**Data Dictionary**

* survival: Survival (0 = No, 1 = Yes)
* pclass: Ticket class(1 = 1st, 2 = 2nd, 3 = 3rd)
* sex: Sex
* Age: Age in years
* sibsp: Number of siblings / spouses aboard the Titanic
* parch: Number of parents / children aboard the Titanic
* ticket: Ticket number
* fare:Passenger fare
* cabin:Cabin number
* embarked:Port of Embarkation(C = Cherbourg, Q = Queenstown, S = Southampton)

**pclass**: A proxy for socio-economic status (SES) -1st = Upper -2nd = Middle -3rd = Lower

**Age**: Age is fractional if less than 1. If the age is estimated, it is in the form of xx.5

**sibsp**: The dataset defines family relations in this way...

* Sibling = brother, sister, stepbrother, stepsister
* Spouse = husband, wife (mistresses and fiancés were ignored)

**parch**: The dataset defines family relations in this way...

* Parent = mother, father
* Child = daughter, son, stepdaughter, stepson
* Some children travelled only with a nanny, therefore parch=0 for them

## Import Libraries

In [252]:

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**from** **pandas** **import** Series, DataFrame

**import** **matplotlib** **as** **mpl**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

%matplotlib inline

*# Set default matplot figure size*

#pylab.rcParams['figure.figsize'] = (10.0, 8.0)

## Reading Data Set using Pandas

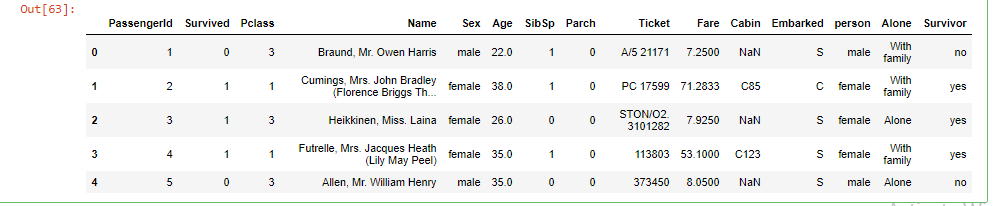
In [62]: titanic\_df = pd.read\_csv('train.csv')

## Analysis

In [63]:

*# Check the first 5 rows of the data frame*

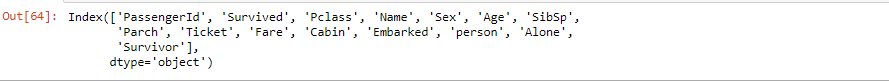
titanic\_df.head()



In [64]

*# Column names*

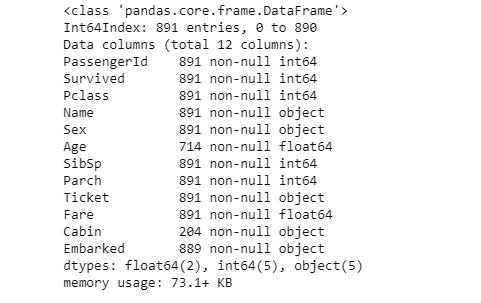
titanic\_df.columns



In [138]:

*# Information about the data set*

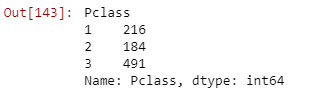
titanic\_df.info()



In [143]:

*# Number of passengers in each class*

titanic\_df.groupby('Pclass')['Pclass'].count()

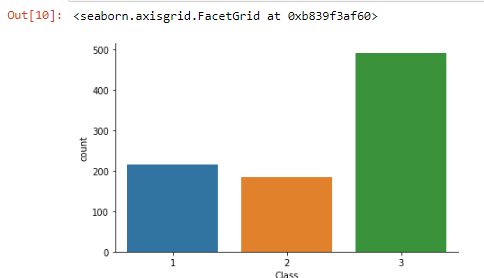


In [205]:

*# Instead of a group by, use seaborn to plot the count of passengers in each class*

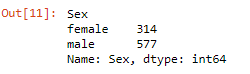
fg = sns.factorplot('Pclass', data=titanic\_df, kind='count', aspect=1.5)

fg.set\_xlabels('Class')



In [29]:

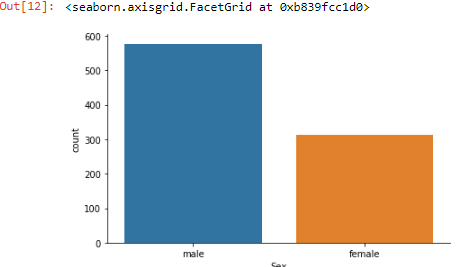
titanic\_df.groupby('Sex')['Sex'].count()



In [199]:

*# Instead of a group by, use seaborn to plot the number of males and females*

sns.factorplot('Sex', data=titanic\_df, kind='count', aspect=1.5)

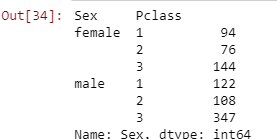


There are almost two times males as much as there were females.

In [34]:

*# Number of men and women in each of the passenger class*

titanic\_df.groupby(['Sex', 'Pclass'])['Sex'].count()

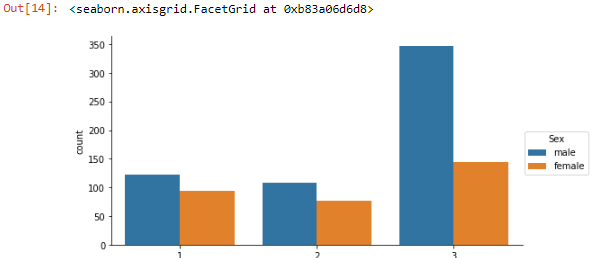


In [207]:

*# Again use saeborn to group by Sex and class*

g = sns.factorplot('Pclass', data=titanic\_df, hue='Sex', kind='count', aspect=1.75)

g.set\_xlabels('Class')

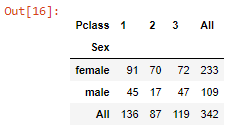


As shown in the figure above, there are more than two times males than females in class 3. However, in classes 1 and 2, the ratio of male to female is almost 1.

In [79]:

*# Number of passengers who survived in each class grouped by sex. Also total was found for each class grouped by sex.*

titanic\_df.pivot\_table('Survived', 'Sex', 'Pclass', aggfunc=np.sum, margins=True)



In [65]:

not\_survived = titanic\_df[titanic\_df['Survived']==0]

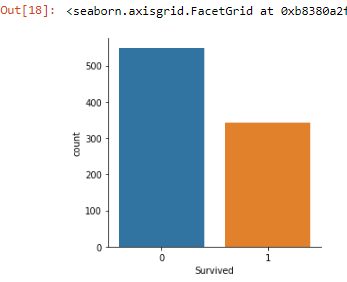
len(not\_survived)

out :549

In [357]:

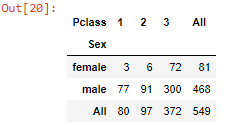
*# Factor plot of those who survived vs. who didn't*

sns.factorplot('Survived', data=titanic\_df, kind='count')



*# Number of passengers who did not survive in each class grouped by sex.*

not\_survived.pivot\_table('Survived', 'Sex', 'Pclass', aggfunc=len, margins=True)

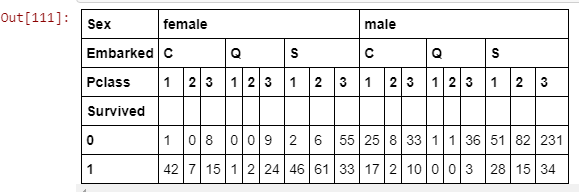


In [215]:

*# Passengers who survived and who didn't survive grouped by class and sex*

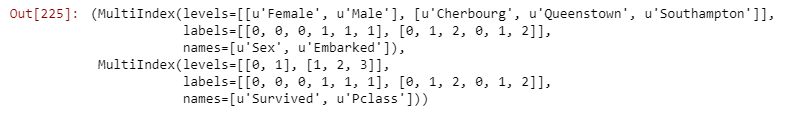
table = pd.crosstab(index=[titanic\_df.Survived,titanic\_df.Pclass], columns=[titanic\_df.Sex,titanic\_df.Embarked])

table.unstack()



In [225]:

table.columns, table.index



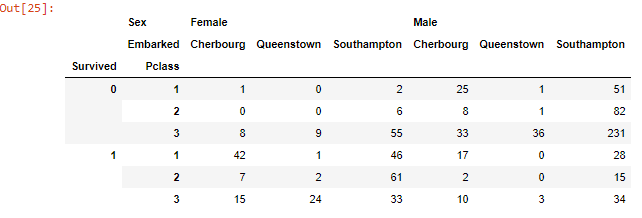
In [224]:

*# Change name of columns*

table.columns.set\_levels(['Female', 'Male'], level=0, inplace=True)

table.columns.set\_levels(['Cherbourg','Queenstown','Southampton'], level=1, inplace=True)

table



In [241]:

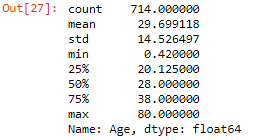
**print**('Average and median age of passengers are %0.f and %0.f years old, respectively'%(titanic\_df.Age.mean(),

titanic\_df.Age.median()))

out :Average and median age of passengers are 30 and 28 years old, respectively

In [246]:

titanic\_df.Age.describe()



In [314]:

*# Drop missing values for the records in which age passenger is missing*

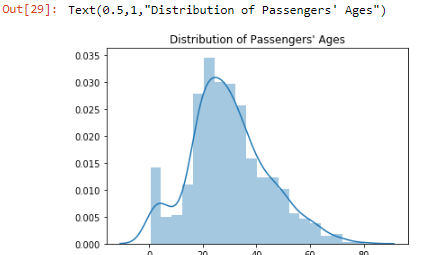
age = titanic\_df['Age'].dropna()

In [347]:

*# Distribution of age, with an overlay of a density plot*

age\_dist = sns.distplot(age)

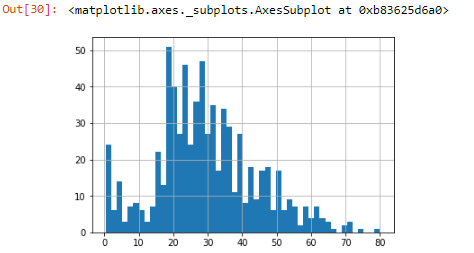
age\_dist.set\_title("Distribution of Passengers' Ages")



In [348]:

*# Another way to plot a histogram of ages is shown below*

titanic\_df['Age'].hist(bins=50)



In [327]:

titanic\_df['Parch'].dtype, titanic\_df['SibSp'].dtype, len(titanic\_df.Cabin.dropna())

Out[327]:

(dtype('int64'), dtype('int64'), 204)

In [331]:

*# Create a function to define those who are children (less than 16)*

**def** male\_female\_child(passenger):

age, sex = passenger

**if** age < 16:

**return** 'child'

**else**:

**return** sex

In [332]:

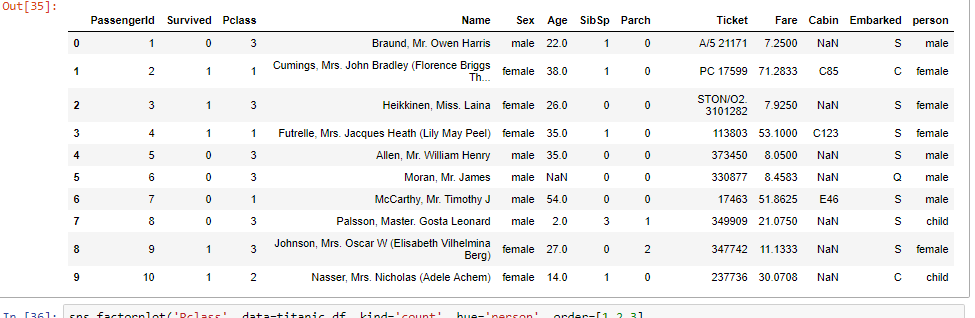
titanic\_df['person'] = titanic\_df[['Age', 'Sex']].apply(male\_female\_child, axis=1)

In [338]:

*# Lets have a look at the first 10 rows of the data frame*

titanic\_df[:10]

Out[338]:

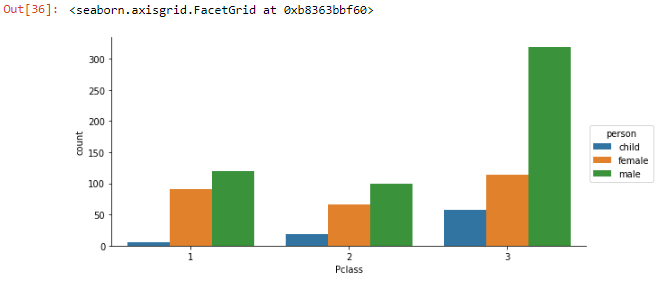


In [454]:

*# Lets do a factorplot of passengers splitted into sex, children and class*

sns.factorplot('Pclass', data=titanic\_df, kind='count', hue='person', order=[1,2,3],

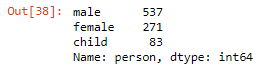
hue\_order=['child','female','male'], aspect=2)



In [354]:

*# Count number of men, women and children*

titanic\_df['person'].value\_counts()



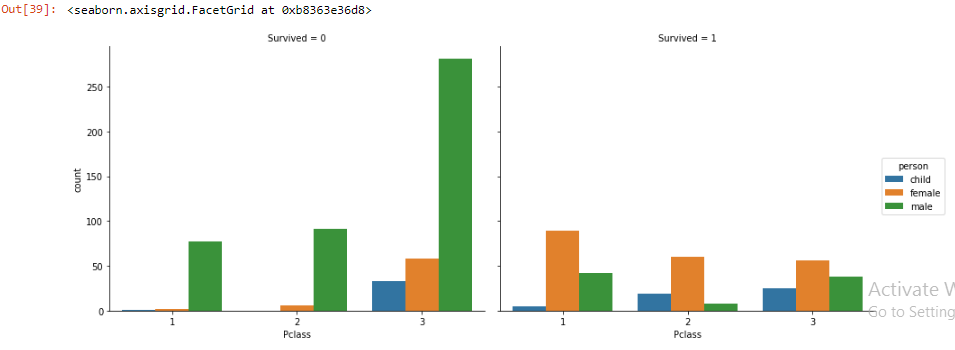
In [353]:

*# Do the same as above, but split the passengers into either survived or not*

sns.factorplot('Pclass', data=titanic\_df, kind='count', hue='person', col='Survived', order=[1,2,3],

hue\_order=['child','female','male'], aspect=1.25, size=5)

Out[353]:



There are much more children in third class than there are in first and second class. However, one may expect that there woould be more children in 1st and 2nd class than there are in 3rd class.

### kde plot, Distribution of Passengers' Ages

#### Grouped by Gender

In [364]:

fig = sns.FacetGrid(titanic\_df, hue='Sex', aspect=4)

fig.map(sns.kdeplot, 'Age', shade=True)

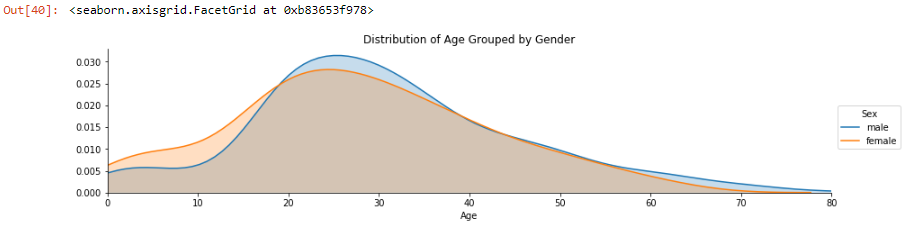
oldest = titanic\_df['Age'].max()

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

fig.set(title='Distribution of Age Grouped by Gender')

fig.add\_legend()

Out[364]:



In [366]:

fig = sns.FacetGrid(titanic\_df, hue='person', aspect=4)

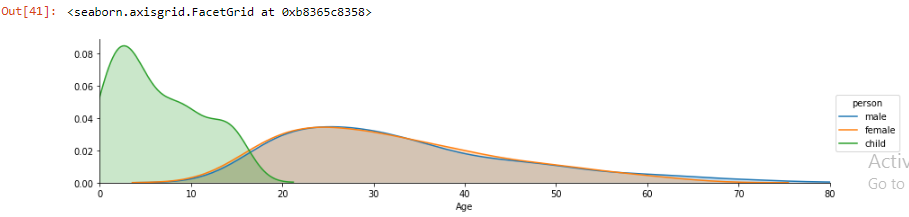
fig.map(sns.kdeplot, 'Age', shade=True)

oldest = titanic\_df['Age'].max()

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

fig.add\_legend()

Out[366]:



#### Grouped by Class

In [367]:

fig = sns.FacetGrid(titanic\_df, hue='Pclass', aspect=4)

fig.map(sns.kdeplot, 'Age', shade=True)

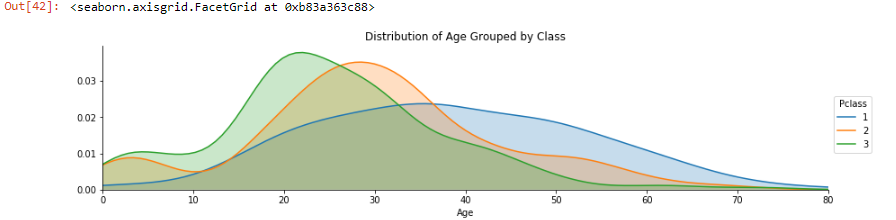
oldest = titanic\_df['Age'].max()

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

fig.set(title='Distribution of Age Grouped by Class')

fig.add\_legend()

Out[367]:



From the plot above, class 1 has a normal distribution. However, classes 2 and 3 have a skewed distribution towards 20 and 30-year old passengers.

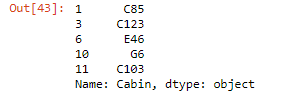
#### What cabins did the Passengers stay in?

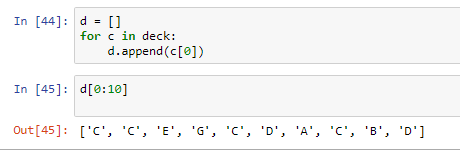
In [384]:

deck = titanic\_df['Cabin'].dropna()

deck.head()

Out[384]:





In [398]:

**from** **collections** **import** Counter

Counter(d)

Out[398]:

Counter({'C': 59, 'B': 47, 'D': 33, 'E': 32, 'A': 15, 'F': 13, 'G': 4, 'T': 1})

In [410]:

*# Now lets factorplot the cabins. First transfer the d list into a data frame. Then rename the column Cabin*

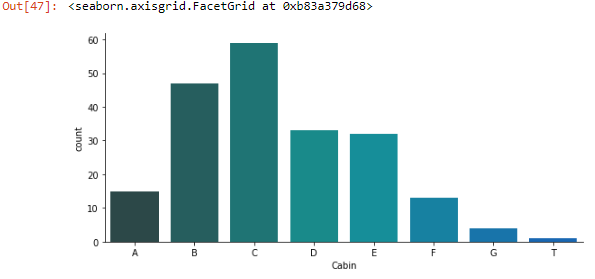
cabin\_df = DataFrame(d)

cabin\_df.columns=['Cabin']

sns.factorplot('Cabin', data=cabin\_df, kind='count', order=['A','B','C','D','E','F','G','T'], aspect=2,

palette='winter\_d')

Out[410]:



In [411]:

*# Drop the 'T' cabin*

cabin\_df = cabin\_df[cabin\_df['Cabin'] != 'T']

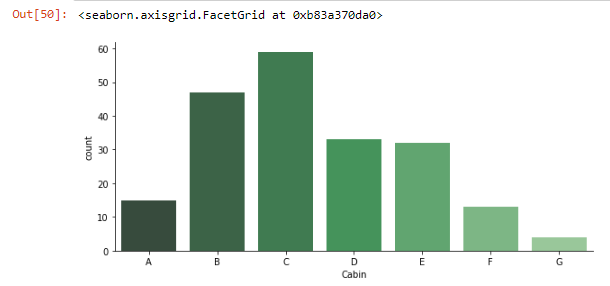
In [433]:

*# Then replot the Cabins factorplot as above*

sns.factorplot('Cabin', data=cabin\_df, kind='count', order=['A','B','C','D','E','F','G'], aspect=2,

palette='Greens\_d')

Out[433]:



In [434]:

*# Below is a link to the list of matplotlib colormaps*

url = 'http://matplotlib.org/api/pyplot\_summary.html?highlight=colormaps#matplotlib.pyplot.colormaps'

**import** **webbrowser**

webbrowser.open(url)

Out[434]:

True

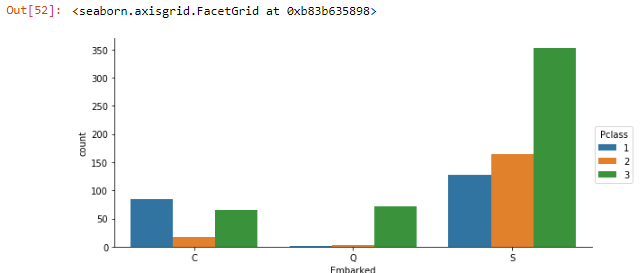
#### Where did the passengers come from i.e. Where did the passengers land into the ship from?

In [476]:

sns.factorplot('Embarked', data=titanic\_df, kind='count', hue='Pclass', hue\_order=range(1,4), aspect=2,

order = ['C','Q','S'])

Out[476]:

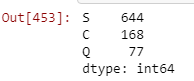


From the figure above, one may conclude that almost all of the passengers who boarded from Queenstown were in third class. On the other hand, many who boarded from Cherbourg were in first class. The biggest portion of passengers who boarded the ship came from Southampton, in which 353 passengers were in third class, 164 in second class and 127 passengers were in first class. In such cases, one may need to look at the economic situation at these different towns at that period of time to understand why most passengers who boarded from Queenstown were in third class for example.

In [453]:

titanic\_df.Embarked.value\_counts()

Out[453]:



In [470]:

*# For tabulated values, use crosstab pandas method instead of the factorplot in seaborn*

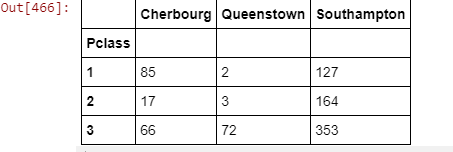
port = pd.crosstab(index=[titanic\_df.Pclass], columns=[titanic\_df.Embarked])

port.columns = [['Cherbourg','Queenstown','Southampton']]

In [466]:

port

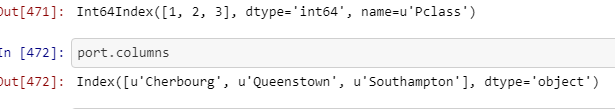
Out[466]:



In [471]:

port.index

Out[471]:



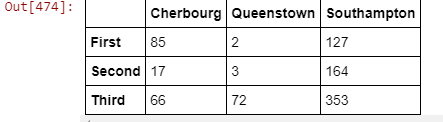
In [473]:

port.index=[['First','Second','Third']]

In [474]:

port

Out[474]:

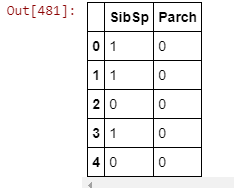


#### Who was alone and who was with parents or siblings?

In [481]:

titanic\_df[['SibSp','Parch']].head()

Out[481]:



In [552]:

*# Alone dataframe i.e. the passenger has no siblings or parents*

alone\_df = titanic\_df[(titanic\_df['SibSp'] == 0) & (titanic\_df['Parch']==0)]

*# Add Alone column*

alone\_df['Alone'] = 'Alone'

*# Not alone data frame i.e. the passenger has either a sibling or a parent.*

not\_alone\_df = titanic\_df[(titanic\_df['SibSp'] != 0) | (titanic\_df['Parch']!=0)]

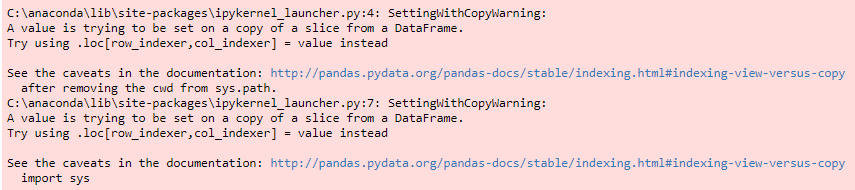
not\_alone\_df['Alone'] = 'With family'

*# Merge the above dataframes*

comb = [alone\_df, not\_alone\_df]

*# Merge and sort by index*

titanic\_df = pd.concat(comb).sort\_index()



In [519]:

[len(alone\_df), len(not\_alone\_df)]

Out[519]:

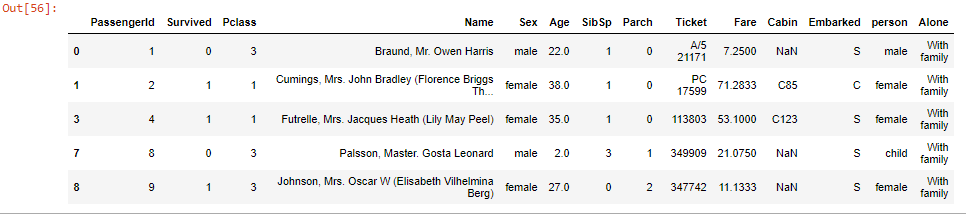
[537, 354]

In [520]:

*# Show the first five records of the alone data frame*

alone\_df.head()

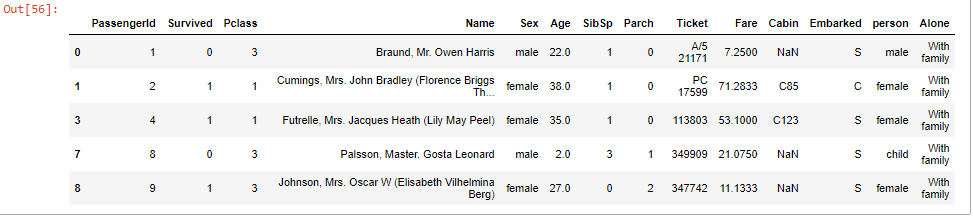
Out[520]:



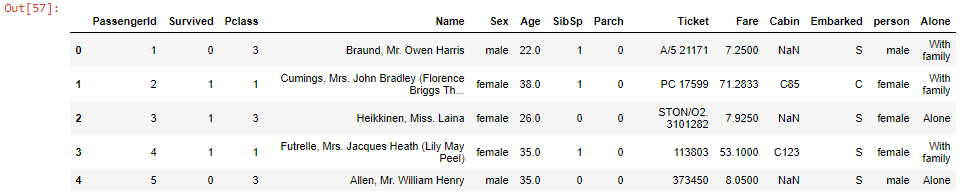
In [521]:

*# Show the first five rows of the not alone data frame*

not\_alone\_df.head()

In [553]:

titanic\_df.head()

In [539]:

*""" Another way to perform the above*

*titanic\_df['Alone'] = titanic\_df.SibSp + titanic\_df.Parch*

*titanic\_df['Alone'].loc[titanic\_df['Alone']>0] = 'With family'*

*titanic\_df['Alone'].loc[titanic\_df['Alone']==0] = 'Alone'"""*

Out[539]:

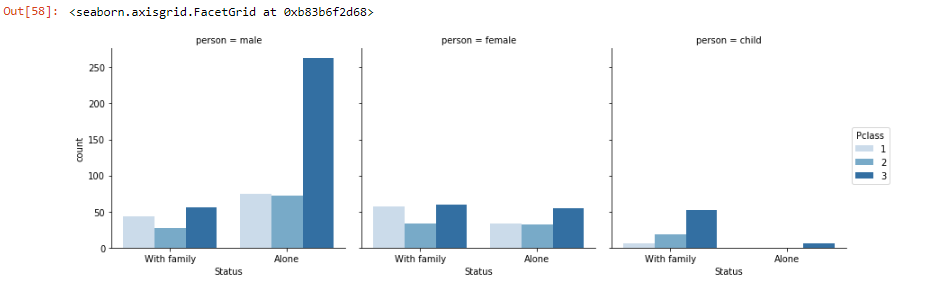
" Another way to perform the above\ntitanic\_df['Alone'] = titanic\_df.SibSp + titanic\_df.Parch\n\ntitanic\_df['Alone'].loc[titanic\_df['Alone']>0] = 'With family'\ntitanic\_df['Alone'].loc[titanic\_df['Alone']==0] = 'Alone'"

In [551]:

fg=sns.factorplot('Alone', data=titanic\_df, kind='count', hue='Pclass', col='person', hue\_order=range(1,4),

palette='Blues')

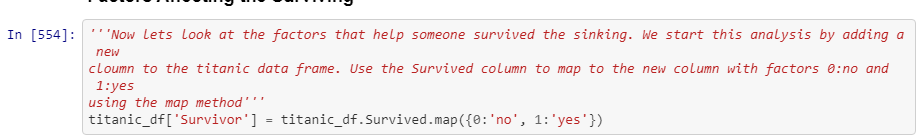
fg.set\_xlabels('Status')



From the figure above, it is clear that most children traveled with family in third class. For men, most traveled alone in third class. On the other hand, the number of female passengers who traveled either with family or alone among the second and third class is comparable. However, more women traveled with family than alone in first class.

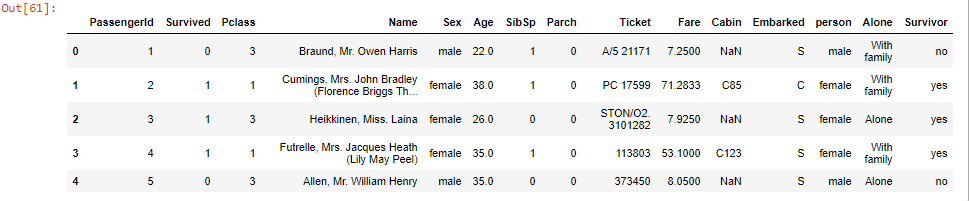
### Factors Affecting the Surviving

In [554]:



In [555]:

titanic\_df.head()

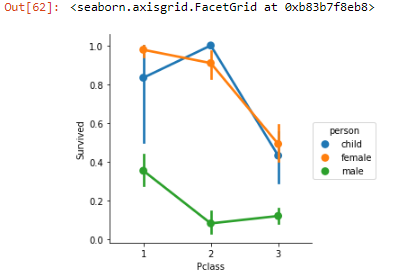


#### Class Factor

In [570]:

*# Survived vs. class Grouped by gender*

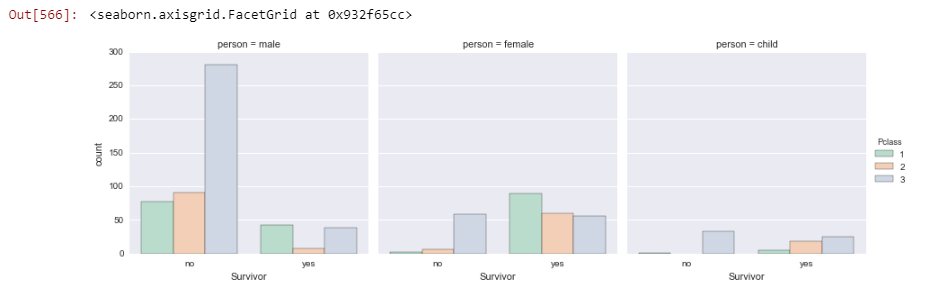
sns.factorplot('Pclass','Survived', hue='person', data=titanic\_df, order=range(1,4), hue\_order = ['child','female','male')



From the figure above, being a male or a third class reduce the chance for one to survive.

In [566]:

sns.factorplot('Survivor', data=titanic\_df, hue='Pclass', kind='count', palette='Pastel2', hue\_order=range(1,4), col='person')

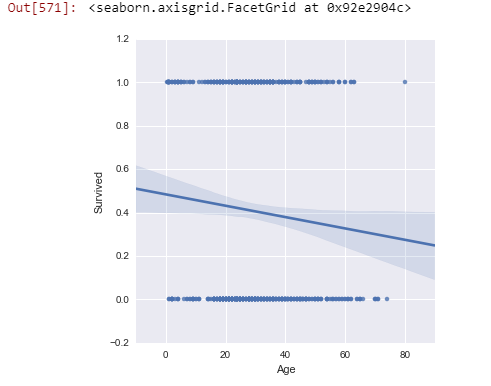


### Age Factor

In [571]:

*# Linear plot of age vs. survived*

sns.lmplot('Age', 'Survived', data=titanic\_df)

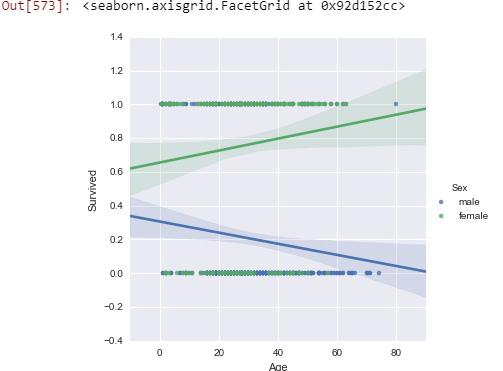


There seems to be a general linear trend between age and the survived field. The plot shows that the older the passenger is, the less chance he/she would survive.

In [573]:

*# Survived vs. Age grouped by Sex*

sns.lmplot('Age', 'Survived', data=titanic\_df, hue='Sex')

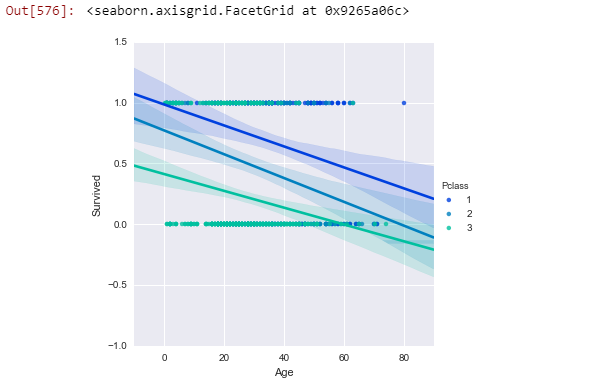


Older women have higher rate of survival than older men as shown in the figure above. Also, older women has higher rate of srvival than younger women; an opposite trend to the one for the male passengers.

In [576]:

*# Survived vs. Age gruped by class*

sns.lmplot('Age', 'Survived', hue='Pclass', data=titanic\_df, palette='winter', hue\_order=range(1,4))



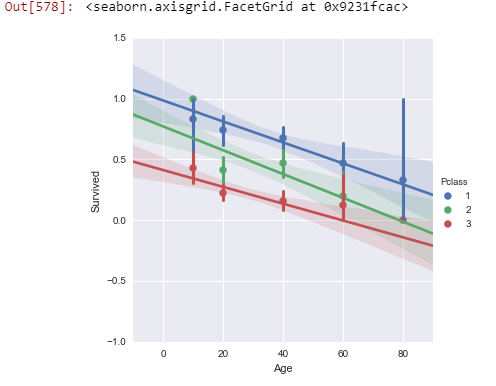
In all three classes, the chance to survive reduced as the passengers got older.

n [578]:

*# Create a generation bin*

generations = [10,20,40,60,80]

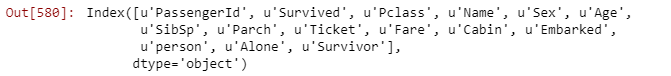
sns.lmplot('Age','Survived',hue='Pclass',data=titanic\_df,x\_bins=generations, hue\_order=[1,2,3])



#### Deck Factor

In [580]:

titanic\_df.columns



In [606]:

titanic\_DF = titanic\_df.dropna(subset=['Cabin'])

In [596]:

d[0:10]

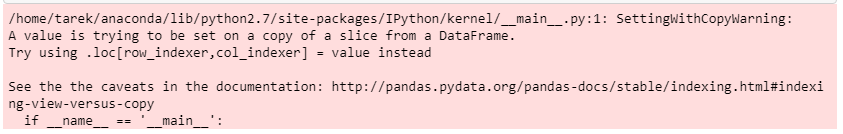
Out[596]:

['C', 'C', 'E', 'G', 'C', 'D', 'A', 'C', 'B', 'D']

In [607]:

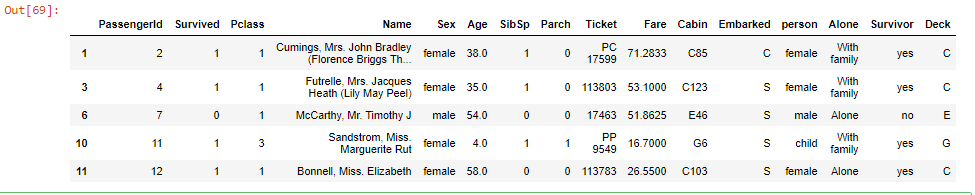
len(titanic\_DF), len(d)

Out[607]: (204, 204)



In [613]: titanic\_DF = titanic\_DF[titanic\_DF.Deck != 'T']

In [614]: titanic\_DF.head()



In [616]: sns.factorplot('Deck', 'Survived', data=titanic\_DF, order=['A','B','C','D','E','F','G'])

Out[616]:

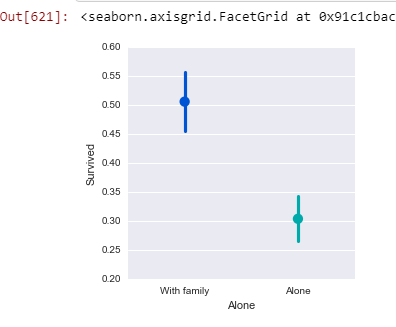


There does not seem to be any relation between deck and the survival rate as shown in the above figure!

#### Family Status Factor

In [621]: sns.factorplot('Alone', 'Survived', data=titanic\_df, palette='winter') *#hue='person',*

*#hue\_order=['child', 'female', 'male'])*

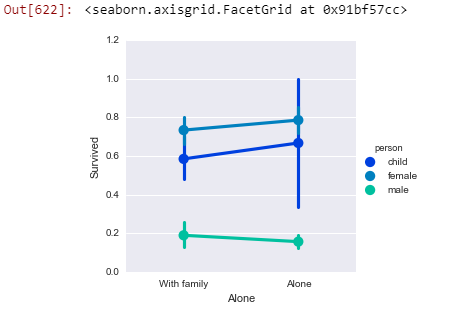


There seems that the survival rate diminishes significantly for those who were alone. However, lets check if a gender or age play a factor. From the figure below, one may conclude that the survival rate for women and children are much higher than that of men, as was concluded previously and as anticipated. However, the survival rate is not significant for either gender or for children who were with family versus who were alone. Moreover, the survival rate for women and children increases for those who were alone. For men, the survival rate diminishes slightly for those who were alone versus for those who were with family.

In [622]:

sns.factorplot('Alone', 'Survived', data=titanic\_df, palette='winter', hue='person',

hue\_order=['child', 'female', 'male'])



In [626]:

*# Lets split it by class now!*

sns.factorplot('Alone', 'Survived', data=titanic\_df, palette='summer', hue='person',

hue\_order=['child', 'female', 'male'], col='Pclass', col\_order=[1,2,3])

