Surviving the Titanic EDA

In [1]:
```
#import neccessary library
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

set up Titanic csv file as dataframe

In [2]:
```
titanic_df=pd.read_csv(r'C:\Users\Dell\Downloads\train.csv')
#let's see a preview of a data
titanic_df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

In [3]:
```
#we could also get overall info for dataset
titanic_df.info()
```

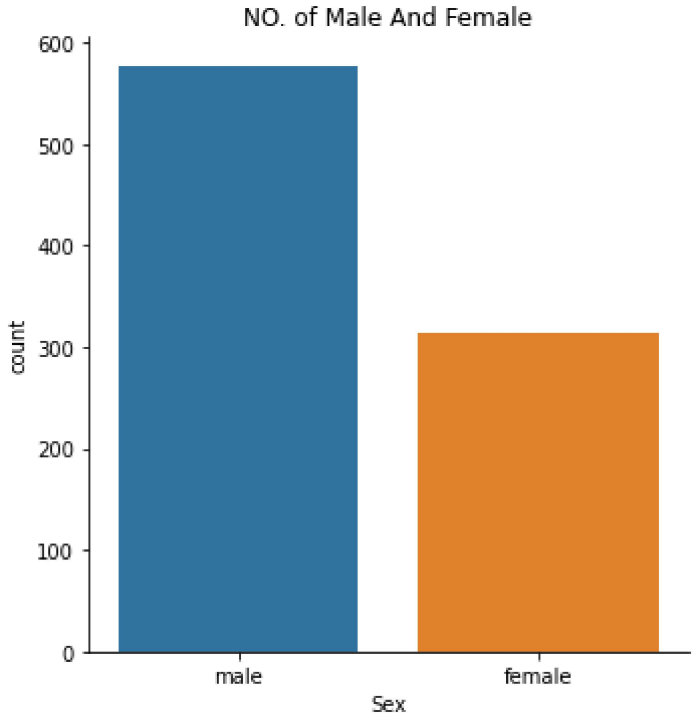
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    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
```

let's try to answer a quetions

1) who were the passengers on the titanic?(age,Gender,Class..etc) 2) What deck were the passengers on and how does that relate to their class? 3) where did the passengers come from? 4) who was alone and who was with family? 5) what factors helped somone survived the sinking?

In [8]:
```
#let's start with first question who were the passengers on the titanic?
#let's check gender
sns.factorplot('Sex',kind='count',data=titanic_df)
plt.title('NO. of Male And Female')
```

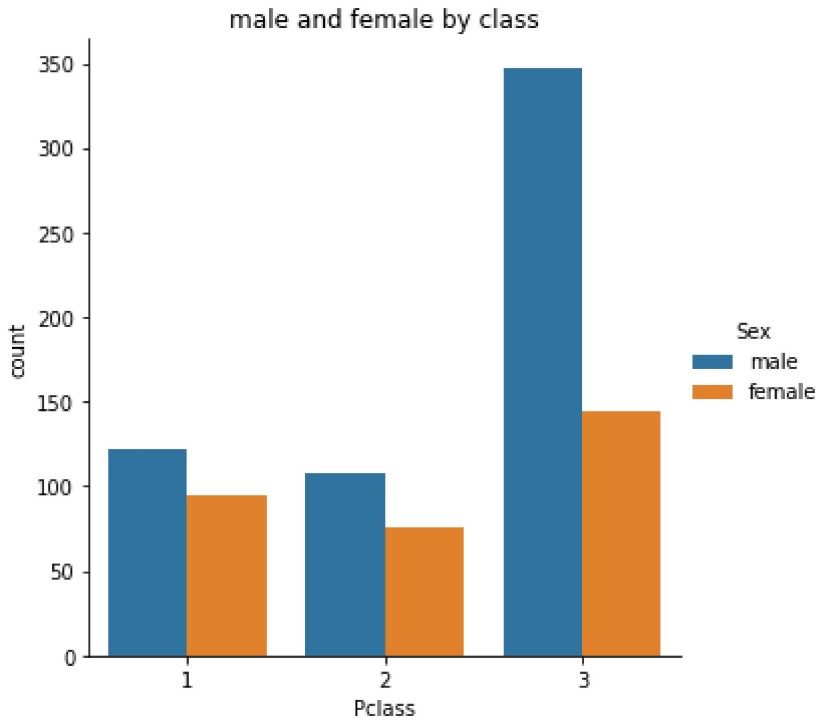
Out[8]: Text(0.5, 1.0, 'NO. of Male And Female')



it clear that about 350 female and 580 male.

In [10]:
```
#now let's separate the genders by classes.
sns.factorplot('Pclass',kind='count',data=titanic_df,hue='Sex')
plt.title('male and female by class')
```

Out[10]: Text(0.5, 1.0, 'male and female by class')



more male and female in class third it might be useful to know the split between males,female and children

In [12]:
```
# We'll trate anyone as under 16 as a child, and then use the apply techniqu
def male_female_child(passenger):
    age,sex=passenger
    if age<16:
        return 'child'
    else:
        return sex

# we will define new column called 'person'
titanic_df['person']=titanic_df[['Age','Sex']].apply(male_female_child,axis=
```

In [13]:
```
1 titanic_df.head(10)
```

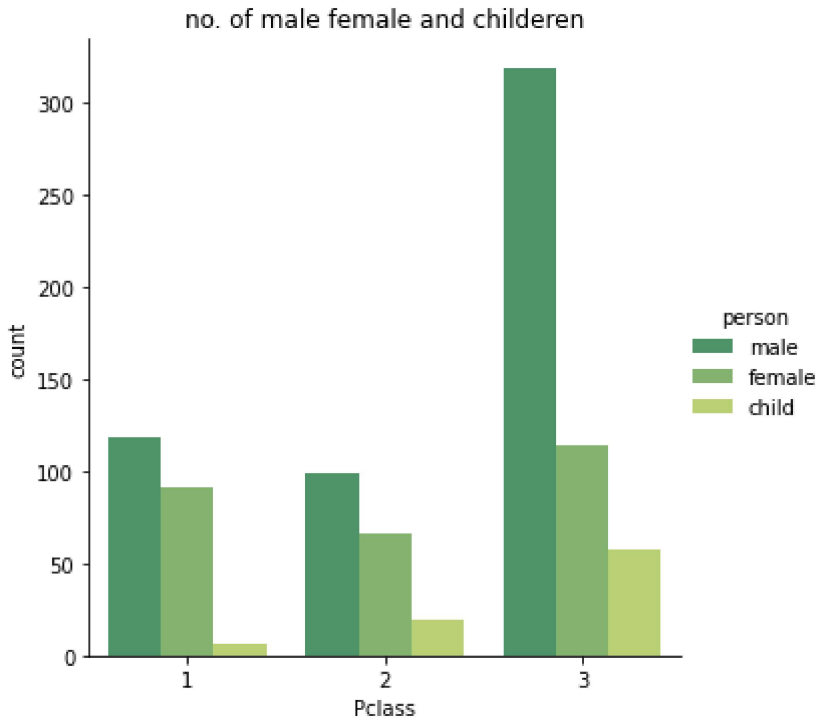
Out[13]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	Na
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E4
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Na
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Na
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Na

now we have separated the passengers between female,male and child

In [14]:
```
#let's try to factor plot again!
sns.factorplot('Pclass',kind='count',data=titanic_df,hue='person',palette='s
plt.title('no. of male female and childeren')
```

Out[14]: Text(0.5, 1.0, 'no. of male female and childeren')




```
children in 3rd class and not so many in 1st! how about  we create a
distriution of the ages to get a  more precise picture of the who the
passengers were.
```

In [15]:
```
#we could also getoverall comparison of male, female and child
titanic_df['person'].value_counts()

```

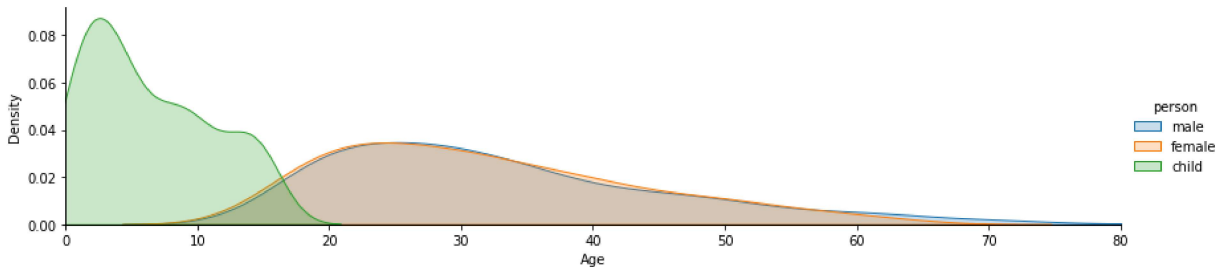
Out[15]:
```
male      537
female    271
child      83
Name: person, dtype: int64
```

In [16]:
```
#mean age
titanic_df['Age'].mean()
```

Out[16]: 29.69911764705882

In [18]:
```
#Another way to vusualize the data is to use FacetGrid to plot multiple kdep
fig=sns.FacetGrid(titanic_df,hue='person',aspect=4)
#next use map to plot all the possible kdeplots for the age column by the hu
fig.map(sns.kdeplot,'Age',shade=True)
oldest=titanic_df['Age'].max()
fig.set(xlim=(0,oldest))
fig.add_legend()
```

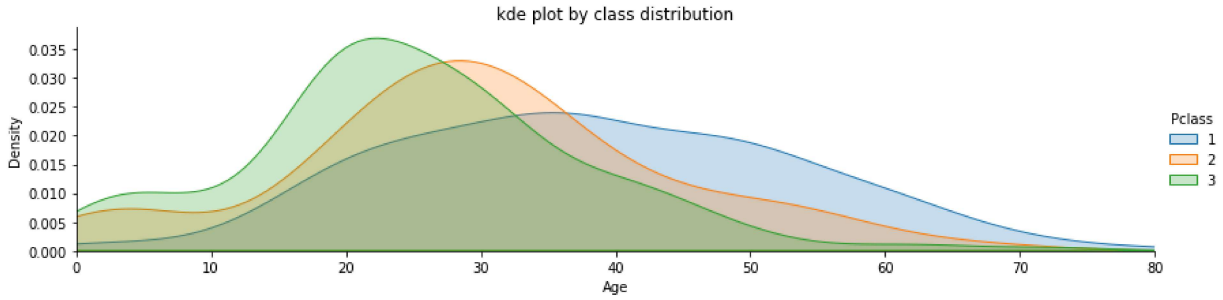
Out[18]: <seaborn.axisgrid.FacetGrid at 0x204c946bf70>



it show age distribution male female and child.

In [21]:
```
#let's same for class by changing hue argument:
fig=sns.FacetGrid(titanic_df,hue='Pclass',aspect=4)
fig.map(sns.kdeplot,'Age',shade=True)
oldest=titanic_df['Age'].max()
fig.set(xlim=(0,oldest))
fig.add_legend()
plt.title('kde plot by class distribution')
```

Out[21]: Text(0.5, 1.0, 'kde plot by class distribution')



we've gotten pretty good picture of who the passengers were based on sex, Age, and Class So let's move 2nd Question: what deck were the passengers on and how does that relate to their class?

In [22]:
```
1 titanic_df.head()
```

Out[22]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

So we can see that the cabin column has infomation on the deck, but it has several null values, so we will have to drop them.

