In [633]:
```
import pandas as pd
import numpy as np

df = pd.read_csv("data_set_cgpa_as_percentage/students_data_cgpa_percen.csv")
```

In [634]:
```
df.head()
```

Out[634]:

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
0	31401.0	Lois	H	Walker	F	4	76	54	56	64	62.50	89.0
1	31402.0	Brenda	S	Robinson	F	4	64	41	97	51	63.25	95.0
2	31403.0	Joe	W	Robinson	M	4	81	46	56	64	61.75	70.0
3	31404.0	Diane	I	Evans	F	4	48	60	60	44	53.00	61.0
4	NaN	Benjamin	R	Russell	M	4	55	58	78	47	59.50	64.0

In [635]:
```
df.dtypes
```

Out[635]:
```
Roll Number       float64
First Name         object
Middle Initial     object
Last Name          object
Gender             object
No of subjects      int64
CC                  int64
WT                  int64
AI                  int64
DSBDA               int64
CGPA              float64
Attendance        float64
dtype: object
```

In [636]:
```
df['Roll Number'] = df['Roll Number'].astype(float).astype("Int64")
df.dtypes
```

Out[636]:
```
Roll Number         Int64
First Name         object
Middle Initial     object
Last Name          object
Gender             object
No of subjects      int64
CC                  int64
WT                  int64
AI                  int64
DSBDA               int64
CGPA              float64
```

```
Attendance        float64
dtype: object
```

In [637]:
```
df.head(100)
```

Out[637]:

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
0	31401	Lois	H	Walker	F	4	76	54	56	64	62.50	89.0
1	31402	Brenda	S	Robinson	F	4	64	41	97	51	63.25	95.0
2	31403	Joe	W	Robinson	M	4	81	46	56	64	61.75	70.0
3	31404	Diane	I	Evans	F	4	48	60	60	44	53.00	61.0
4	<NA>	Benjamin	R	Russell	M	4	55	58	78	47	59.50	64.0
...	...	...	...	...	...	...	...	...	...	...	...	...
95	31496	Jose	K	Hill	M	4	61	48	50	99	64.50	84.0
96	31497	Harold	Z	Nelson	M	4	97	41	90	63	72.75	NaN
97	31498	Nicole	O	Ward	F	4	92	80	53	48	68.25	86.0
98	31499	Theresa	R	Murphy	F	4	53	68	69	82	68.00	NaN
99	31500	Tammy	B	Young	F	4	100	45	72	74	72.75	62.0

100 rows × 12 columns

In [638]:
```
df['CGPA']
```

Out[638]:
```
0     62.50
1     63.25
2     61.75
3     53.00
4     59.50
      ...
95    64.50
96    72.75
97    68.25
98    68.00
99    72.75
Name: CGPA, Length: 100, dtype: float64
```

