We will follow the charttypeinfographics pdf and analyse first the single columns and then go for the relationships between the different multiple variables.

What is the ratio of males vs females in the organization?

Who has spent a longer duration in the company male or female?

Which skill set is paid highest and lowest?

What is the maximum difference of earnings between the 2018 and 2019 and to which stream it belongs to?

What is the minimum difference of earnings between the 2018 and 2019 and to which stream it belongs to?

What is the impact of the given skill on the work life balance?

What is the impact of given skill on performance rating?

What is the percentage of people falling in the average rating? What is the percentage of highly performing and poorly performing people?

What is the impact of the given skill on the years at the company?

Does the Gender impact your performance rating?

Does the Gender impact your work life balance?

Does the Gender impact the pay hike?

What is the population in the different level of experiences in the workforce?

Plot the relation between the years of experience and the income.

Plot the relation between the total years of experience and the years in the company.

If the person is already having overall higher experience is it likely that he stays longer with the given company? Indirectly is the switch rate lower at the higher level total experience.

Which gender is more involved in which stream?

Skill wise we can try to analyse the impact of exp on the salary

Impact of rating on the salary hike🡪 for this you will need a new column which will hold the difference of the monthly income of two columns Use the below code from the titanic dataset

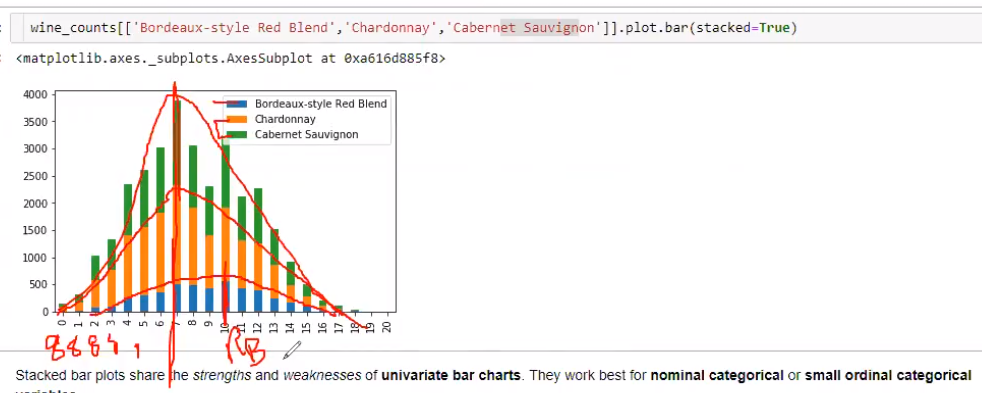
We have to build the story for our data on our own.

Add another column to the data which will map the rating to a subjective description like 4-Role model, 3-Average etc..

Map the Work Life balance rating with a new column call it as WorkLife Rating and map the numbers with category like poor,fair,good,excellent

Notes:

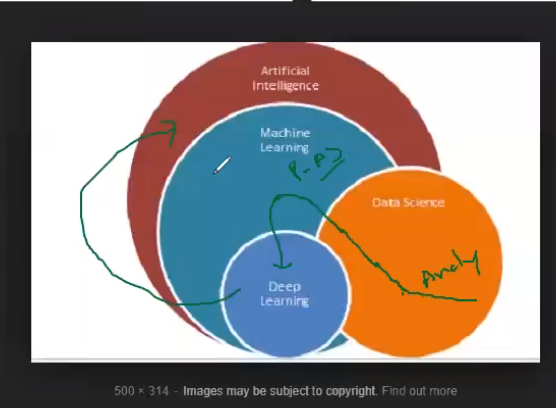
1. Compare the valid range of data. If you are comparing the huge variant amount of data you need to break down and then compare within the category. Compare apples to apples.
2. On bar plot we are required to explicitly sort on index and then plot the graph but if we plot on histogram it automatically sorts and then plots.
3. A two-dimensional plot can contain multiple attributes in the same plot. We should try to capture maximum of 5 attributes of data in the given single plot.
4. The Scatter plot can reveal us the best results when both the axis have numeric and continuous values.
5. In order to understand the concentration of data towards a range we can use the hexbin or the heatmaps.
6. Line chart could be best used for the time based indexed data.
7. Stacked plots example comparison of the company revenues in different quarters.



1. Creating a new Dataframe from the columns of existing dataframe

dataframe=employee\_data.filter(['Gender','Total Working Years'],axis=1)

1. Bokeh can help us draw the user interactive charts and for more information we can refer to the site demo.bokeh.org/movies
2. If we have a labelled data column for example the survived column in a titanic case which is actually dependent on lot of other parameters in the data then we cannot replace the missing values in such case we need to either get it sourced from the right source or get rid of that particular records.
3. If we want to focus on just the visualization and charting we can go using learning Tabeulu after completing the Machine Learning-1. We can add Tabeulu to your profile.



Imports done in the project

import numpy as np # Implemennts milti-dimensional array and matrices

import pandas as pd # For data manipulation and analysis

import pandas\_profiling

import matplotlib.pyplot as plt # Plotting library for Python programming language and it's numerical mathematics extension NumPy

import seaborn as sns # Provides a high level interface for drawing attractive and informative statistical graphics

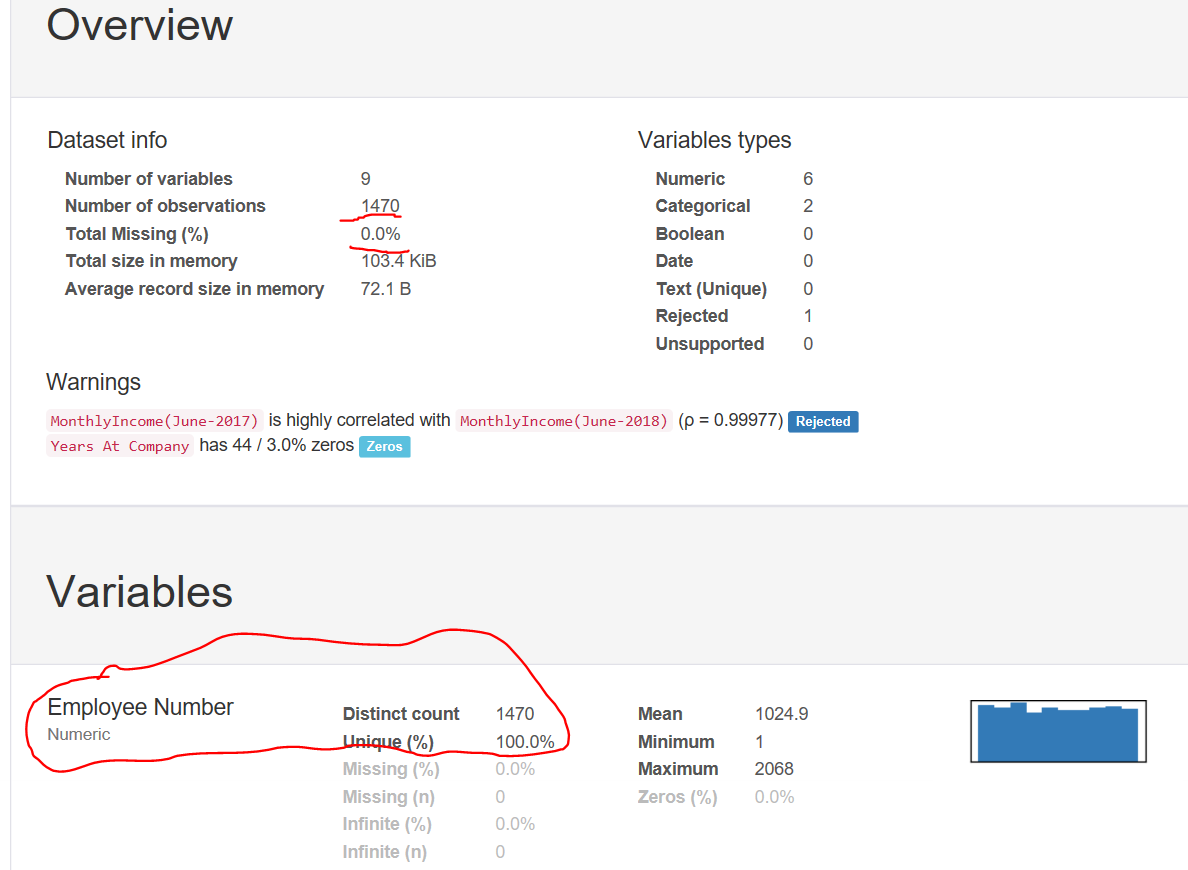
%matplotlib inline

sns.set()

from subprocess import check\_output

# Key Points noted from the Pre profiling of the Data

1. There is no missing value.
2. There are no outliers in any of the columns.
3. Employee Number is the key (index column) as it has all the unique values.



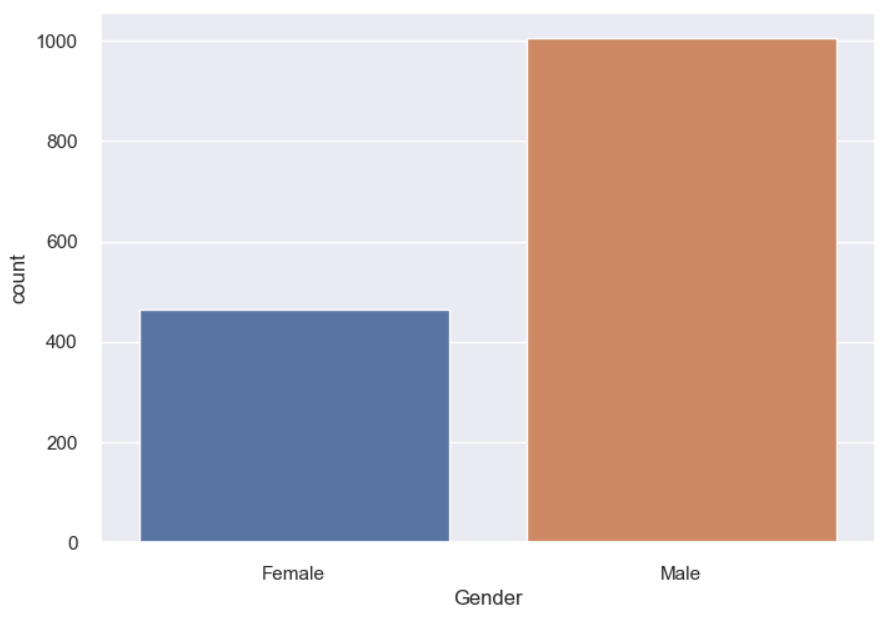
Index(['Gender', 'Employee Number', 'Skills', 'Total Working Years', 'Work Life Balance', 'Performance Rating', 'Years At Company', 'MonthlyIncome(June-2018)', 'MonthlyIncome(June-2017)'], dtype='object')

# For the description we can refer to the tiatic notebook file description section. And for our given data we have to build a background story of our own.

# What is the ratio of males vs females in the organization?

sns.set(style="darkgrid")

ax=sns.countplot(x="Gender",data=employee\_data)



This clearly shows that the ratio of male in the organisation is far more than the females.

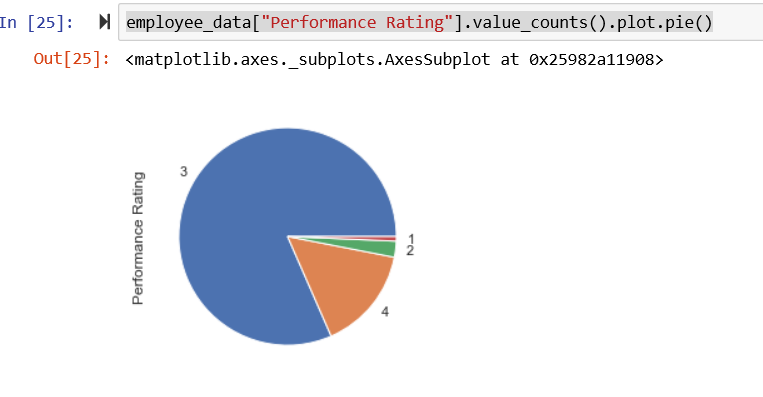
# How many unique skills we have in the organization?

employee\_data.Skills.nunique()

7

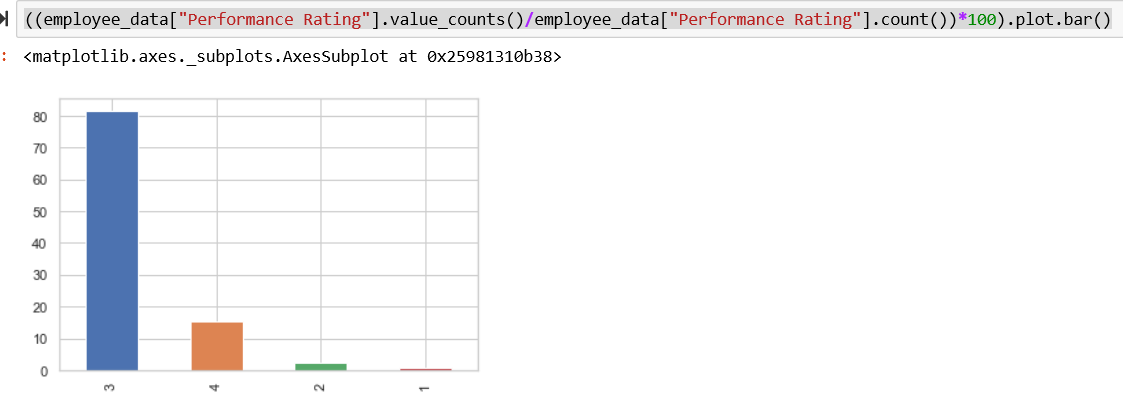
# What are the ratios of people falling under the different buckets of performance ratings?

employee\_data["Performance Rating"].value\_counts().plot.pie()



The same details in terms of % of people falling under the different rate criterias

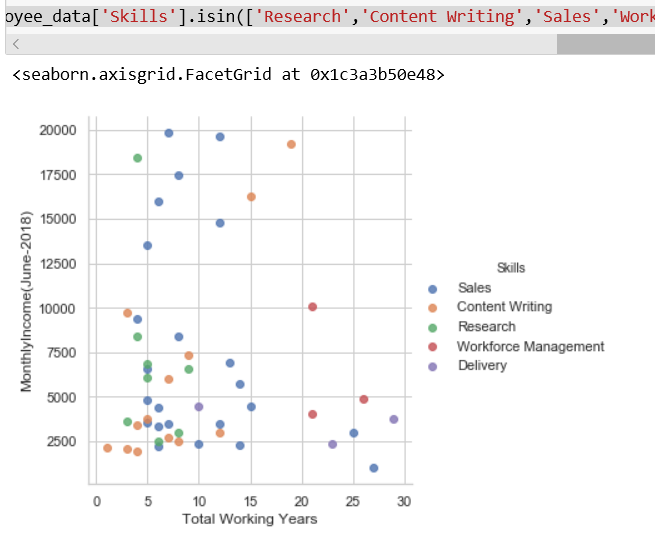
((employee\_data["Performance Rating"].value\_counts()/employee\_data["Performance Rating"].count())\*100).plot.bar()



TODO: We need to put up the axis and labels and figures in the percentages.

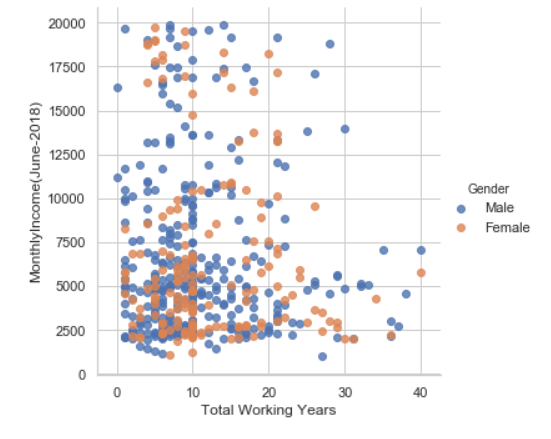
# Relation between the years and the income based on the skill set

sns.lmplot(x='Total Working Years',y='MonthlyIncome(June-2018)',hue='Skills',data=employee\_data.loc[employee\_data['Skills'].isin(['Research','Content Writing','Sales','Workforce Management','Delivery'])].sample(50),fit\_reg=False)



# Relation between the years and the income based on the Gender

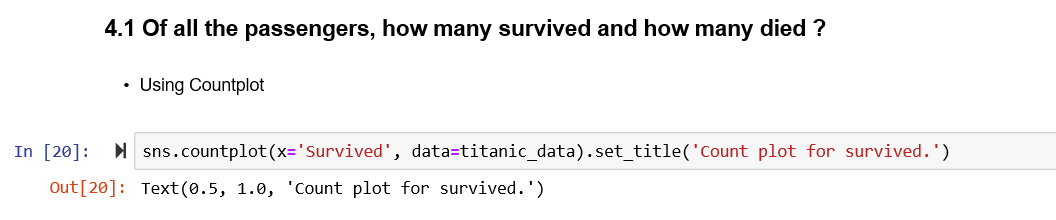
sns.lmplot(x='Total Working Years',y='MonthlyIncome(June-2018)',hue='Gender',data=employee\_data.sample(500),fit\_reg=False)



So from the plot its quite clear that there is no bias on the basis of gender in terms of income. But we can see more blue dots as compared to the brown ones because we have more male population.

# OBSERVATIONS FROM THE TITANIC THAT WE CAN REFER TO INTO OUR PROJECT TOO.

We can use the below logic for the Gender segregation

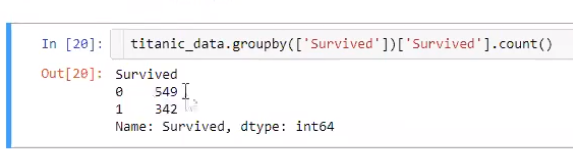


We can also calculate the % of hike as a new column for the better analysis.

titanic\_data['FamilySize'] = titanic\_data['SibSp'] + titanic\_data['Parch']+1

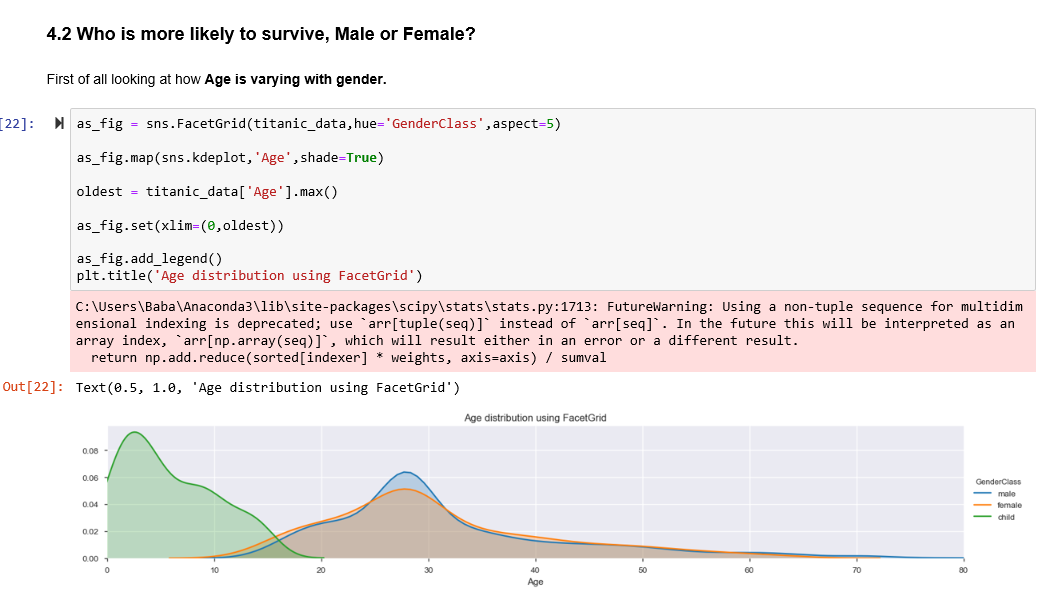
titanic\_data['GenderClass'] = titanic\_data.apply(lambda x: 'child' if x['Age'] < 15 else x['Sex'],axis=1)

Getting the frequency count based on the grouping of data we can also get the % segregation and also write down the conclusions in all the sections after the plots



who is likely to get better rating/hike/years in the organization based on the gender and the skill sets

The below is plotting the frequency of age based on the gender



What is the % of different people in the different skill sets you can do it on basis of gender as well



We can also plot pie for this

f,ax = plt.subplots(1,3,figsize=(20,7))

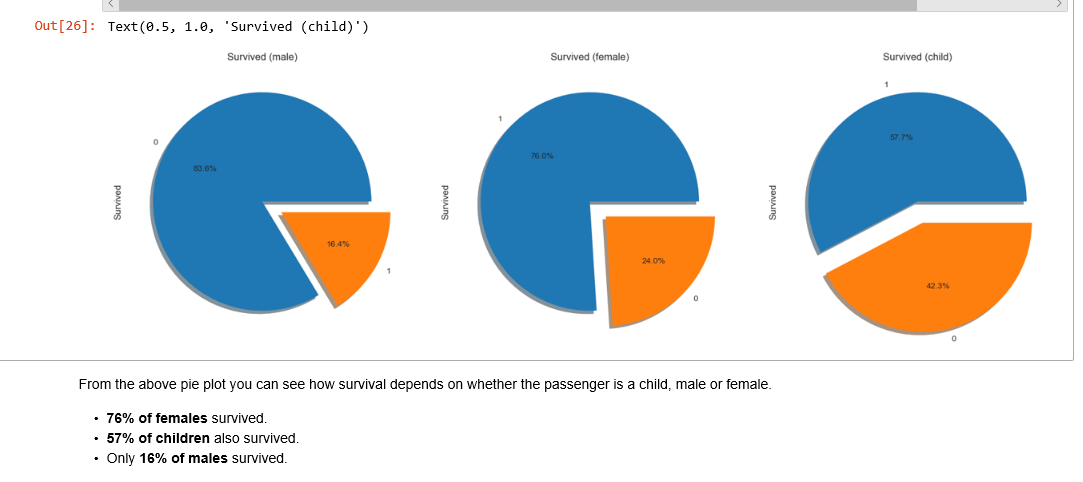
titanic\_data['Survived'][titanic\_data['GenderClass'] == 'male'].value\_counts().plot.pie(explode=[0,0.2],autopct='%1.1f%%',ax=ax[0],shadow=True)

titanic\_data['Survived'][titanic\_data['GenderClass'] == 'female'].value\_counts().plot.pie(explode=[0,0.2],autopct='%1.1f%%',ax=ax[1],shadow=True)

titanic\_data['Survived'][titanic\_data['GenderClass'] == 'child'].value\_counts().plot.pie(explode=[0,0.2],autopct='%1.1f%%',ax=ax[2],shadow=True)

ax[0].set\_title('Survived (male)')

ax[1].set\_title('Survived (female)')

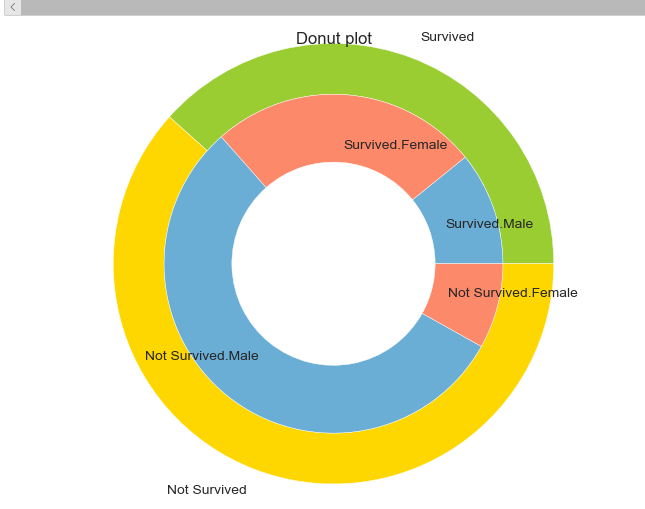
ax[2].set\_title('Survived (child)') 

COMPARING THE COLUMNS FOR A GIVEN FACTOR



Donut pie chart we can use for the segregation based on the skills and the gender.





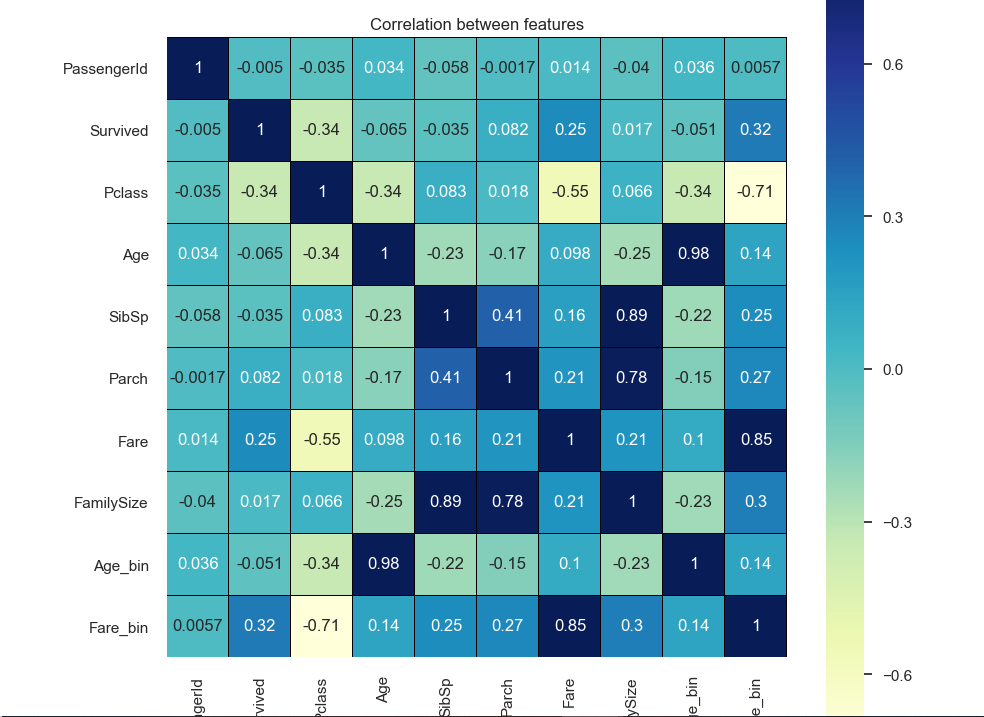
**4.11 Establish coorelation between all the features using heatmap.**

corr = titanic\_data.corr()

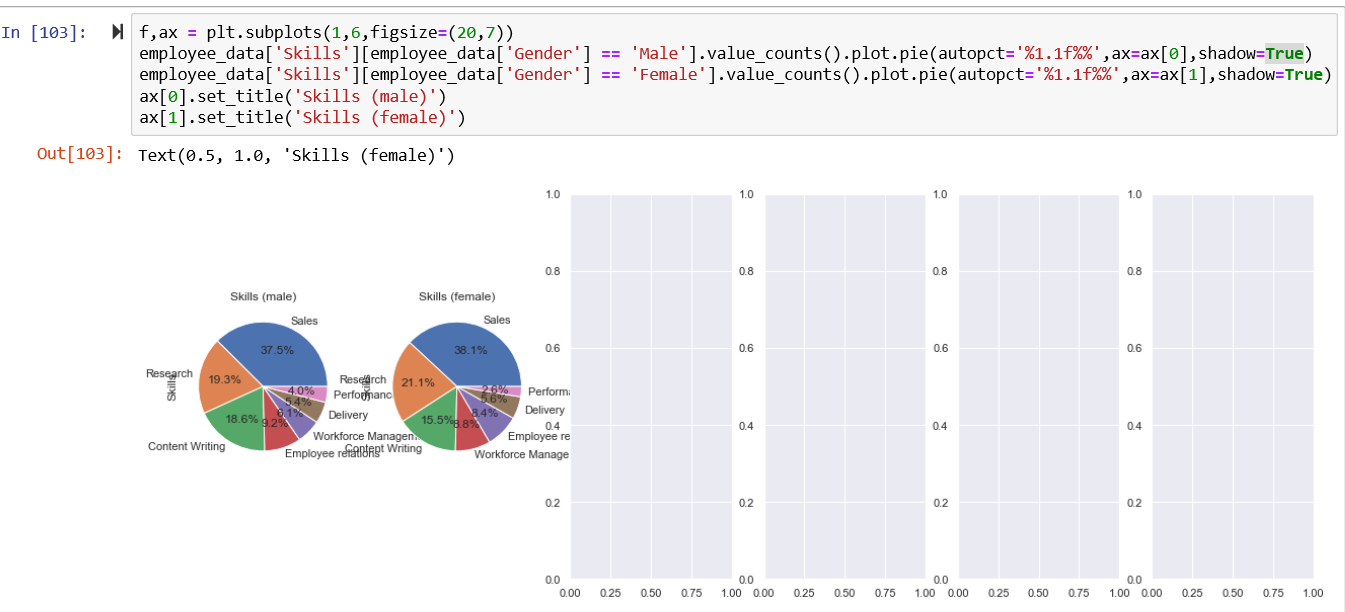
plt.figure(figsize=(10,10))

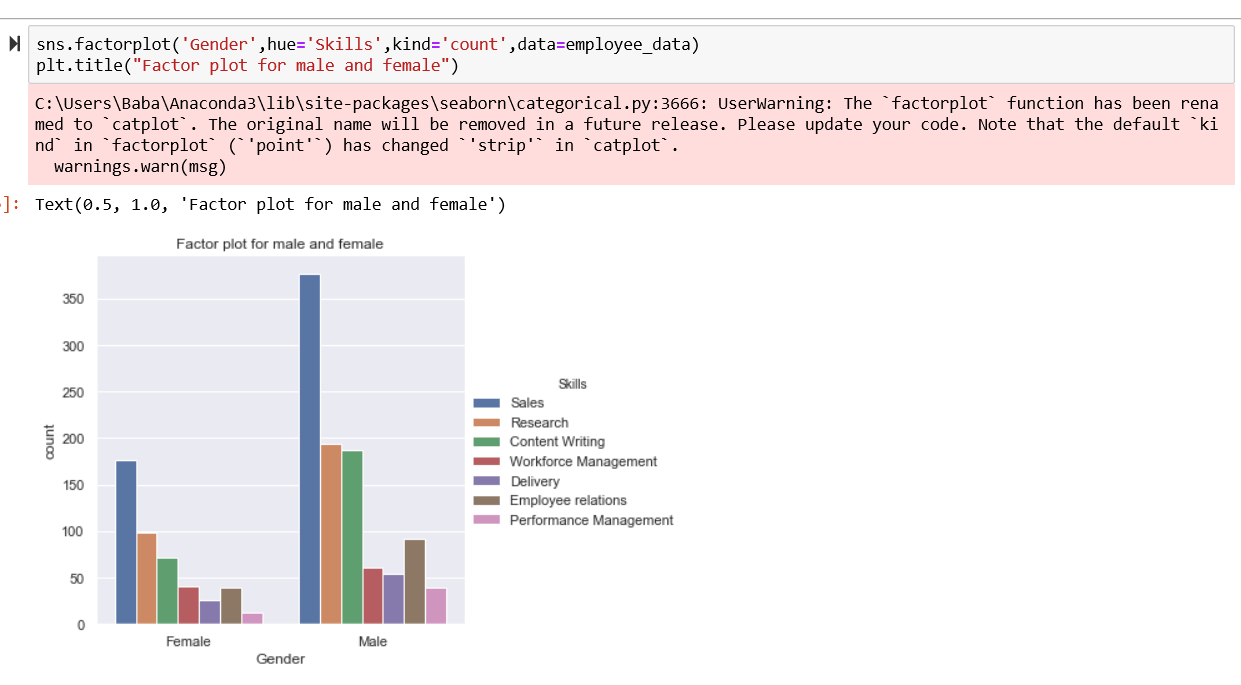
sns.heatmap(corr,vmax=.8,linewidth=.01, square = True, annot = True,cmap='YlGnBu',linecolor ='black')

plt.title('Correlation between features')



The maximum number of employees are lying under the sales dept for both the males and the females which is quite evident from the below given pie and the factor charts





Who has got better ratings is it men or woman?



# Donut Pie chart for different skills in two given genders

(employee\_data.Gender=='Female').sum()

465

(employee\_data.Gender=='Male').sum()

1005