In [1]:

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

import Data

In [2]:

```python
df = pd.read_csv('titanic_data.csv')
```

In [3]:

```python
df.head()
```

Out[3]:

	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 [4]:

```python
df.describe()
```

Out[4]:

	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

In [5]:

```python
df.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
```

analysis question

What factors made people more likely to survive?

remove columns that we don't need

In [6]:

```python
#Name and Ticket does not add any valuble information
#cabin has alot of Nan values which will not be useful
```

```
df.drop(columns = ['Name','Ticket','Cabin'], inplace=True)
df.head()
```

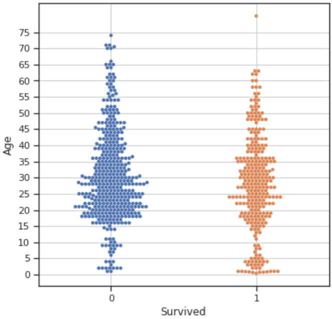
Out[6]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	0	3	male	22.0	1	0	7.2500	S
1	2	1	1	female	38.0	1	0	71.2833	C
2	3	1	3	female	26.0	0	0	7.9250	S
3	4	1	1	female	35.0	1	0	53.1000	S
4	5	0	3	male	35.0	0	0	8.0500	S

how age affects survival

In [7]:

```
sns.set(rc={"figure.figsize":(6, 6)}, style="ticks")
sns.swarmplot(x=df['Survived'], y=df['Age'],size=4);
plt.yticks(range(0,80,5))
plt.grid()
plt.show()
```

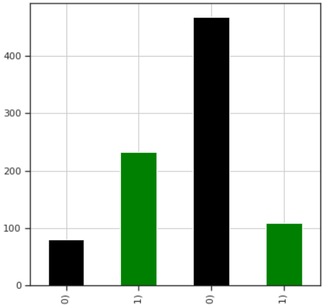


- it appears that for children < 6 the survival rate is high
- for people between ages 15 to 35 the survival rate is low

how sex affects survival

In [8]:

```
Survive_or_not= df.groupby(['Sex','Survived'])['PassengerId'].count()
Survive_or_not.plot(kind='bar',color=['black','green']);
plt.grid()
```



In [9]:

```
Survive_or_not
```

Out[9]:

```
Sex     Survived
female  0            81
        1           233
male    0           468
        1           109
Name: PassengerId, dtype: int64
```

In [10]:

```
percentage_of_females_survived = (Survive_or_not[1])/(Survive_or_not[0]+Survive_or_not[1])*100
percentage_of_females_survived
```

Out[10]:

```
74.20382165605095
```

In [11]:

```
percentage_of_males_survived = (Survive_or_not[3])/(Survive_or_not[2]+Survive_or_not[3])*100
percentage_of_males_survived
```

Out[11]:

```
18.890814558058924
```

- Clearly more females survived than males
- more than 74% of females survived
- only around 19% of males survived

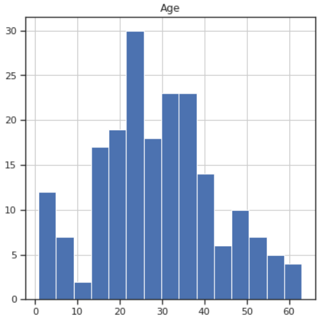
ages of females that survived

In [12]:

```
def get_ages(df, gender):
    gender_survived = df[ (df['Survived']==1) & (df['Sex'] == gender) ][['Age']]
    return gender_survived
```

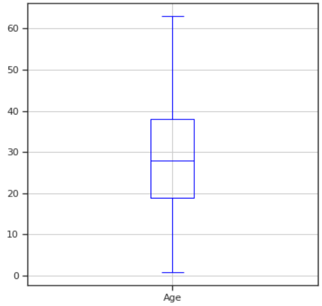
In [13]:

```
age_female_survived = get_ages(df, 'female')
age_female_survived.hist(bins = 15);
```



In [14]:

```
age_female_survived.boxplot(color='blue');
```

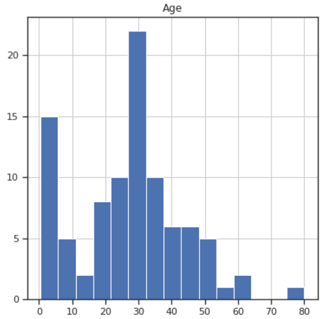


- most of the females survived were between ages 19 and 48

ages of males that survived

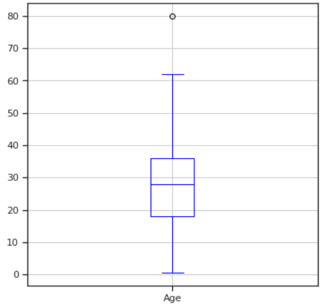
In [15]:

```
age_male_survived = get_ages(df, 'male')
age_male_survived.hist(bins = 15);
```



In [16]:

```
age_male_survived.boxplot(color='blue');
```

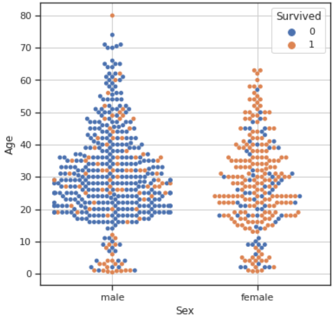


- most of the males survived were around 30 or children less than 6
- one male of age 80 managed to survive

age and sex swarmplot

In [17]:

```
sns.swarmplot(x=df['Sex'], y=df['Age'],hue=df['Survived'],size=5);
plt.grid()
plt.show()
```



- for ages between 19 and 60 females have much higher survival rate

Pclass effect on survival

In [18]:

```
df[['Survived','Pclass']].head()
```

Out[18]:

	Survived	Pclass
0	0	3
1	1	1
2	1	3
3	1	1
4	0	3

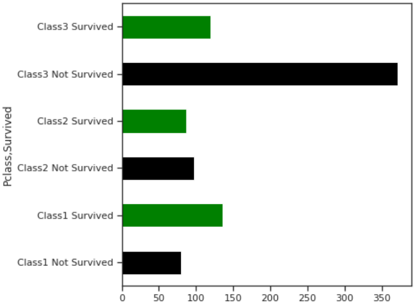
In [19]:

```
#function that plots bar graph based on given categorical column and grouped by survived column
def bar_graph_col_and_survive(col_name, labels):
    col_name_survive_group = df.groupby([col_name,'Survived'])['PassengerId'].count()
    print(col_name_survive_group)
    col_name_survive_group.plot(kind='barh',color=['black','green']);
    plt.yticks(ticks=np.arange(0,6), labels=labels);
    return col_name_survive_group
```

In [20]:

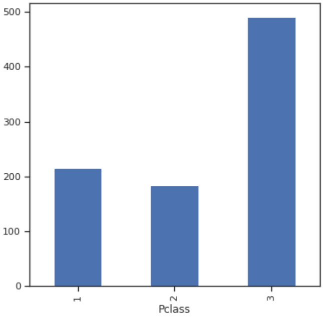
```
labels=["Class1 Not Survived", "Class1 Survived", "Class2 Not Survived", "Class2 Survived","Class3 Not Survivec
Survive_pclass = bar_graph_col_and_survive('Pclass', labels)
```

```
Pclass  Survived
1       0            80
        1           136
2       0            97
        1            87
3       0           372
        1           119
Name: PassengerId, dtype: int64
```



In [21]:

```
df.groupby(['Pclass'])['PassengerId'].count().plot(kind='bar');
```



- alot of class 3 did not survive but we can also see that there were more passengers in class 3 than the other classes
- class 1 has the highest survival rate

Embarked effect on survival

In [22]:

```
df.head()
```

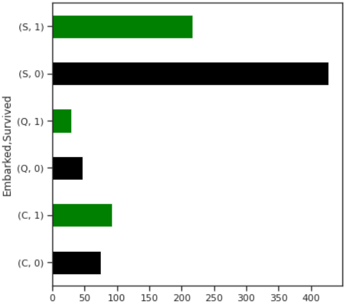
Out[22]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	0	3	male	22.0	1	0	7.2500	S
1	2	1	1	female	38.0	1	0	71.2833	C
2	3	1	3	female	26.0	0	0	7.9250	S
3	4	1	1	female	35.0	1	0	53.1000	S
4	5	0	3	male	35.0	0	0	8.0500	S

In [23]:

```
Survive_embarked = bar_graph_col_and_survive("Embarked",None)
```

```
Embarked  Survived
C         0            75
          1            93
Q         0            47
          1            30
S         0           427
          1           217
Name: PassengerId, dtype: int64
```

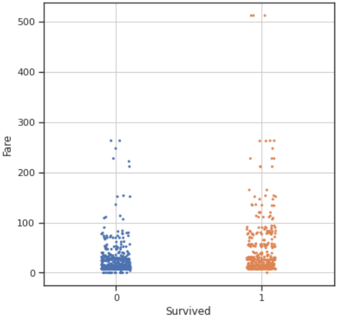


- it seems S embarked passengers had lowest survival rate

fare price effect survival

In [24]:

```
sns.stripplot(x=df['Survived'], y=df['Fare'], size=3);
plt.grid()
```



- we can see the three people with highest fare price survived
- we can remove outliners for better visual of the rest of the figure

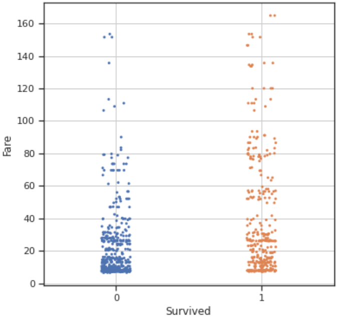
In [25]:

```
#remove outliner
q_low = df["Fare"].quantile(0.02)
q_hi  = df["Fare"].quantile(0.98)

df_filtered = df[(df["Fare"] < q_hi) & (df["Fare"] > q_low)]
```

In [26]:

```
sns.stripplot(x=df_filtered['Survived'], y=df['Fare'], size=3);
plt.grid()
```



- for fare price more than 100 people had better survival rate
- for fare price below 20 the survival rate is low
- for fare price between 20 and 60 the distributions are close

parch

number of parents / children aboard the Titanic

In [27]:
```
df[['Survived','Parch']]
```

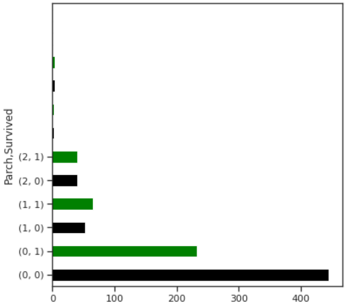
Out[27]:

	Survived	Parch
0	0	0
1	1	0
2	1	0
3	1	0
4	0	0
...	...	...
886	0	0
887	1	0
888	0	2
889	1	0
890	0	0

891 rows × 2 columns

In [28]:
```
Survived_parch = bar_graph_col_and_survive("Parch",None)
```

```
Parch  Survived
0      0           445
       1           233
1      0            53
       1            65
2      0            40
       1            40
3      0             2
       1             3
4      0             4
5      0             4
       1             1
6      0             1
Name: PassengerId, dtype: int64
```



check percentage of survival for parch 0, 1 and 2

In [29]:

```
Survived_parch_df = pd.DataFrame(Survived_parch)
Survived_parch_df.index
#calculate percentage of people in with certain parch that survived
def percentage_survived(df, parch):
    total_parch = df.loc[parch].sum()
    parch_survive = (df.loc[parch,1]/total_parch)*100
    return parch_survive.values[0]

parch_0 = percentage_survived(Survived_parch_df,0)
parch_1 = percentage_survived(Survived_parch_df,1)
parch_2 = percentage_survived(Survived_parch_df,2)
print("Of people with 0 parent/child around ",str(round(parch_0)) +"%", " survived")
print("Of people with 1 parent/child around ",str(round(parch_1)) +"%", " survived")
print("Of people with 2 parent/child        ",str(parch_2) +"%", " survived")
```

```
Of people with 0 parent/child around  34%  survived
Of people with 1 parent/child around  55%  survived
Of people with 2 parent/child         50.0%  survived
```

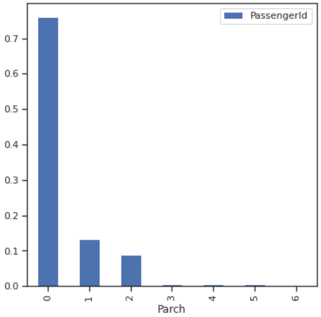
check percentage of passenger for each parch

In [30]:

```
parch_percentage = (df.groupby('Parch')[['PassengerId']].count())/df['PassengerId'].count()
parch_percentage.plot(kind='bar')
```

Out[30]:

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



- Around 75% of people traveled alone and they had low survival rate only around 34% of all people traveling alone survived
- Of all people who had one parent/child abroad 55% survived
- Of all people who had two parent/child abroad 50% survived

SibSp effect of survival

number of siblings / spouses aboard the Titanic

In [31]:

```
df[['Survived','SibSp']].head()
```

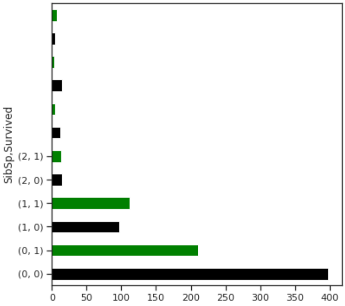
Out[31]:

	Survived	SibSp
0	0	1
1	1	1
2	1	0
3	1	1
4	0	0

In [32]:

```
Survived_SibSp = bar_graph_col_and_survive("SibSp",None)
```

```
SibSp  Survived
0      0           398
       1           210
1      0            97
       1           112
2      0            15
       1            13
3      0            12
       1             4
4      0            15
       1             3
5      0             5
8      0             7
Name: PassengerId, dtype: int64
```



In [33]:

```
Survived_sib_df = pd.DataFrame(Survived_SibSp)

sibsp_0 = percentage_survived(Survived_sib_df,0)
sibsp_1 = percentage_survived(Survived_sib_df,1)
sibsp_2 = percentage_survived(Survived_sib_df,2)
print("Of people with 0 Sibling/Spouse around ",str(round(sibsp_0)) +"%", " survived")
print("Of people with 1 Sibling/Spouse around ",str(round(sibsp_1)) +"%", " survived")
print("Of people with 2 Sibling/Spouse around",str(round(sibsp_2)) +"%", " survived")
```

```
Of people with 0 Sibling/Spouse around  35%  survived
Of people with 1 Sibling/Spouse around  54%  survived
Of people with 2 Sibling/Spouse around 46%  survived
```

- people who had one sibling/spouse had highest rate of survival of 54% followed by people who had two Sibling/Spouse on board

Summary

- Females had higher rate of survival
- Most children < 6 survived
- Most of class 3 (lower class) didn't survive while class 1 (Upper class) had highest chance of survival
- S embarked passengers had lowest survival rate
- people who traveled alone had lowest survival rate