In [639]:
```
df['CGPA'].min()
```

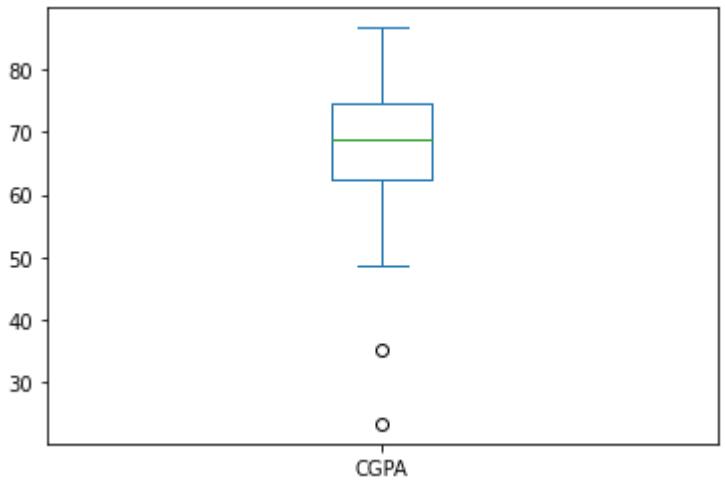
Out[639]:
```
23.3
```

In [640]:
```
df['CGPA'].max()
```

Out[640]:
```
86.75
```

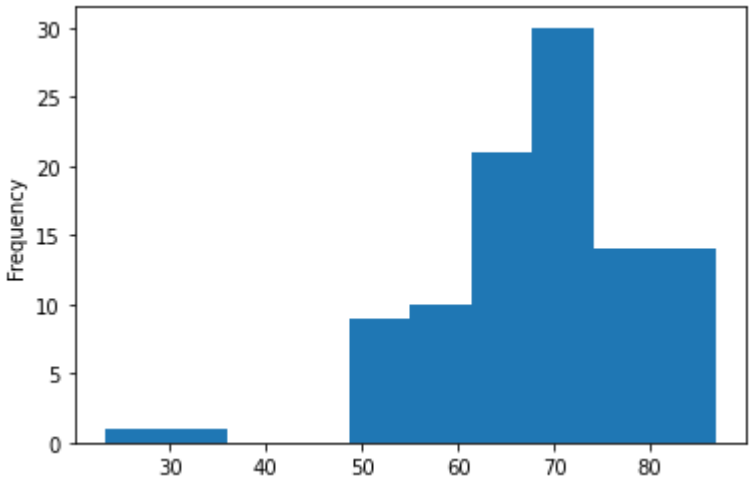
In [641]:
```
df['CGPA'].plot(kind='box')
```

Out[641]: <AxesSubplot:>



In [642]:
```
df['CGPA'].plot(kind='hist')
```

Out[642]: <AxesSubplot:ylabel='Frequency'>



In [643]:
```
df.head(100)
```

Out[643]:

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
0	31401	Lois	H	Walker	F	4	76	54	56	64	62.50	89.0
1	31402	Brenda	S	Robinson	F	4	64	41	97	51	63.25	95.0
2	31403	Joe	W	Robinson	M	4	81	46	56	64	61.75	70.0
3	31404	Diane	I	Evans	F	4	48	60	60	44	53.00	61.0
4	<NA>	Benjamin	R	Russell	M	4	55	58	78	47	59.50	64.0
...	...	...	...	...	...	...	...	...	...	...	...	...
95	31496	Jose	K	Hill	M	4	61	48	50	99	64.50	84.0
96	31497	Harold	Z	Nelson	M	4	97	41	90	63	72.75	NaN
97	31498	Nicole	O	Ward	F	4	92	80	53	48	68.25	86.0

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
98	31499	Theresa	R	Murphy	F	4	53	68	69	82	68.00	NaN
99	31500	Tammy	B	Young	F	4	100	45	72	74	72.75	62.0

100 rows × 12 columns

Finding total count and indexes of missing values in attendance fields and replacing them with mean value.

In [644]:
```
print('Total Rows with missing attendance :
',format(df['Attendance'].head(100).isna().sum()))
indices_attendance = []
for index, row in df.iterrows():
     if(pd.isnull(row['Attendance'])):
         indices_attendance.append(index)
print('Indices : ',indices_attendance)
```

```
Total Rows with missing attendance :  8
Indices :  [5, 14, 26, 30, 69, 89, 96, 98]
```

In [645]:
```
att_mean = df['Attendance'].mean()
df['Attendance'].fillna(att_mean, inplace= True)
df['Attendance'] = df['Attendance'].round(1)
```

In [646]:
```
df.head(100)
```

Out[646]:

