**OBJECTIVE**- This is the Applied Question related to College dataset. In this data our Objective is that Visualize data using some libraries in python and Solve some que.

**Q.1) Using Python read the dataset using some libraries ?**

# importing dataset

import numpy as np

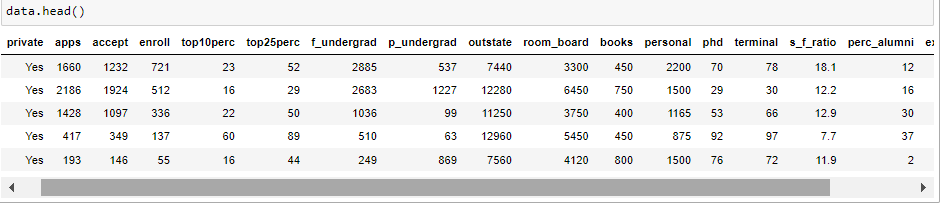
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

data=pd.read\_csv("data.csv")

data



• Private : Public/private indicator .

• Apps : Number of applications received.

• Accept : Number of applicants accepted.

• Enroll : Number of new students enrolled.

• Top10perc : New students from top 10 % of high school class.

• Top25perc : New students from top 25 % of high school class.

• F.Undergrad : Number of full-time undergraduates.

• P.Undergrad : Number of part-time undergraduates.

• Outstate : Out-of-state tuition

• Room.Board : Room and board costs

• Books : Estimated book costs

• Personal : Estimated personal spending

• PhD : Percent of faculty with Ph.D.’s

• Terminal : Percent of faculty with terminal degree

• S.F.Ratio : Student/faculty ratio

• perc.alumni : Percent of alumni who donate

• Expend : Instructional expenditure per student

• Grad.Rate : Graduation rate.

**Q.2) Obtain The first Column in the dataset?**

rown=data['private']

rown

0 Yes

1 Yes

2 Yes

3 Yes

4 Yes

...

772 No

773 Yes

774 Yes

775 Yes

776 Yes

Name: private, Length: 777, dtype: object.

In the first column there is yes/No for private and

public.

**Q.3) Obtain the summaries of the data?**

data.describe()

It gives all the descriptive statistics like mean, median, Q1,Q2,Q3.

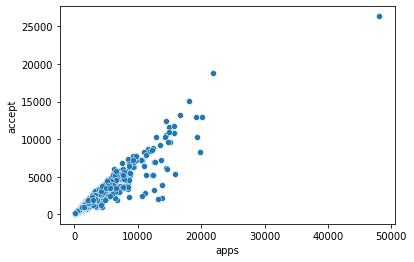
**Q.4) Visualization of the dataset using Different graphs**

A) scatterplot for two variables.

x='apps'

y='accept'

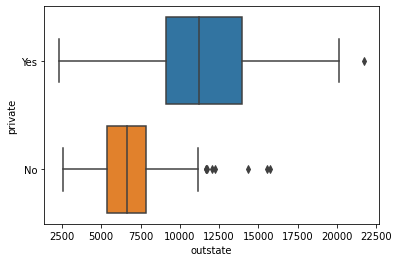
sns.scatterplot(x,y,data=data)



There is positive linear relationship between apps and accept and 1 outlier is present their.

**B) obtain side by side boxplot of outstate vs private**

sns.boxplot(x=data['outstate'],y=data['private'])

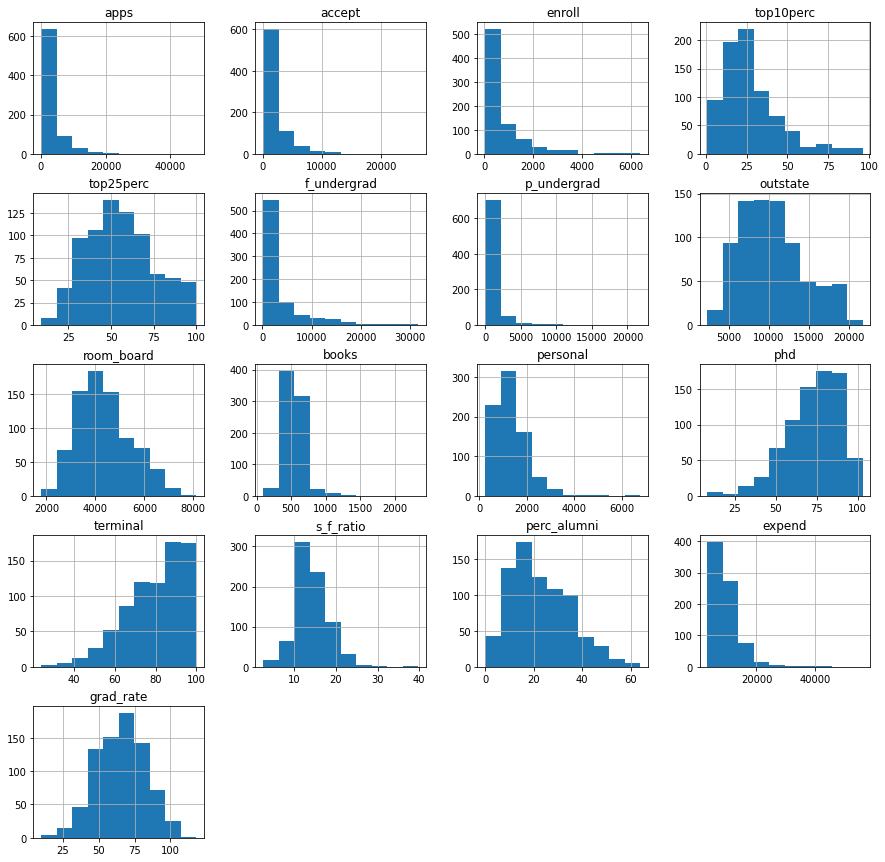


In the given boxplot there is private sector in which outliers present in yes/no indicator.

**C) Obtain histogram of all corresponding variables.**

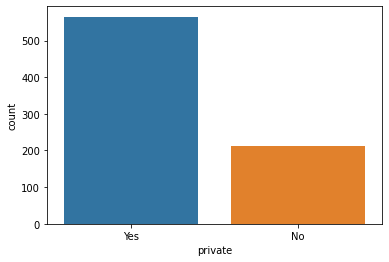
data.hist(figsize=(15,15))

plt.show()



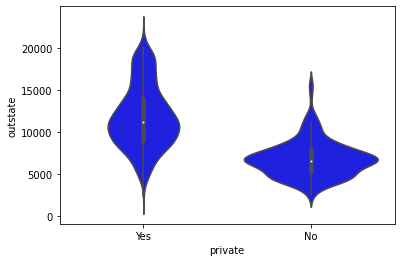
**E) Obtain Countplot for categorical variable.**

sns.countplot(x=data['private'],data=data)



It gives barplot for the categorical variable. Most of students belongs to private sector.

**F) Obtain Violin graph for one or several distribution.**



Violint plot gives distribution of one or several variables. The white dot represents median.The thick grey bar represents interquartile range.

**Q.2) Fitting Linear Regression on the dataset of Hours and Scores. In which we predict Marks of the students is depend on how many hr they studied.**

# Importing Libraries

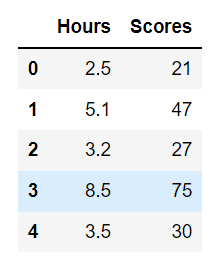
import numpy as np

import seaborn as sns

import pandas as pd

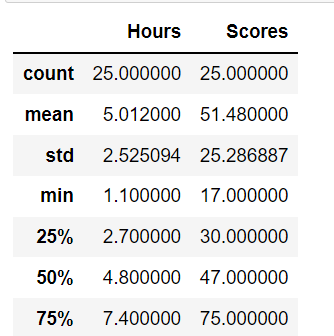
import matplotlib.pyplot as plt

**#Loading the dataset.**



**# For Discriptive Statistics**

data.describe()

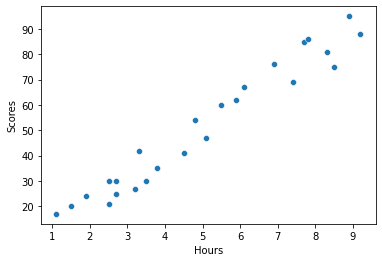


1. To see the relationship between predictor and response.

y='Scores'

x='Hours'

sns.scatterplot(y='Scores',x='Hours',data=data)



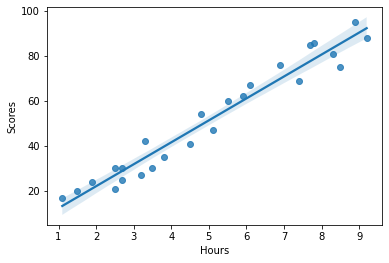
There is Positive relationship between Students Scores and no of hr they studied.

**B) ploting Regression plot.**

y='Scores'

x='Hours'

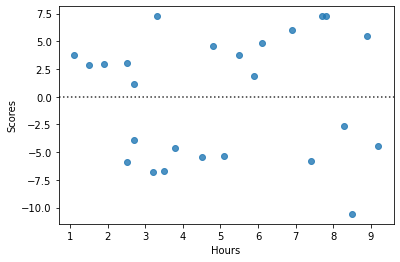
sns.regplot(y='Scores',x='Hours',data=data)



This is the plot for Linear Regression model fit. It means that linear Regression is fits to this data.

**c) Plotting Residual plot.**

sns.residplot(y='Scores',x='Hours',data=data)



### In the residual plot hours is independent on X\_axis and Scores of student is dependent on Y\_axis.In this particular plot residuals are equally and randomly spaced around center line Hence Linear Regression is Best Choice.

**% -----This command is used for location of x and y.----**

x=data['Hours']

y=data['Scores']

x= data.iloc[:, :-1].values

y= data.iloc[:, 1].values

**%----------For training and testing data-----------------**

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,random\_state=0)

**%-----------Fitting Linear Regression ----------------**

from sklearn.linear\_model import LinearRegression

model=LinearRegression()

LinearRegression()

model.fit(x\_train,y\_train)

16.88414476, 33.73226078, 75.357018, 26.79480124, 60.491038

**%---------------B1 and B0----------------**

print(model.coef\_) #B1

If the no.of hours study is increase by 1 then average scores of the students is also increase by 9.9106.

print(model.intercept\_) #B0

Average Scores of the students is 2.0181 when x=0

**%---------------For Confidence Interval---------------**

import scipy.stats as st

st.t.interval(alpha=0.95,df=len(data)-1,loc=np.mean(data), scale=st.sem(data))

In these dataset the no.of hour study of student is lies between the interval(3.96,41.04) and no.of score obtained by studets is lie within the interval (6.04,61.91)