Breast Cancer Analysis - EDA and Python implention

DATA EXPLORING AND VISUALITION

LOAD A LIBRIARIES

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

LOAD A DATASET

In [2]:
```
import pandas as pd
breast_cancer=pd.read_csv("https://raw.githubusercontent.com/gayathriravi2111997/Assessment/main/breast_cancer.csv")
breast_cancer
```

Out[2]:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histology	ER status	PR status	HER2 status	Surgery_type	Date_of_Surgery	Date_of_Last_Visit	Patient_Status
0	TCGA-D8-A1XD	36	FEMALE	0.080353	0.42638	0.54715	0.273680	III	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Modified Radical Mastectomy	15-Jan-17	19-Jun-17	Alive
1	TCGA-EW-A1OX	43	FEMALE	-0.420320	0.57807	0.61447	-0.031505	II	Mucinous Carcinoma	Positive	Positive	Negative	Lumpectomy	26-Apr-17	09-Nov-18	Dead
2	TCGA-A8-A079	69	FEMALE	0.213980	1.31140	-0.32747	-0.234260	III	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	08-Sep-17	09-Jun-18	Alive
3	TCGA-D8-A1XR	56	FEMALE	0.345090	-0.21147	-0.19304	0.124270	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Modified Radical Mastectomy	25-Jan-17	12-Jul-17	Alive
4	TCGA-BH-A0BF	56	FEMALE	0.221550	1.90680	0.52045	-0.311990	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	06-May-17	27-Jun-19	Dead
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
329	TCGA-AN-A04A	36	FEMALE	0.231800	0.61804	-0.55779	-0.517350	III	Infiltrating Ductal Carcinoma	Positive	Positive	Positive	Simple Mastectomy	11-Nov-19	09-Feb-20	Dead
330	TCGA-A8-A085	44	MALE	0.732720	1.11170	-0.26952	-0.354920	II	Infiltrating Lobular Carcinoma	Positive	Positive	Negative	Other	01-Nov-19	04-Mar-20	Dead
331	TCGA-A1-A0SG	61	FEMALE	-0.719470	2.54850	-0.15024	0.339680	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Lumpectomy	11-Nov-19	18-Jan-21	Dead
332	TCGA-A2-A0EU	79	FEMALE	0.479400	2.05590	-0.53136	-0.188480	I	Infiltrating Ductal Carcinoma	Positive	Positive	Positive	Lumpectomy	21-Nov-19	19-Feb-21	Dead
333	TCGA-B6-A40B	76	FEMALE	-0.244270	0.92556	-0.41823	-0.067848	I	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Lumpectomy	11-Nov-19	05-Jan-21	Dead

334 rows × 16 columns

In [3]:
```
breast_cancer.head()
```

Out[3]:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histology	ER status	PR status	HER2 status	Surgery_type	Date_of_Surgery	Date_of_Last_Visit	Patient_Status
0	TCGA-D8-A1XD	36	FEMALE	0.080353	0.42638	0.54715	0.273680	III	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Modified Radical Mastectomy	15-Jan-17	19-Jun-17	Alive
1	TCGA-EW-A1OX	43	FEMALE	-0.420320	0.57807	0.61447	-0.031505	II	Mucinous Carcinoma	Positive	Positive	Negative	Lumpectomy	26-Apr-17	09-Nov-18	Dead
2	TCGA-A8-A079	69	FEMALE	0.213980	1.31140	-0.32747	-0.234260	III	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	08-Sep-17	09-Jun-18	Alive
3	TCGA-D8-A1XR	56	FEMALE	0.345090	-0.21147	-0.19304	0.124270	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Modified Radical Mastectomy	25-Jan-17	12-Jul-17	Alive
4	TCGA-BH-A0BF	56	FEMALE	0.221550	1.90680	0.52045	-0.311990	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	06-May-17	27-Jun-19	Dead

In [4]:
```
breast_cancer.tail()
```

Out[4]:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histology	ER status	PR status	HER2 status	Surgery_type	Date_of_Surgery	Date_of_Last_Visit	Patient_Status
329	TCGA-AN-A04A	36	FEMALE	0.23180	0.61804	-0.55779	-0.517350	III	Infiltrating Ductal Carcinoma	Positive	Positive	Positive	Simple Mastectomy	11-Nov-19	09-Feb-20	Dead
330	TCGA-A8-A085	44	MALE	0.73272	1.11170	-0.26952	-0.354920	II	Infiltrating Lobular Carcinoma	Positive	Positive	Negative	Other	01-Nov-19	04-Mar-20	Dead
331	TCGA-A1-A0SG	61	FEMALE	-0.71947	2.54850	-0.15024	0.339680	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Lumpectomy	11-Nov-19	18-Jan-21	Dead
332	TCGA-A2-A0EU	79	FEMALE	0.47940	2.05590	-0.53136	-0.188480	I	Infiltrating Ductal Carcinoma	Positive	Positive	Positive	Lumpectomy	21-Nov-19	19-Feb-21	Dead
333	TCGA-B6-A40B	76	FEMALE	-0.24427	0.92556	-0.41823	-0.067848	I	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Lumpectomy	11-Nov-19	05-Jan-21	Dead

data preprocessing

In [5]:
```
breast_cancer.shape
```

Out[5]:
```
(334, 16)
```

In [6]:
```
breast_cancer.columns
```

Out[6]:
```
Index(['Patient_ID', 'Age', 'Gender', 'Protein1', 'Protein2', 'Protein3',
       'Protein4', 'Tumour_Stage', 'Histology', 'ER status', 'PR status',
       'HER2 status', 'Surgery_type', 'Date_of_Surgery', 'Date_of_Last_Visit',
       'Patient_Status'],
      dtype='object')
```

In [7]:
```
breast_cancer.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 334 entries, 0 to 333
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Patient_ID          334 non-null    object
 1   Age                 334 non-null    int64
 2   Gender              334 non-null    object
 3   Protein1            334 non-null    float64
 4   Protein2            334 non-null    float64
 5   Protein3            334 non-null    float64
 6   Protein4            334 non-null    float64
 7   Tumour_Stage        334 non-null    object
 8   Histology           334 non-null    object
 9   ER status           334 non-null    object
 10  PR status           334 non-null    object
 11  HER2 status         334 non-null    object
 12  Surgery_type        334 non-null    object
 13  Date_of_Surgery     334 non-null    object
 14  Date_of_Last_Visit  317 non-null    object
 15  Patient_Status      321 non-null    object
dtypes: float64(4), int64(1), object(11)
memory usage: 41.9+ KB
```

In [8]:
```
breast_cancer.describe()
```

Out[8]:

	Age	Protein1	Protein2	Protein3	Protein4
count	334.000000	334.000000	334.000000	334.000000	334.000000
mean	58.886228	-0.029991	0.946896	-0.090204	0.009819
std	12.961212	0.563588	0.911637	0.585175	0.629055
min	29.000000	-2.340900	-0.978730	-1.627400	-2.025500
25%	49.000000	-0.358888	0.362173	-0.513748	-0.377090
50%	58.000000	0.006129	0.992805	-0.173180	0.041768
75%	68.000000	0.343598	1.627900	0.278353	0.425630
max	90.000000	1.593600	3.402200	2.193400	1.629900

In [9]:
```
breast_cancer.describe(include='all')
```

Out[9]:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histology	ER status	PR status	HER2 status	Surgery_type	Date_of_Surgery	Date_of_Last_Visit	Pat
count	334	334.000000	334	334.000000	334.000000	334.000000	334.000000	334	334	334	334	334	334	334	317	
unique	334	NaN	2	NaN	NaN	NaN	NaN	3	3	1	1	2	4	181	285	
top	TCGA-D8-A1XD	NaN	FEMALE	NaN	NaN	NaN	NaN	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	06-Nov-18	20-Feb-20	
freq	1	NaN	330	NaN	NaN	NaN	NaN	189	233	334	334	305	105	5	3	
mean	NaN	58.886228	NaN	-0.029991	0.946896	-0.090204	0.009819	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
std	NaN	12.961212	NaN	0.563588	0.911637	0.585175	0.629055	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
min	NaN	29.000000	NaN	-2.340900	-0.978730	-1.627400	-2.025500	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
25%	NaN	49.000000	NaN	-0.358888	0.362173	-0.513748	-0.377090	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
50%	NaN	58.000000	NaN	0.006129	0.992805	-0.173180	0.041768	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
75%	NaN	68.000000	NaN	0.343598	1.627900	0.278353	0.425630	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
max	NaN	90.000000	NaN	1.593600	3.402200	2.193400	1.629900	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

In [10]:
```
breast_cancer.isnull().sum()
```

Out[10]:
```
Patient_ID             0
Age                    0
Gender                 0
Protein1               0
Protein2               0
Protein3               0
Protein4               0
Tumour_Stage           0
Histology              0
ER status              0
PR status              0
HER2 status            0
Surgery_type           0
Date_of_Surgery        0
Date_of_Last_Visit    17
Patient_Status        13
dtype: int64
```

In [11]:
```
breast_cancer.nunique()
```

Out[11]:
```
Patient_ID            334
Age                    57
Gender                  2
Protein1              333
Protein2              334
Protein3              334
Protein4              333
Tumour_Stage            3
Histology               3
ER status               1
PR status               1
HER2 status             2
Surgery_type            4
Date_of_Surgery       181
Date_of_Last_Visit    285
Patient_Status          2
dtype: int64
```

In [12]:
```
breast_cancer.Gender.unique()
```

Out[12]:
```
array(['FEMALE', 'MALE'], dtype=object)
```

In [13]:
```
breast_cancer.Gender.value_counts()
```

Out[13]:
```
FEMALE    330
MALE        4
Name: Gender, dtype: int64
```

In [14]:
```
breast_cancer.Age.unique()
```

Out[14]:
```
array([36, 43, 69, 56, 84, 53, 50, 77, 40, 71, 72, 75, 52, 41, 37, 59, 62,
       74, 87, 45, 55, 79, 47, 61, 68, 49, 48, 46, 81, 89, 44, 78, 85, 60,
       57, 54, 76, 58, 67, 63, 82, 65, 73, 51, 83, 80, 39, 66, 42, 38, 64,
       29, 32, 70, 90, 35, 88], dtype=int64)
```

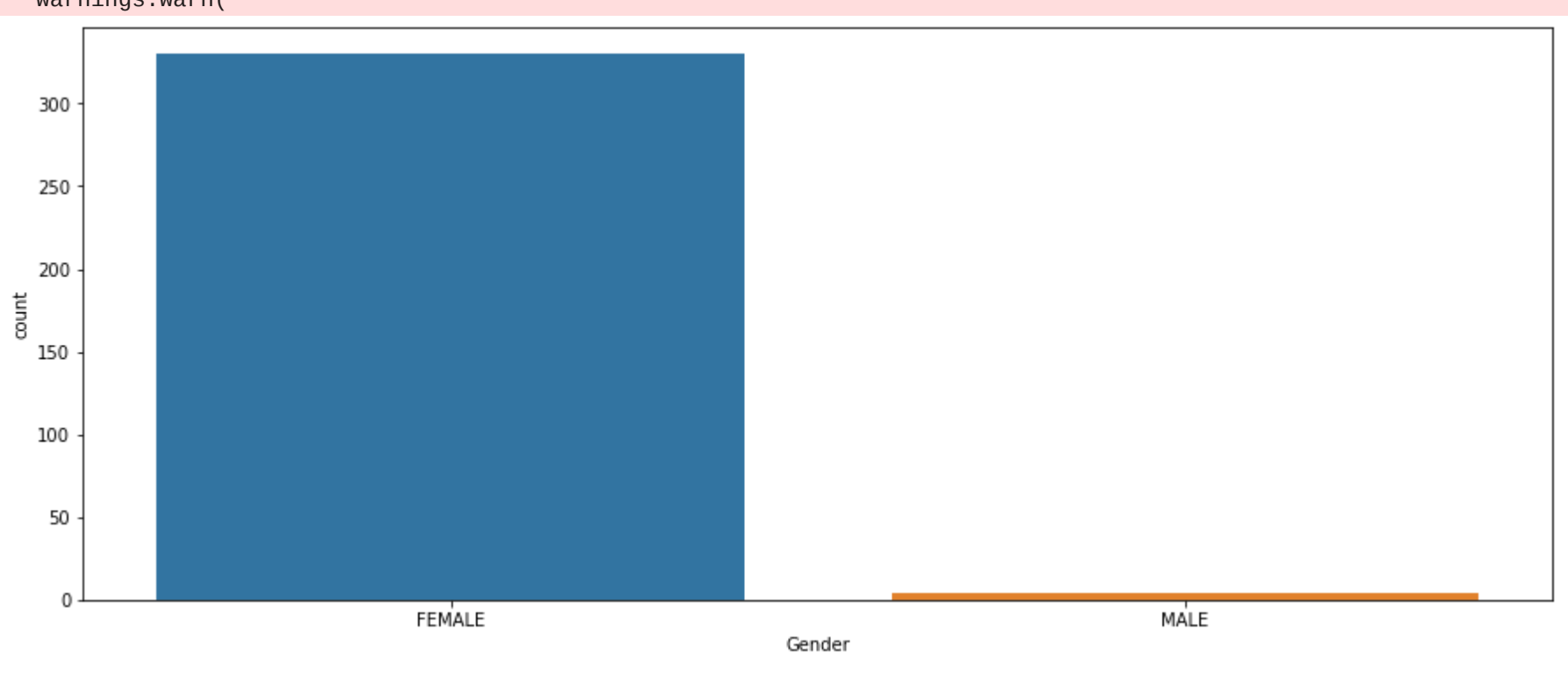
In [15]:
```
breast_cancer.Age.value_counts()
```

Out[15]:
```
59    15
50    14
63    14
54    13
56    13
46    12
62    11
53    10
68    10
60     9
49     9
47     9
45     9
51     9
52     8
61     8
77     8
66     8
71     7
48     7
41     6
74     6
58     6
64     6
65     6
79     6
40     6
57     6
44     5
76     5
42     5
80     4
78     4
85     4
84     4
69     4
55     4
36     3
39     3
38     3
73     3
88     3
67     3
43     3
75     3
72     3
82     2
83     2
87     2
37     2
29     2
70     2
89     1
32     1
90     1
35     1
81     1
Name: Age, dtype: int64
```

DATA VISULLATION

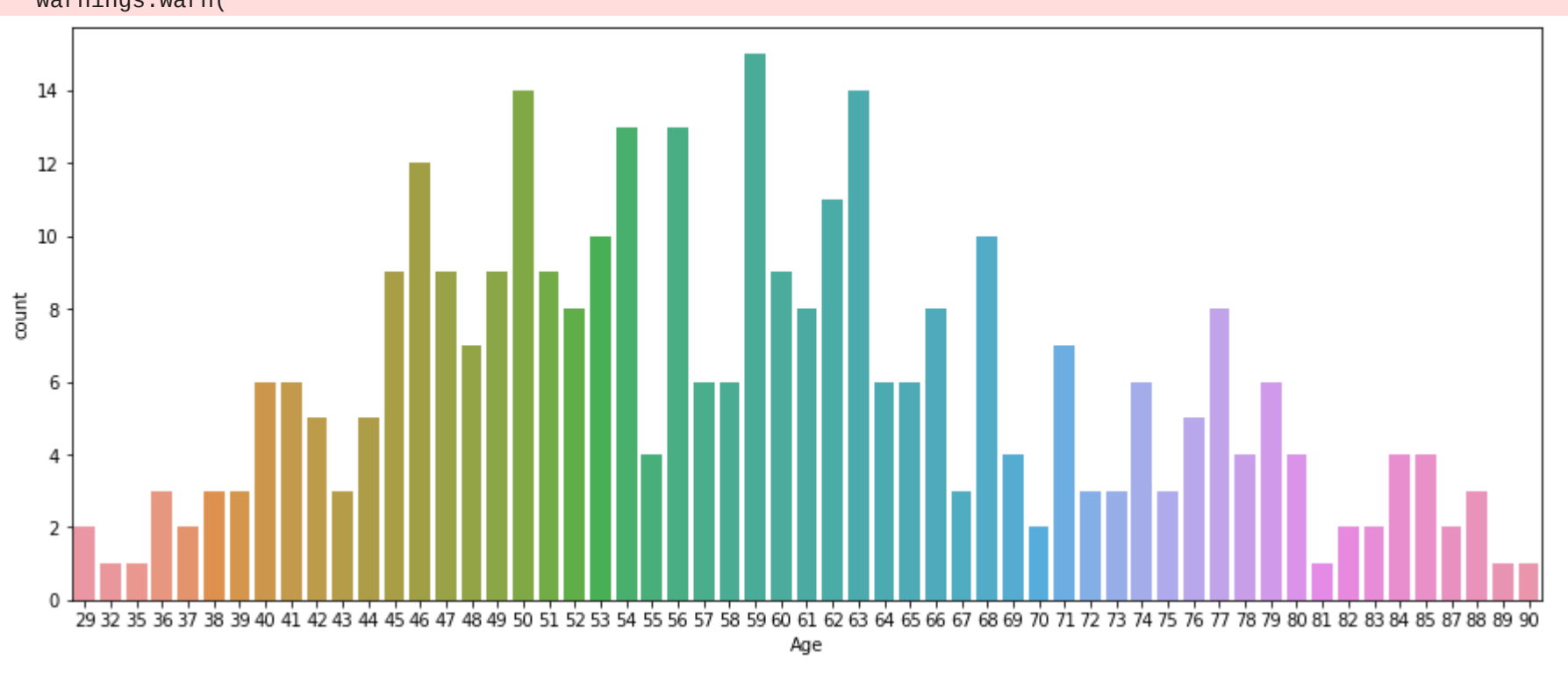
In [18]:
```
import matplotlib. pyplot as plt
import seaborn as sns
plt.figure(figsize=(15,6))
sns.countplot('Gender',data=breast_cancer)
plt.xticks(rotation=0)
plt.show()
```

```
C:\Users\SRIRAM\anaconda3\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(
```



In [19]:
```
plt.figure(figsize=(15,6))
sns.countplot('Age',data=breast_cancer)
plt.xticks(rotation=0)
plt.show()
```

```
C:\Users\SRIRAM\anaconda3\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(
```

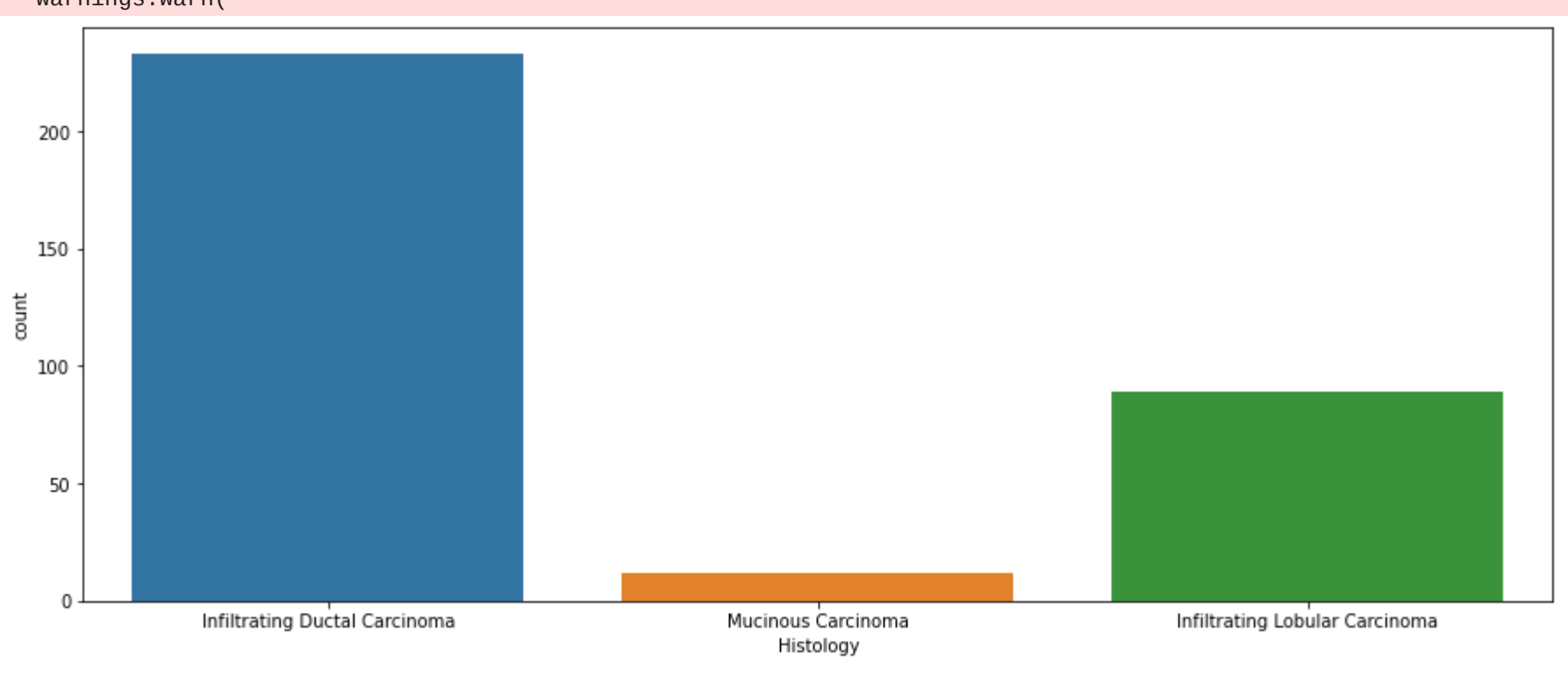


In [20]:
```
breast_cancer.Histology.unique()
```

Out[20]:
```
array(['Infiltrating Ductal Carcinoma', 'Mucinous Carcinoma',
       'Infiltrating Lobular Carcinoma'], dtype=object)
```

In [21]:
```
plt.figure(figsize=(15,6))
sns.countplot('Histology',data=breast_cancer)
plt.xticks(rotation=0)
plt.show()
```

```
C:\Users\SRIRAM\anaconda3\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(
```



In [22]:
```
breast_cancer.Surgery_type.unique()
```

Out[22]:
```
array(['Modified Radical Mastectomy', 'Lumpectomy', 'Other',
       'Simple Mastectomy'], dtype=object)
```

In [23]:
```
plt.figure(figsize=(15,6))
sns.countplot('Surgery_type',data=breast_cancer)
plt.xticks(rotation=0)
plt.show()
```

```
C:\Users\SRIRAM\anaconda3\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(
```

In [ ]: