WEEK - 05

Data Preprocessing

- Data preprocessing is the concept of changing the raw data into a clean data set
- The dataset is preprocessed in order to check missing values, noisy data

Importance of data preprocessing

- It improves accuracy and reliability
- changing the raw data into a clean data set

In [56]:
```
import numpy as np
import pandas as pd
stu=pd.read_csv(r'C:\Users\God\Desktop\sample student details.csv')
stu
```

Out[56]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail
0	ram	111.0	142.0	76	9.687000e+09	mandya	CS	ram12@gmail.com
1	sham	112.0	143.0	38	7.338521e+09	maddur	CS	sham456@gmail.com
2	jan	113.0	144.0	56	9.611344e+09	channapatna	EEE	jan123@gmail.com
3	jaamu	114.0	NaN	94	2.344880e+09	kanakapura	EC	jaamu34@gmail.com
4	sashi	115.0	146.0	45	9.876676e+09	sathanur	ME	sashi675@gmail.com
5	madhu	116.0	147.0	24	NaN	ramanagara	CP	madhu23@gmail.com
6	poo	117.0	148.0	63	2.693746e+09	maagadi	EC	poo9143@gmail.com
7	poorvika	NaN	149.0	15	9.594039e+09	banglore	CS	poorvika33@gmail.com
8	sirisha	119.0	150.0	45	4.675679e+09	belagaavi	CP	NaN
9	navya	120.0	151.0	76	9.887654e+09	bidadi	CP	navya@gmail.com
10	nanmay	121.0	152.0	45	7.624919e+09	tumakur	ME	nanmay@gmail.com

In [32]:
```
stu.describe()
```

Out[32]:

	stu_ID	clg_code	stu_marks	stu_cnum
count	10.000000	10.000000	11.000000	1.000000e+01
mean	115.800000	147.200000	52.454545	7.333446e+09
std	3.425395	3.425395	23.551491	3.025661e+09
min	111.000000	142.000000	15.000000	2.344880e+09
25%	113.250000	144.500000	41.500000	5.341389e+09
50%	115.500000	147.500000	45.000000	8.609479e+09
75%	118.500000	149.750000	69.500000	9.668086e+09
max	121.000000	152.000000	94.000000	9.887654e+09

In [33]:
```
stu.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11 entries, 0 to 10
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   stu_name     11 non-null     object 
 1   stu_ID       10 non-null     float64
 2   clg_code     10 non-null     float64
 3   stu_marks    11 non-null     int64  
 4   stu_cnum     10 non-null     float64
 5   stu_address  11 non-null     object 
 6   stu_branch   11 non-null     object 
 7   stu_mail     10 non-null     object 
dtypes: float64(3), int64(1), object(4)
memory usage: 832.0+ bytes
```

In [34]:
```
stu.isnull().sum()
```

Out[34]:
```
stu_name       0
stu_ID         1
clg_code       1
stu_marks      0
stu_cnum       1
stu_address    0
stu_branch     0
stu_mail       1
dtype: int64
```

In [35]:
```
stu.fillna(00)
```

Out[35]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail
0	ram	111.0	142.0	76	9.687000e+09	mandya	CS	ram12@gmail.com
1	sham	112.0	143.0	38	7.338521e+09	maddur	CS	sham456@gmail.com
2	jan	113.0	144.0	56	9.611344e+09	channapatna	EEE	jan123@gmail.com
3	jaamu	114.0	0.0	94	2.344880e+09	kanakapura	EC	jaamu34@gmail.com
4	sashi	115.0	146.0	45	9.876676e+09	sathanur	ME	sashi675@gmail.com
5	madhu	116.0	147.0	24	0.000000e+00	ramanagara	CP	madhu23@gmail.com
6	poo	117.0	148.0	63	2.693746e+09	maagadi	EC	poo9143@gmail.com
7	poorvika	0.0	149.0	15	9.594039e+09	banglore	CS	poorvika33@gmail.com
8	sirisha	119.0	150.0	45	4.675679e+09	belagaavi	CP	0
9	navya	120.0	151.0	76	9.887654e+09	bidadi	CP	navya@gmail.com
10	nanmay	121.0	152.0	45	7.624919e+09	tumakur	ME	nanmay@gmail.com

In [36]:
```
stu.dropna()
```

Out[36]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail
0	ram	111.0	142.0	76	9.687000e+09	mandya	CS	ram12@gmail.com
1	sham	112.0	143.0	38	7.338521e+09	maddur	CS	sham456@gmail.com
2	jan	113.0	144.0	56	9.611344e+09	channapatna	EEE	jan123@gmail.com
4	sashi	115.0	146.0	45	9.876676e+09	sathanur	ME	sashi675@gmail.com
6	poo	117.0	148.0	63	2.693746e+09	maagadi	EC	poo9143@gmail.com
9	navya	120.0	151.0	76	9.887654e+09	bidadi	CP	navya@gmail.com
10	nanmay	121.0	152.0	45	7.624919e+09	tumakur	ME	nanmay@gmail.com

In [37]:
```
stu.dropna(how='all')
```

Out[37]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail
0	ram	111.0	142.0	76	9.687000e+09	mandya	CS	ram12@gmail.com
1	sham	112.0	143.0	38	7.338521e+09	maddur	CS	sham456@gmail.com
2	jan	113.0	144.0	56	9.611344e+09	channapatna	EEE	jan123@gmail.com
3	jaamu	114.0	NaN	94	2.344880e+09	kanakapura	EC	jaamu34@gmail.com
4	sashi	115.0	146.0	45	9.876676e+09	sathanur	ME	sashi675@gmail.com
5	madhu	116.0	147.0	24	NaN	ramanagara	CP	madhu23@gmail.com
6	poo	117.0	148.0	63	2.693746e+09	maagadi	EC	poo9143@gmail.com
7	poorvika	NaN	149.0	15	9.594039e+09	banglore	CS	poorvika33@gmail.com
8	sirisha	119.0	150.0	45	4.675679e+09	belagaavi	CP	NaN
9	navya	120.0	151.0	76	9.887654e+09	bidadi	CP	navya@gmail.com
10	nanmay	121.0	152.0	45	7.624919e+09	tumakur	ME	nanmay@gmail.com

In [38]:
```
stu.isna()
```

Out[38]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	True	False	False	False	False	False
4	False	False	False	False	False	False	False	False
5	False	False	False	False	True	False	False	False
6	False	False	False	False	False	False	False	False
7	False	True	False	False	False	False	False	False
8	False	False	False	False	False	False	False	True
9	False	False	False	False	False	False	False	False
10	False	False	False	False	False	False	False	False

Outliers

...It is an observation that lies an abnormal distance from other values in a random sample from a population ...It can essentially be much smaller or much larger than values in the dataset

In [39]:
```
import numpy as np
import pandas as pd
age=[20,25,30,35,40,45,50]
height=[150,155,160,165,170,175,180]
d_frame=pd.DataFrame(list(zip(age,height)),columns=['Age','Height(cm)'])
d_frame
```

Out[39]:

	Age	Height(cm)
0	20	150
1	25	155
2	30	160
3	35	165
4	40	170
5	45	175
6	50	180

In [40]:
```
import scipy.stats
scipy.stats.zscore(d_frame)
```

Out[40]:

	Age	Height(cm)
0	-1.5	-1.5
1	-1.0	-1.0
2	-0.5	-0.5
3	0.0	0.0
4	0.5	0.5
5	1.0	1.0
6	1.5	1.5

Reduction

Statisrical technique of reducing the amount of random variable in problem by obtaining a set of principle variables.

In [54]:
```
stu.drop_duplicates(subset=None,keep='first',inplace=True)
print(stu)
```

```
    stu_name  stu_ID  clg_code  stu_marks      stu_cnum  stu_address  \
0        ram   111.0     142.0        NaN  9.687000e+09       mandya
1       sham   112.0     143.0     0.0070  7.338521e+09       maddur
2        jan   113.0     144.0     0.0070  9.611344e+09  channapatna
3      jaamu   114.0       NaN        NaN  2.344880e+09   kanakapura
4      sashi   115.0     146.0        NaN  9.876676e+09     sathanur
5      madhu   116.0     147.0     0.0068           NaN   ramanagara
6        poo   117.0     148.0     0.0068  2.693746e+09      maagadi
7   poorvika     NaN     149.0     0.0067  9.594039e+09     banglore
8    sirisha   119.0     150.0     0.0067  4.675679e+09    belagaavi
9      navya   120.0     151.0     0.0066  9.887654e+09       bidadi
10    nanmay   121.0     152.0     0.0066  7.624919e+09       tumakur

   stu_branch              stu_mail  stu_log
0          CS       ram12@gmail.com      NaN
1          CS     sham456@gmail.com      NaN
2         EEE      jan123@gmail.com      0.0
3          EC     jaamu34@gmail.com      NaN
4          ME    sashi675@gmail.com      NaN
5          CP     madhu23@gmail.com      NaN
6          EC     poo9143@gmail.com      0.0
7          CS  poorvika33@gmail.com      0.0
8          CP                   NaN      0.0
9          CP       navya@gmail.com      0.0
10         ME      nanmay@gmail.com      0.0
```

Redundancy

Data redundancy often refers to having two identical samples of data in two different places within your database

In [11]:
```
stu.duplicated(subset=None, keep='first').sum()
```

Out[11]:
```
0
```

Data transformation

The technical processof converting data from one format, standard, or structure to another-without changing the content of the datasets

In [52]:
```
stu=stu.round(4)
stu.head()
```

Out[52]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail	stu_lo
0	ram	111.0	142.0	NaN	9.687000e+09	mandya	CS	ram12@gmail.com	Na
1	sham	112.0	143.0	0.007	7.338521e+09	maddur	CS	sham456@gmail.com	Na
2	jan	113.0	144.0	0.007	9.611344e+09	channapatna	EEE	jan123@gmail.com	0
3	jaamu	114.0	NaN	NaN	2.344880e+09	kanakapura	EC	jaamu34@gmail.com	Na
4	sashi	115.0	146.0	NaN	9.876676e+09	sathanur	ME	sashi675@gmail.com	Na

