eda-on-titanic-dataset

June 3, 2024

[36]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[37]:
```
df = sns.load_dataset('titanic')
```

[38]:
```
df.shape
```

[38]:
```
(891, 15)
```

[39]:
```
df.head()
```

[39]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	

	who	adult_male	deck	embark_town	alive	alone
0	man	True	NaN	Southampton	no	False
1	woman	False	C	Cherbourg	yes	False
2	woman	False	NaN	Southampton	yes	True
3	woman	False	C	Southampton	yes	False
4	man	True	NaN	Southampton	no	True

[40]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
```

```
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

[41]:
```
pd.isnull(df).sum()
```

[41]:
```
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

[42]:
```
df.drop(['deck'],axis=1, inplace=True)
```

[43]:
```
df.head()
```

[43]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	

	who	adult_male	embark_town	alive	alone
0	man	True	Southampton	no	False
1	woman	False	Cherbourg	yes	False

```
2  woman       False  Southampton   yes   True
3  woman       False  Southampton   yes  False
4    man        True  Southampton    no   True
```

[44]:
```
df.describe()
```

[44]:
```
          survived      pclass         age       sibsp       parch        fare
count  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean     0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std      0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min      0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%      0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%      0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%      1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max      1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

[45]:
```
df.fillna({'age':29.699118},inplace=True)
```

[46]:
```
pd.isnull(df).sum()
```

[46]:
```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
embarked       2
class          0
who            0
adult_male     0
embark_town    2
alive          0
alone          0
dtype: int64
```

[48]:
```
embar_town=df['embark_town'].mode()
```

[51]:
```
pd.isnull(df).sum()
```

[51]:
```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
```

```
embarked       2
class          0
who            0
adult_male     0
embark_town    2
alive          0
alone          0
dtype: int64
```

[53]:
```
df.fillna({'embark_town':'embar_town'},inplace=True)
```

[54]:
```
pd.isnull(df).sum()
```

[54]:
```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
embarked       2
class          0
who            0
adult_male     0
embark_town    0
alive          0
alone          0
dtype: int64
```

[55]:
```
emba=df['embark_town'].mode()
```

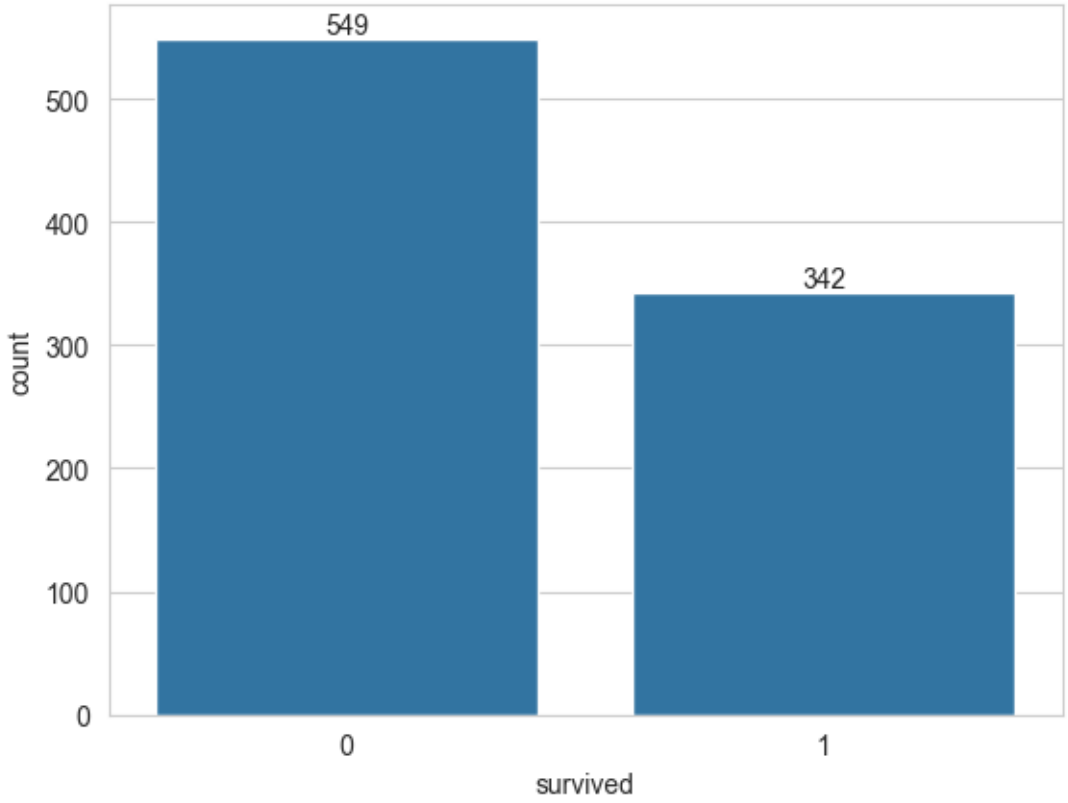
[56]:
```
df.fillna({'embarked':'emba'},inplace=True)
```

[57]:
```
pd.isnull(df).sum()
```

[57]:
```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
embarked       0
class          0
who            0
adult_male     0
embark_town    0
```

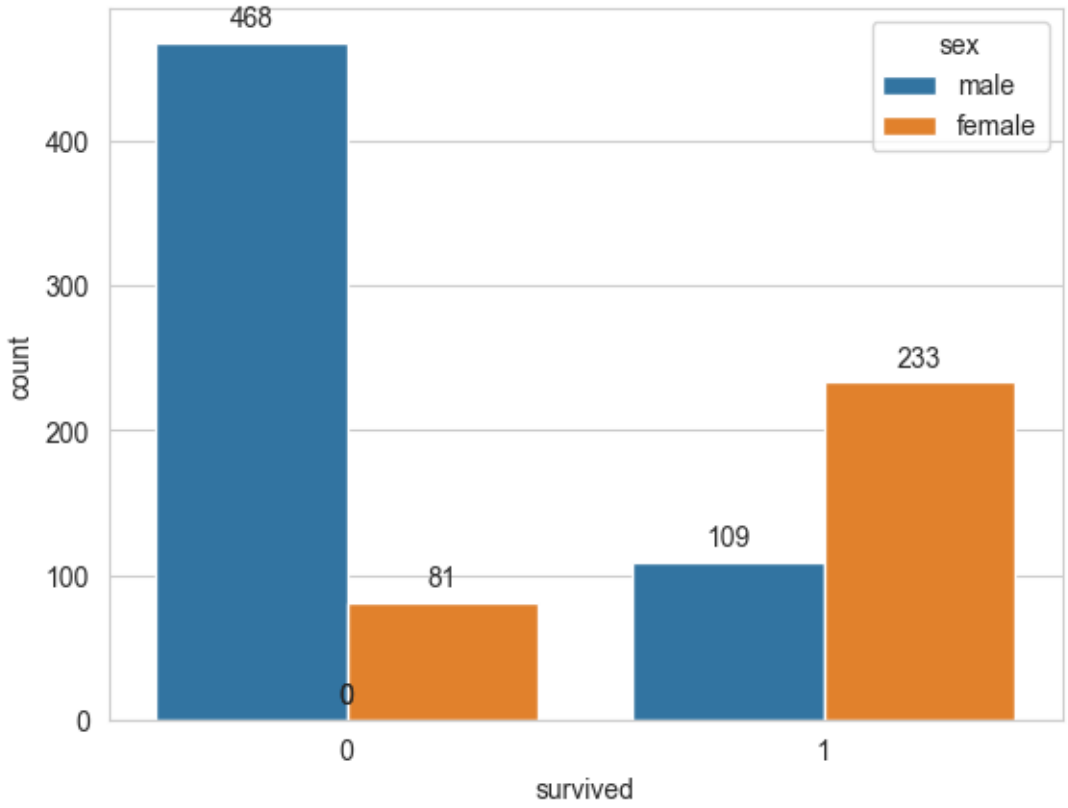
```
alive          0
alone          0
dtype: int64
```

[64]:
```
ax=sns.countplot(x='survived',data=df)
for bars in ax.containers: ax.bar_label(bars)
```



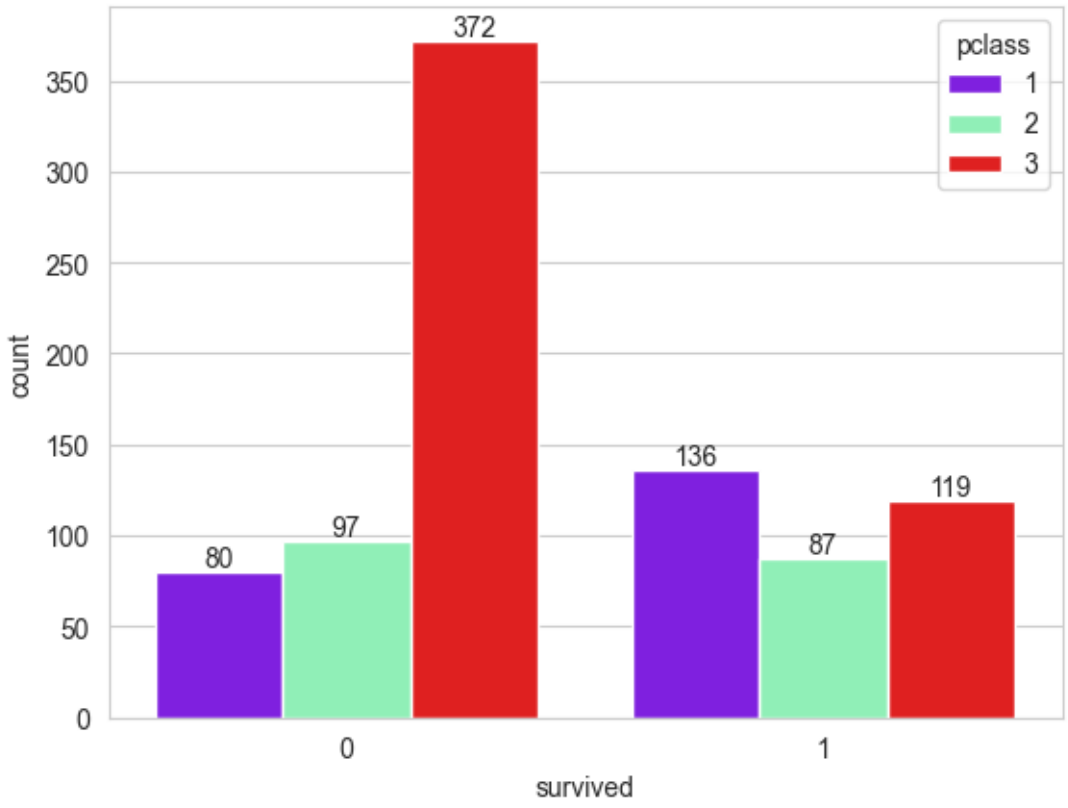
[ ]:
```
We can see from the numbers that 549 died and 342 persons survived. So it means␣
↪that the survival rate is 38.3%.
```

[71]:
```
sns.set_style('whitegrid')
count_plot=sns.countplot(x='survived', hue='sex', data=df)
for p in count_plot.patches:
    count_plot.annotate(format(p.get_height(), '.0f'),
                        (p.get_x() + p.get_width() / 2., p.get_height()),
                        ha = 'center', va = 'center',
                        xytext = (0, 9),
                        textcoords = 'offset points')
plt.show()
```



[ ]:
```
Ther persons who have lost their lives. Their ratio in terms of gender are 85.
↪2% for men and for women the percentage is 14.8%.
While for the survived persons. The ratio for men is 31.8% and the survived␣
↪women have 68.2%.
```

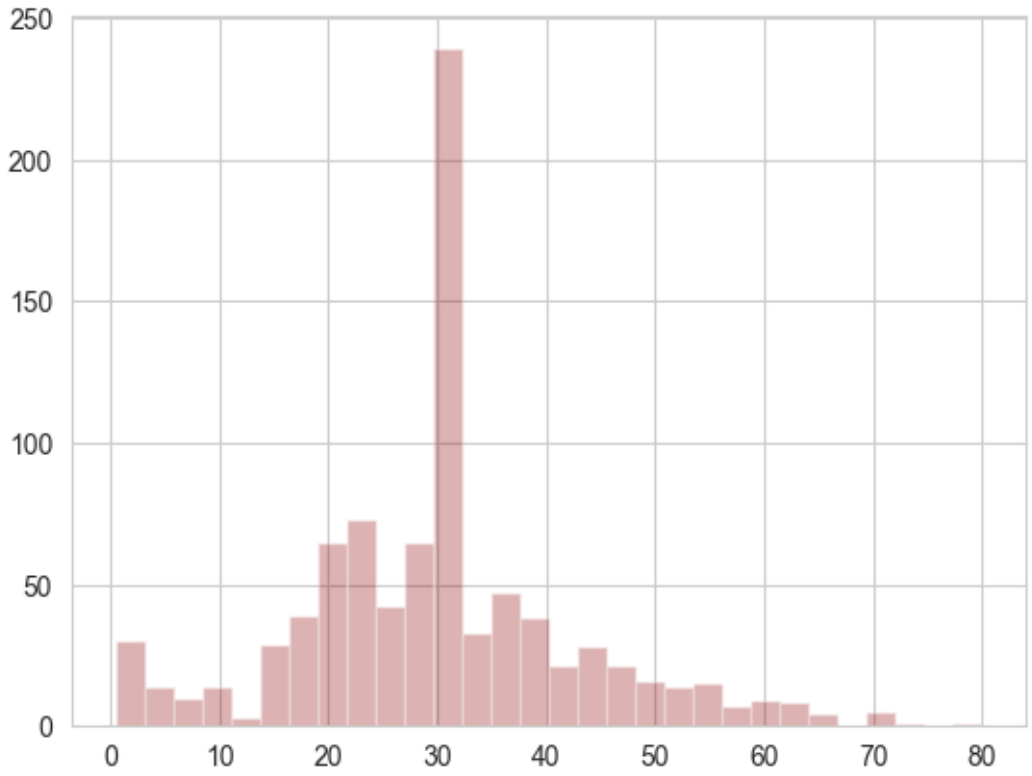
[74]:
```
sns.set_style('whitegrid')
count_plot=sns.countplot(x='survived', hue='pclass', data=df,palette='rainbow')
for container in count_plot.containers:
    count_plot.bar_label(container)
plt.show()
```



[ ]:
```
The above diagram is the reprepresentation of all the passengers in terms of␣
↪Passengers class. Where surprisingly the least survived rate is in mid
passenger class.
```

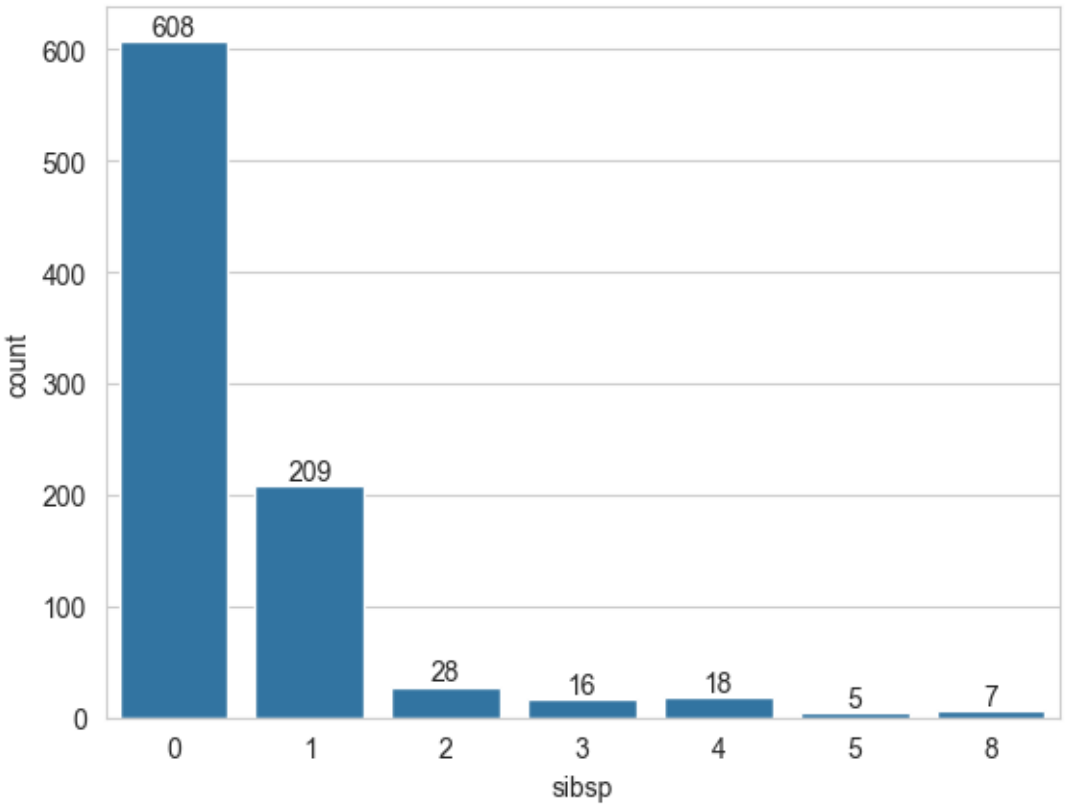
[79]:
```
df['age'].hist(bins=30,color='darkred',alpha=0.3)
```

[79]: <Axes: >



[ ]:
```
Here we can easily figure out that the the max age is between 20 and 30. It
↪could be due to the mean intermulation so we have to look at this fact too
while using this analysis.
```

[82]:
```
ax=sns.countplot(x='sibsp',data=df)
for bars in ax.containers: ax.bar_label(bars)
```



[ ]: Here the maximum amount of passengers are single.