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
0	31401	Lois	H	Walker	F	4	76	54	56	64	62.50	89.0
1	31402	Brenda	S	Robinson	F	4	64	41	97	51	63.25	95.0
2	31403	Joe	W	Robinson	M	4	81	46	56	64	61.75	70.0
3	31404	Diane	I	Evans	F	4	48	60	60	44	53.00	61.0
4	<NA>	Benjamin	R	Russell	M	4	55	58	78	47	59.50	64.0
...	...	...	...	...	...	...	...	...	...	...	...	...
95	31496	Jose	K	Hill	M	4	61	48	50	99	64.50	84.0
96	31497	Harold	Z	Nelson	M	4	97	41	90	63	72.75	79.6
97	31498	Nicole	O	Ward	F	4	92	80	53	48	68.25	86.0
98	31499	Theresa	R	Murphy	F	4	53	68	69	82	68.00	79.6
99	31500	Tammy	B	Young	F	4	100	45	72	74	72.75	62.0

100 rows × 12 columns

Finding total count and indexes of missing values in roll number fields and replacing them with arbritary value.

In [647]:
```
print('Total Rows with missing roll numbers : ',format(df['Roll
Number'].head(100).isna().sum()))
indices_rollN = []
for index, row in df.iterrows():
     if(pd.isnull(row['Roll Number'])):
         indices_rollNrollN.append(index)
print('Indices : ',indices_rollN)
```

```
Total Rows with missing roll numbers :  6
Indices :  [4, 22, 32, 70, 75, 90]
```

In [648]:
```
 for index, row in df.iterrows():
     if(pd.isnull(row['Roll Number'])):
         df.loc[index,'Roll Number'] = (df.at[index+1,'Roll Number'])-1
```

In [649]:
```
df.head(100)
```

Out[649]:

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
0	31401	Lois	H	Walker	F	4	76	54	56	64	62.50	89.0
1	31402	Brenda	S	Robinson	F	4	64	41	97	51	63.25	95.0
2	31403	Joe	W	Robinson	M	4	81	46	56	64	61.75	70.0
3	31404	Diane	I	Evans	F	4	48	60	60	44	53.00	61.0
4	31405	Benjamin	R	Russell	M	4	55	58	78	47	59.50	64.0
...	...	...	...	...	...	...	...	...	...	...	...	...
95	31496	Jose	K	Hill	M	4	61	48	50	99	64.50	84.0
96	31497	Harold	Z	Nelson	M	4	97	41	90	63	72.75	79.6
97	31498	Nicole	O	Ward	F	4	92	80	53	48	68.25	86.0
98	31499	Theresa	R	Murphy	F	4	53	68	69	82	68.00	79.6
99	31500	Tammy	B	Young	F	4	100	45	72	74	72.75	62.0

100 rows × 12 columns

Scaling CGPA values from the scale of 100 to the scale of 0 -> 10 using MinMaxScaler(min,max)

In [650]:
```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler(feature_range=(0,10))
df[['CGPA']] = scaler.fit_transform(df[['CGPA']])
df['CGPA'] = df['CGPA'].round(1)
```

In [651]:
```
df.head(100)
```

Out[651]:

	Roll Number	First Name	Middle Initial	Last Name	Gender	No of subjects	CC	WT	AI	DSBDA	CGPA	Attendance
0	31401	Lois	H	Walker	F	4	76	54	56	64	6.2	89.0
1	31402	Brenda	S	Robinson	F	4	64	41	97	51	6.3	95.0
2	31403	Joe	W	Robinson	M	4	81	46	56	64	6.1	70.0
3	31404	Diane	I	Evans	F	4	48	60	60	44	4.7	61.0
4	31405	Benjamin	R	Russell	M	4	55	58	78	47	5.7	64.0
...	...	...	...	...	...	...	...	...	...	...	...	...
95	31496	Jose	K	Hill	M	4	61	48	50	99	6.5	84.0
96	31497	Harold	Z	Nelson	M	4	97	41	90	63	7.8	79.6
97	31498	Nicole	O	Ward	F	4	92	80	53	48	7.1	86.0
98	31499	Theresa	R	Murphy	F	4	53	68	69	82	7.0	79.6
99	31500	Tammy	B	Young	F	4	100	45	72	74	7.8	62.0

100 rows × 12 columns