In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```
df_train = pd.read_csv('train (3).csv')
```

In [3]:

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

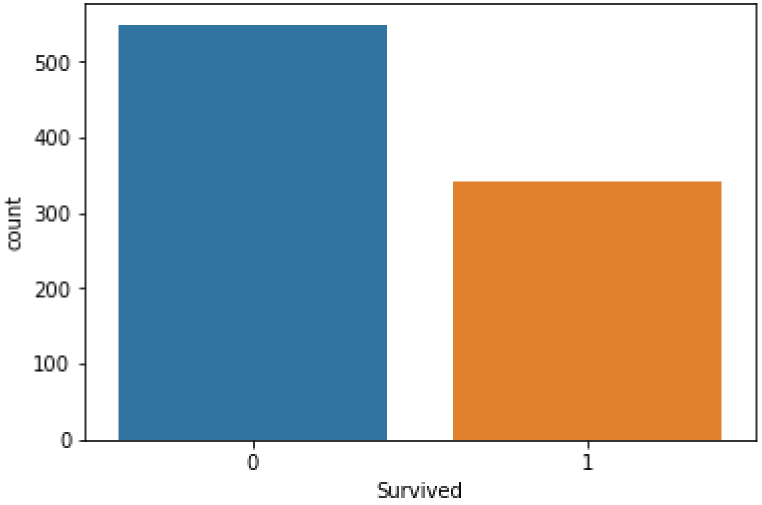
EDA

In [5]:

```
sns.countplot(x = 'Survived',data = df_train)
df_train['Survived'].value_counts()/df_train.shape[0]
```

Out[5]:

```
0    0.616162
1    0.383838
Name: Survived, dtype: float64
```



Questions to ask from our dataset

1. Our demography and whether that has an effect on survival probability
2. If class does have an effect of survival probability
3. Cabin/Seat against survival probability
4. Embarked port vs Survival probability
5. Whether travelling alone affected survival probability

Univariate analytics

In [7]:

```
df_train.columns
```

Out[7]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [8]:

```
df_train.head()
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [9]:

```
df_train.drop(['PassengerId','Name'],1,inplace = True)
```

In [11]:

```
df_train.head()
```

Out[11]:

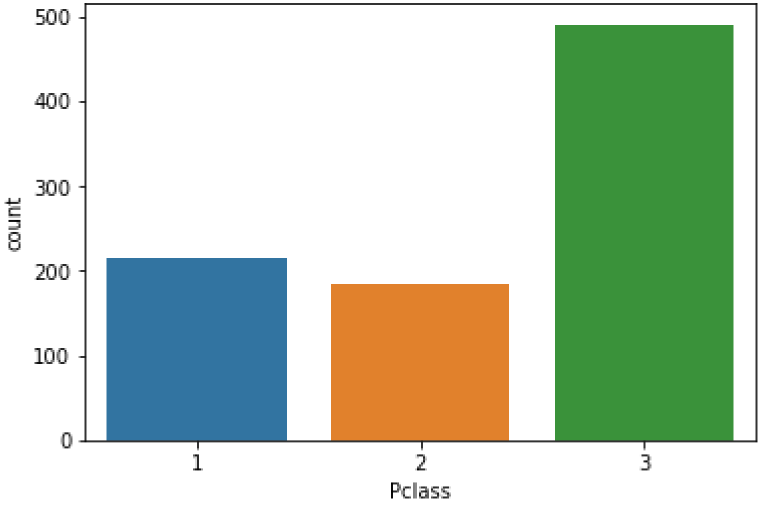
	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	0	3	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	1	1	female	38.0	1	0	PC 17599	71.2833	C85	C
2	1	3	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	1	1	female	35.0	1	0	113803	53.1000	C123	S
4	0	3	male	35.0	0	0	373450	8.0500	NaN	S

In [12]:

```
sns.countplot(x = 'Pclass',data = df_train)
df_train['Pclass'].value_counts()/df_train.shape[0]
```

Out[12]:

```
3    0.551066
1    0.242424
2    0.206510
Name: Pclass, dtype: float64
```

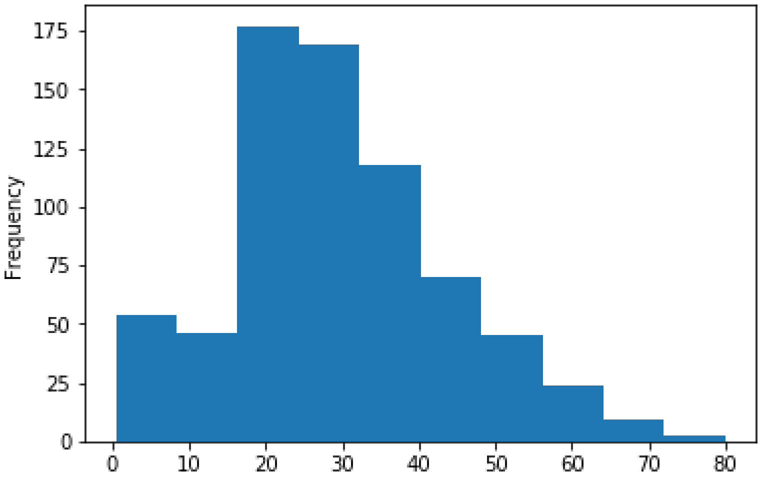


In [15]:

```
df_train['Age'].plot(kind = 'hist')
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x26f7e7a4808>
```

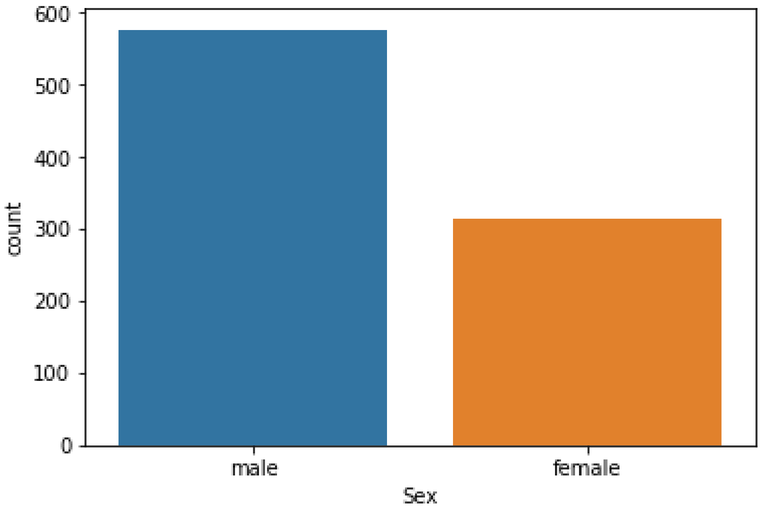


In [17]:
```
df_train['Age'].describe()
```

Out[17]:
```
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
25%       20.125000
50%       28.000000
75%       38.000000
max       80.000000
Name: Age, dtype: float64
```

In [18]:
```
sns.countplot(x = 'Sex',data = df_train)
df_train['Sex'].value_counts()/df_train.shape[0]
```

Out[18]:
```
male      0.647587
female    0.352413
Name: Sex, dtype: float64
```

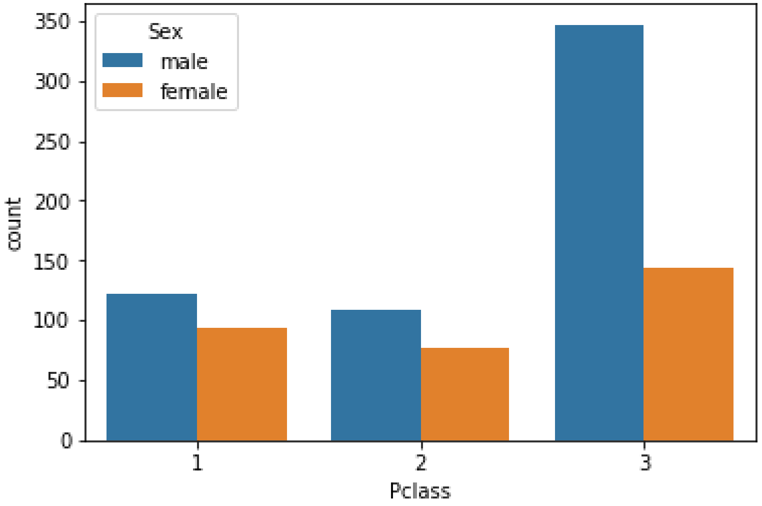


In [19]:

```python
sns.countplot(x = 'Pclass',data = df_train,hue = 'Sex')
df_train['Pclass'].value_counts()/df_train.shape[0]
```

Out[19]:

```
3    0.551066
1    0.242424
2    0.206510
Name: Pclass, dtype: float64
```



In [23]:

```python
def Adult_Child(Age):
    if Age>18:
        return "Adult"
    else:
        return "Minor"

```

In [24]:

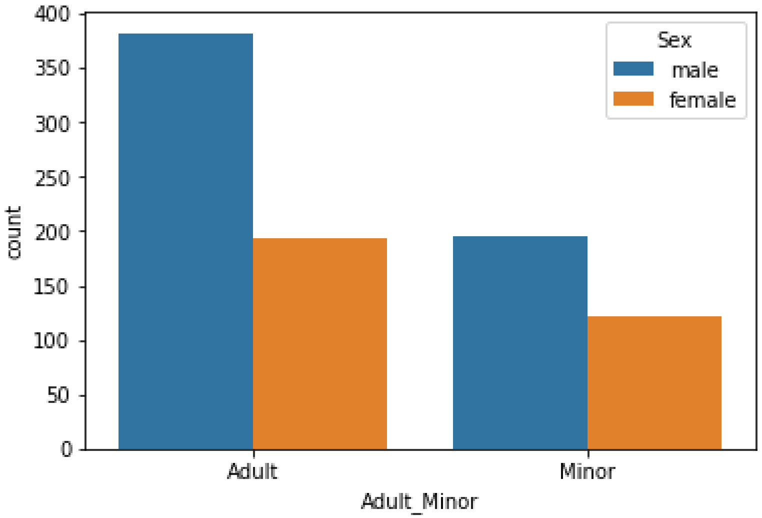
```python
df_train['Adult_Minor'] = df_train['Age'].apply(Adult_Child)
```

In [25]:

```
sns.countplot(x = 'Adult_Minor',data = df_train,hue = 'Sex')
df_train['Adult_Minor'].value_counts()/df_train.shape[0]
```

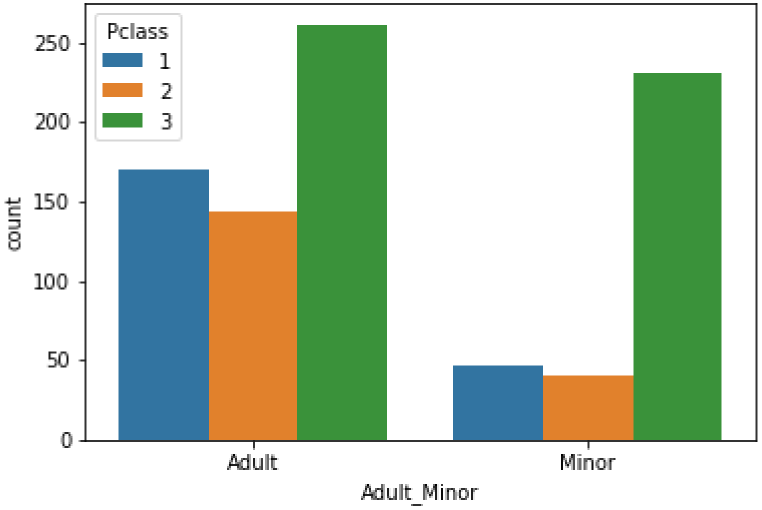
Out[25]:

```
Adult    0.645342
Minor    0.354658
Name: Adult_Minor, dtype: float64
```



In [26]:
```
sns.countplot(x = 'Adult_Minor',data = df_train,hue = 'Pclass')
df_train['Adult_Minor'].value_counts()/df_train.shape[0]
```

Out[26]:
```
Adult    0.645342
Minor    0.354658
Name: Adult_Minor, dtype: float64
```

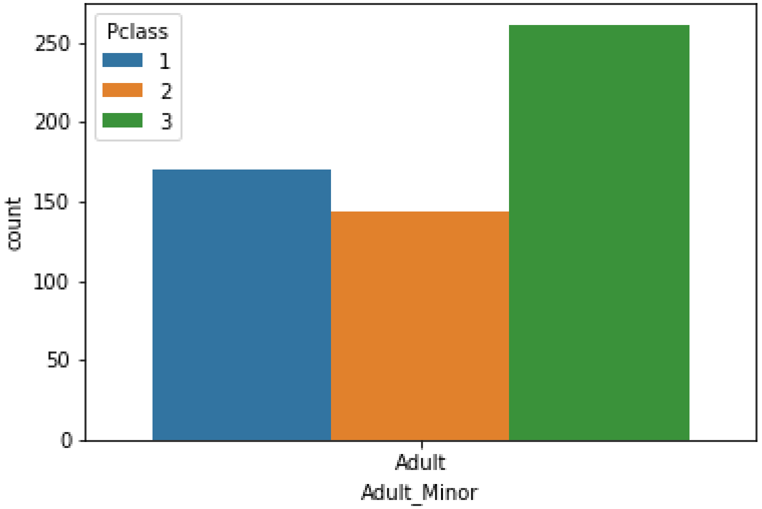


In [34]:

```
temp = df_train[df_train['Adult_Minor'] == "Adult" ]
sns.countplot(x = 'Adult_Minor',data = temp,hue = 'Pclass')
temp['Pclass'].value_counts()/temp.shape[0]
```

Out[34]:

```
3    0.453913
1    0.295652
2    0.250435
Name: Pclass, dtype: float64
```

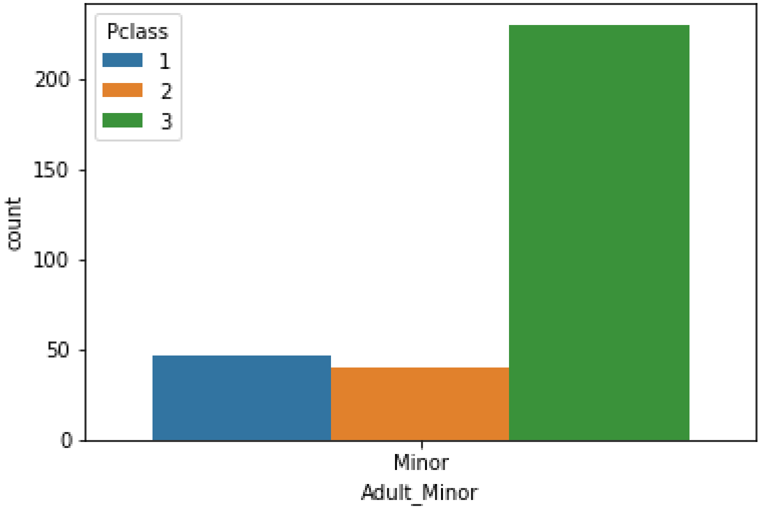


In [35]:

```
temp = df_train[df_train['Adult_Minor'] == "Minor" ]
sns.countplot(x = 'Adult_Minor',data = temp,hue = 'Pclass')
temp['Pclass'].value_counts()/temp.shape[0]
```

Out[35]:

```
3    0.727848
1    0.145570
2    0.126582
Name: Pclass, dtype: float64
```



In [36]:

```
def Person(Passenger):
    Sex,Age = Passenger
    if Age < 18:
        return 'Minor'
    else:
        return Sex
```

In [37]:

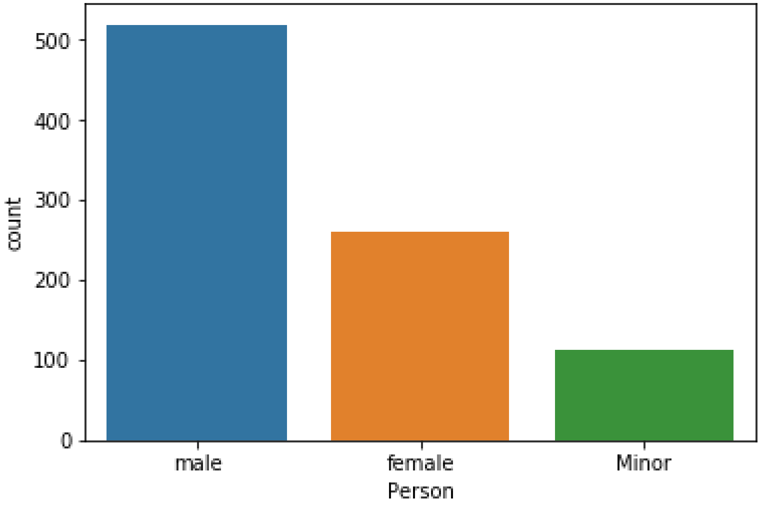
```
df_train['Person'] = df_train[['Sex','Age']].apply(Person,axis = 1)
```

In [39]:

```
sns.countplot(x = 'Person',data = df_train)
df_train['Pclass'].value_counts()/df_train.shape[0]
```

Out[39]:

```
3    0.551066
1    0.242424
2    0.206510
Name: Pclass, dtype: float64
```

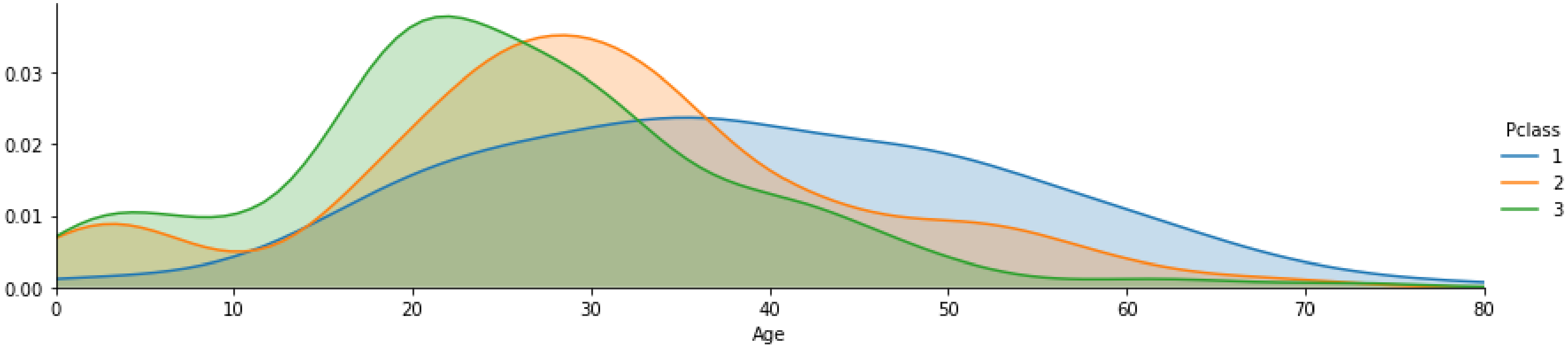


In [41]:

```
fig  = sns.FacetGrid(df_train,hue = 'Pclass',aspect = 4)
fig.map(sns.kdeplot,'Age',shade = True)

oldest = df_train['Age'].max()

fig.set(xlim = (0,oldest))

fig.add_legend()
```

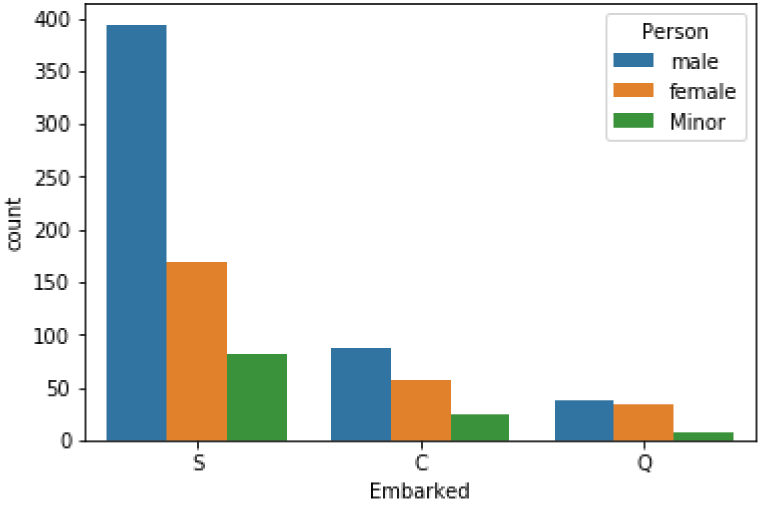
Out[41]:

```
<seaborn.axisgrid.FacetGrid at 0x26f514d3b48>
```



In [43]:
```
sns.countplot(x = 'Embarked',data = df_train,hue = 'Person')
df_train['Embarked'].value_counts()/df_train.shape[0]
```

Out[43]:
```
S    0.722783
C    0.188552
Q    0.086420
Name: Embarked, dtype: float64
```

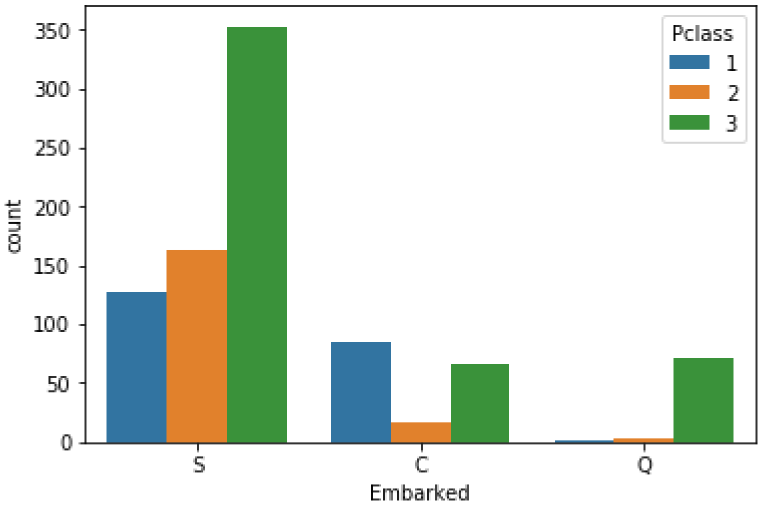


In [44]:

```
sns.countplot(x = 'Embarked',data = df_train,hue = 'Pclass')
df_train['Embarked'].value_counts()/df_train.shape[0]
```

Out[44]:

```
S    0.722783
C    0.188552
Q    0.086420
Name: Embarked, dtype: float64
```



In [ ]:

```
 
```

In [48]:
```
df_train.isna().sum()/df_train.shape[0]
```

Out[48]:
```
Survived       0.000000
Pclass         0.000000
Sex            0.000000
Age            0.198653
SibSp          0.000000
Parch          0.000000
Ticket         0.000000
Fare           0.000000
Cabin          0.771044
Embarked       0.002245
Adult_Minor    0.000000
Person         0.000000
dtype: float64
```

In [52]:
```
df_train['Cabin'].unique()

```

Out[52]:
```
array([nan, 'C85', 'C123', 'E46', 'G6', 'C103', 'D56', 'A6',
       'C23 C25 C27', 'B78', 'D33', 'B30', 'C52', 'B28', 'C83', 'F33',
       'F G73', 'E31', 'A5', 'D10 D12', 'D26', 'C110', 'B58 B60', 'E101',
       'F E69', 'D47', 'B86', 'F2', 'C2', 'E33', 'B19', 'A7', 'C49', 'F4',
       'A32', 'B4', 'B80', 'A31', 'D36', 'D15', 'C93', 'C78', 'D35',
       'C87', 'B77', 'E67', 'B94', 'C125', 'C99', 'C118', 'D7', 'A19',
       'B49', 'D', 'C22 C26', 'C106', 'C65', 'E36', 'C54',
       'B57 B59 B63 B66', 'C7', 'E34', 'C32', 'B18', 'C124', 'C91', 'E40',
       'T', 'C128', 'D37', 'B35', 'E50', 'C82', 'B96 B98', 'E10', 'E44',
       'A34', 'C104', 'C111', 'C92', 'E38', 'D21', 'E12', 'E63', 'A14',
       'B37', 'C30', 'D20', 'B79', 'E25', 'D46', 'B73', 'C95', 'B38',
       'B39', 'B22', 'C86', 'C70', 'A16', 'C101', 'C68', 'A10', 'E68',
       'B41', 'A20', 'D19', 'D50', 'D9', 'A23', 'B50', 'A26', 'D48',
       'E58', 'C126', 'B71', 'B51 B53 B55', 'D49', 'B5', 'B20', 'F G63',
       'C62 C64', 'E24', 'C90', 'C45', 'E8', 'B101', 'D45', 'C46', 'D30',
       'E121', 'D11', 'E77', 'F38', 'B3', 'D6', 'B82 B84', 'D17', 'A36',
       'B102', 'B69', 'E49', 'C47', 'D28', 'E17', 'A24', 'C50', 'B42',
       'C148'], dtype=object)
```

In [53]:

```
cabin = df_train['Cabin'].dropna()
levels = []

for i in cabin:
    levels.append(i[0])

```

In [55]:

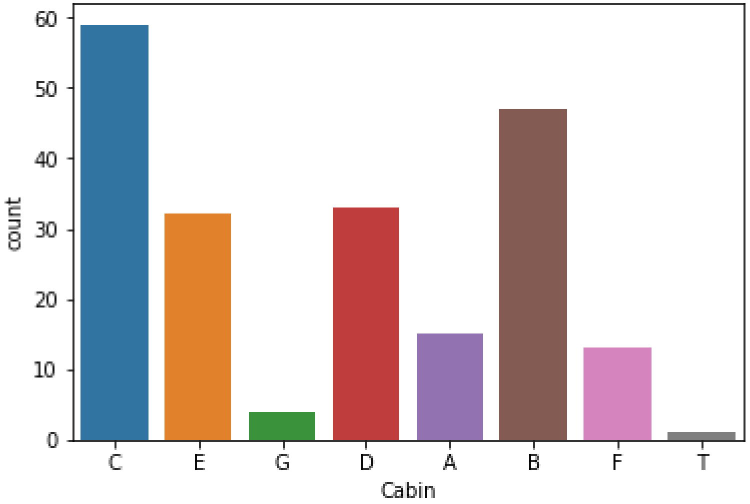
```
cabin = pd.DataFrame(levels,columns = ['Cabin'])
```

In [56]:

```
sns.countplot(x = 'Cabin',data = cabin)
```

Out[56]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x26f51387388>
```



In [59]:

```
df_train[['SibSp','Parch']]


df_train['People_acc'] = df_train['SibSp'] + df_train['Parch']
```

In [63]:

```
df_train['People_acc'].loc[df_train['People_acc']>0] = 'With Family'
df_train['People_acc'].loc[df_train['People_acc'] == 0] = 'Alone'
```

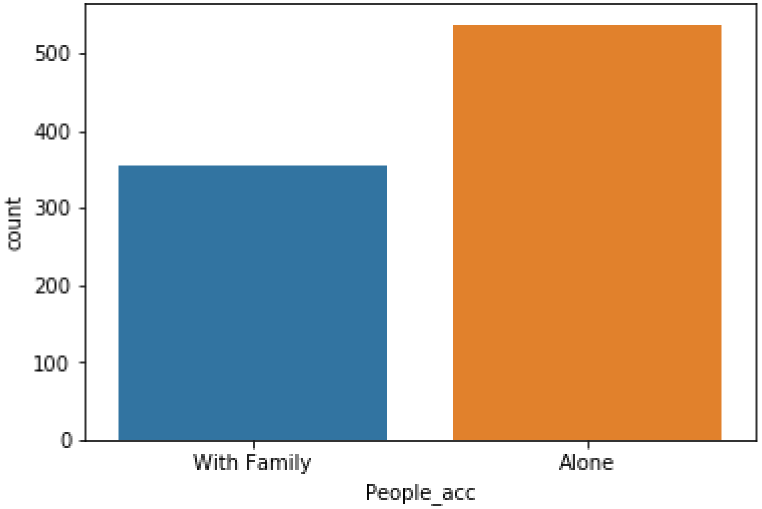
```
C:\Users\yashm\anaconda3\lib\site-packages\pandas\core\indexing.py:670: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a
-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-co
py)
  iloc._setitem_with_indexer(indexer, value)
```

In [64]:

```
sns.countplot('People_acc',data = df_train)
```

Out[64]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x26f51dddec8>
```



In [ ]:

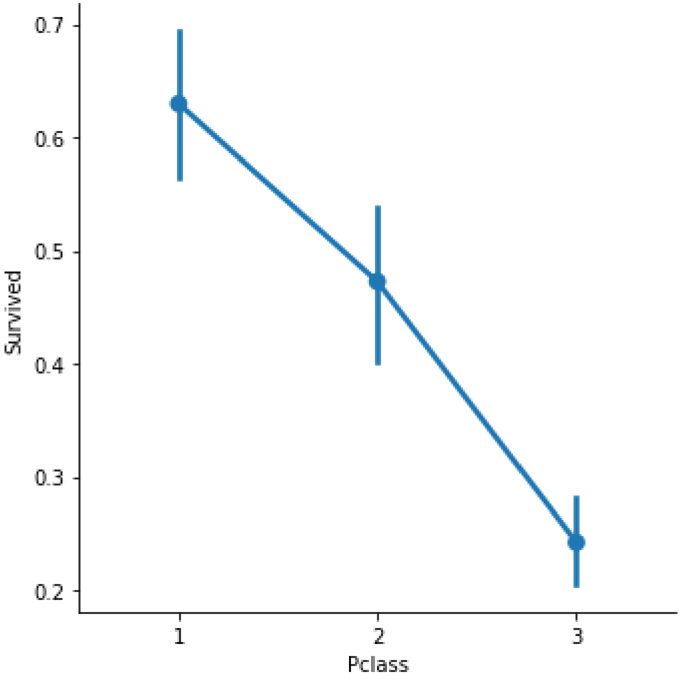
```

```

EDA wrt our target variable

In [65]:
```
sns.factorplot(x = 'Pclass',y = 'Survived',data = df_train)
```

```
C:\Users\yashm\anaconda3\lib\site-packages\seaborn\categorical.py:3669: UserWarning: The `factorplot` function has been
renamed to `catplot`. The original name will be removed in a future release. Please update your code. Note that the def
ault `kind` in `factorplot` (`'point'`) has changed `'strip'` in `catplot`.
  warnings.warn(msg)
```

Out[65]:
```
<seaborn.axisgrid.FacetGrid at 0x26f52304e48>
```



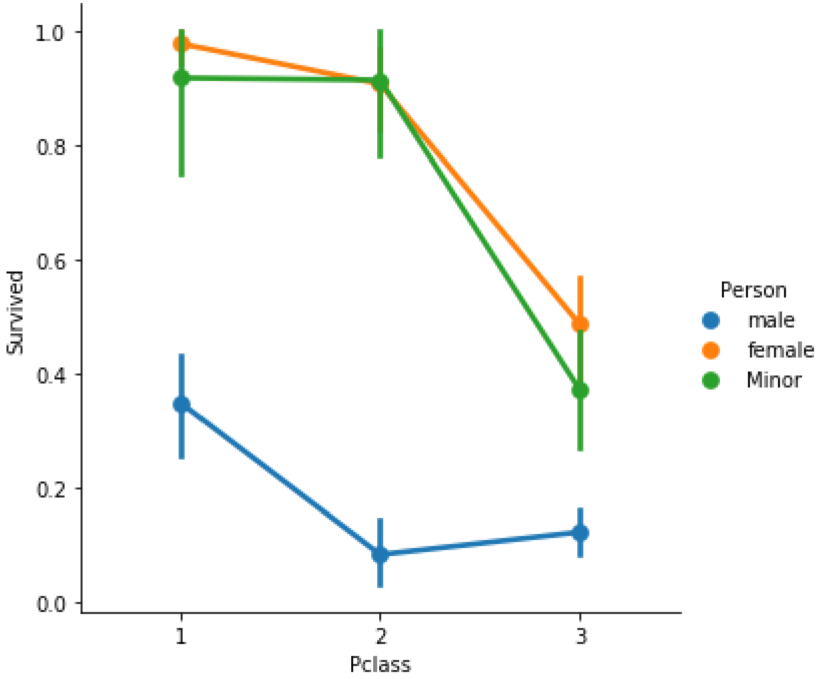
In [74]:

```python
sns.factorplot(x = 'Pclass',y = 'Survived',data = df_train,hue = 'Person')
```

```
C:\Users\yashm\anaconda3\lib\site-packages\seaborn\categorical.py:3669: UserWarning: The `factorplot` function has been
renamed to `catplot`. The original name will be removed in a future release. Please update your code. Note that the def
ault `kind` in `factorplot` (`'point'`) has changed `'strip'` in `catplot`.
  warnings.warn(msg)
```

Out[74]:

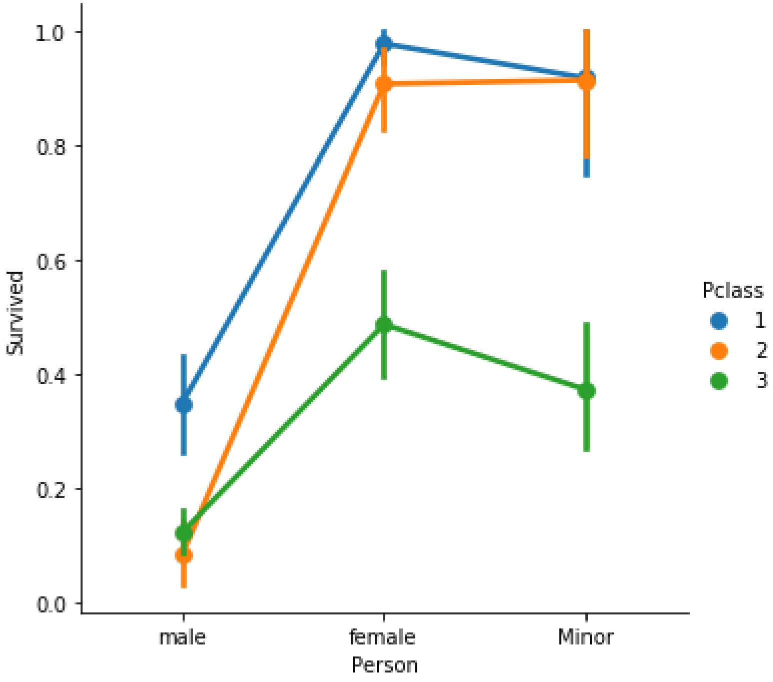
```
<seaborn.axisgrid.FacetGrid at 0x26f514ca888>
```



In [77]:
```
sns.factorplot(x = 'Person',y = 'Survived',data = df_train,hue='Pclass')
```

```
C:\Users\yashm\anaconda3\lib\site-packages\seaborn\categorical.py:3669: UserWarning: The `factorplot` function has been
renamed to `catplot`. The original name will be removed in a future release. Please update your code. Note that the def
ault `kind` in `factorplot` (`'point'`) has changed `'strip'` in `catplot`.
  warnings.warn(msg)
```

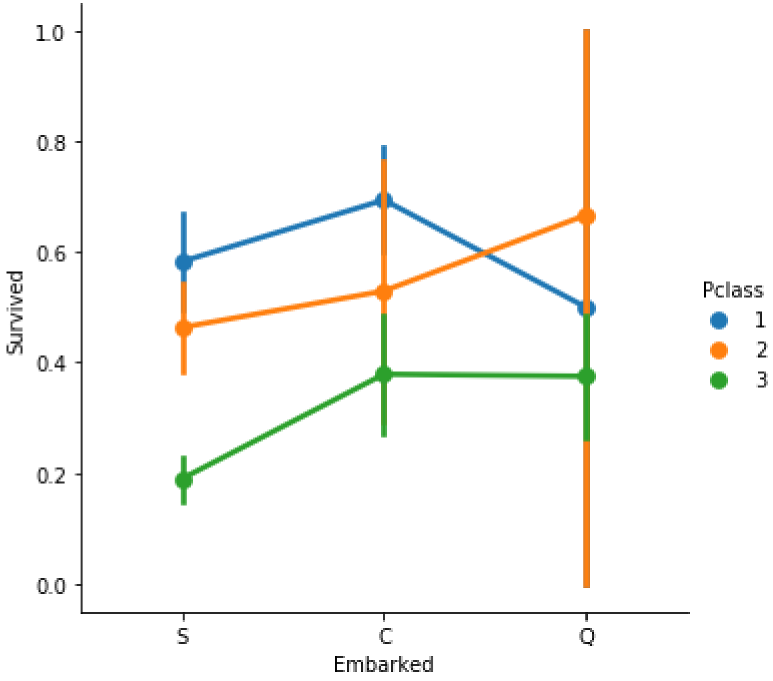
Out[77]:
```
<seaborn.axisgrid.FacetGrid at 0x26f5208d688>
```



In [68]:
```
sns.factorplot(x = 'Embarked',y = 'Survived',data = df_train,hue = 'Pclass')
```

```
C:\Users\yashm\anaconda3\lib\site-packages\seaborn\categorical.py:3669: UserWarning: The `factorplot` function has been
renamed to `catplot`. The original name will be removed in a future release. Please update your code. Note that the def
ault `kind` in `factorplot` (`'point'`) has changed `'strip'` in `catplot`.
  warnings.warn(msg)
```

Out[68]:
```
<seaborn.axisgrid.FacetGrid at 0x26f52568508>
```

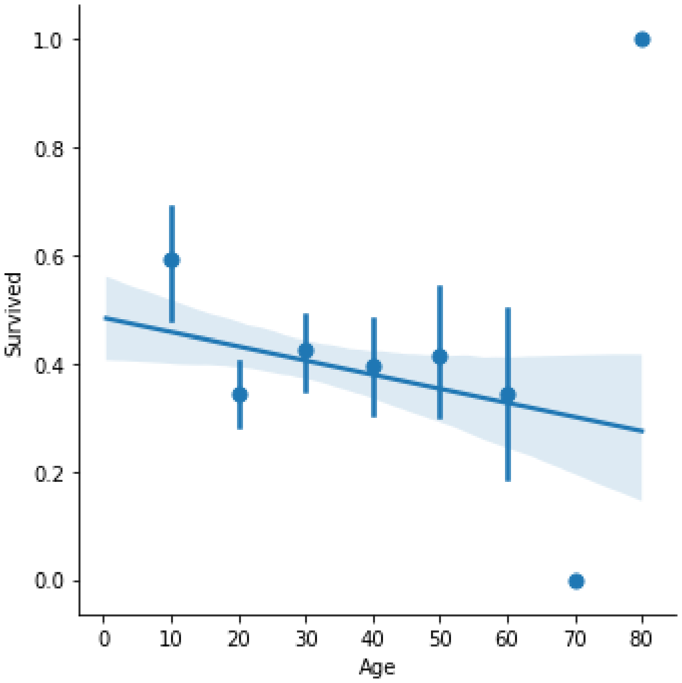


In [71]:

```
sns.lmplot('Age','Survived',data = df_train,x_bins = [10,20,30,40,50,60,70,80])
```

Out[71]:

```
<seaborn.axisgrid.FacetGrid at 0x26f52159188>
```

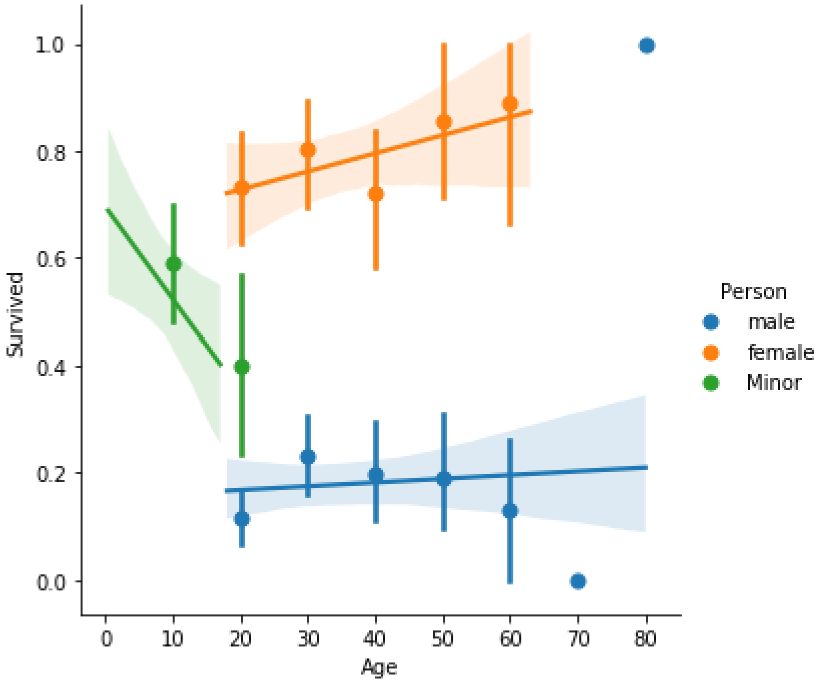


In [79]:

```
sns.lmplot('Age','Survived',data = df_train,x_bins = [10,20,30,40,50,60,70,80],hue = 'Person')
```

Out[79]:

```
<seaborn.axisgrid.FacetGrid at 0x26f52a57048>
```

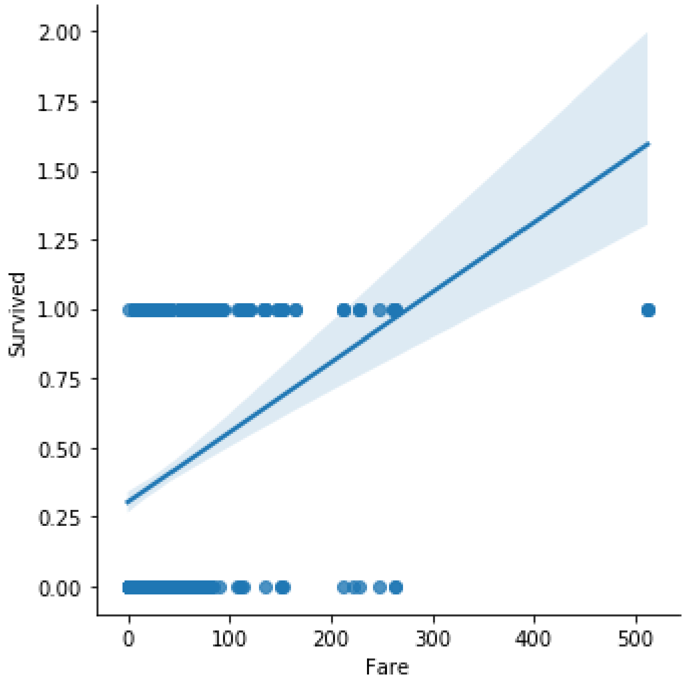


In [82]:

```
sns.lmplot('Fare','Survived',data = df_train)
```

Out[82]:

```
<seaborn.axisgrid.FacetGrid at 0x26f52b302c8>
```



In [86]:

```
df_train['Fare'].sort_values()
```

Out[86]:

```
271      0.0000
597      0.0000
302      0.0000
633      0.0000
277      0.0000
         ...  
438    263.0000
341    263.0000
737    512.3292
258    512.3292
679    512.3292
Name: Fare, Length: 891, dtype: float64
```

In [ ]:

```

```