Homework 1: Data Preprocessing and Analyze by pivoting features of Titanic dataset

- Uses train.csv and test.csv
- Sections will be separated by similar questions.

Questions 1 - 6: Basic Data Descriptors

1. In **training** set, which features are **available**?
2. In **training** set, which features are **categorical**?
3. In **training** set, which features are **numerical** (e.g., discrete, continuous, or time series based)?
4. In **training** set, which features are mixed data types (Carbin is not mixed data type: [C19 C18 C17], Ticket is a mixed data type: 342354, SA/12434324)?
5. In **training** set, which features contain blank, null or empty values? In **test** set, which features contain blank, null or empty values?
6. In **training** set, what are the data types (e.g., integer, floats or strings) for various features?

- ***Package Imports and File Input***

In [3]:
```
#packages for data pre-processing, analysis, and visualization
#NumPy allows users to perform many different types of dataframe analysis
import numpy as np
#Pandas, built on top of NumPy, allows users to pre-process and clean data, along wit
import pandas as pd
#seaborn allows users to perform multiple visualization techniques
import seaborn as sns
%matplotlib inline
sns.set_theme(style="darkgrid")
#pyplot, part of matplotlib, allows users to plot various types of data
import matplotlib.pyplot as plt

train_df = pd.read_csv('train.csv')
test_df = pd.read_csv('test.csv')
combine = [train_df, test_df]
```

1. In **training** set, which features are **available**?

In [4]:
```
print(train_df.columns.values)
```
```
['PassengerId' 'Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch'
 'Ticket' 'Fare' 'Cabin' 'Embarked']
```

1. In **training** set, which features are **categorical**?
2. In **training** set, which features are **numerical** (e.g., discrete, continuous, or time series based)?
3. In **training** set, which features are mixed data types (Carbin is not mixed data type: [C19 C18 C17], Ticket is a mixed data type: 342354, SA/12434324)?
4. In **training** set, which features contain blank, null or empty values? In **test** set, which features contain blank, null or empty values?
5. In **training** set, what are the data types (e.g., integer, floats or strings) for various features?

In [5]:
```
train_df.describe
train_df.tail()
```

Out[5]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

In [6]:
```
test_df.describe
test_df.head()
```

Out[6]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

1. Training Set:

- (Question 2) Categorical: Survived, Sex, Embarked, Pclass
- (Question 3) Numerical:
  1. (Discrete): SibSp, Parch
  2. (Continuous): Age, Fare
- (Question 4) Mixed Data Types: Ticket
- (Question 5) Blank, Null, or Empty: Survived, Age, Cabin, Embarked

1. Test Set:

- (Question 5) Blank, Null, or Empty: Age, Cabin

1. In **training** set, what are the data types (e.g., integer, floats or strings) for various features?

In [7]:
```
train_df.dtypes
```

Out[7]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

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

Questions 7 - 13: Distributions and Correlations

1. In training set, to understand the distribution of numerical feature values across the samples, please list the properties, including count, mean, std, min, 25% percentile, 50% percentile, 75% percentile, max, of numerical features?
2. In training set, to understand the distribution of categorical features, we define: count is the total number of categorical values per column; unique is the total number of unique categorical values per column; top is the most frequent categorical value; freq is the total number of the most frequent categorical value. Please list the properties, including count, unique, top, freq, of categorical features?
3. In training set, can you observe significant correlation (average survivied ratio>0.5) among the group of Pclass=1 and Survived? If Pclass has significant correlation with Survivied, we should include this feature in the predictive model. Based on your computation, will you include this feature in the predictive model?
4. In training set, are Women (Sex=female) were more likely to have survived?
5. In training set, let us start by understanding correlations between a numeric feature (Age) and our predictive goal (Survived). A histogram chart is useful for analyzing continuous numerical variables like Age where banding or ranges will help identify useful patterns. The histogram can indicate distribution of samples using automatically defined bins or equally ranged bands. This helps us answer questions relating to specific bands (e.g., infants, old). Please plot the histograms between ages and Survived (Figure 1 is an example), and answer the following questions:

- Do infants (Age <=4) have high survival rate?
- Do oldest passengers (Age = 80) survive?
- Do large number of 15-25 year olds not survive?

Based on your analysis of the figures,

- Should we consider Age in our model training? (If yes, then we should complete the Age feature for null values.)
- Should we should band age groups?

1. In training set, we can combine three features (age, Pclass, and survivied) for identifying correlations using a single plot. This can be done with numerical and categorical features which have numeric values. Please plot the plot using python, and answer the following questions:

- Does Pclass=3 have most passengers, however most did not survive?
- Do infant passengers in Pclass=2 and Pclass=3 mostly survive?
- Do most passengers in Pclass=1 survive?
- Does Pclass vary in terms of Age distribution of passengers?
- Should we consider Pclass for model training?

1. In training set, we want to correlate categorical features (with non-numeric values) and numeric features. We can consider correlating Embarked (Categorical non-numeric), Sex (Categorical nonnumeric), Fare (Numeric continuous), with Survived (Categorical numeric). Please plot a figure to illustrate the correlations of Embarked, Sex, Fare, and Survivied. And answer the following questions:

- Do higher fare paying passengers have better survival?
- Should we consider banding fare feature?

1. In training set, to understand the distribution of numerical feature values across the samples, please list the properties, including count, mean, std, min, 25% percentile, 50% percentile, 75% percentile, max, of numerical features?

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

Out[8]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

1. In **training** set, to understand the distribution of categorical features, we define:

- count is the total number of categorical values per column
- unique is the total number of unique categorical values per column
- top is the most frequent categorical value
- freq is the total number of the most frequent categorical value.

Please list the properties, including count, unique, top, freq, of categorical features.

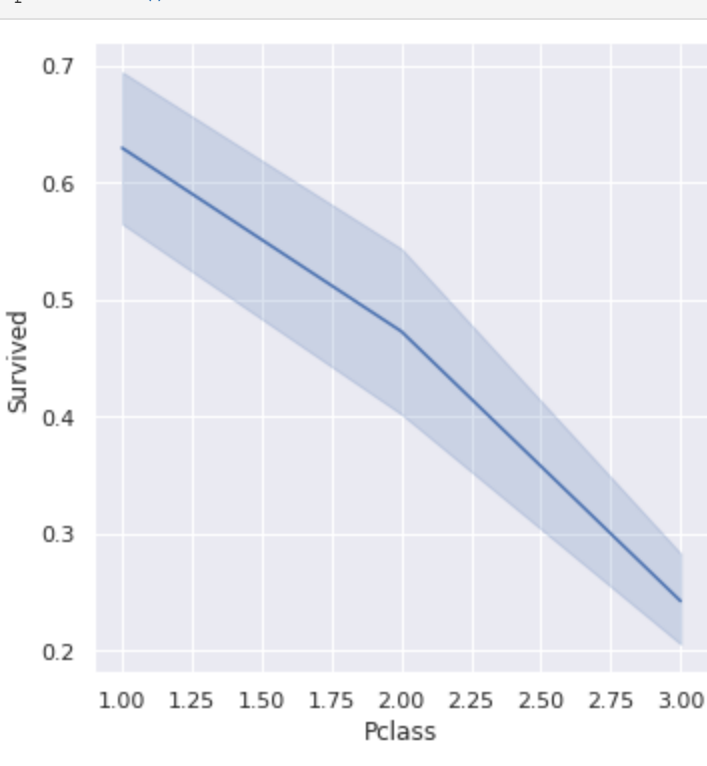
In [9]:
```
# Categorical Features of Training Set: Survived, Sex, Embarked, Pclass

train_df.info()
train_df.describe(include=['O'])
```
```
<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
```

Out[9]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Radeff, Mr. Alexander	male	1601	B96 B98	S
freq	1	577	7	4	644

1. In training set, can you observe significant correlation (average survivied ratio>0.5) among the group of Pclass=1 and Survived? If Pclass has significant correlation with Survivied, we should include this feature in the predictive model. Based on your computation, will you include this feature in the predictive model?

In [10]:
```
sns.relplot(x = "Pclass", y = "Survived", kind="line", data = train_df)
plt.show()
```

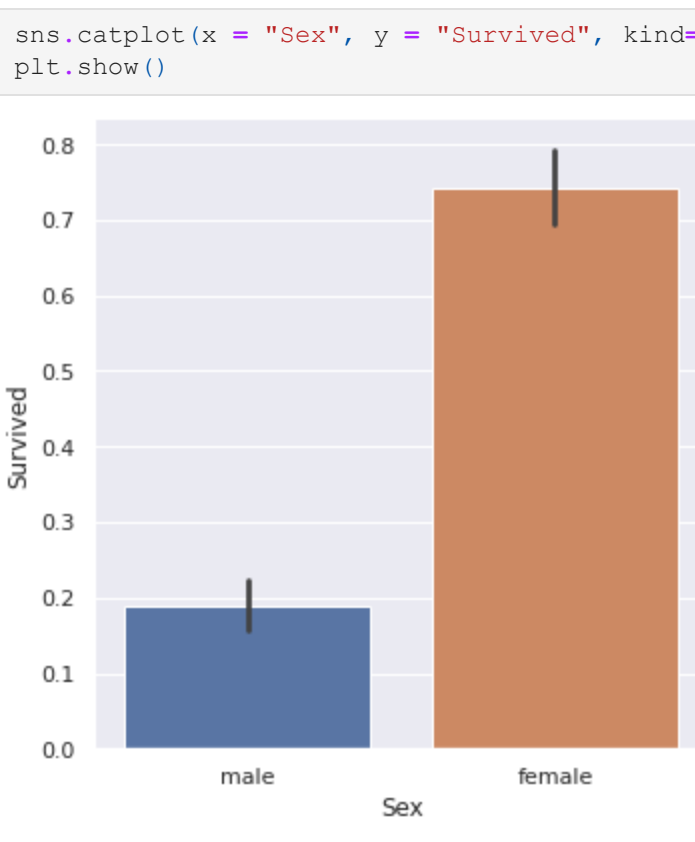


Correlation is significant between Pclass = 1, and those who survived (which is about 65%).

Data should be included.

1. In training set, are Women (Sex=female) were more likely to have survived?

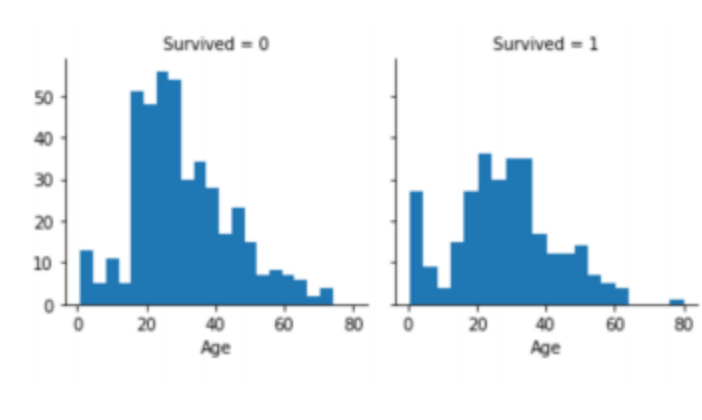
In [11]:
```
sns.catplot(x = "Sex", y = "Survived", kind="bar", data = train_df)
plt.show()
```



Yes, women had about a 55% higher rate of survival.

1. In training set, let us start by understanding correlations between a numeric feature (Age) and our predictive goal (Survived). A histogram chart is useful for analyzing continuous numerical variables like Age where banding or ranges will help identify useful patterns. The histogram can indicate distribution of samples using automatically defined bins or equally ranged bands. This helps us answer questions relating to specific bands (e.g., infants, old). Please plot the histograms between ages and Survived (Figure 1 is an example), and answer the following questions:

- Do infants (Age <=4) have high survival rate?
- Do oldest passengers (Age = 80) survive?
- Do large number of 15-25 year olds not survive?

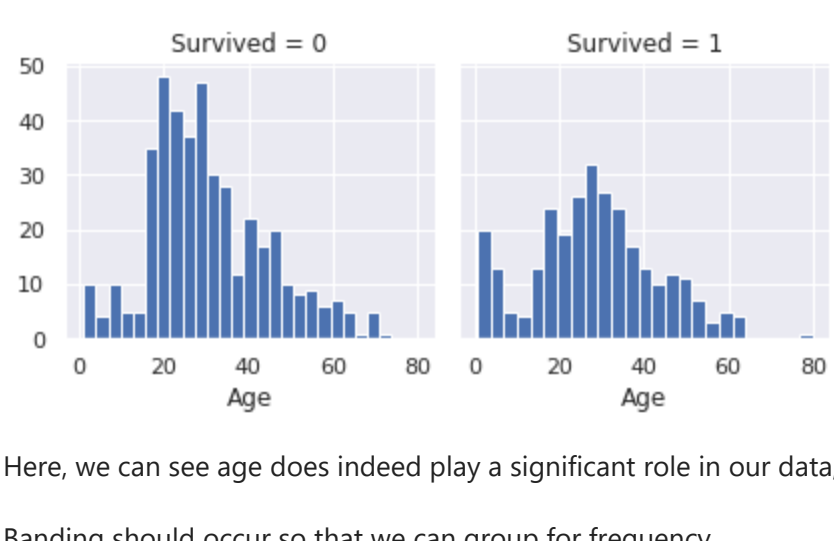


Based on your analysis of the figures,

- Should we consider Age in our model training? (If yes, then we should complete the Age feature for null values.)
- Should we should band age groups?

In [12]:
```
g = sns.FacetGrid(train_df, col='Survived')
g.map(plt.hist, 'Age', bins=25)
```

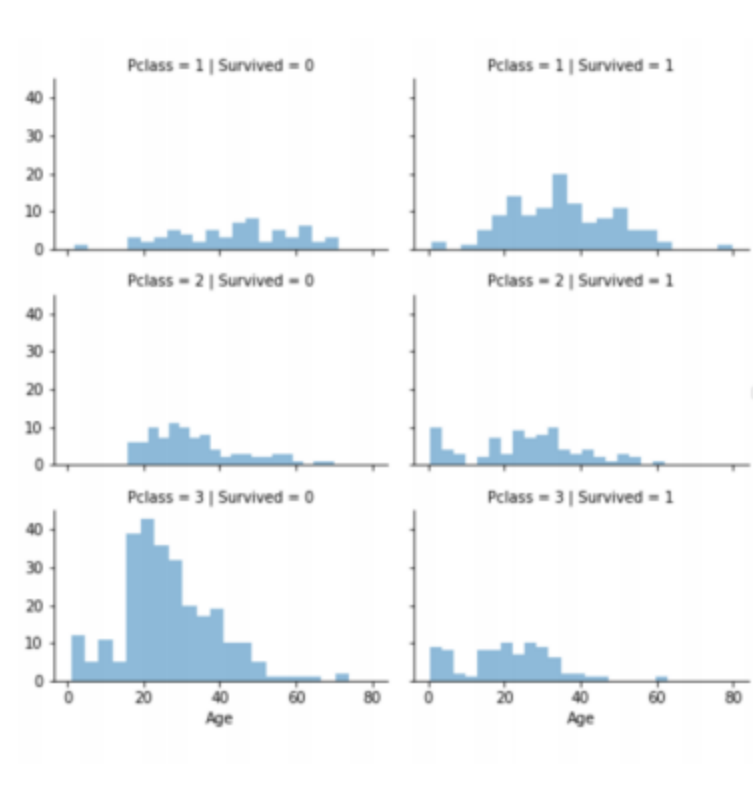
Out[12]: <seaborn.axisgrid.FacetGrid at 0x7f8c50e19e80>



Here, we can see age does indeed play a significant role in our data, and should be used in model training.

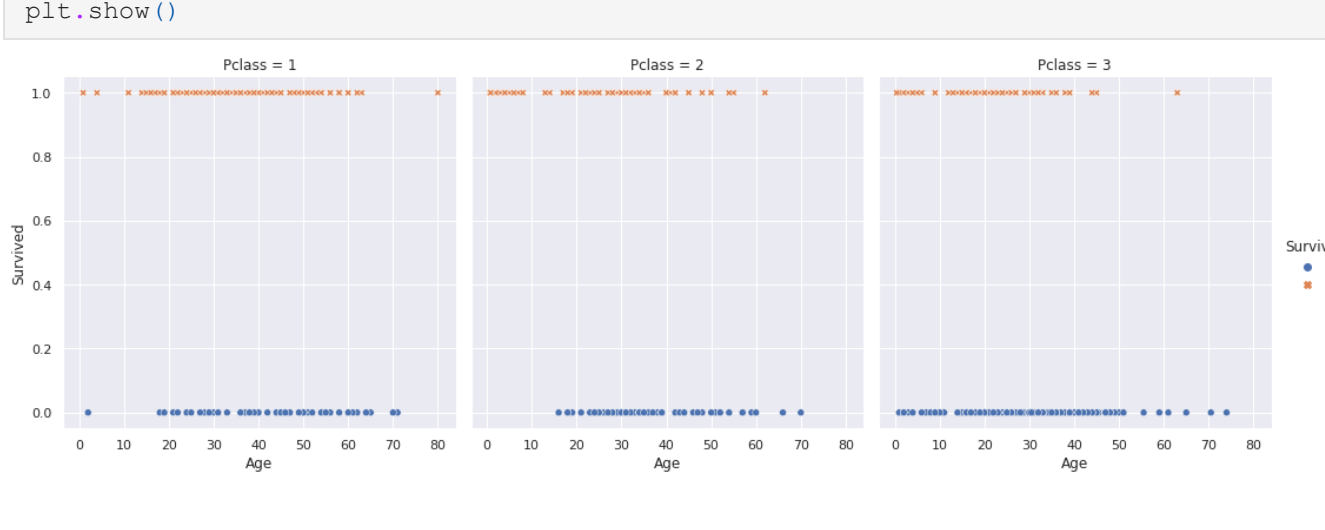
Banding should occur so that we can group for frequency.

1. In training set, we can combine three features (age, Pclass, and survivied) for identifying correlations using a single plot. This can be done with numerical and categorical features which have numeric values. Please plot the plot using python, and answer the following questions:

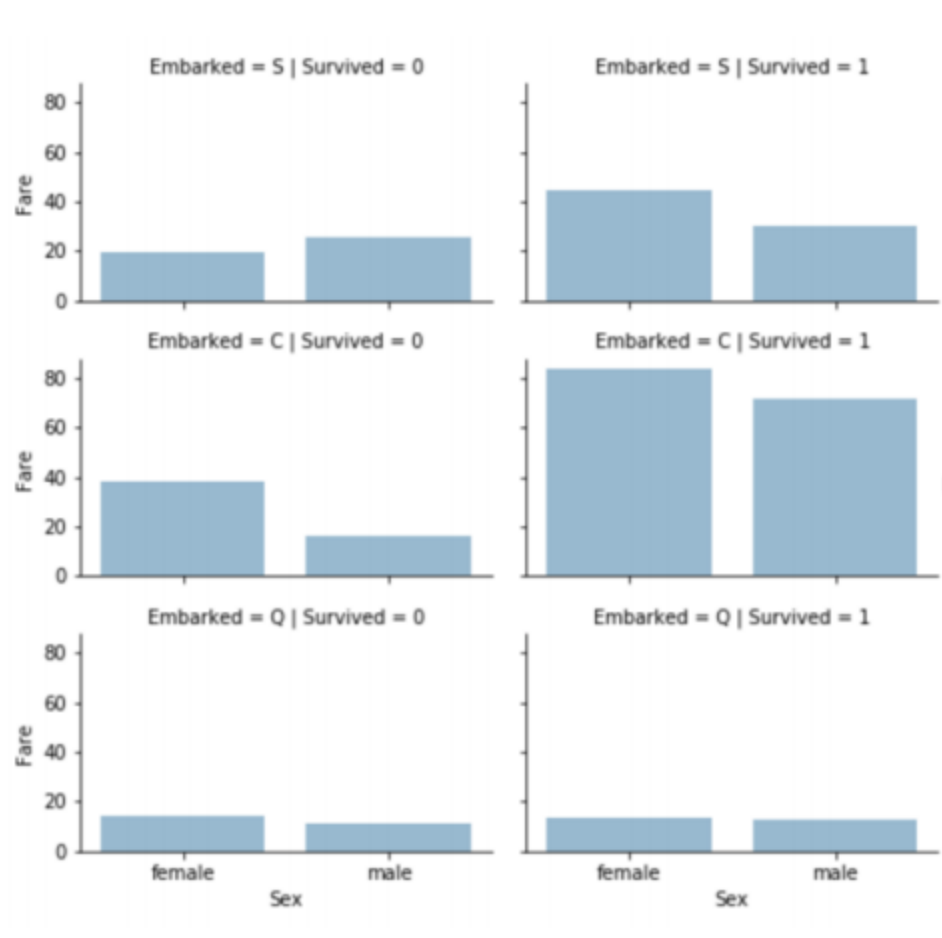


- Does Pclass=3 have most passengers, however most did not survive?
- Do infant passengers in Pclass=2 and Pclass=3 mostly survive?
- Do most passengers in Pclass=1 survive?
- Does Pclass vary in terms of Age distribution of passengers?
- Should we consider Pclass for model training?

In [13]:
```
sns.relplot(x = "Age", y = "Survived", hue="Survived", col="Pclass", style="Survived",
plt.show()
```

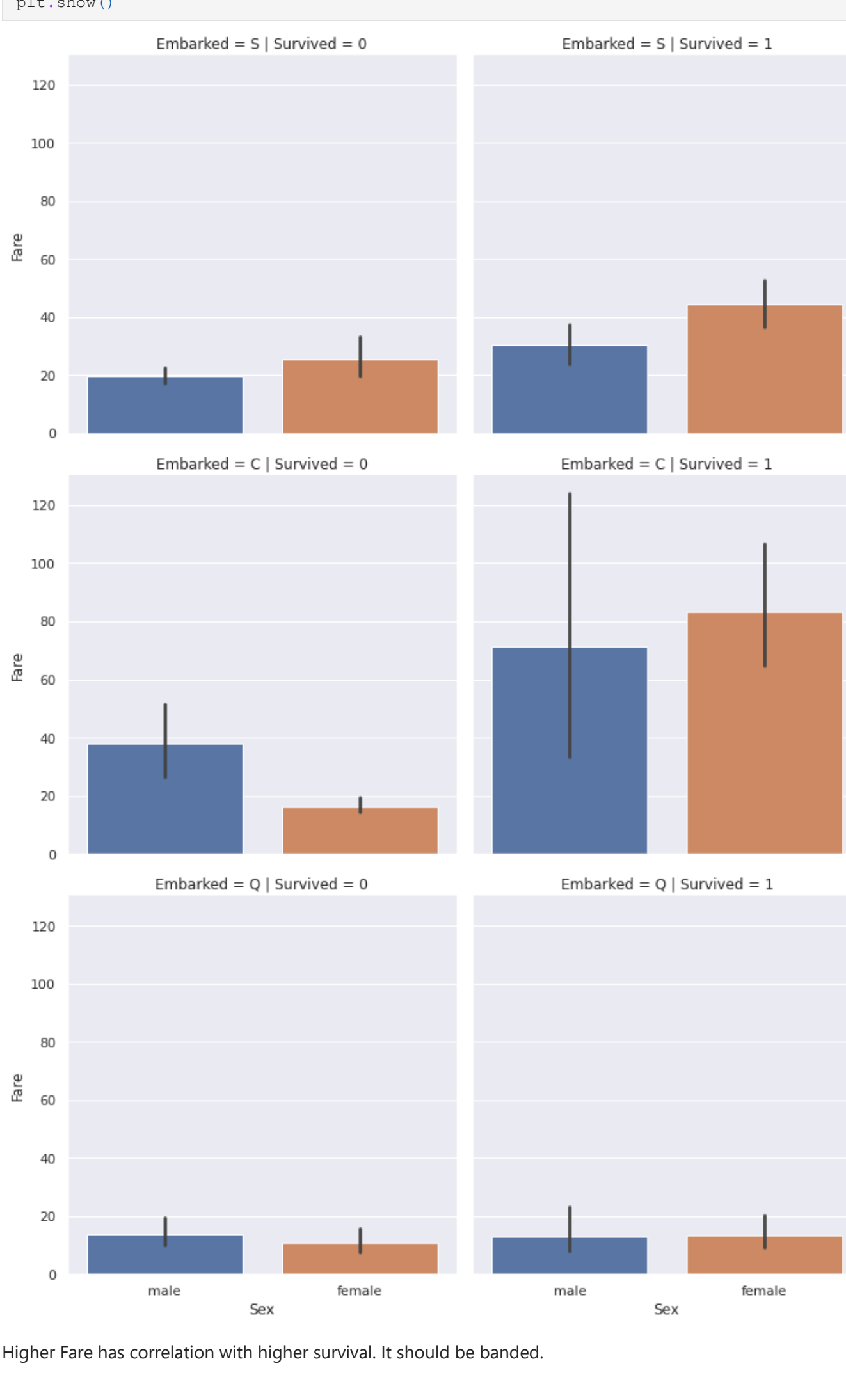


1. In training set, we want to correlate categorical features (with non-numeric values) and numeric features. We can consider correlating Embarked (Categorical non-numeric), Sex (Categorical nonnumeric), Fare (Numeric continuous), with Survived (Categorical numeric). Please plot a figure to illustrate the correlations of Embarked, Sex, Fare, and Survivied. And answer the following questions:



- Do higher fare paying passengers have better survival?
- Should we consider banding fare feature?

In [23]:
```
sns.catplot(x = "Sex", y = "Fare", row="Embarked", col="Survived", kind="bar", data =
plt.show()
```



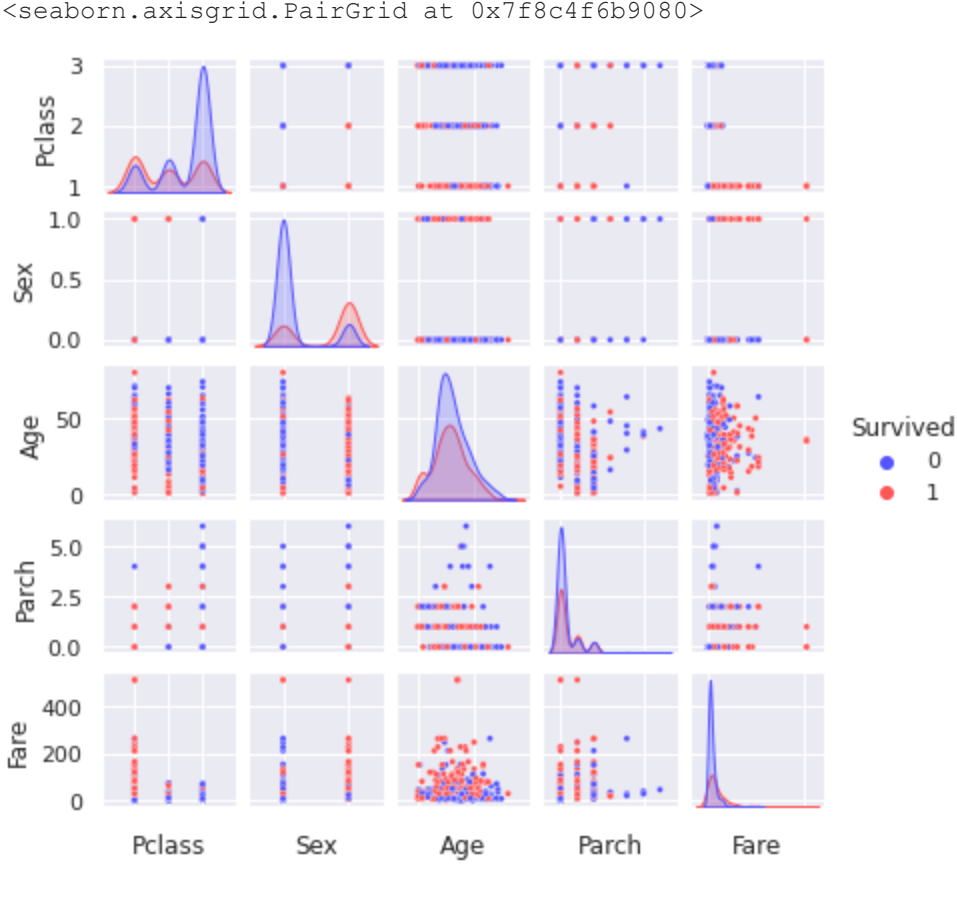
Higher Fare has correlation with higher survival. It should be banded.

Questions 14 - 20: Data Pre-Processing and Analysis

1. In training set, what is the rate of duplicates for the Ticket feature? Is there a correlation between Ticket and survival? Should we drop the Ticket feature?
2. In the training set, Is the Cabin feature complete? How many null values there are in the Cabin features of the combined dataset of training and test dataset? Should we drop the Cabin feature?
3. In the training set, we can convert features which contain strings to numerical values. This is required by most model algorithms. Doing so will also help us in achieving the feature completing goal. In this question ,please convert Sex feature to a new feature called Gender where female=1 and male=0.
4. In the training set, we start estimating and completing features with missing or null values. We will first do this for the Age feature. We can consider three methods to complete a numerical continuous feature. A simple way is to generate random numbers between mean and standard deviation. More accurate way of guessing missing values is to use the K-Nearest Neighbor algorithm to select the top-K most similar data points, and then use the top-K most similar data points to impute the missing values of ages.
5. In the training set, complete a categorical feature: Embarked feature takes S, Q, C values based on port of embarkation. Our training dataset has some missing values. Please simply fill these with the most common occurrences.
6. In the training set, complete and convert a numeric feature. Please complete the Fare feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature.
7. In the training set, convert the Fare feature to ordinal values based on the FareBand (defined in Assignment PDF).

In [43]:
```
g = sns.pairplot(train_df[[u'Survived', u'Pclass', u'Sex', u'Age', u'Parch', u'Fare',
                 diag_kws=dict(shade=True),plot_kws=dict(s=10) )
g.set(xticklabels=[])
```
```
/usr/local/lib/python3.6/dist-packages/seaborn/axisgrid.py:1969: UserWarning: The `siz
e` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[43]: <seaborn.axisgrid.PairGrid at 0x7f8c4f6b9080>



1. In training set, what is the rate of duplicates for the Ticket feature? Is there a correlation between Ticket and survival? Should we drop the Ticket feature?

In [25]:
```
train_df.describe(include=['O'])
#ticket uniqueness ratio = 891 - 681 = 210/891 * 100% = 23.6%
```

Out[25]:

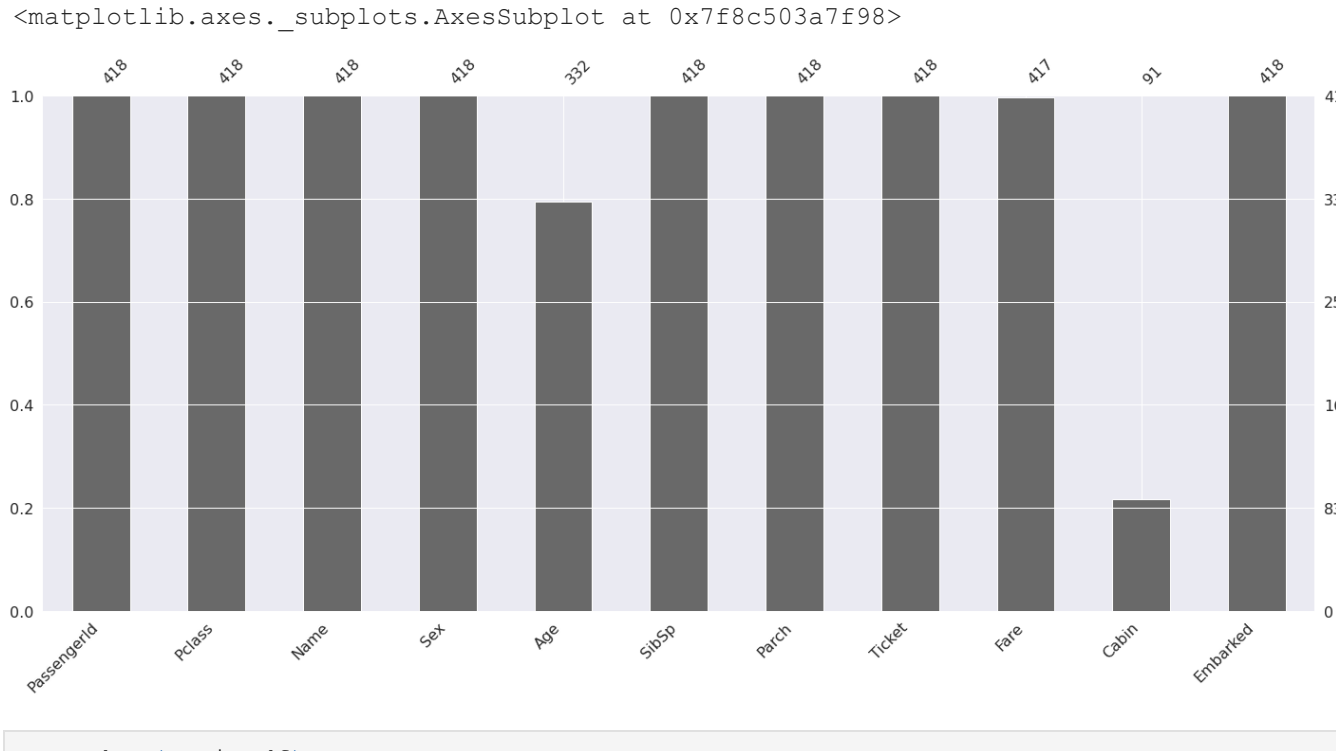
	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Radeff, Mr. Alexander	male	1601	B96 B98	S
freq	1	577	7	4	644

- Ticket uniqueness ratio = 891 - 681 = 210/891 * 100% = 23.6%
- Tickets should be dropped
- Correlation is not found between ticket and survival

1. In the training set, Is the Cabin feature complete? How many null values there are in the Cabin features of the combined dataset of training and test dataset? Should we drop the Cabin feature?

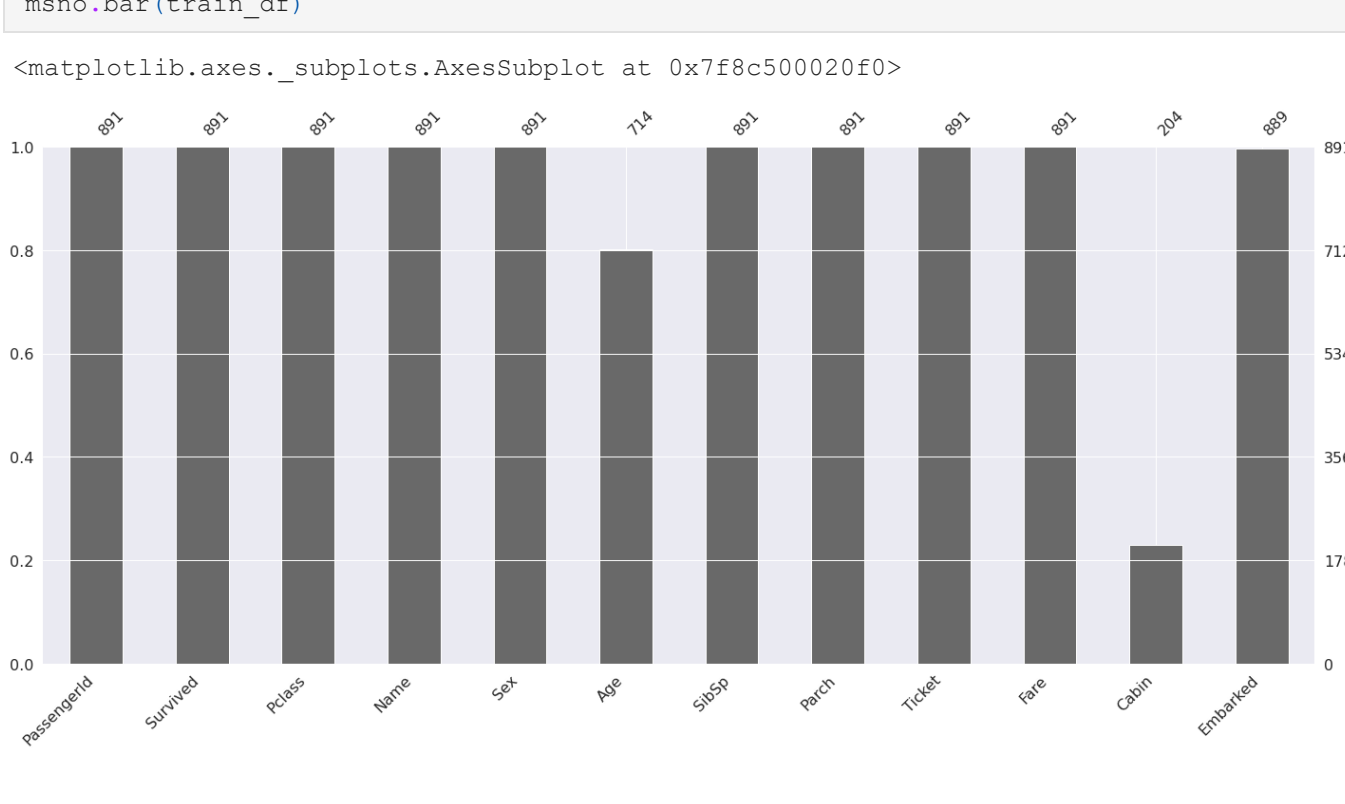
In [29]:
```
import missingno as msno
msno.bar(test_df)
#train_df.describe(include=['O'])
#cabin completeness ratio = 204/891 * 100% = 22.9%
```

Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8c503a7f98>



In [31]:
```
msno.bar(train_df)
```

Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8c500020f0>



- Cabin completeness ratio (training) = 204/891 * 100% = 22.9%
- Cabin completeness ratio (test) = 91/891 * 100% = 10.2%
- Cabin completeness ratio (combined) = 151/891 * 100% = 16.9%
- Cabin should be dropped

1. In the training set, we can convert features which contain strings to numerical values. This is required by most model algorithms. Doing so will also help us in achieving the feature completing goal. In this question ,please convert Sex feature to a new feature called Gender where female=1 and male=0.

In [72]:
```
combine = [train_df, test_df]
#for dataset in combine:
dataset.replace({'sex':{'female':0, 'male':1}}, inplace=True)
g = sns.heatmap(dataset[["Age","Sex","SibSp","Parch","Pclass"]].corr(),cmap="BrBG",ann
```

1. In the training set, we start estimating and completing features with missing or null values. We will first do this for the Age feature. We can consider three methods to complete a numerical continuous feature. A simple way is to generate random numbers between mean and standard deviation. More accurate way of guessing missing values is to use the K-Nearest Neighbor algorithm to select the top-K most similar data points, and then use the top-K most similar data points to impute the missing values of ages.

In [74]:

Out[74]:

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

1. In the training set, complete a categorical feature: Embarked feature takes S, Q, C values based on port of embarkation. Our training dataset has some missing values. Please simply fill these with the most common occurrences.

In [76]:
```
highest_embark = train_df.Embarked.dropna().mode()[0]
for dataset in combine:
  dataset['Embarked'] = dataset['Embarked'].fillna(highest_embark)
train_df[['Embarked', 'Survived']].groupby(['Embarked'], as_index=False).mean().sort_
```

Out[76]:

	Embarked	Survived
1	1	0.553571
2	2	0.389610
0	0	0.339009

1. In the training set, complete and convert a numeric feature. Please complete the Fare feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature.

In [81]:
```
test_df['Fare'].fillna(test_df['Fare'].dropna().median(), inplace=True)
train_df['BandedF'] = pd.qcut(train_df['Fare'], 4)
train_df[['BandedF', 'Survived']].groupby(['BandedF'], as_index=False).mean().sort_val
```

Out[81]:

	BandedF	Survived
0	(-0.001, 7.91]	0.197309
1	(7.91, 14.454]	0.303571
2	(14.454, 31.0]	0.454955
3	(31.0, 512.329]	0.581081

1. In the training set, convert the Fare feature to ordinal values based on the FareBand (defined in Assignment PDF).

In [93]:
```
#train_df = train_df.drop(['BandedF'], axis=1)
for dataset in combine:
  dataset.loc[ dataset['Fare'] <= 7.91, 'Fare'] = 0
  dataset.loc[(dataset['Fare'] > 7.91) & (dataset['Fare'] <= 14.454), 'Fare'] = 1
  dataset.loc[(dataset['Fare'] > 14.454) & (dataset['Fare'] <= 31), 'Fare'] = 2
  dataset.loc[ dataset['Fare'] > 31, 'Fare'] = 3
  dataset['Fare'] = dataset['Fare'].astype(int)
combine = [train_df, test_df]
train_df.head(50)
```

Out[93]:

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

Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500		S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500		S
14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0000		S
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.1250		Q
18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.0	1	0	345763	18.0000		S
20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250		C
21	0	2	Fynney, Mr. Joseph J	male	35.0	0	0	239865	26.0000		S
22	1	3	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.0000	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15.0	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28.0	0	0	113788	35.5000	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8.0	3	1	349909	21.0750		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38.0	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250		C
28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0000	C23 C25 C27	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lalio	male	NaN	0	0	349216	7.8958		S
31	0	1	Uruchurtu, Don. Manuel E	male	40.0	0	0	PC 17601	27.7208		C
32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NaN	1	0	PC 17569	146.5208	B78	C
33	1	3	Glynn, Miss. Mary Agatha	female	NaN	0	0	335677	7.7750		Q
34	0	2	Wheadon, Mr. Edward H	male	66.0	0	0	C.A. 24579	10.5000		S
35	0	1	Meyer, Mr. Edgar Joseph	male	28.0	1	0	PC 17604	82.1708		C
36	0	1	Holverson, Mr. Alexander Oskar	male	42.0	1	0	113789	52.0000		S
37	1	3	Mamee, Mr. Hanna	male	NaN	0	0	2677	7.2292		C
38	0	3	Cann, Mr. Ernest Charles	male	21.0	0	0	A./5. 2152	8.0500		S
39	0	3	Vander Planke, Miss. Augusta Maria	female	18.0	2	0	345764	18.0000		S
40	1	3	Nicola-Yarred, Miss. Jamila	female	14.0	1	0	2651	11.2417		C
41	0	3	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	female	40.0	1	0	7546	9.4750		S
42	0	2	Turpin, Mrs. William John Robert (Dorothy Ann Wonnacott)	female	27.0	1	0	11668	21.0000		S
43	0	3	Kraeff, Mr. Theodor	male	NaN	0	0	349253	7.8958		C
44	1	2	Laroche, Miss. Simonne Marie Anne Andree	female	3.0	1	2	SC/Paris 2123	41.5792		C
45	1	3	Devaney, Miss. Margaret Delia	female	19.0	0	0	330958	7.8792		Q
46	0	3	Rogers, Mr. William John	male	NaN	0	0	S.C./A.4. 23567	8.0500		S
47	1	3	Lennon, Mr. Denis	male	NaN	1	0	370371	15.5000		Q
48	1	3	O'Driscoll, Miss. Bridget	female	NaN	0	0	14311	7.7500		Q
49	0	3	Samaan, Mr. Youssef	male	2.0	2	0	2662	21.6792		C
50	0	3	Arnold-Franchi, Mrs. Josef (Josefine Franchi)	female	18.0	1	0	A/5 3540	17.8000		S
51	0	3	Panula, Master. Juha Niilo	male	7.0	4	1	3101295	39.6875		S
52	0	3	Nosworthy, Mr. Richard Cater	male	21.0	0	0	A/4. 39886	7.8000		S
53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun Bennett)	female	49.0	1	0	PC 17572	76.7292	D33	C
54	1	2	Faunthorpe, Mrs. Lizzie (Elizabeth Anne Wilkinson)	female	29.0	1	0	2926	26.0000		S
55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.9792	B30	C
56	1	1	Woolner, Mr. Hugh	male	NaN	0	0	19947	35.5000	A21	S
57	1	2	Rugg, Miss. Emily	female	21.0	0	0	C.A. 31026	10.5000		S
58	0	3	Novel, Mr. Mansouer	male	28.5	0	0	2697	7.2292		C
59	1	2	West, Miss. Constance Mirium	female	5.0	1	2	C.A. 34651	27.7500		S
60	0	3	Goodwin, Master. William Frederick	male	11.0	5	2	CA 2144	46.9000		S
61	0	3	Sirayanian, Mr. Orsen	male	22.0	0	0	2669	7.2292		C
62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0000	B28	
63	0	1	Harris, Mr. Henry Birkhardt	male	45.0	1	0	36973	83.4750	C83	S
64	0	3	Skoog, Master. Harald	male	4.0	3	2	347088	27.9000		S
65	0	1	Stewart, Mr. Albert A	male	NaN	0	0	PC 17605	27.7208		C
66	1	3	Moubarek, Master. Gerios	male	NaN	1	1	2661	15.2458		C
67	1	2	Nye, Mrs. (Elizabeth Ramell)	female	29.0	0	0	C.A. 29395	10.5000	F33	S
68	0	3	Crease, Mr. Ernest James	male	19.0	0	0	S.P. 3464	8.1583		S
69	1	3	Andersson, Miss. Erna Alexandra	female	17.0	4	2	3101281	7.9250		S
70	0	3	Kink, Mr. Vincenz	male	26.0	2	0	315151	8.6625		S
71	0	2	Jenkin, Mr. Stephen Curnow	male	32.0	0	0	C.A. 33111	10.5000		S
72	0	3	Goodwin, Miss. Lillian Amy	female	16.0	5	2	CA 2144	46.9000		S
73	0	2	Hood, Mr. Ambrose Jr	male	21.0	0	0	S.O.C. 14879	73.5000		S
74	0	3	Chronopoulos, Mr. Apostolos	male	NaN	1	0	2680	14.4542		C
75	1	3	Bing, Mr. Lee	male	32.0	0	0	1601	56.4958		S
76	0	3	Moen, Mr. Sigurd Hansen	male	25.0	0	0	348123	7.6500	F G73	S
77	0	3	Staneff, Mr. Ivan	male	NaN	0	0	349208	7.8958		S
78	0	3	Moutal, Mr. Rahamin Haim	male	NaN	0	0	374746	8.0500		S
79	1	2	Caldwell, Master. Alden Gates	male	0.83	0	2	248738	29.0000		S
80	1	3	Dowdell, Miss. Elizabeth	female	30.0	0	0	364516	12.4750		S
81	0	3	Waelens, Mr. Achille	male	22.0	0	0	345767	9.0000		S
82	1	3	Sheerlinck, Mr. Jan Baptist	male	29.0	0	0	345779	9.5000		S
83	1	3	McDermott, Miss. Brigdet Delia	female	NaN	0	0	330932	7.7875		Q
84	0	2	Carrau, Mr. Francisco M	male	28.0	0	0	113059	47.1000		S
85	1	2	Ilett, Miss. Bertha	female	17.0	0	0	SO/C 14885	10.5000		S
86	1	3	Backstrom, Mrs. Karl Alfred (Maria Mathilda Gustafsson)	female	33.0	3	0	3101278	15.8500		S
87	0	3	Ford, Mr. William Neal	male	16.0	1	3	W./C. 6608	34.3750		S
88	0	3	Slocovski, Mr. Selman Francis	male	NaN	0	0	SOTON/OQ 392086	8.0500		S
89	1	1	Fortune, Miss. Mabel Helen	female	23.0	3	2	19950	263.0000	C23 C25 C27	S
90	0	3	Celotti, Mr. Francesco	male	24.0	0	0	343275	8.0500		S
91	0	3	Christmann, Mr. Emil	male	29.0	0	0	343276	8.0500		S
92	0	3	Andreasson, Mr. Paul Edvin	male	20.0	0	0	347466	7.8542		S
93	0	1	Chaffee, Mr. Herbert Fuller	male	46.0	1	0	W.E.P. 5734	61.1750	E31	S
94	0	3	Dean, Mr. Bertram Frank	male	26.0	1	2	C.A. 2315	20.5750		S
95	0	3	Coxon, Mr. Daniel	male	59.0	0	0	364500	7.2500		S
96	0	3	Shorney, Mr. Charles Joseph	male	NaN	0	0	374910	8.0500		S
97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C
98	1	1	Greenfield, Mr. William Bertram	male	23.0	0	1	PC 17759	63.3583	D10 D12	C
99	1	2	Doling, Mrs. John T (Ada Julia Bone)	female	34.0	0	1	231919	23.0000		S
100	0	2	Kantor, Mr. Sinai	male	34.0	1	0	244367	26.0000		S