In [23]:
```
1 #drop NaN values
2 deck=titanic_df['Cabin'].dropna()
```

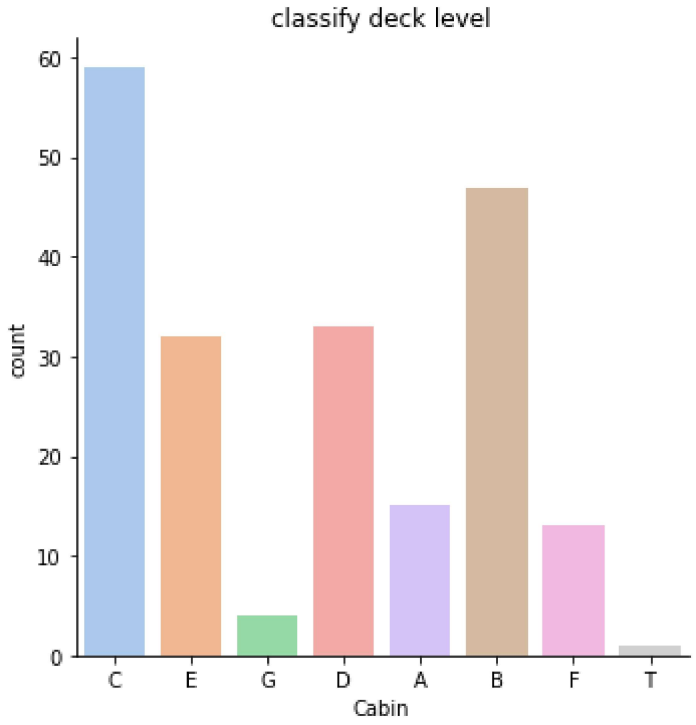
In [24]:
```
1 #quick Preview of the decks
2 deck.head()
```

Out[24]:
```
1      C85
3     C123
6      E46
10      G6
11    C103
Name: Cabin, dtype: object
```

Notice We only need the first letter of the deck to classify its level(e.g. A,B,C,D,E,F,G)

In [34]:
```
#so lei's grab that letter for the deck level with a simple for loop
levels=[]
for level in deck:
    levels.append(level[0])
#Reset DataFrame and use factor plot
cabin_df=pd.DataFrame(levels)
cabin_df.columns=['Cabin']
sns.factorplot('Cabin',kind='count',data=cabin_df,palette='pastel')
plt.title('classify deck level')
```

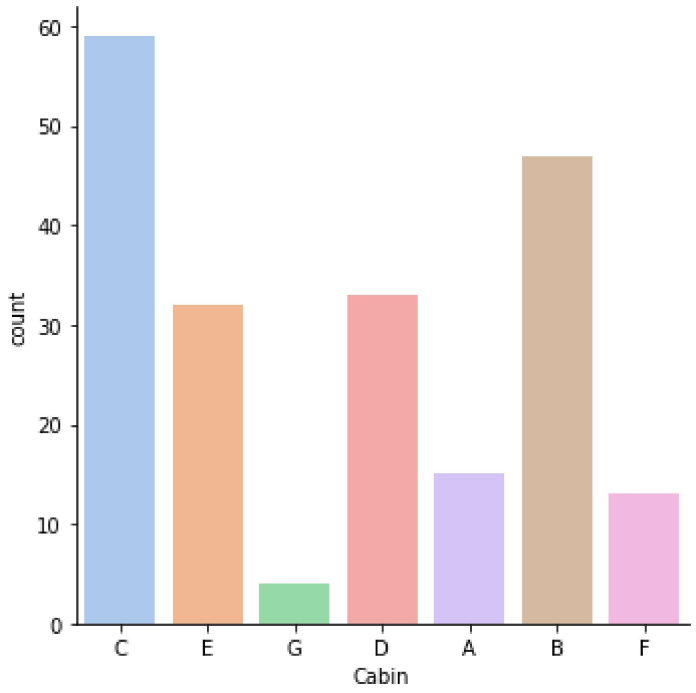
Out[34]:
```
Text(0.5, 1.0, 'classify deck level')
```



interesting to note we have a 'T' deck value ther which doesn't make sense, we can drop it out with the following code.

In [36]:
```
#Redefine cabin_df as everything but where the row was not equal 'T'
cabin_df=cabin_df[cabin_df.Cabin!='T']
sns.factorplot('Cabin',kind='count',data=cabin_df,palette='pastel')
```

Out[36]: <seaborn.axisgrid.FacetGrid at 0x204cfccecb0>



we analysed the distribution by decks

3) where did the passengers come from?

In [37]:
```
#let's take another look at our original data
titanic_df.head()
```

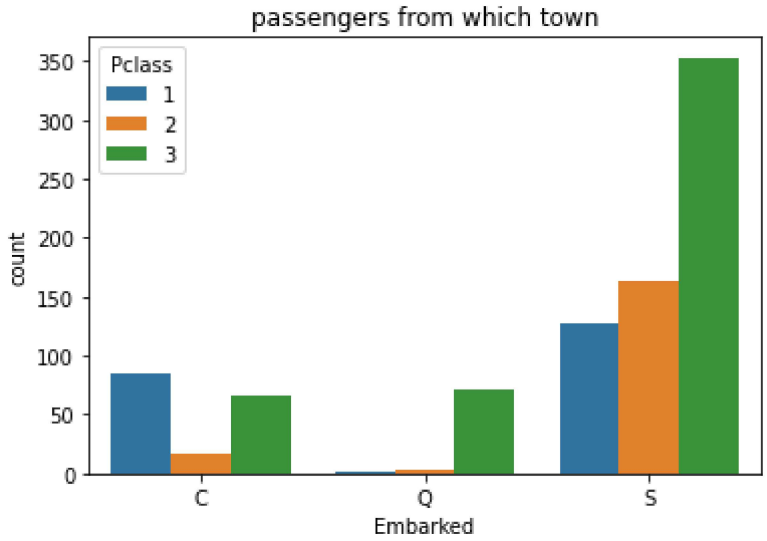
Out[37]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

Embarked column has C,Q and S values.

In [42]:
```
#now we can make a quick plot to check out the results.

sns.countplot(x='Embarked',data=titanic_df,hue='Pclass',order=['C','Q','S'])
plt.title('passengers from which town')
```

Out[42]:
```
Text(0.5, 1.0, 'passengers from which town')
```



An interesting find here is that in queenstown,almost all the passengers that boarded there were 3rd class.It would we intersting to look at the economics of that town in that time period for further investigation.

4) Who was alone and whoe was with Family?

In [43]:
```
#let's start adding new column to define alone

titanic_df['Alone']=titanic_df.Parch+titanic_df.SibSp
titanic_df['Alone']
```

Out[43]:
```
0      1
1      1
2      0
3      1
4      0
      ..
886    0
887    0
888    3
889    0
890    0
Name: Alone, Length: 891, dtype: int64
```

Now we know that if the alone column is anything but 0, then the passenger had famil aboard and wasn't alone. so let's change the column now that if the value is greater than 0, we know the passenger was with his/her family, otherwise they were alone.

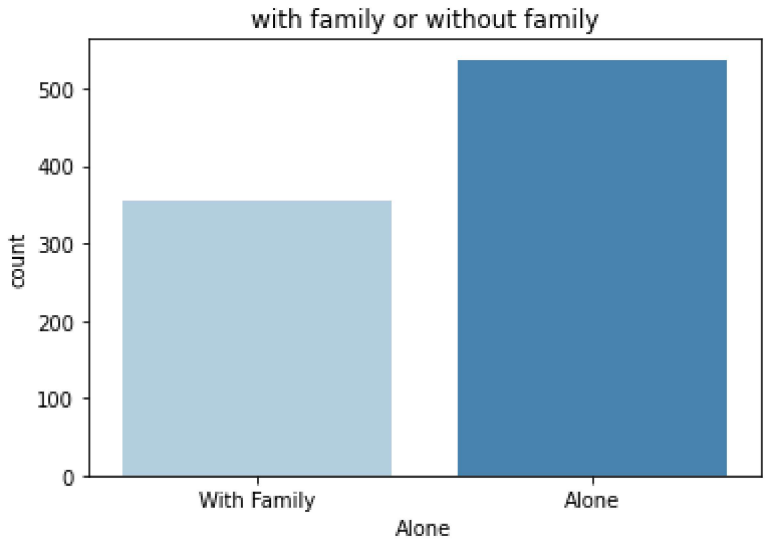
In [44]:
```
titanic_df['Alone'].loc[titanic_df['Alone']>0]='With Family'
titanic_df['Alone'].loc[titanic_df['Alone']==0]='Alone'

#check to make sure it worked
titanic_df.head()
```

Out[44]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

In [46]:
```
#now let's get a visualization!
sns.countplot(x='Alone',data=titanic_df,palette='Blues')
plt.title('with family or without family')
```

Out[46]:
```
Text(0.5, 1.0, 'with family or without family')
```



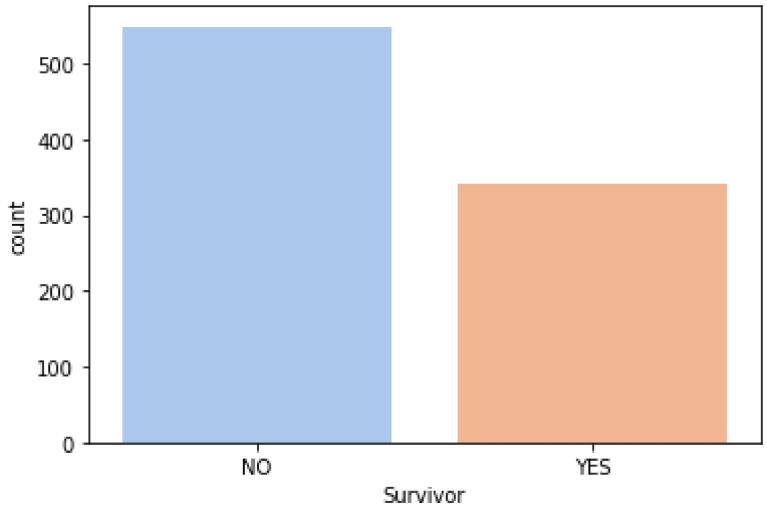
In [47]:
```
1 titanic_df['Alone'].value_counts()
```

Out[47]:
```
Alone          537
With Family    354
Name: Alone, dtype: int64
```

5) What the factor helped somone survived the sinking?

In [51]:
```
1 #let's start by creating a new column for legibility purposes through mappin
2 titanic_df['Survivor']=titanic_df.Survived.map({0:"NO",1:"YES"})
3
4 #let's just get a quick overall view of survived vs died.
5 sns.countplot('Survivor',data=titanic_df,palette='pastel')
```

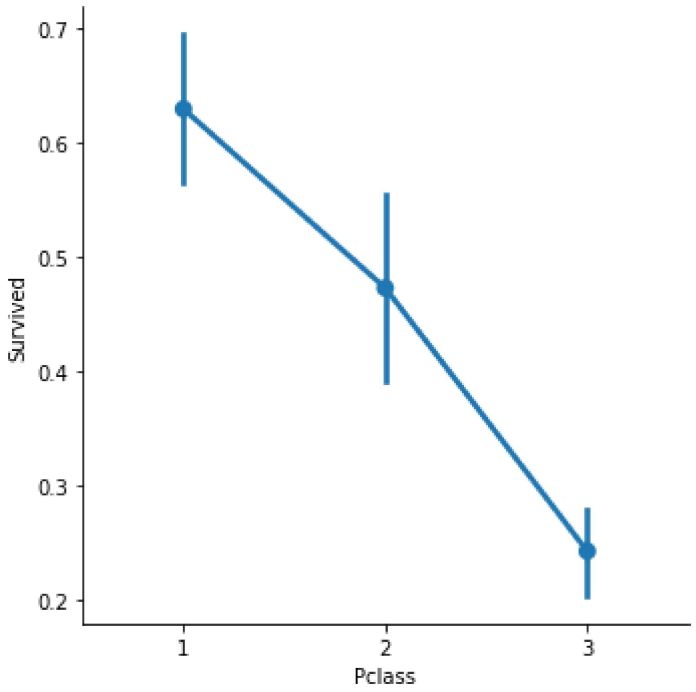
Out[51]:
```
<AxesSubplot:xlabel='Survivor', ylabel='count'>
```



So quite a few more people died than who survived. let's see of the class of the passengers had an effect on their survival rate.

In [52]:
```
# let's use factor plot
sns.factorplot('Pclass','Survived',data=titanic_df)
```

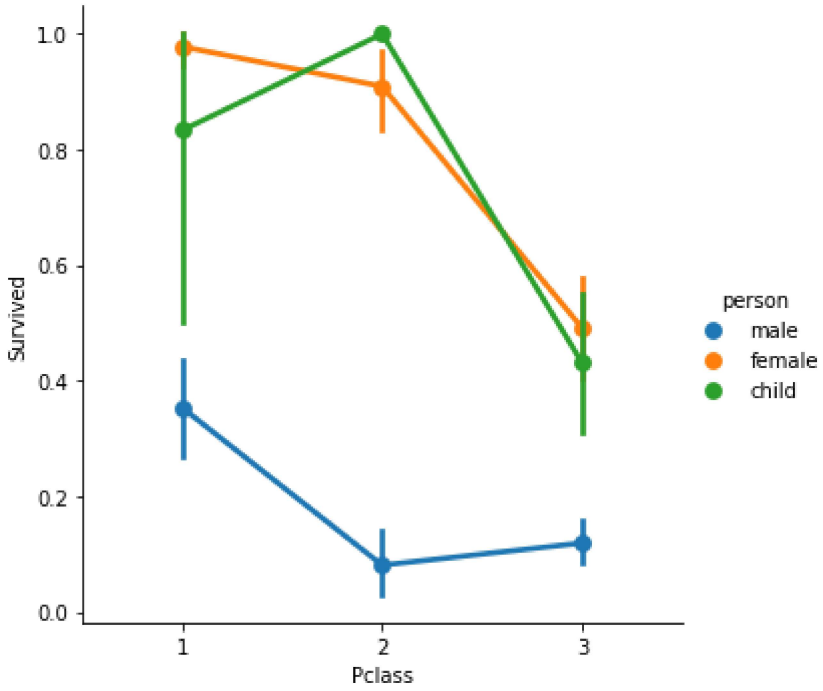
Out[52]: <seaborn.axisgrid.FacetGrid at 0x204d326bbb0>



look like survival rated for 3rd class are substantially lower! But maybe this effect caused by the large amount of men in 3rd class in combination with the women and children first policy. let's use the hue get cleared picture on this.

In [54]:
```
1 sns.factorplot('Pclass','Survived',hue='person',data=titanic_df)
```

Out[54]:
```
<seaborn.axisgrid.FacetGrid at 0x204d28eeb60>
```

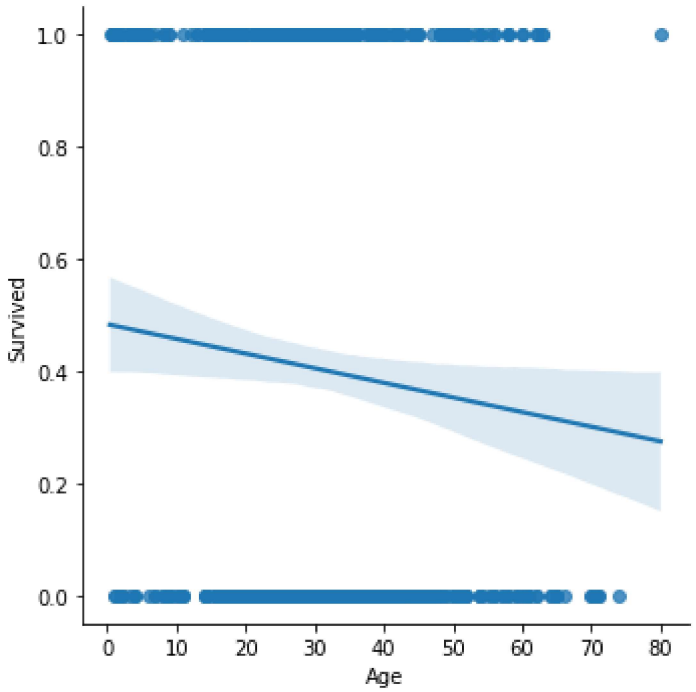


from this data it look like being a male or being in 3rd class were both not favourable for survival. Even regardless of class the result of being a male in any class dramitically decreases your chances of survival.

But what about age? did being younger or older have an effect on survival rate?

In [57]:
```
#age effect on survival rate.
sns.lmplot('Age','Survived',data=titanic_df)
```

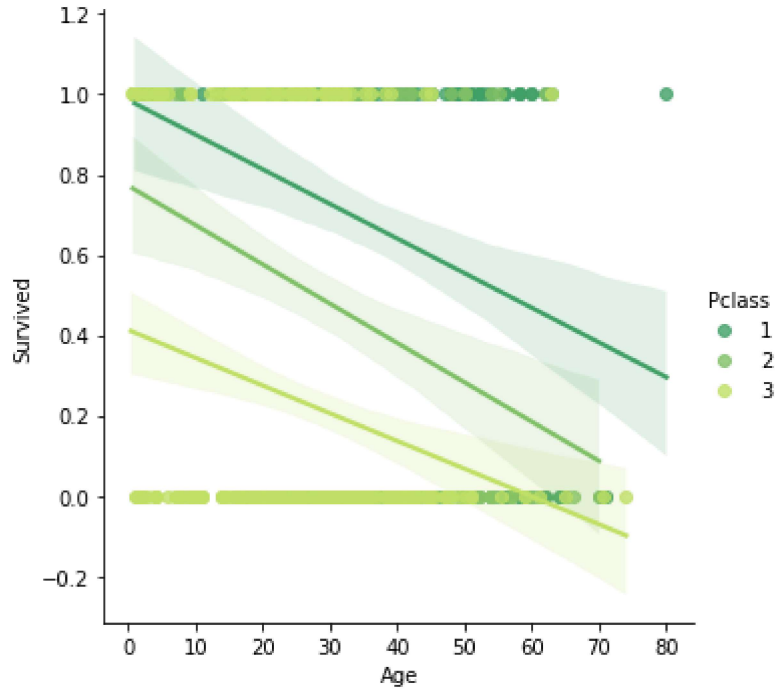
Out[57]: <seaborn.axisgrid.FacetGrid at 0x204d3295570>



Looks like there is a genral trens that the older the passenger was the less likely they survived. Let's go ahed and use to take a look at the effect of class and age.

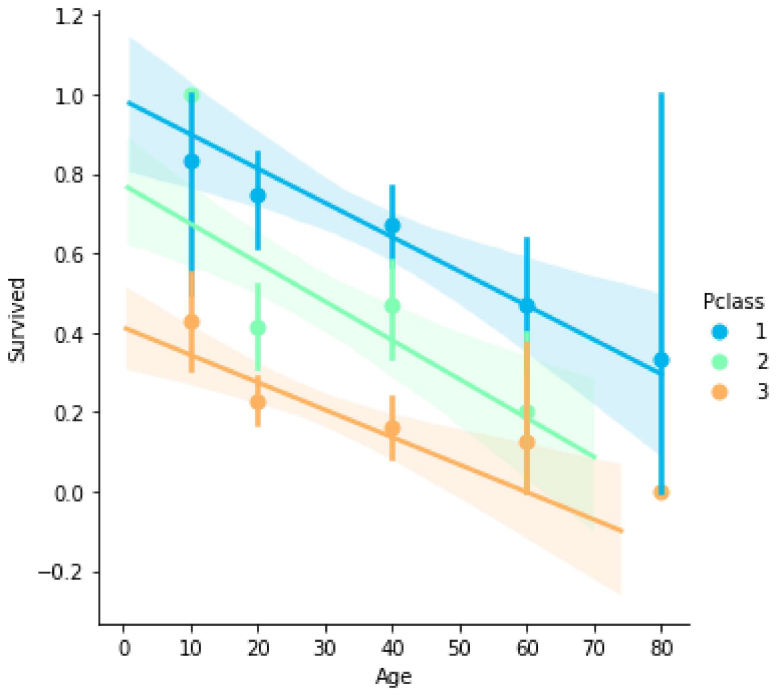
In [58]:
```
1 sns.lmplot('Age','Survived',hue='Pclass',data=titanic_df, palette='summer')
```

Out[58]: <seaborn.axisgrid.FacetGrid at 0x204d3214f70>



In [61]:
```
# let's use a linear plot on age versus survival using hue for class seperat
generations=[10,20,40,60,80]
sns.lmplot('Age','Survived',hue='Pclass',data=titanic_df,palette='rainbow',x
```

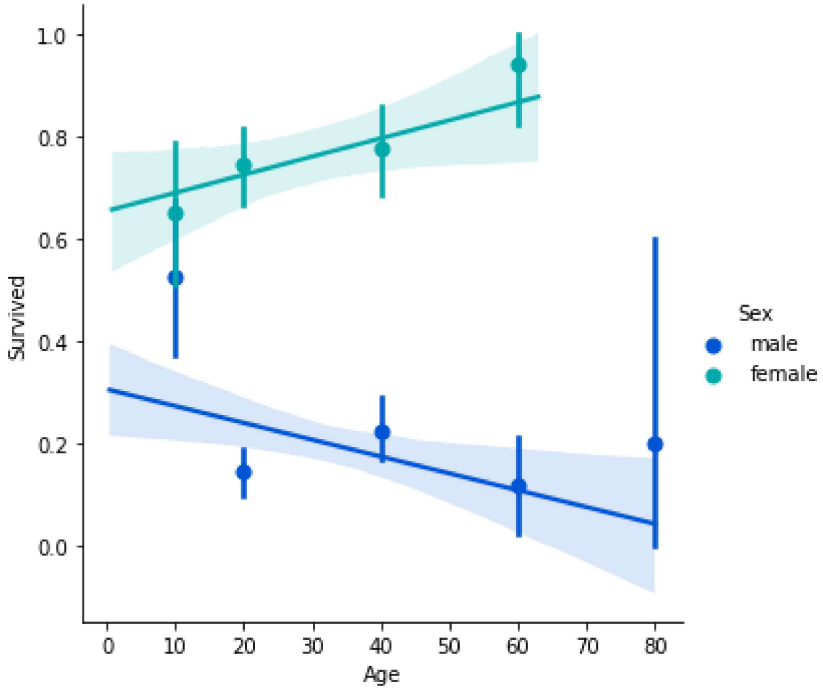
Out[61]: <seaborn.axisgrid.FacetGrid at 0x204d31d6860>



what about if we relate gender and age with survival set?

In [64]:
```
sns.lmplot('Age','Survived',hue='Sex',data=titanic_df,palette='winter',x_bin
```

Out[64]: <seaborn.axisgrid.FacetGrid at 0x204d2867ac0>



We have gotten some really great insights on how gender,age and class all related to a

passengers chance of survival.

In [ ]:
```
1
```