In [50]:
```
import pandas as pd
```

In [51]:
```
import numpy as np
```

In [52]:
```
import matplotlib.pyplot as plt
```

In [53]:
```
df=pd.read_csv("dataset.csv.csv")
```

In [54]:
```
print(df)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04         A
2        M          KW       KuwaIT    lowerlevel    G-04         A
3        M          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father           15                16
1           IT        F   Father           20                20
2           IT        F   Father           10                 7
3           IT        F   Father           30                25
4           IT        F   Father           40                50
..         ...      ...      ...          ...               ...
475  Chemistry        S   Father            5                 4
476    Geology        F   Father           50                77
477    Geology        S   Father           55                74
478    History        F   Father           30                17
479    History        S   Father           35                14

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 ...         ...                   ...
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       ...                ...   ...
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[480 rows x 17 columns]
```

In [55]:
```
df.isnull()
```

Out[55]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester
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
...	...	...	...	...	...	...	...	...
475	False	False	False	False	False	False	False	False
476	False	False	False	False	False	False	False	False
477	False	False	False	False	False	False	False	False
478	False	False	False	False	False	False	False	False
479	False	False	False	False	False	False	False	False

480 rows × 17 columns

In [56]:
```
df.dtypes
```

Out[56]:
```
gender                     object
NationalITy                object
PlaceofBirth               object
StageID                    object
GradeID                    object
SectionID                  object
Topic                      object
Semester                   object
Relation                   object
raisedhands                 int64
VisITedResources            int64
AnnouncementsView           int64
Discussion                  int64
ParentAnsweringSurvey      object
ParentschoolSatisfaction   object
StudentAbsenceDays         object
Class                      object
dtype: object
```

In [57]:
```
df.isnull().sum()
```

Out[57]:
```
gender                     0
NationalITy                0
PlaceofBirth               0
StageID                    0
GradeID                    0
SectionID                  0
Topic                      0
Semester                   0
Relation                   0
raisedhands                0
VisITedResources           0
AnnouncementsView          0
Discussion                 0
ParentAnsweringSurvey      0
ParentschoolSatisfaction   0
StudentAbsenceDays         0
Class                      0
dtype: int64
```

In [58]:
```
df.describe()
```

Out[58]:

	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	480.000000	480.000000	480.000000	480.000000
mean	46.775000	54.797917	37.918750	43.283333
std	30.779223	33.080007	26.611244	27.637735
min	0.000000	0.000000	0.000000	1.000000
25%	15.750000	20.000000	14.000000	20.000000
50%	50.000000	65.000000	33.000000	39.000000
75%	75.000000	84.000000	58.000000	70.000000
max	100.000000	99.000000	98.000000	99.000000