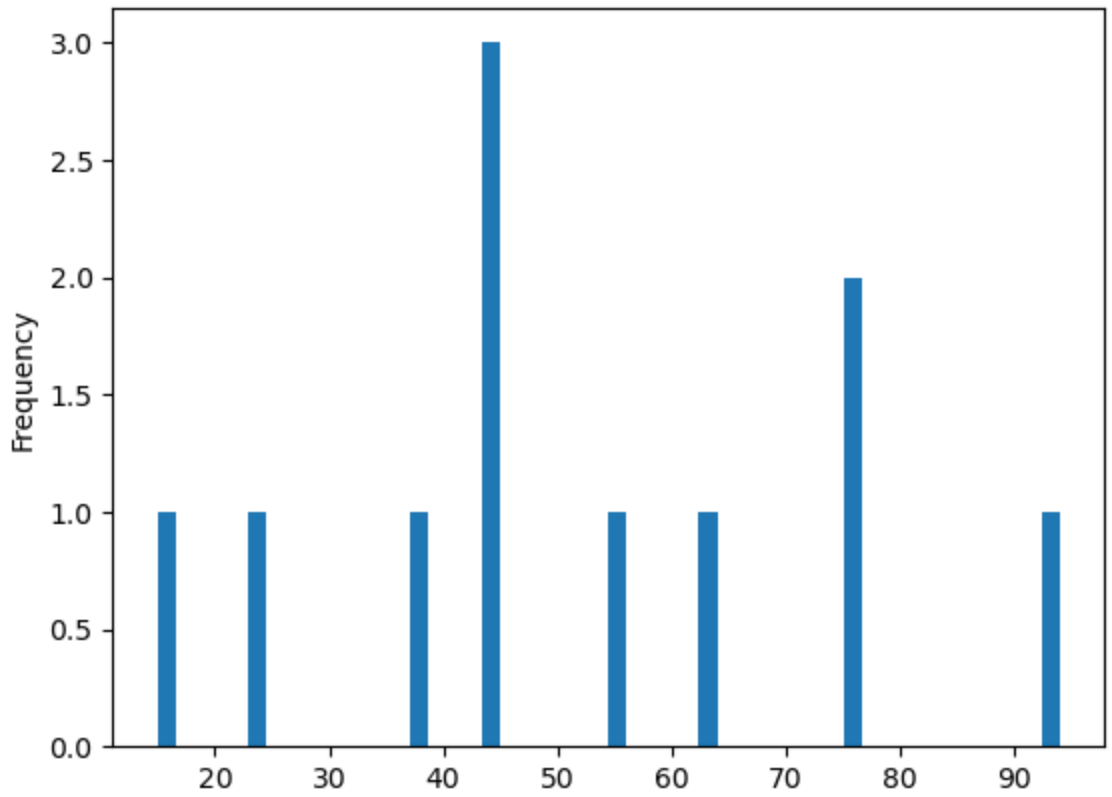
In [48]:
```
import numpy as np
stu['stu_marks']=np.log(stu.clg_code).diff()
stu['stu_log']=stu.stu_marks.rolling(2).std()
```

Out[48]:

	stu_name	stu_ID	clg_code	stu_marks	stu_cnum	stu_address	stu_branch	stu_mail	stu
0	ram	111.0	142.0	NaN	9.687000e+09	mandya	CS	ram12@gmail.com	
1	sham	112.0	143.0	0.007018	7.338521e+09	maddur	CS	sham456@gmail.com	
2	jan	113.0	144.0	0.006969	9.611344e+09	channapatna	EEE	jan123@gmail.com	0.00
3	jaamu	114.0	NaN	NaN	2.344880e+09	kanakapura	EC	jaamu34@gmail.com	
4	sashi	115.0	146.0	NaN	9.876676e+09	sathanur	ME	sashi675@gmail.com	
5	madhu	116.0	147.0	0.006826	NaN	ramanagara	CP	madhu23@gmail.com	
6	poo	117.0	148.0	0.006780	2.693746e+09	maagadi	EC	poo9143@gmail.com	0.00
7	poorvika	NaN	149.0	0.006734	9.594039e+09	banglore	CS	poorvika33@gmail.com	0.00
8	sirisha	119.0	150.0	0.006689	4.675679e+09	belagaavi	CP	NaN	0.00
9	navya	120.0	151.0	0.006645	9.887654e+09	bidadi	CP	navya@gmail.com	0.00

In [57]:
```
import matplotlib.pyplot as plt
stu.stu_marks.plot(kind='hist',bins=50)
```

Out[57]:
```
<AxesSubplot:ylabel='Frequency'>
```



In [ ]: