In [29]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [30]:

```
data=pd.read_csv('Datasets\\student-mat.csv')
```

In [31]:

```
data.head(10)
```

Out[31]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	f
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	
5	GP	M	16	U	LE3	T	4	3	services	other	...	5	
6	GP	M	16	U	LE3	T	2	2	other	other	...	4	
7	GP	F	17	U	GT3	A	4	4	other	teacher	...	4	
8	GP	M	15	U	LE3	A	3	2	services	other	...	4	
9	GP	M	15	U	GT3	T	3	4	other	other	...	5	

10 rows × 33 columns

In [32]:

```
data.tail(10)
```

Out[32]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel
385	MS	F	18	R	GT3	T	2	2	at_home	other	...	5
386	MS	F	18	R	GT3	T	4	4	teacher	at_home	...	4
387	MS	F	19	R	GT3	T	2	3	services	other	...	5
388	MS	F	18	U	LE3	T	3	1	teacher	services	...	4
389	MS	F	18	U	GT3	T	1	1	other	other	...	1
390	MS	M	20	U	LE3	A	2	2	services	services	...	5
391	MS	M	17	U	LE3	T	3	1	services	services	...	2
392	MS	M	21	R	GT3	T	1	1	other	other	...	5
393	MS	M	18	R	LE3	T	3	2	services	other	...	4
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3

10 rows × 33 columns

In [33]:

```
data.describe()
```

Out[33]:

	age	Medu	Fedu	traveltime	studytime	failures	famrel
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000

In [34]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64 
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64 
 7   Fedu        395 non-null    int64 
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64 
 13  studytime   395 non-null    int64 
 14  failures    395 non-null    int64 
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64 
 24  freetime    395 non-null    int64 
 25  goout       395 non-null    int64 
 26  Dalc        395 non-null    int64 
 27  Walc        395 non-null    int64 
 28  health      395 non-null    int64 
 29  absences    395 non-null    int64 
 30  G1          395 non-null    int64 
 31  G2          395 non-null    int64 
 32  G3          395 non-null    int64 
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
```

In [35]:

```
data.isnull().sum()
```

Out[35]:

```
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

In [36]:

```
data.isnull()
```

Out[36]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	f
0	False	False	False	False	False	False	False	False	False	False	...	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	
390	False	False	False	False	False	False	False	False	False	False	...	False	
391	False	False	False	False	False	False	False	False	False	False	...	False	
392	False	False	False	False	False	False	False	False	False	False	...	False	
393	False	False	False	False	False	False	False	False	False	False	...	False	
394	False	False	False	False	False	False	False	False	False	False	...	False	

395 rows × 33 columns

In [37]:

```
data.shape
```

Out[37]:

```
(395, 33)
```

Handling with missing data

isnull() notnull() dropna() fillna() replace()

In [38]:

```
data.isnull()
```

Out[38]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	f
0	False	False	False	False	False	False	False	False	False	False	...	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	
390	False	False	False	False	False	False	False	False	False	False	...	False	
391	False	False	False	False	False	False	False	False	False	False	...	False	
392	False	False	False	False	False	False	False	False	False	False	...	False	
393	False	False	False	False	False	False	False	False	False	False	...	False	
394	False	False	False	False	False	False	False	False	False	False	...	False	

395 rows × 33 columns

In [39]:

```
data.notnull()
```

Out[39]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	free
0	True	True	True	True	True	True	True	True	True	True	...	True	
1	True	True	True	True	True	True	True	True	True	True	...	True	
2	True	True	True	True	True	True	True	True	True	True	...	True	
3	True	True	True	True	True	True	True	True	True	True	...	True	
4	True	True	True	True	True	True	True	True	True	True	...	True	
...	...	...	...	...	...	...	...	...	...	...	...	...	
390	True	True	True	True	True	True	True	True	True	True	...	True	
391	True	True	True	True	True	True	True	True	True	True	...	True	
392	True	True	True	True	True	True	True	True	True	True	...	True	
393	True	True	True	True	True	True	True	True	True	True	...	True	
394	True	True	True	True	True	True	True	True	True	True	...	True	

395 rows × 33 columns

In [40]:

```
data.isnull().sum
```

Out[40]:

```
<bound method NDFrame._add_numeric_operations.<locals>.sum of      school
sex    age  address  famsize  Pstatus   Medu   Fedu   Mjob  \
0     False  False  False    False    False    False  False  False  False
1     False  False  False    False    False    False  False  False  False
2     False  False  False    False    False    False  False  False  False
3     False  False  False    False    False    False  False  False  False
4     False  False  False    False    False    False  False  False  False
..      ...    ...    ...      ...      ...      ...    ...    ...    ...
390   False  False  False    False    False    False  False  False  False
391   False  False  False    False    False    False  False  False  False
392   False  False  False    False    False    False  False  False  False
393   False  False  False    False    False    False  False  False  False
394   False  False  False    False    False    False  False  False  False

      Fjob  ...  famrel  freetime  goout   Dalc   Walc  health  absences
\
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     False  False  False  False   False     False
4    False  ...   False     False  False  False  False   False     False
..     ...  ...     ...       ...    ...    ...    ...     ...       ...
390  False  ...   False     False  False  False  False   False     False
391  False  ...   False     False  False  False  False   False     False
392  False  ...   False     False  False  False  False   False     False
393  False  ...   False     False  False  False  False   False     False
394  False  ...   False     False  False  False  False   False     False

        G1     G2     G3
0    False  False  False
1    False  False  False
2    False  False  False
3    False  False  False
4    False  False  False
..     ...    ...    ...
390  False  False  False
391  False  False  False
392  False  False  False
393  False  False  False
394  False  False  False

[395 rows x 33 columns]>
```

In [41]:

```
data.dropna()
```

Out[41]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4
3	GP	F	15	U	GT3	T	4	2	health	services	...	3
4	GP	F	16	U	GT3	T	3	3	other	other	...	4
...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	...	5
391	MS	M	17	U	LE3	T	3	1	services	services	...	2
392	MS	M	21	R	GT3	T	1	1	other	other	...	5
393	MS	M	18	R	LE3	T	3	2	services	other	...	4
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3

395 rows × 33 columns

In [42]:

```
data.fillna(-99)
```

Out[42]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4
3	GP	F	15	U	GT3	T	4	2	health	services	...	3
4	GP	F	16	U	GT3	T	3	3	other	other	...	4
...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	...	5
391	MS	M	17	U	LE3	T	3	1	services	services	...	2
392	MS	M	21	R	GT3	T	1	1	other	other	...	5
393	MS	M	18	R	LE3	T	3	2	services	other	...	4
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3

395 rows × 33 columns

In [15]:

```
dup_df=pd.DataFrame(data)
```

Filling missing values using:

1.mean 2.median 3.standard deviation

In [43]:

```
dup_df['G1']=dup_df['G1'].fillna(dup_df['G1'].mean())
dup_df['G2']=dup_df['G2'].fillna(dup_df['G2'].median())
dup_df['G3']=dup_df['G3'].fillna(dup_df['G3'].std())
```

In [44]:

```
dup_df.isnull().sum
```

Out[44]:

```
<bound method NDFrame._add_numeric_operations.<locals>.sum of      school
sex    age  address  famsize  Pstatus   Medu   Fedu   Mjob  \
0     False  False  False    False    False    False  False  False  False
1     False  False  False    False    False    False  False  False  False
2     False  False  False    False    False    False  False  False  False
3     False  False  False    False    False    False  False  False  False
4     False  False  False    False    False    False  False  False  False
..      ...    ...    ...      ...      ...      ...    ...    ...    ...
390   False  False  False    False    False    False  False  False  False
391   False  False  False    False    False    False  False  False  False
392   False  False  False    False    False    False  False  False  False
393   False  False  False    False    False    False  False  False  False
394   False  False  False    False    False    False  False  False  False

      Fjob  ...  famrel  freetime  goout   Dalc   Walc  health  absences
\
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     False  False  False  False   False     False
4    False  ...   False     False  False  False  False   False     False
..     ...  ...     ...       ...    ...    ...    ...     ...       ...
390  False  ...   False     False  False  False  False   False     False
391  False  ...   False     False  False  False  False   False     False
392  False  ...   False     False  False  False  False   False     False
393  False  ...   False     False  False  False  False   False     False
394  False  ...   False     False  False  False  False   False     False

        G1     G2     G3
0    False  False  False
1    False  False  False
2    False  False  False
3    False  False  False
4    False  False  False
..     ...    ...    ...
390  False  False  False
391  False  False  False
392  False  False  False
393  False  False  False
394  False  False  False

[395 rows x 33 columns]>
```

In [18]:

```
dup_df.isnull()
```

Out[18]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	f
0	False	False	False	False	False	False	False	False	False	False	...	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	
390	False	False	False	False	False	False	False	False	False	False	...	False	
391	False	False	False	False	False	False	False	False	False	False	...	False	
392	False	False	False	False	False	False	False	False	False	False	...	False	
393	False	False	False	False	False	False	False	False	False	False	...	False	
394	False	False	False	False	False	False	False	False	False	False	...	False	

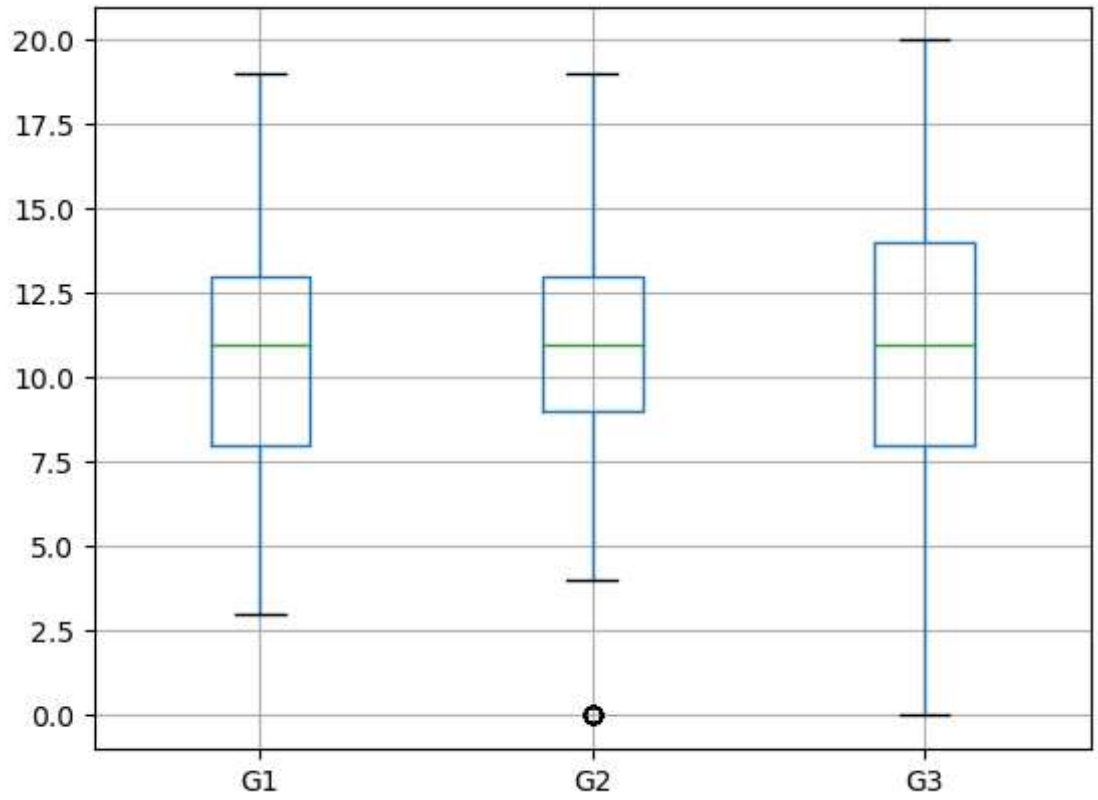
395 rows × 33 columns

In [19]:

```
#cols = ['G1','G2','G3']
data.boxplot(column=['G1', 'G2', 'G3'])
```

Out[19]:

<Axes: >



Handling with outliers

In [45]:

```
Q1 = data['G2'].quantile(0.25)
```

Type *Markdown* and LaTeX: α^2

In [46]:

```
Q3 = data['G2'].quantile(0.75)
```

In [47]:

```
IQR = Q3-Q1
```

In [48]:

```
lower_limit = Q1-1.5*IQR
```

In [49]:

```
upper_limit = Q3+1.5*IQR
```

In [50]:

```
print("Q1:",Q1,"\nQ3:",Q3,"\nIQR:",IQR,"\nlower_limit:",lower_limit,"\nupper_limit:",uppe
```

```
Q1: 9.0
Q3: 13.0
IQR: 4.0
lower_limit: 3.0
upper_limit: 19.0
```

In [51]:

```
data.boxplot(column=['G2'])
```

Out[51]:

```
<Axes: >
```

In [57]:

```
data=data[(data['G3']>lower_limit)&(data['G3']<upper_limit)]
data[60:70]
```

Out[57]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel
62	GP	F	16	U	LE3	T	1	2	other	services	...	4
63	GP	F	16	U	GT3	T	4	3	teacher	health	...	3
64	GP	F	15	U	LE3	T	4	3	services	services	...	4
65	GP	F	16	U	LE3	T	4	3	teacher	services	...	5
66	GP	M	15	U	GT3	A	4	4	other	services	...	1
67	GP	F	16	U	GT3	T	3	1	services	other	...	4
68	GP	F	15	R	LE3	T	2	2	health	services	...	4
69	GP	F	15	R	LE3	T	3	1	other	other	...	4
70	GP	M	16	U	GT3	T	3	1	other	other	...	4
71	GP	M	15	U	GT3	T	4	2	other	other	...	3

10 rows × 33 columns

In [58]:

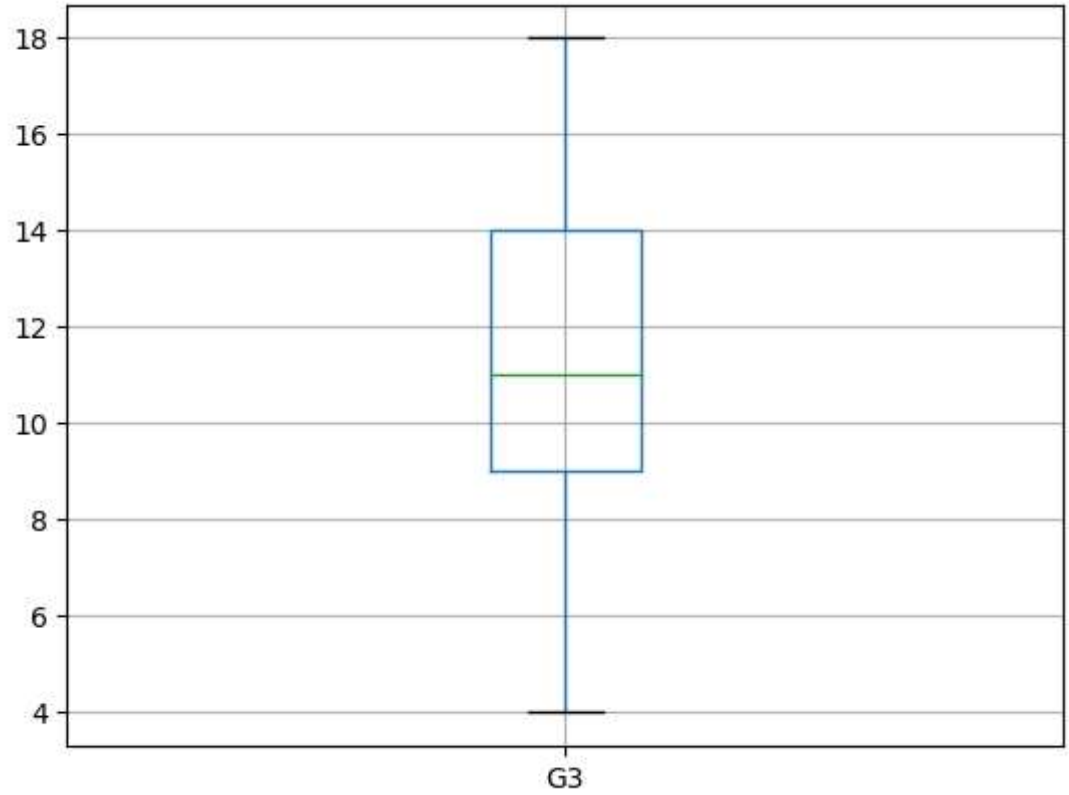
```
data.boxplot(column='G3')
```

Out[58]:

```
<Axes: >
```

In [59]:

```
plt.show()
```



In [ ]:

In [ ]:

In [ ]: