['Sex'])
```

```
[56]: <AxesSubplot:xlabel='Survived', ylabel='Age'>
```



[52]:
```
sns.jointplot(x = df['Pclass'], y = df['Age'], kind = 'kde')
```

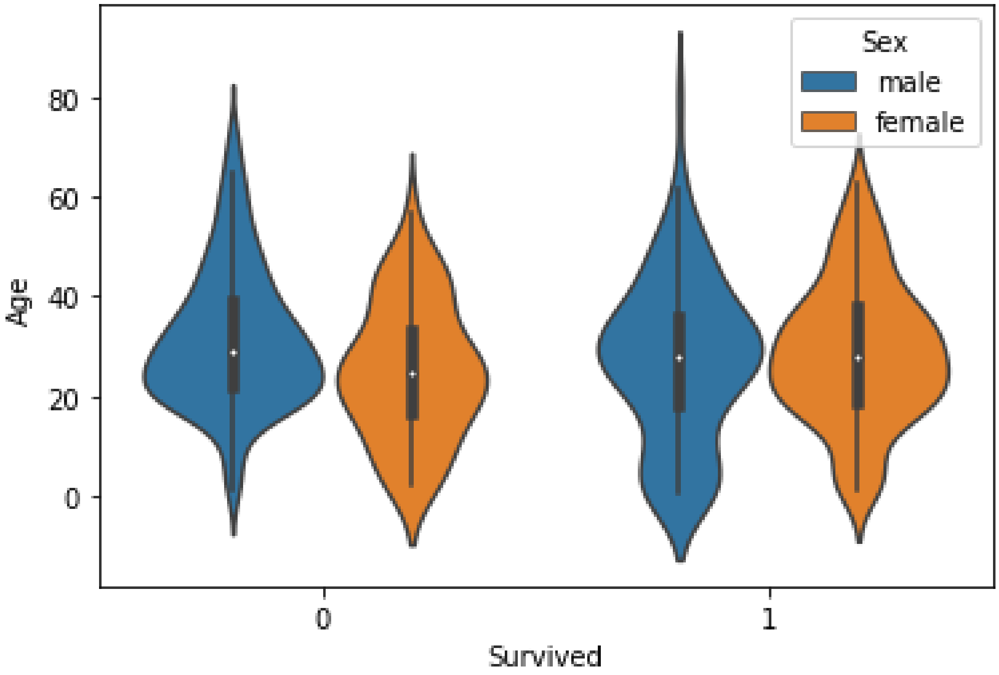
[52]:
```
<seaborn.axisgrid.JointGrid at 0x24682c2e730>
```



[62]:
```
sns.violinplot?
```

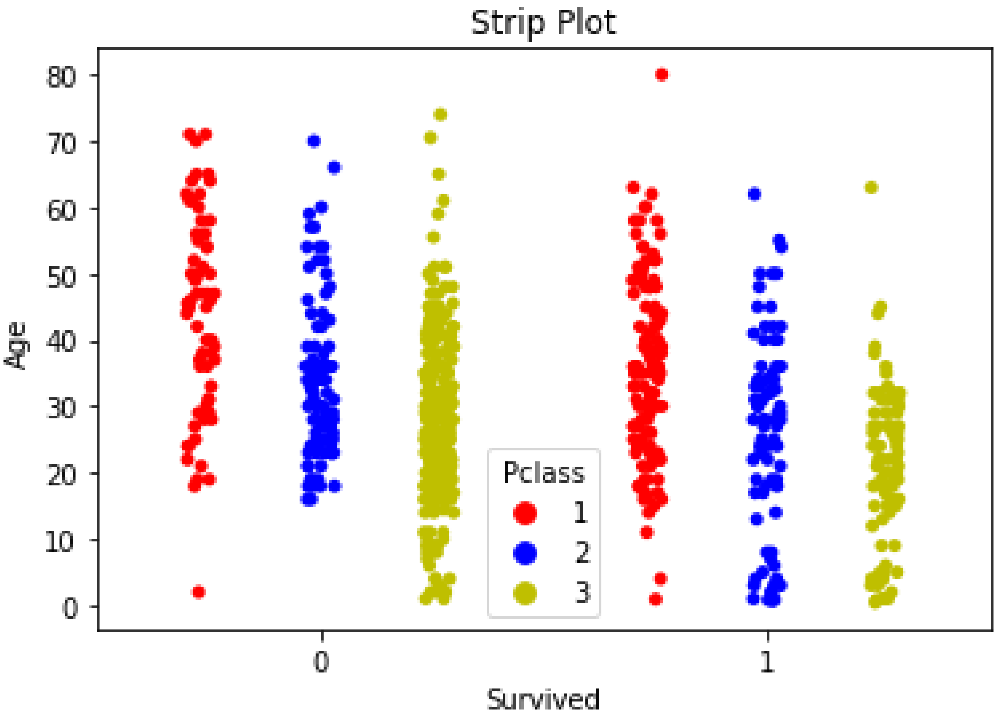
[66]:
```
sns.violinplot(x = df['Survived'], y = df['Age'], hue = df['Sex'], inner =␣
↪'box')
```

[66]:
```
<AxesSubplot:xlabel='Survived', ylabel='Age'>
```



```
[71]: sns.stripplot(x = df['Survived'], y = df['Age'], jitter = True,  palette =␣
       ↪['r','b', 'y'], dodge = True, hue = df['Pclass'])
      plt.title("Strip Plot")
```

```
[71]: Text(0.5, 1.0, 'Strip Plot')
```



[66]:
```
sns.stripplot?
```

[ ]:
```

```