In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

In [4]:

```
breast_cancer = pd.read_csv("breast_cancer.csv")
```

In [5]:

```
breast_cancer.head()
```

Out[5]:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histology
0	TCGA-D8-A1XD	36	FEMALE	0.080353	0.42638	0.54715	0.273680	III	Infiltrating Ductal Carcinoma
1	TCGA-EW-A1OX	43	FEMALE	-0.420320	0.57807	0.61447	-0.031505	II	Mucinous Carcinoma
2	TCGA-A8-A079	69	FEMALE	0.213980	1.31140	-0.32747	-0.234260	III	Infiltrating Ductal Carcinoma
3	TCGA-D8-A1XR	56	FEMALE	0.345090	-0.21147	-0.19304	0.124270	II	Infiltrating Ductal Carcinoma
4	TCGA-BH-A0BF	56	FEMALE	0.221550	1.90680	0.52045	-0.311990	II	Infiltrating Ductal Carcinoma

In [6]:

```
breast_cancer.tail()
```

Out[6]:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histolog
329	TCGA-AN-A04A	36	FEMALE	0.23180	0.61804	-0.55779	-0.517350	III	Infiltratin Ducta Carcinom
330	TCGA-A8-A085	44	MALE	0.73272	1.11170	-0.26952	-0.354920	II	Infiltratin Lobula Carcinom
331	TCGA-A1-A0SG	61	FEMALE	-0.71947	2.54850	-0.15024	0.339680	II	Infiltratin Ducta Carcinom
332	TCGA-A2-A0EU	79	FEMALE	0.47940	2.05590	-0.53136	-0.188480	I	Infiltratin Ducta Carcinom
333	TCGA-B6-A40B	76	FEMALE	-0.24427	0.92556	-0.41823	-0.067848	I	Infiltratin Ducta Carcinom

In [7]:

```
breast_cancer.shape
```

Out[7]:

```
(334, 16)
```

In [8]:

```
breast_cancer.columns
```

Out[8]:

```
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_Vis
it',
       'Patient_Status'],
      dtype='object')
```

In [9]:

```
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 [10]:

```
breast_cancer.describe()
```

Out[10]:

	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 [11]:

```
breast_cancer.isnull().sum()
```

Out[11]:

```
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 [12]:

```
breast_cancer.dropna(inplace = True)
```

In [13]:

```
breast_cancer.isnull().sum()
```

Out[13]:

```
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    0
Patient_Status        0
dtype: int64
```

In [14]:

```
breast_cancer.nunique()
```

Out[14]:

```
Patient_ID            317
Age                    57
Gender                  2
Protein1              316
Protein2              317
Protein3              317
Protein4              316
Tumour_Stage            3
Histology               3
ER status               1
PR status               1
HER2 status             2
Surgery_type            4
Date_of_Surgery       178
Date_of_Last_Visit    285
Patient_Status          2
dtype: int64
```

In [15]:

```
breast_cancer.Gender.unique()
```

Out[15]:

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

In [17]:

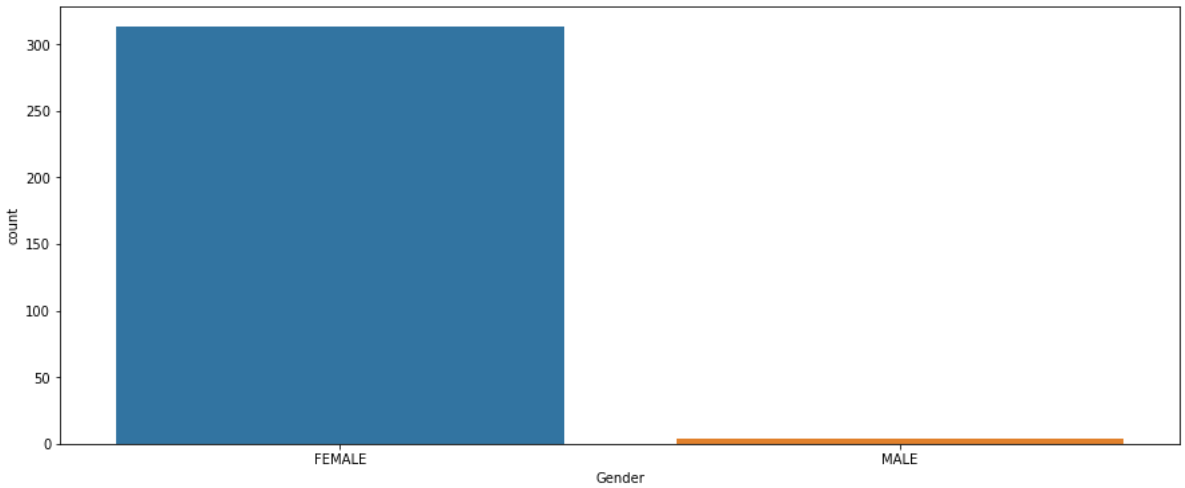
```
breast_cancer.Gender.value_counts()
```

Out[17]:

```
FEMALE    313
MALE        4
Name: Gender, dtype: int64
```

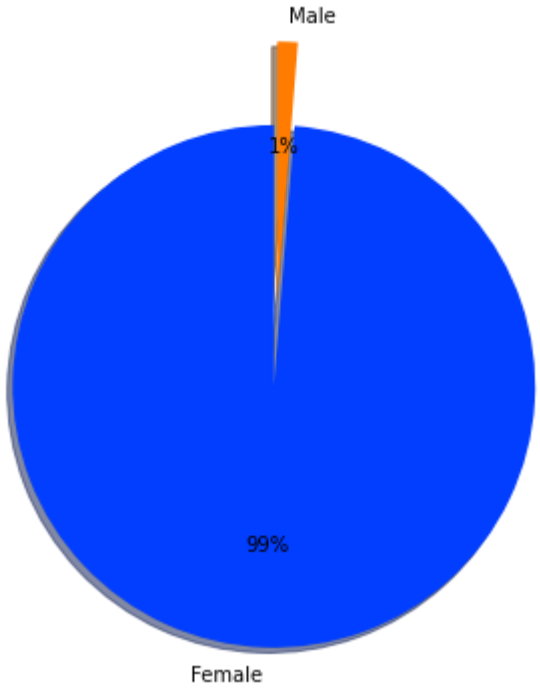
In [18]:

```
plt.figure(figsize=(15,6))
sns.countplot('Gender', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



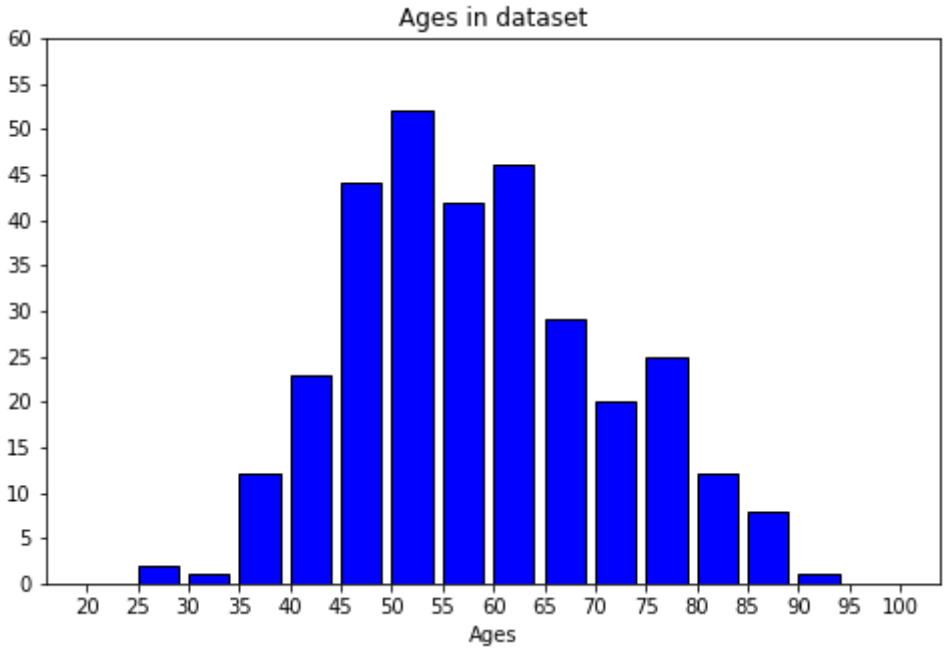
In [21]:

```
plt.figure(figsize=(15,6))
explode = [0.3,0.02]
colors = sns.color_palette('bright')
plt.pie(breast_cancer['Gender'].value_counts(), labels=['Female', 'Male'],
        colors = colors, autopct = '%0.0f%%',  explode = explode, shadow = 'True',
        startangle = 90)
plt.show()
```



In [26]:

```
bins = list(range(20,105,5))
plt.figure(figsize = (8,5))
plt.hist(breast_cancer['Age'].astype(int), width = 4, align = 'mid',
         bins = bins, color = 'blue', edgecolor = 'black')
plt.xticks(bins)
plt.xlabel('Ages')
plt.title('Ages in dataset')
plt.yticks(np.arange(0,65,5))
plt.show()
```



In [27]:

```
breast_cancer.Histology.unique()
```

Out[27]:

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

In [28]:

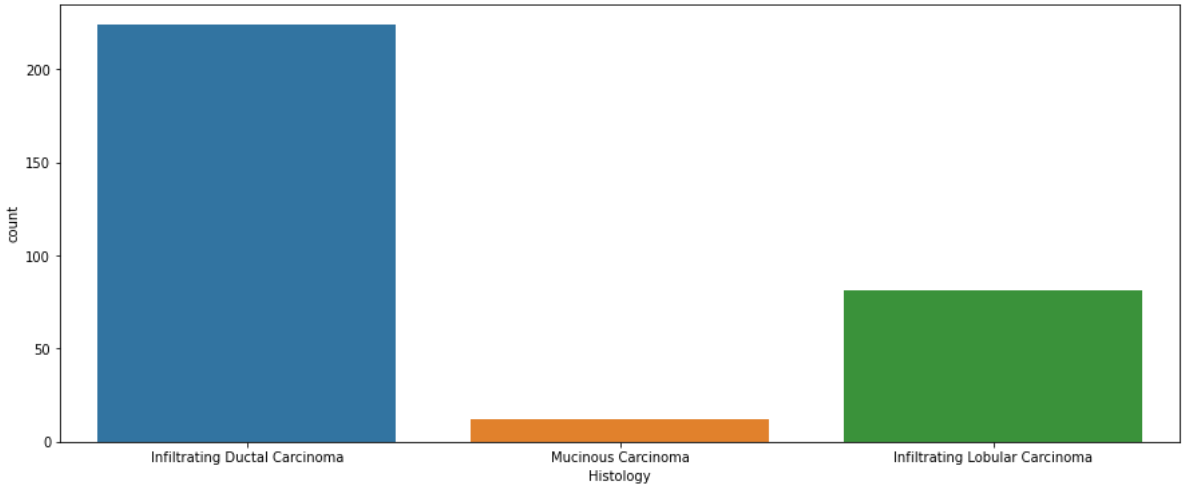
```
breast_cancer.Histology.value_counts()
```

Out[28]:

```
Infiltrating Ductal Carcinoma     224
Infiltrating Lobular Carcinoma     81
Mucinous Carcinoma                 12
Name: Histology, dtype: int64
```

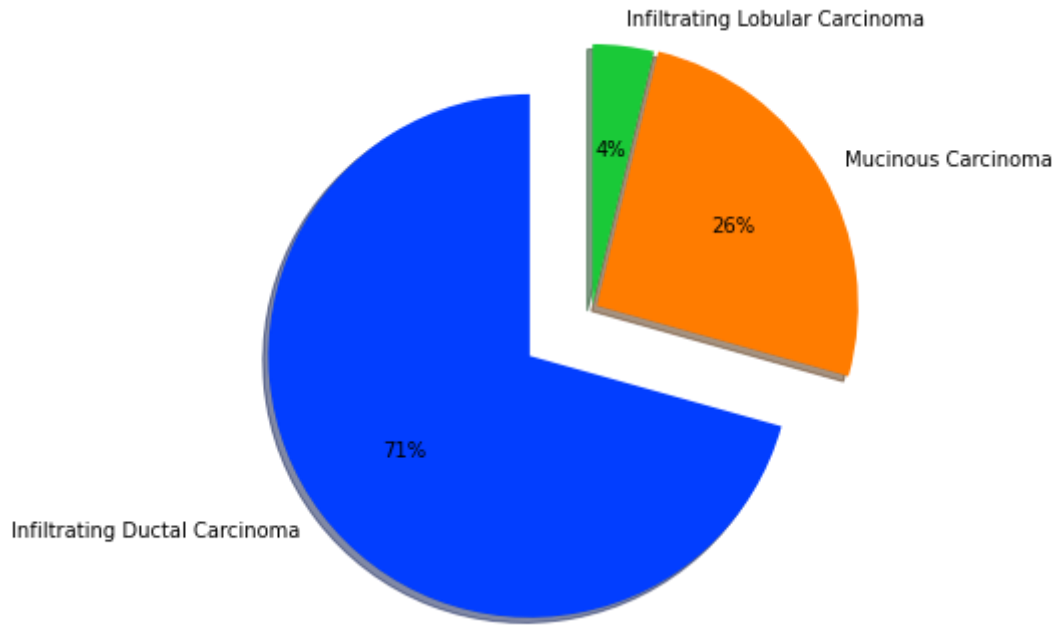
In [29]:

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



In [32]:

```
plt.figure(figsize=(15,6))
explode = [0.3,0.02, 0.01]
colors = sns.color_palette('bright')
plt.pie(breast_cancer['Histology'].value_counts(), labels=['Infiltrating Ductal Carcinor
                                                           'Mucinous Carcinoma',
                                                           'Infiltrating Lobular Carcinc
        colors = colors, autopct = '%0.0f%%',  explode = explode, shadow = 'True',
        startangle = 90)
plt.show()
```



In [33]:

```
breast_cancer.Tumour_Stage.unique()
```

Out[33]:

```
array(['III', 'II', 'I'], dtype=object)
```

In [34]:

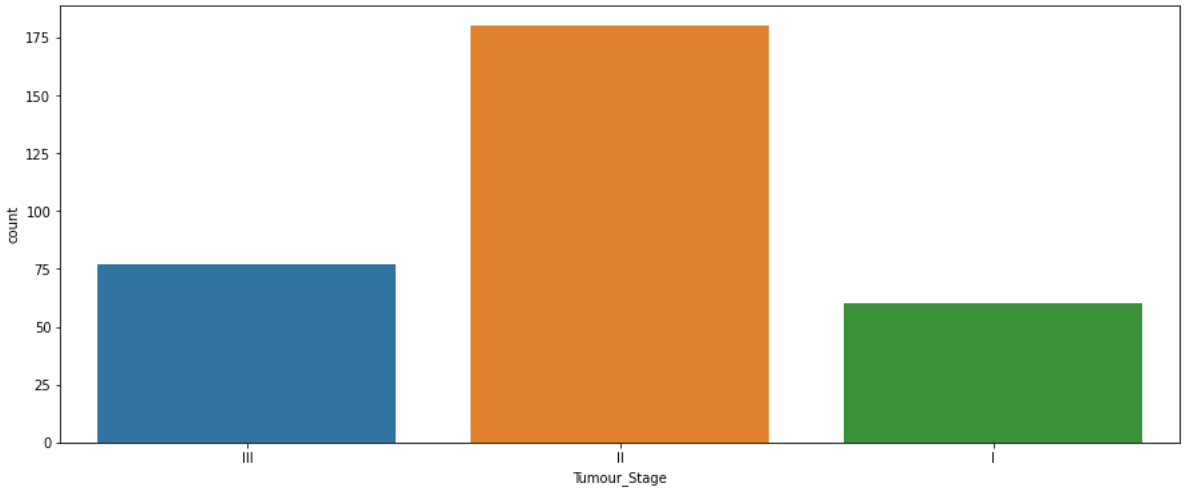
```
breast_cancer.Tumour_Stage.value_counts()
```

Out[34]:

```
II     180
III     77
I       60
Name: Tumour_Stage, dtype: int64
```

In [35]:

```
plt.figure(figsize=(15,6))
sns.countplot('Tumour_Stage', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



In [36]:

```
breast_cancer_type_by_stage = (breast_cancer.groupby(['Histology', 'Tumour_Stage'],
                                                     as_index = False).agg(Total = ('Age
                                                                                    'cou
)
```

In [37]:

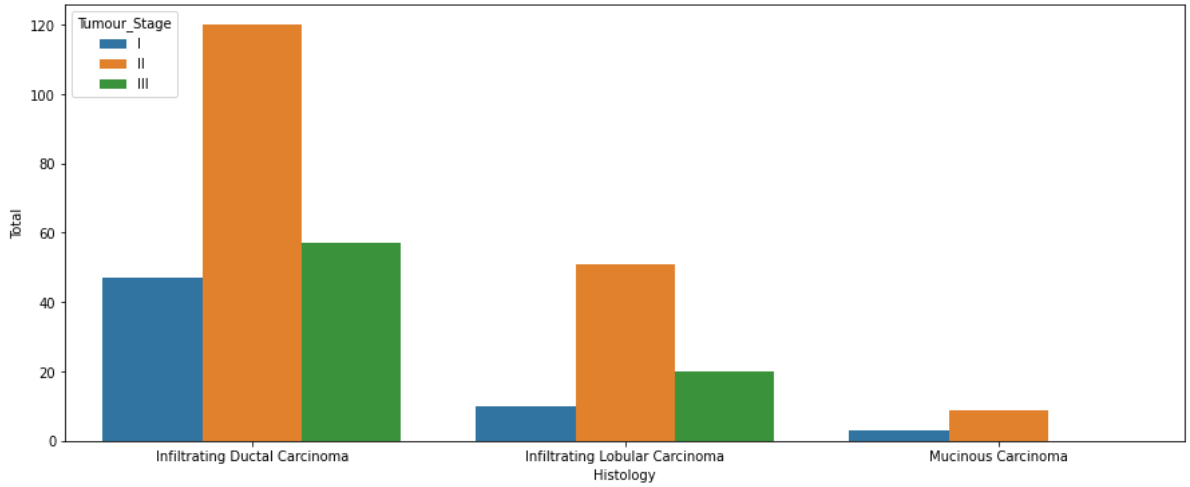
```
breast_cancer_type_by_stage
```

Out[37]:

	Histology	Tumour_Stage	Total
0	Infiltrating Ductal Carcinoma	I	47
1	Infiltrating Ductal Carcinoma	II	120
2	Infiltrating Ductal Carcinoma	III	57
3	Infiltrating Lobular Carcinoma	I	10
4	Infiltrating Lobular Carcinoma	II	51
5	Infiltrating Lobular Carcinoma	III	20
6	Mucinous Carcinoma	I	3
7	Mucinous Carcinoma	II	9

In [38]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Histology', hue = 'Tumour_Stage', y = 'Total',
            data = breast_cancer_type_by_stage)
plt.xticks(rotation = 0)
plt.show()
```



In [44]:

```
breast_cancer['Age'] = pd.cut(breast_cancer['Age'], bins=5)
```

In [46]:

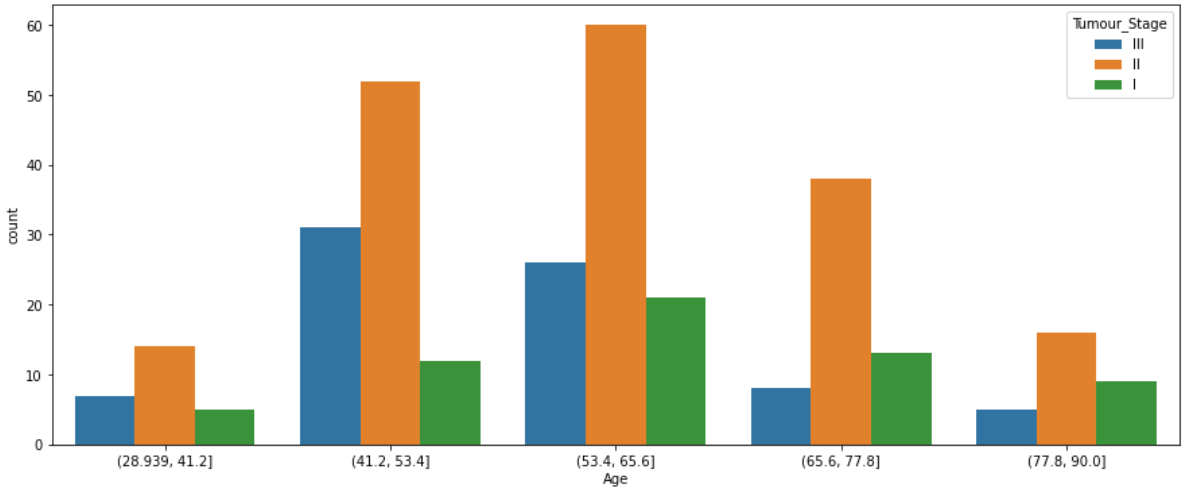
```
breast_cancer['Age'].head()
```

Out[46]:

```
0    (28.939, 41.2]
1      (41.2, 53.4]
2      (65.6, 77.8]
3      (53.4, 65.6]
4      (53.4, 65.6]
Name: Age, dtype: category
Categories (5, interval[float64, right]): [(28.939, 41.2] < (41.2, 53.4] <
(53.4, 65.6] < (65.6, 77.8] < (77.8, 90.0]]
```

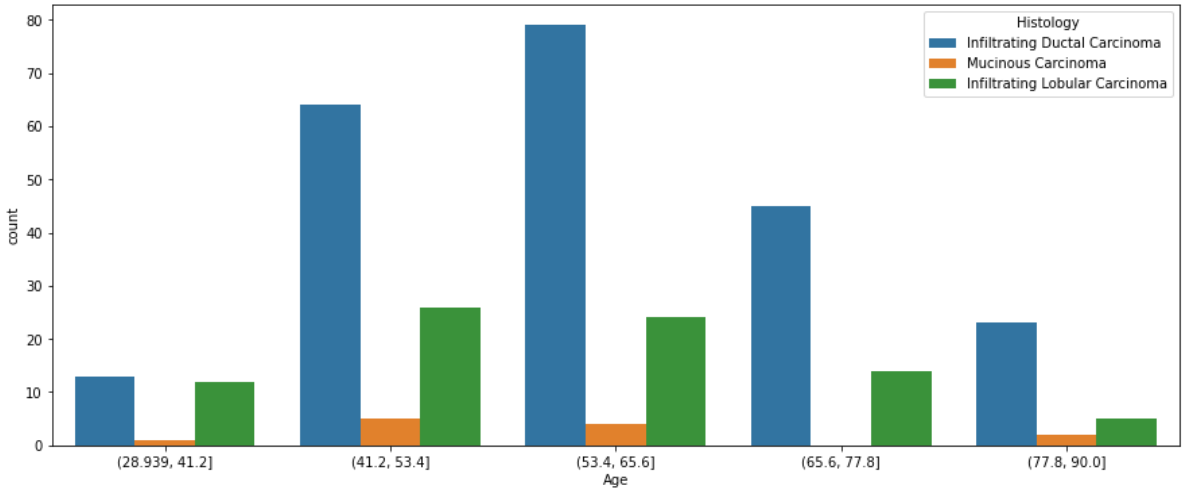
In [48]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'Age', hue = 'Tumour_Stage', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



In [49]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'Age', hue = 'Histology', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```

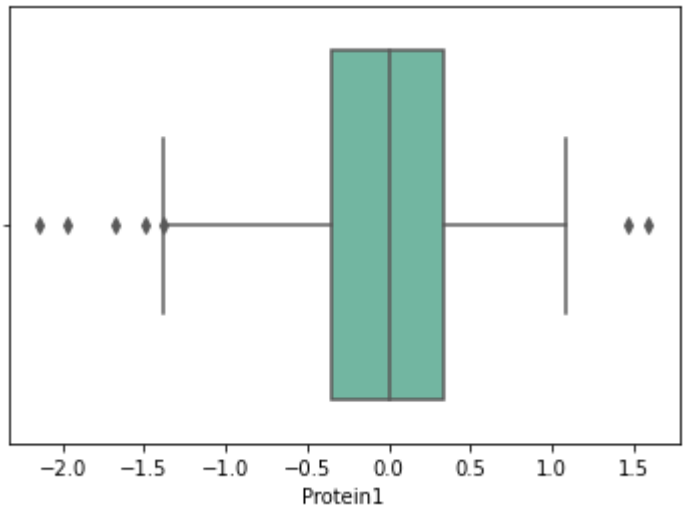


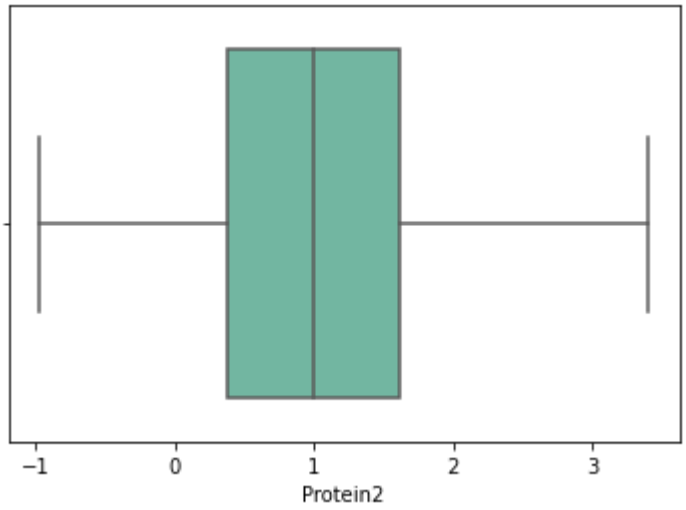
In [50]:

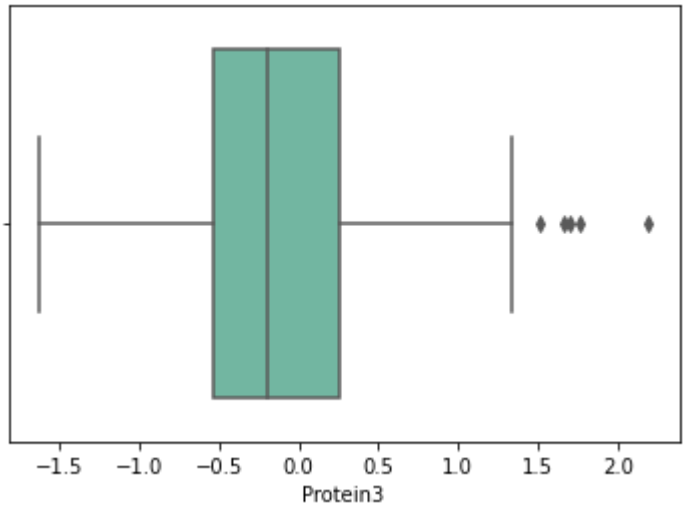
```
protein_types = breast_cancer[['Protein1', 'Protein2', 'Protein3', 'Protein4']]
```

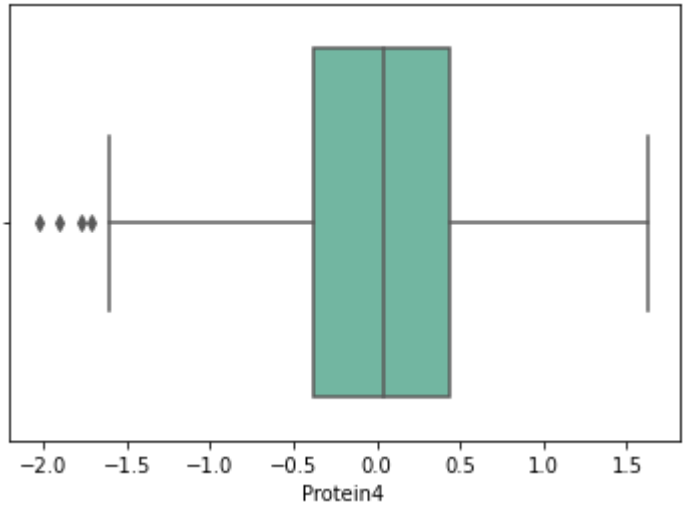
In [51]:

```
for i in protein_types.columns:
    sns.boxplot(x=protein_types[i], orient = 'h', palette = 'Set2')
    plt.show()
```









In [52]:

```
breast_cancer_type_protein = breast_cancer[['Histology', 'Protein1', 'Protein2',
                                            'Protein3', 'Protein4']]
```

In [53]:

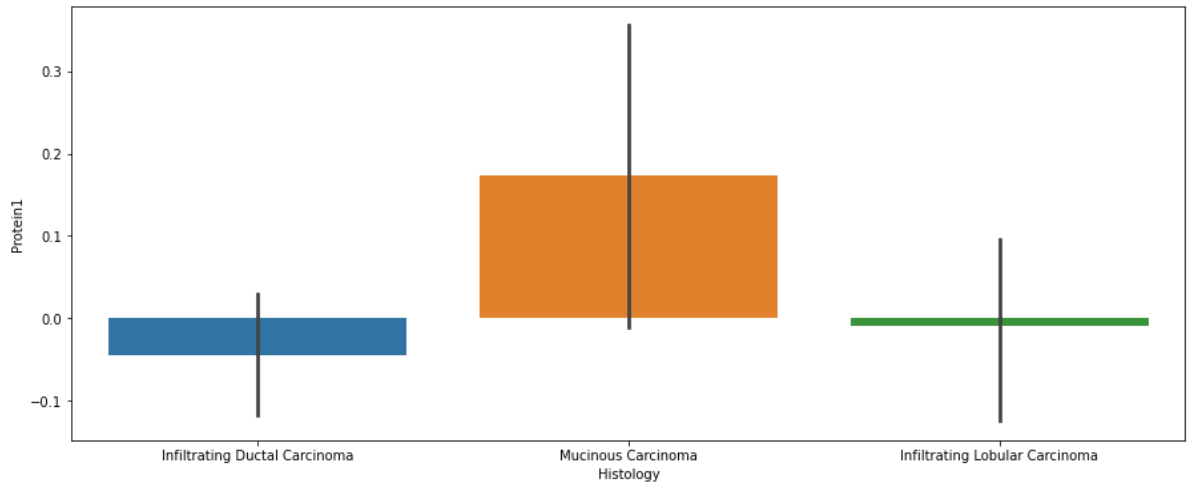
```
breast_cancer_type_protein.head()
```

Out[53]:

	Histology	Protein1	Protein2	Protein3	Protein4
0	Infiltrating Ductal Carcinoma	0.080353	0.42638	0.54715	0.273680
1	Mucinous Carcinoma	-0.420320	0.57807	0.61447	-0.031505
2	Infiltrating Ductal Carcinoma	0.213980	1.31140	-0.32747	-0.234260
3	Infiltrating Ductal Carcinoma	0.345090	-0.21147	-0.19304	0.124270
4	Infiltrating Ductal Carcinoma	0.221550	1.90680	0.52045	-0.311990

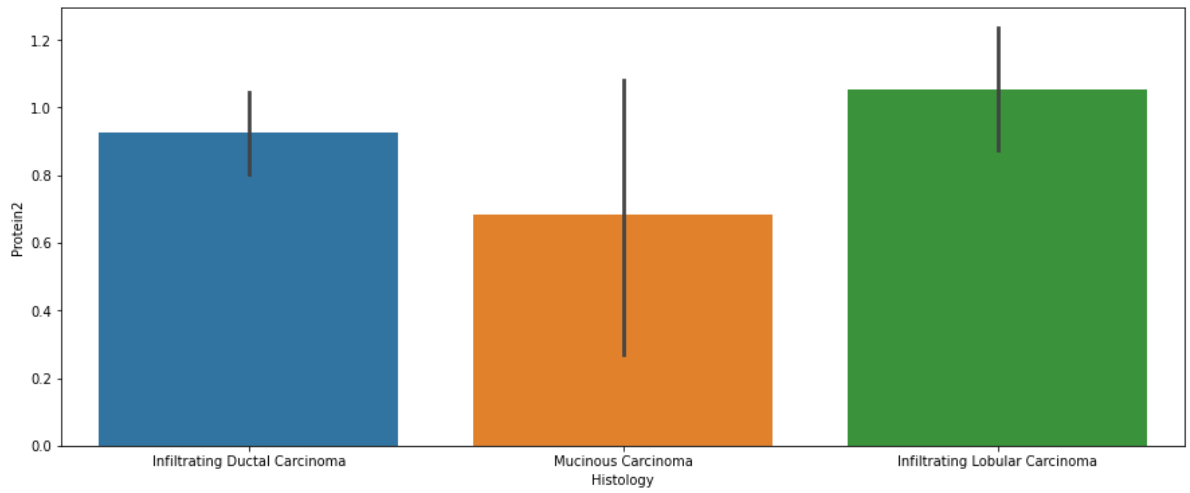
In [58]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Histology', y = 'Protein1', data = breast_cancer_type_protein)
plt.xticks(rotation = 0)
plt.show()
```



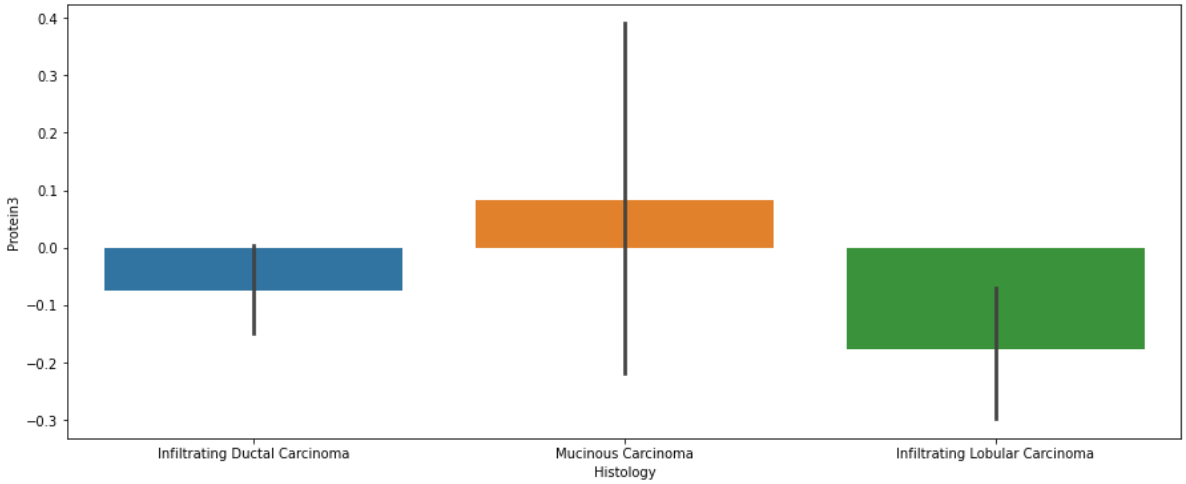
In [59]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Histology', y = 'Protein2', data = breast_cancer_type_protein)
plt.xticks(rotation = 0)
plt.show()
```



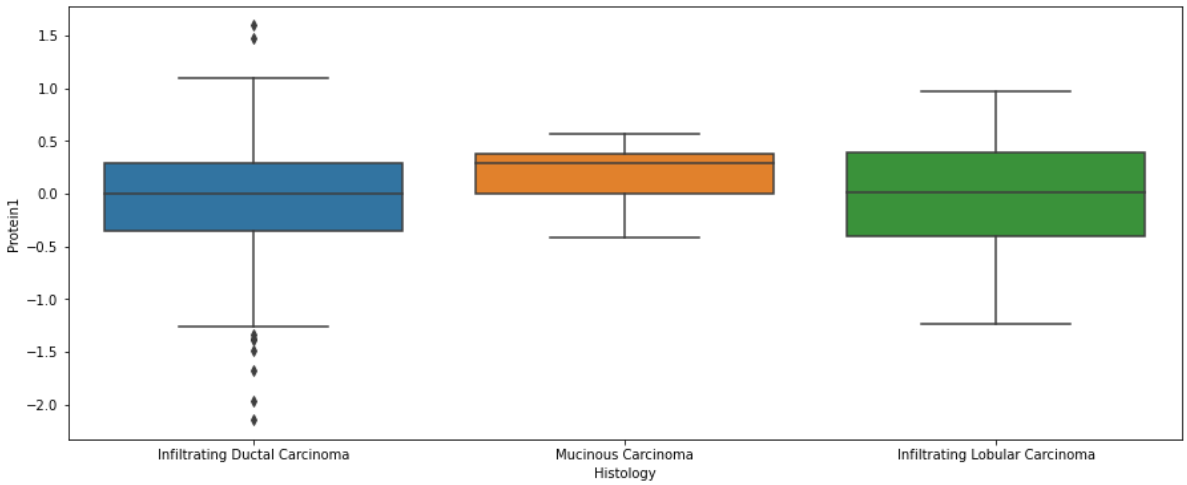
In [60]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Histology', y = 'Protein3', data = breast_cancer_type_protein)
plt.xticks(rotation = 0)
plt.show()
```



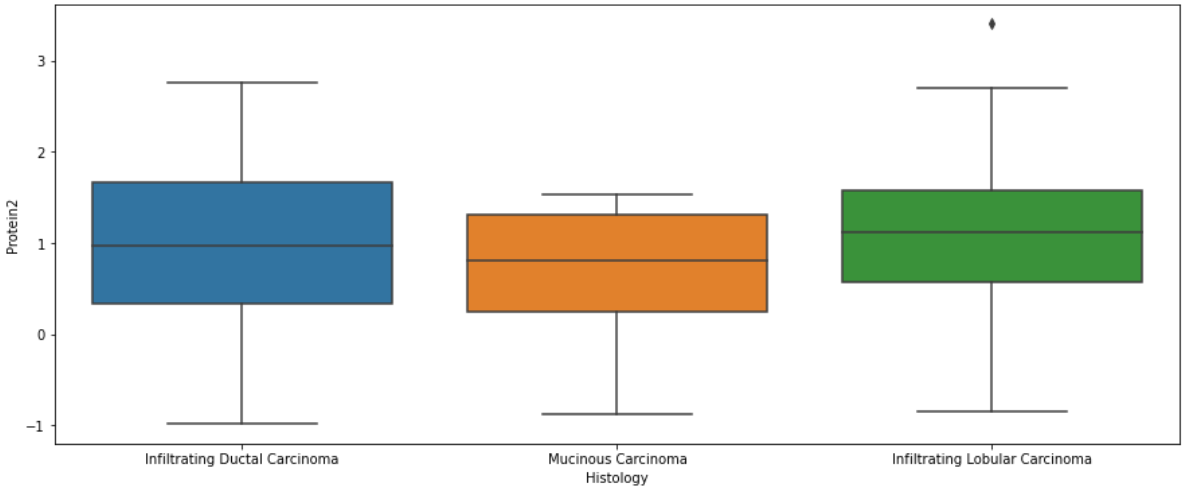
In [63]:

```
plt.figure(figsize=(15,6))
sns.boxplot(x = 'Histology', y = 'Protein1', data = breast_cancer_type_protein)
plt.xticks(rotation = 0)
plt.show()
```



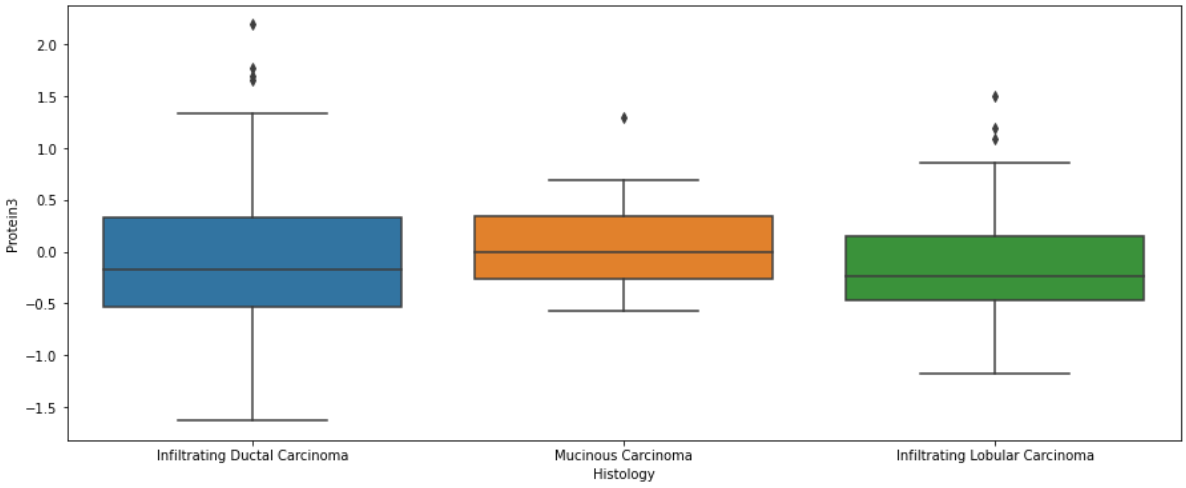
In [64]:

```
plt.figure(figsize=(15,6))
sns.boxplot(x = 'Histology', y = 'Protein2', data = breast_cancer_type_protein)
plt.xticks(rotation = 0)
plt.show()
```



In [65]:

```
plt.figure(figsize=(15,6))
sns.boxplot(x = 'Histology', y = 'Protein3', data = breast_cancer_type_protein)
plt.xticks(rotation = 0)
plt.show()
```



In [66]:

```
breast_cancer_stage_protein = breast_cancer[['Tumour_Stage', 'Protein1', 'Protein2',
                                             'Protein3', 'Protein4']]
```

In [67]:

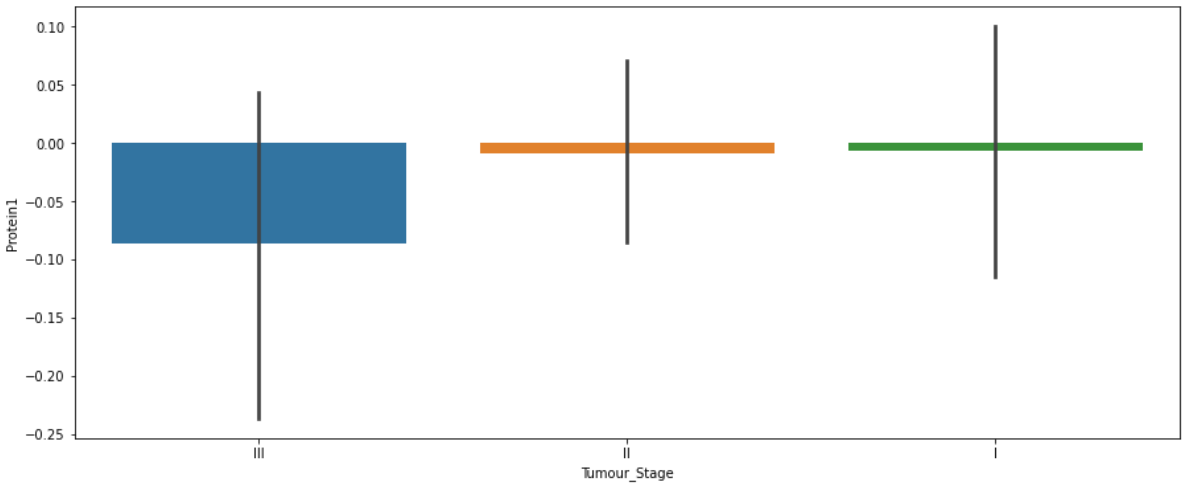
```
breast_cancer_stage_protein.head()
```

Out[67]:

	Tumour_Stage	Protein1	Protein2	Protein3	Protein4
0	III	0.080353	0.42638	0.54715	0.273680
1	II	-0.420320	0.57807	0.61447	-0.031505
2	III	0.213980	1.31140	-0.32747	-0.234260
3	II	0.345090	-0.21147	-0.19304	0.124270
4	II	0.221550	1.90680	0.52045	-0.311990

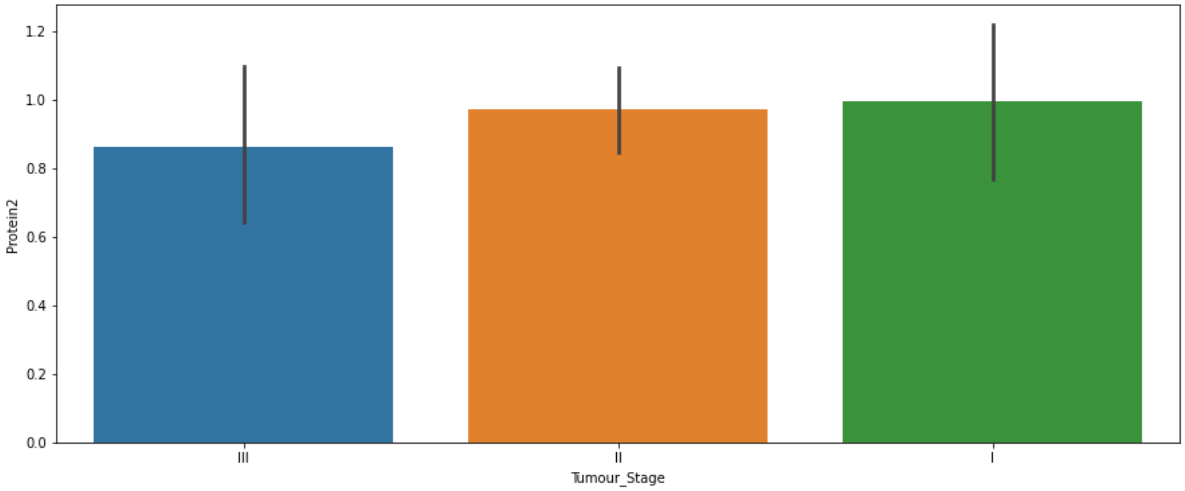
In [69]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Tumour_Stage', y = 'Protein1', data = breast_cancer_stage_protein)
plt.xticks(rotation = 0)
plt.show()
```



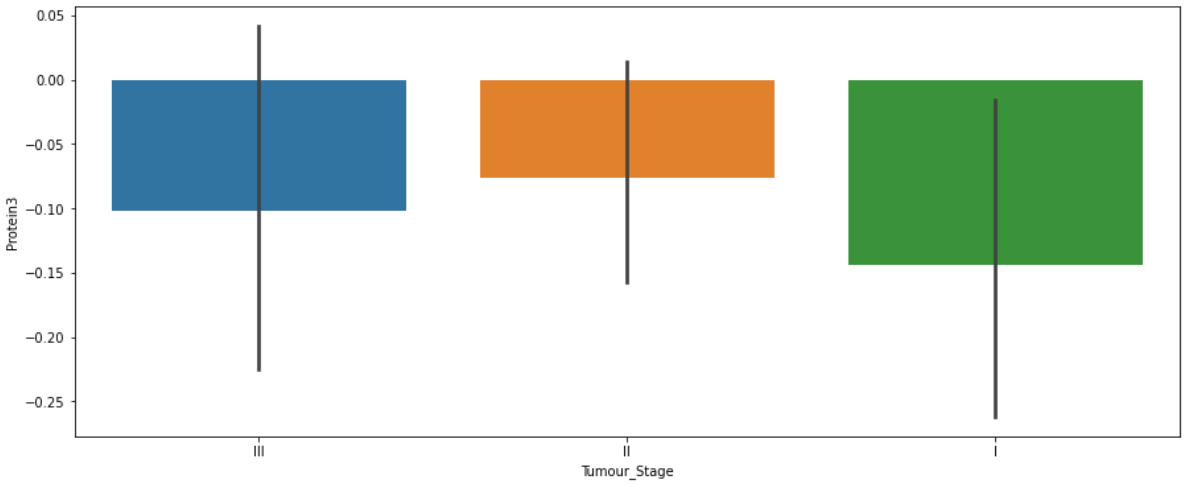
In [70]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Tumour_Stage', y = 'Protein2', data = breast_cancer_stage_protein)
plt.xticks(rotation = 0)
plt.show()
```



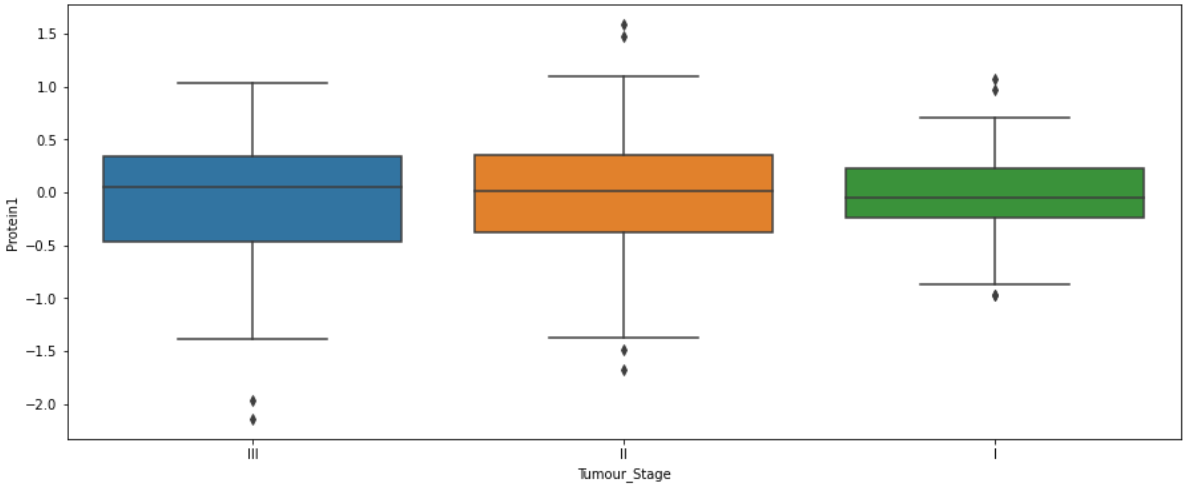
In [71]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Tumour_Stage', y = 'Protein3', data = breast_cancer_stage_protein)
plt.xticks(rotation = 0)
plt.show()
```



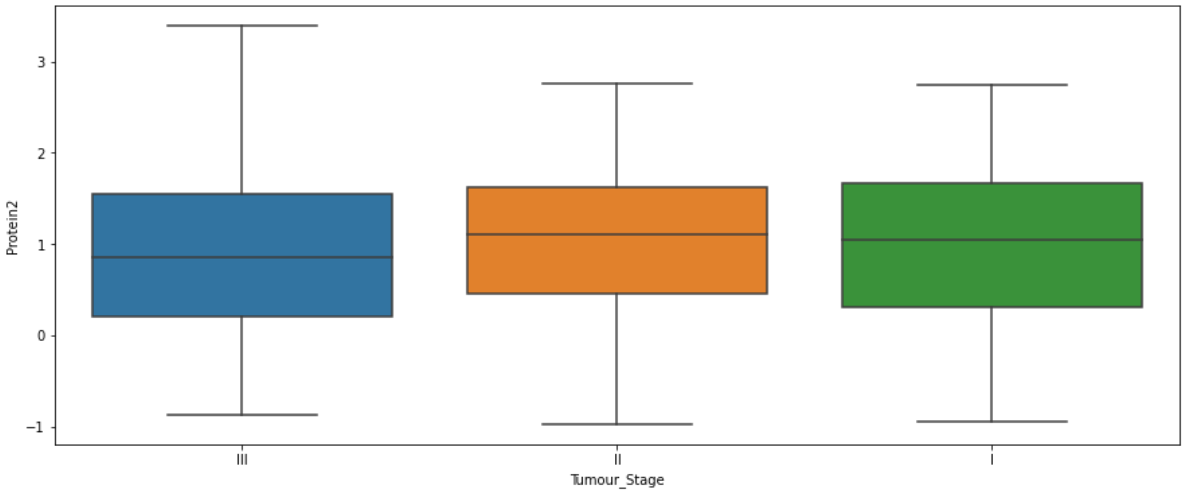
In [72]:

```
plt.figure(figsize=(15,6))
sns.boxplot(x = 'Tumour_Stage', y = 'Protein1', data = breast_cancer_stage_protein)
plt.xticks(rotation = 0)
plt.show()
```



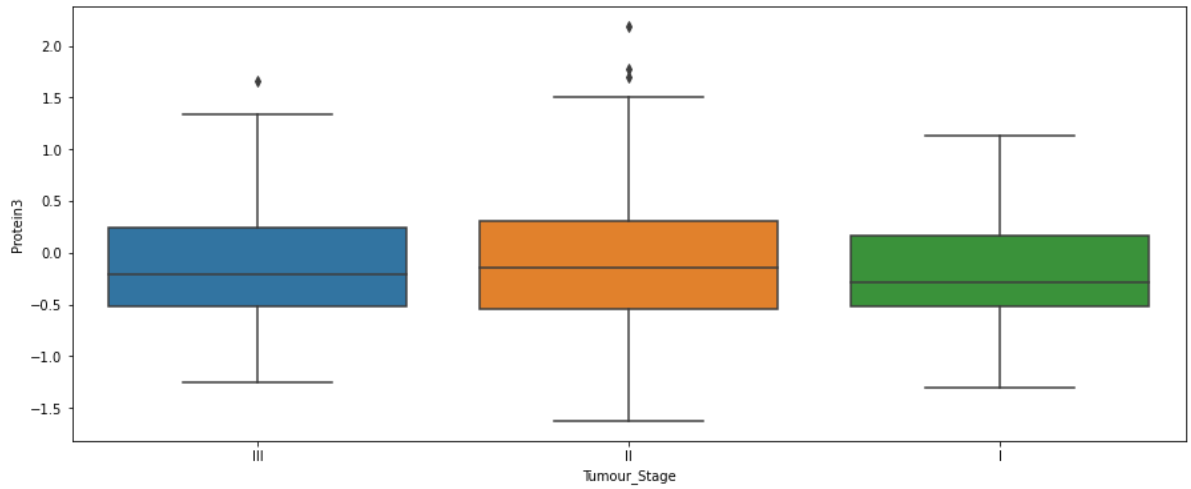
In [73]:

```
plt.figure(figsize=(15,6))
sns.boxplot(x = 'Tumour_Stage', y = 'Protein2', data = breast_cancer_stage_protein)
plt.xticks(rotation = 0)
plt.show()
```



In [74]:

```
plt.figure(figsize=(15,6))
sns.boxplot(x = 'Tumour_Stage', y = 'Protein3', data = breast_cancer_stage_protein)
plt.xticks(rotation = 0)
plt.show()
```



In [75]:

```
breast_cancer_age_protein = breast_cancer[['Age', 'Protein1', 'Protein2',
                                           'Protein3', 'Protein4']]
```

In [76]:

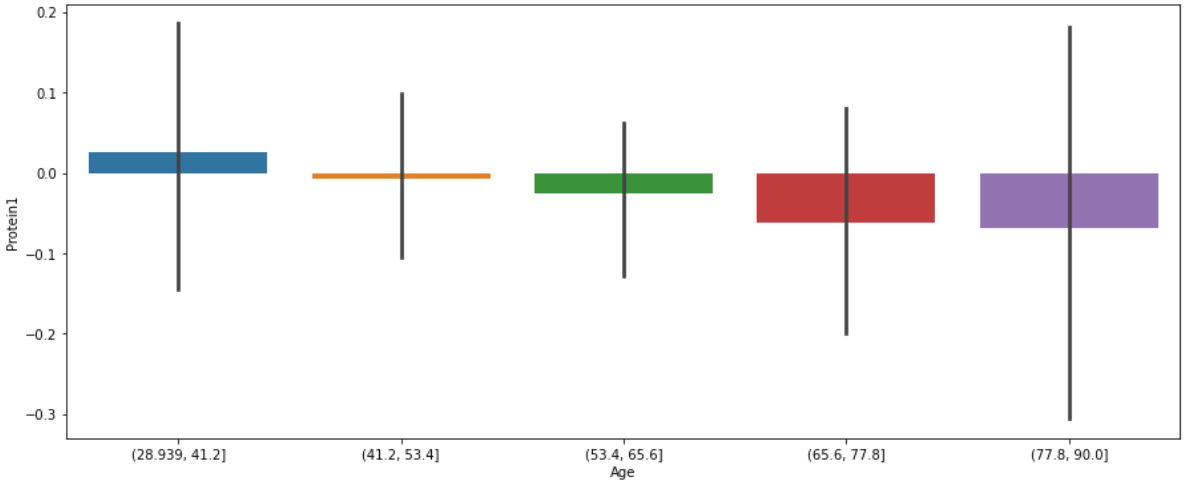
```
breast_cancer_age_protein.head()
```

Out[76]:

	Age	Protein1	Protein2	Protein3	Protein4
0	(28.939, 41.2]	0.080353	0.42638	0.54715	0.273680
1	(41.2, 53.4]	-0.420320	0.57807	0.61447	-0.031505
2	(65.6, 77.8]	0.213980	1.31140	-0.32747	-0.234260
3	(53.4, 65.6]	0.345090	-0.21147	-0.19304	0.124270
4	(53.4, 65.6]	0.221550	1.90680	0.52045	-0.311990

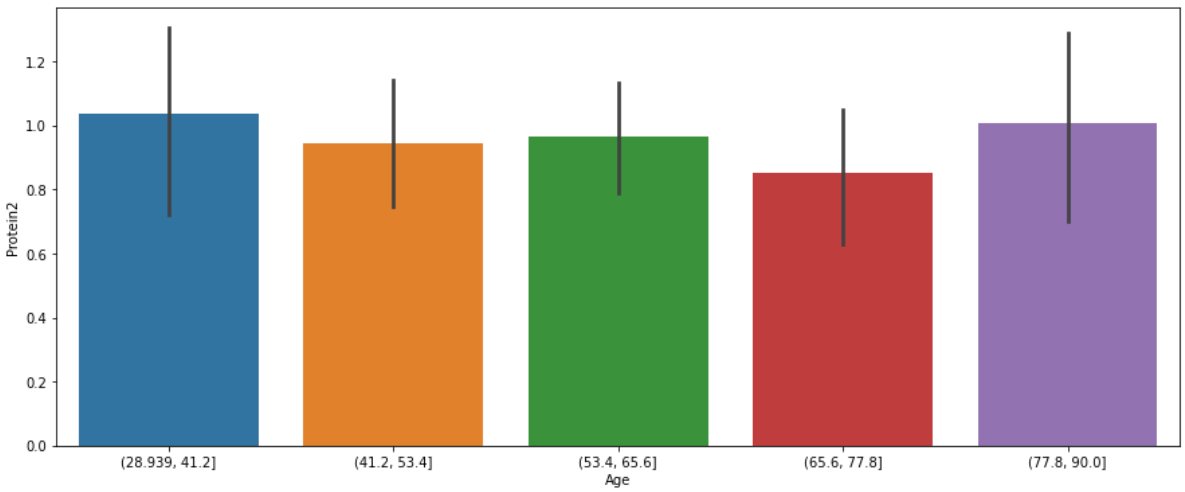
In [81]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Age', y = 'Protein1', data = breast_cancer_age_protein)
plt.xticks(rotation = 0)
plt.show()
```



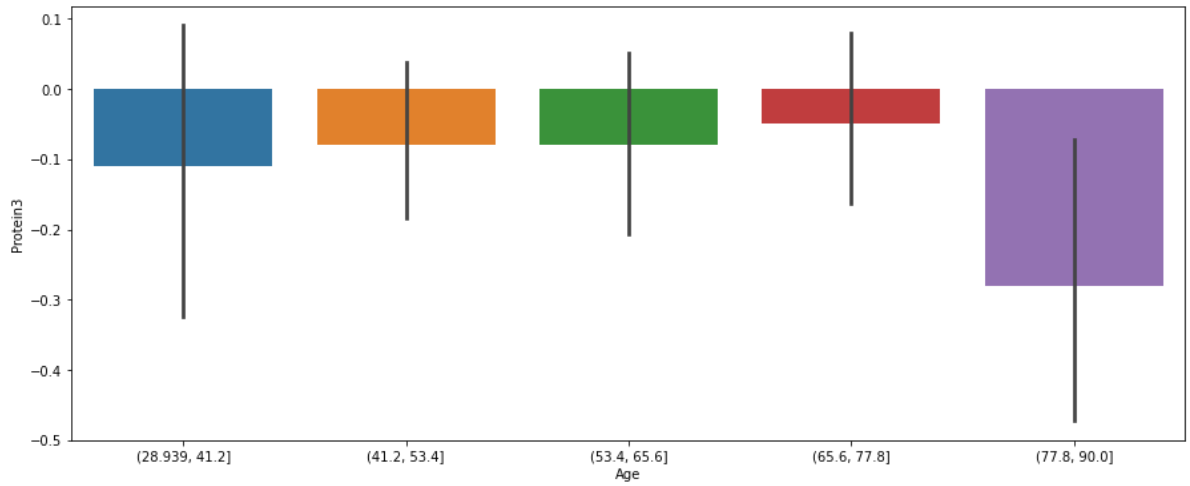
In [78]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Age', y = 'Protein2', data = breast_cancer_age_protein)
plt.xticks(rotation = 0)
plt.show()
```



In [79]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'Age', y = 'Protein3', data = breast_cancer_age_protein)
plt.xticks(rotation = 0)
plt.show()
```



In [83]:

```
n_markers = breast_cancer[['Histology', 'ER status', 'PR status', 'HER2 status']]
```

In [84]:

```
n_markers.head()
```

Out[84]:

	Histology	ER status	PR status	HER2 status
0	Infiltrating Ductal Carcinoma	Positive	Positive	Negative
1	Mucinous Carcinoma	Positive	Positive	Negative
2	Infiltrating Ductal Carcinoma	Positive	Positive	Negative
3	Infiltrating Ductal Carcinoma	Positive	Positive	Negative
4	Infiltrating Ductal Carcinoma	Positive	Positive	Negative

In [85]:

```
n_markers['ER status'].unique()
```

Out[85]:

```
array(['Positive'], dtype=object)
```

In [86]:

```
n_markers['ER status'].value_counts()
```

Out[86]:

```
Positive    317
Name: ER status, dtype: int64
```

In [87]:

```
n_markers['PR status'].value_counts()
```

Out[87]:

```
Positive    317
Name: PR status, dtype: int64
```

In [88]:

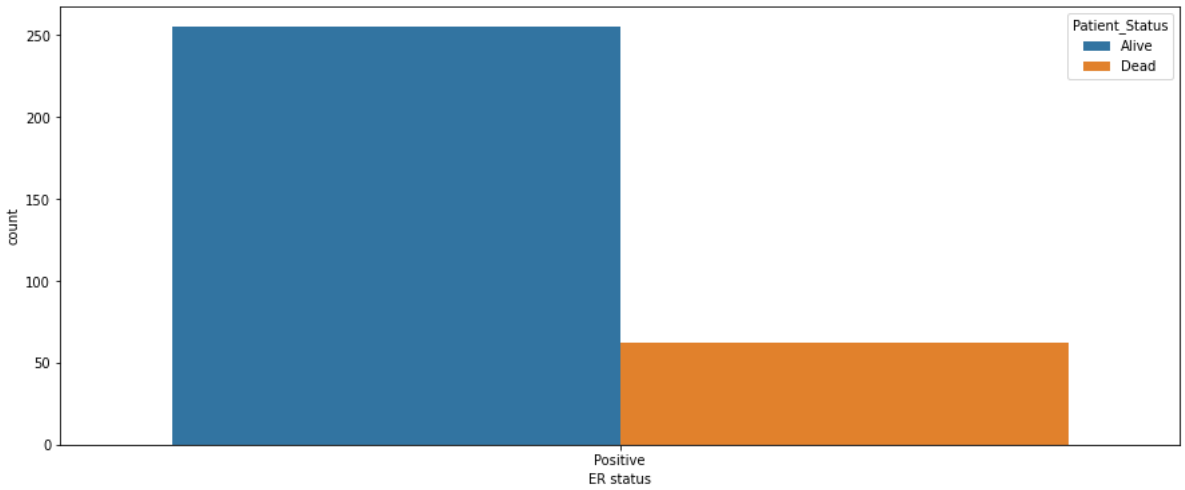
```
n_markers['HER2 status'].value_counts()
```

Out[88]:

```
Negative    288
Positive     29
Name: HER2 status, dtype: int64
```

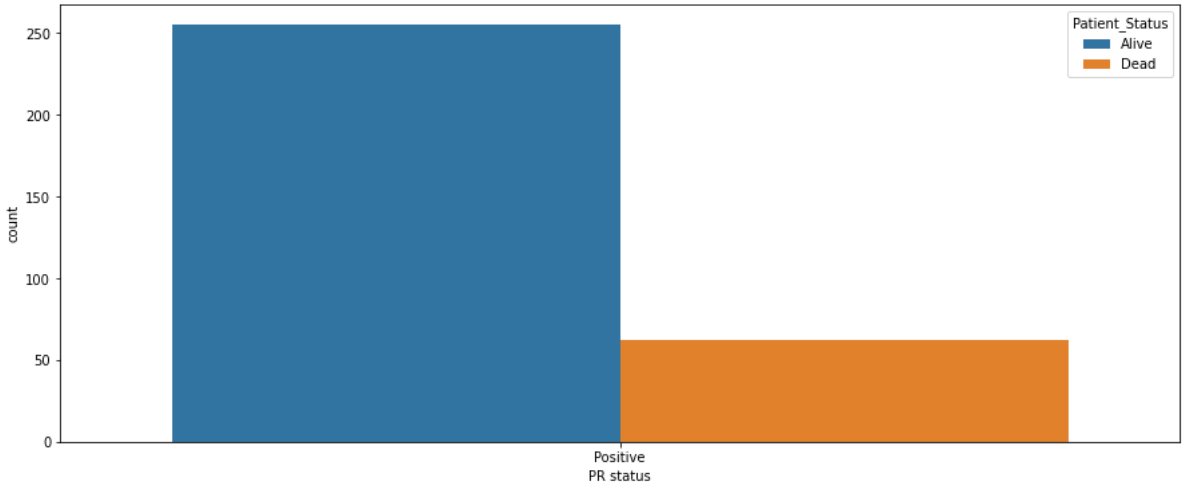
In [89]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'ER status', hue = 'Patient_Status', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



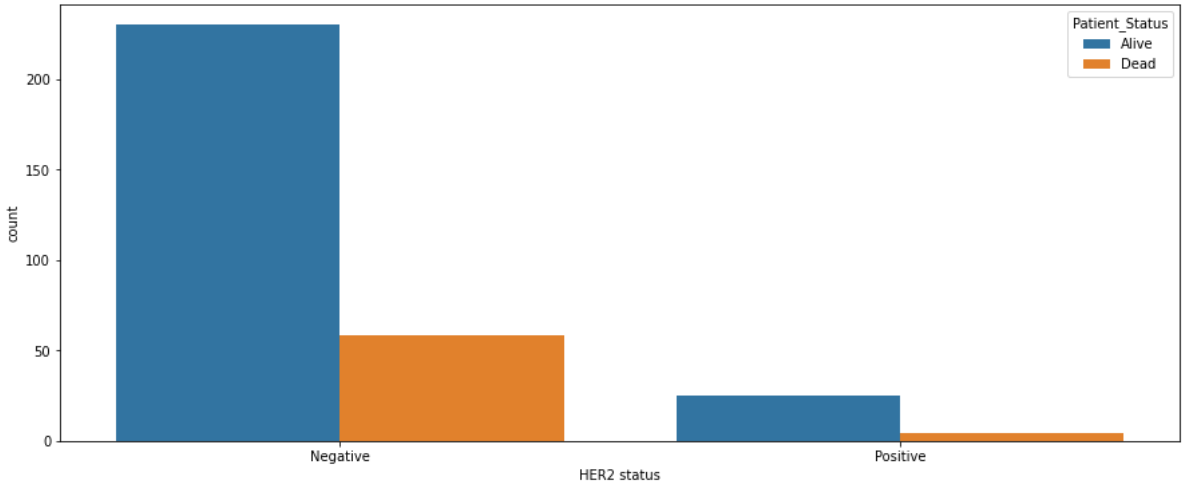
In [90]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'PR status', hue = 'Patient_Status', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



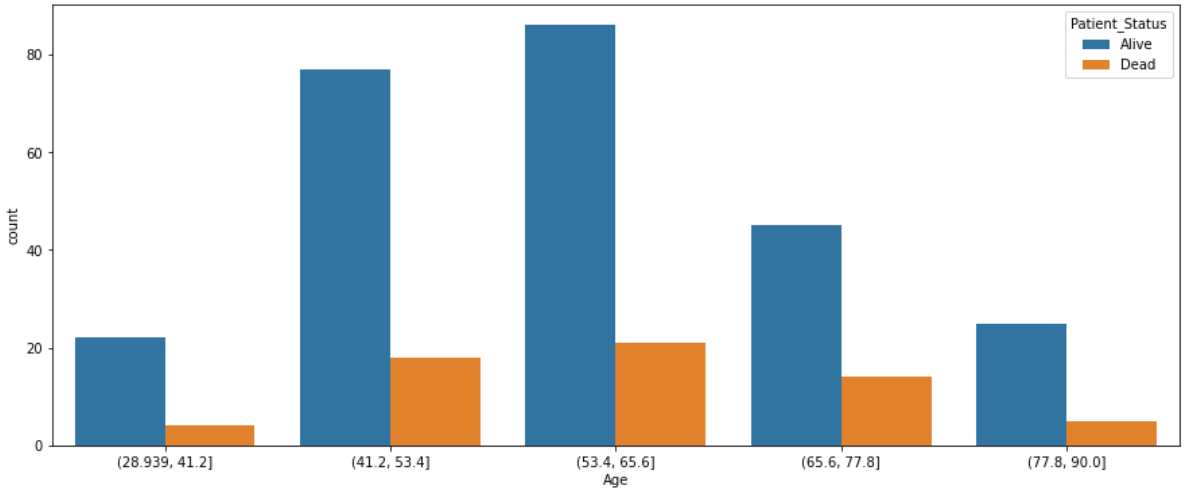
In [91]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'HER2 status', hue = 'Patient_Status', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



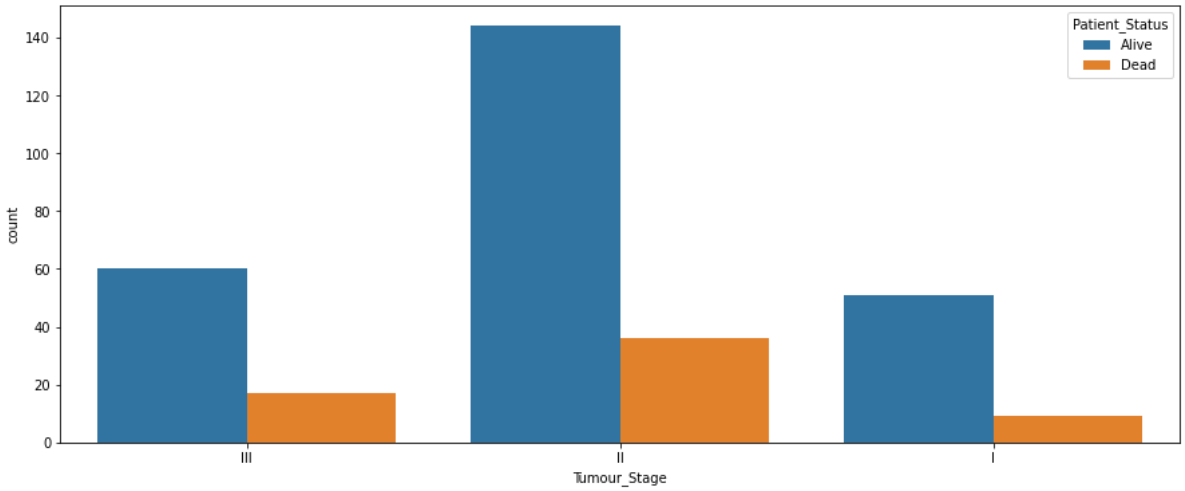
In [92]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'Age', hue = 'Patient_Status', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



In [93]:

```
plt.figure(figsize=(15,6))
sns.countplot(x = 'Tumour_Stage', hue = 'Patient_Status', data = breast_cancer)
plt.xticks(rotation = 0)
plt.show()
```



In [94]:

```
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
```

In [95]:

```
breast_cancer['Tumour_Stage']= label_encoder.fit_transform(breast_cancer['Tumour_Stage'])
```

In [96]:

```
breast_cancer['Histology']= label_encoder.fit_transform(breast_cancer['Histology'])
```

In [97]:

```
breast_cancer['ER status']= label_encoder.fit_transform(breast_cancer['ER status'])
```

In [98]:

```
breast_cancer['PR status']= label_encoder.fit_transform(breast_cancer['PR status'])
```

In [99]:

```
breast_cancer['HER2 status']= label_encoder.fit_transform(breast_cancer['HER2 status'])
```

In [100]:

```
breast_cancer['Surgery_type']= label_encoder.fit_transform(breast_cancer['Surgery_type'])
```

In [101]:

```
breast_cancer['Patient_Status']= label_encoder.fit_transform(breast_cancer['Patient_Stat
```

In [129]:

```
x = breast_cancer.drop(['Patient_ID','Age', 'Gender',
                        'Date_of_Surgery','Date_of_Last_Visit', 'Patient_Status'], axis
y = breast_cancer.Patient_Status
```

In [130]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

In [131]:

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
```

In [132]:

```
model = LogisticRegression()
model.fit(X_train, y_train)
```

Out[132]:

```
LogisticRegression()
```

In [133]:

```
y_pred = model.predict(X_test)
```

In [134]:

```
print("Training Accuracy :", model.score(X_train, y_train))
print("Testing Accuracy :", model.score(X_test, y_test))
```

```
Training Accuracy : 0.8054298642533937
Testing Accuracy : 0.8020833333333334
```