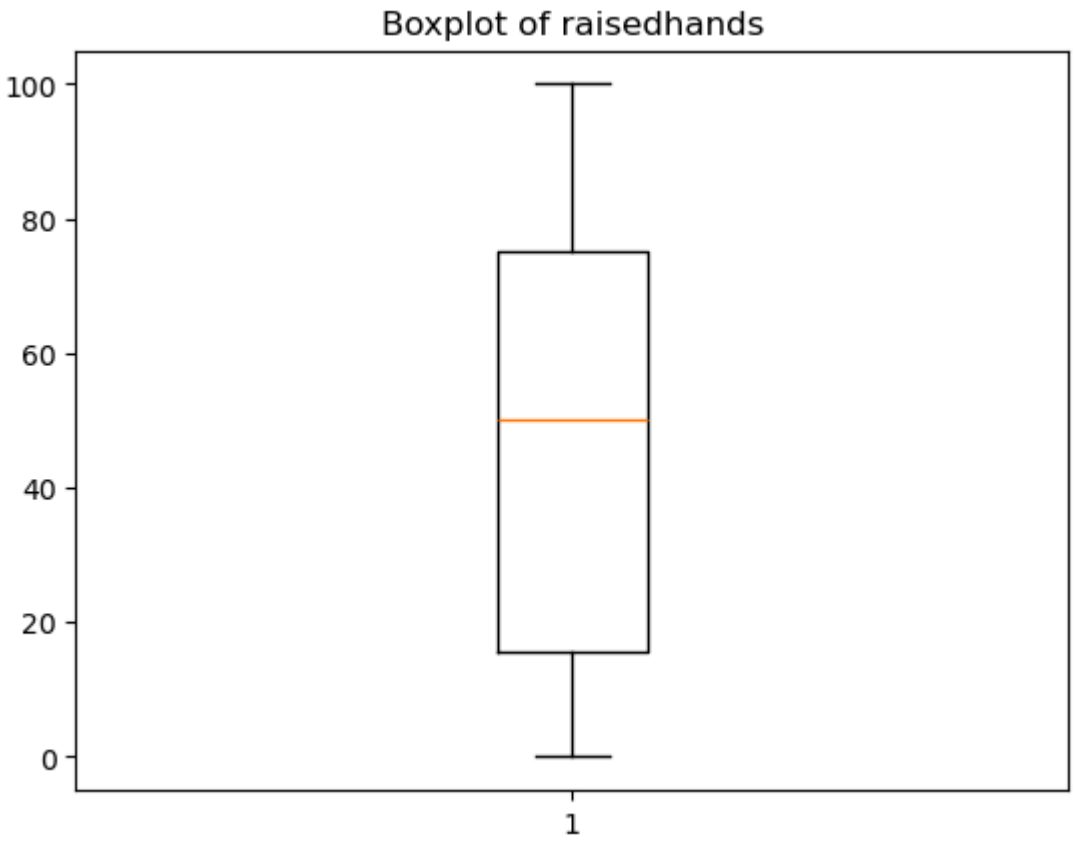
In [59]:
```
for col in numeric_cols:
    df[col] = df[col].fillna(df[col].mean())
```

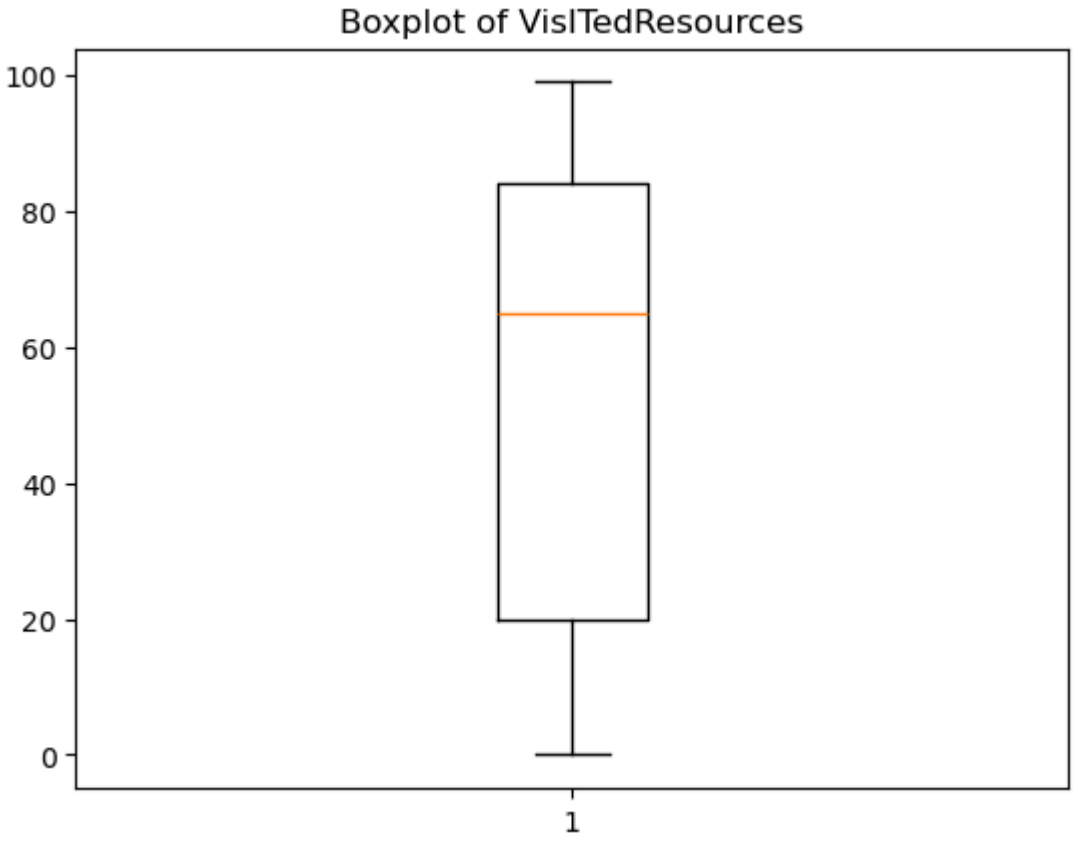
In [60]:
```
df = df[df["raisedhands"] >= 0]
df = df[df["VisITedResources"] >= 0]
df = df[df["AnnouncementsView"] >= 0]
```

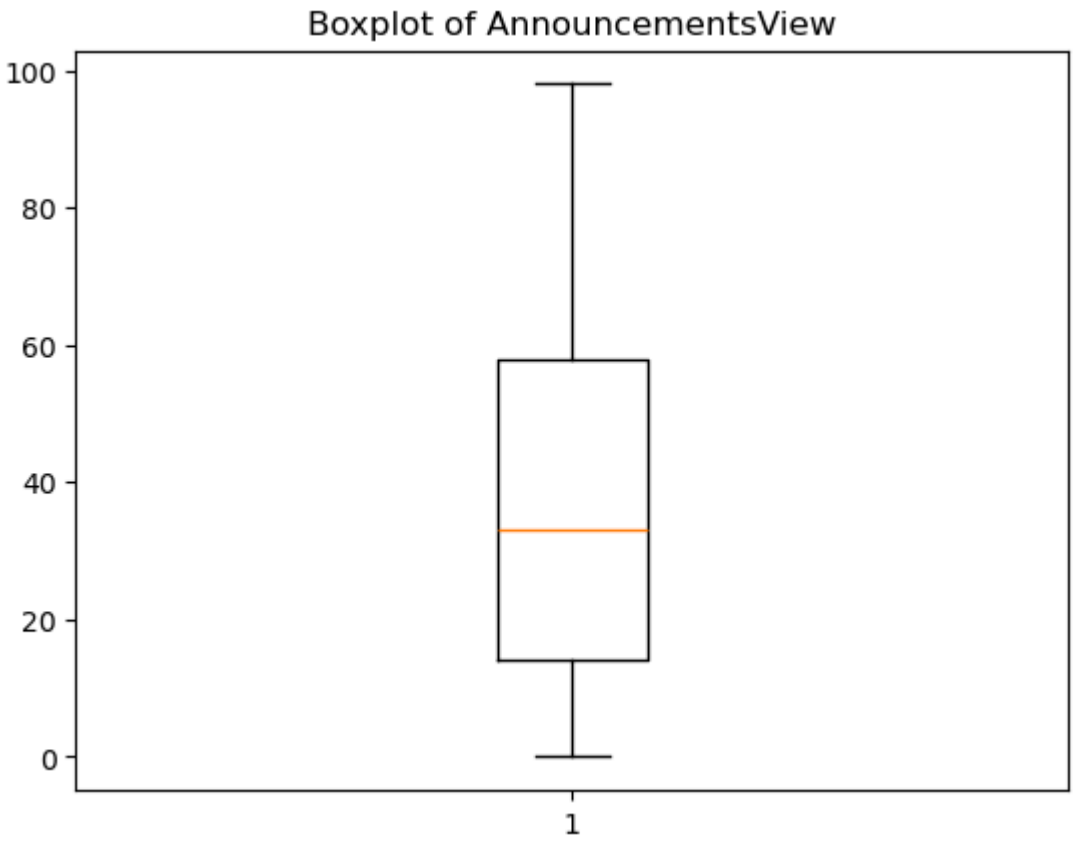
In [61]:
```
numeric_cols
```

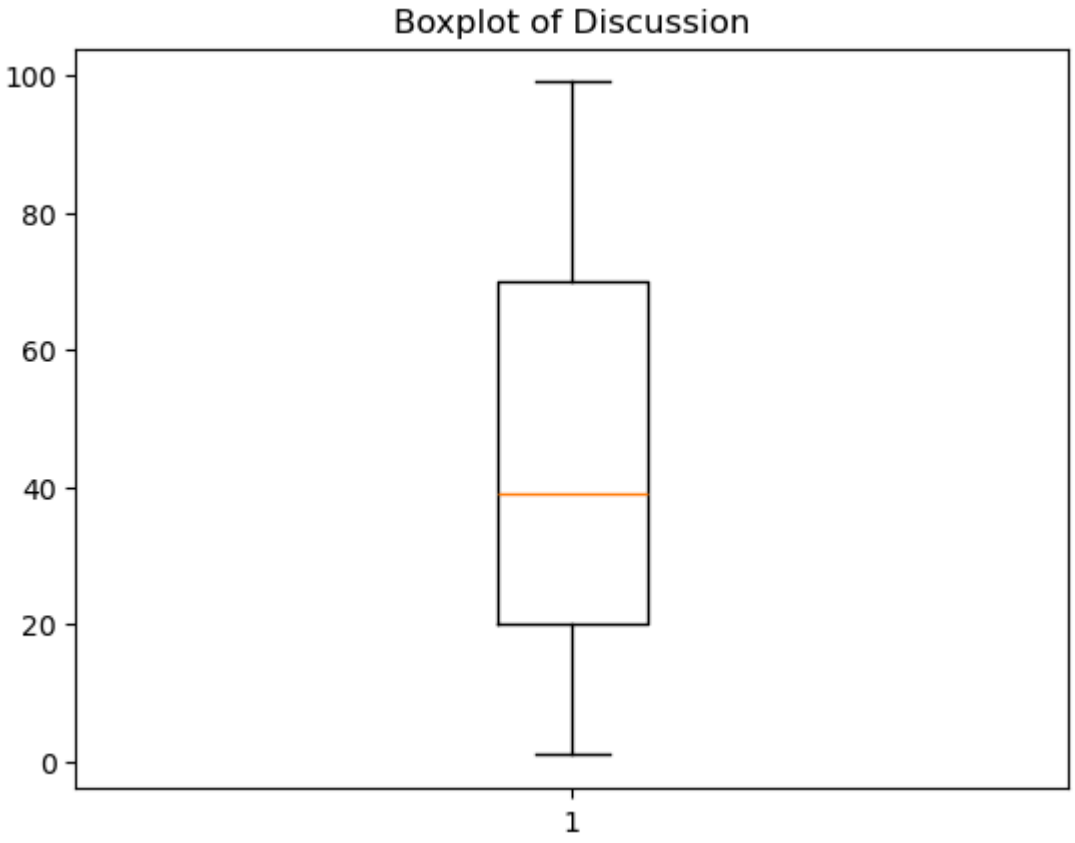
Out[61]:
```
Index(['raisedhands', 'VisITedResources', 'AnnouncementsView', 'Discussion'], d
type='object')
```

In [62]:
```
for col in numeric_cols:
    plt.boxplot(df[col])
    plt.title(f"Boxplot of {col}")
    plt.show()
```


Boxplot of raisedhands
100
80
60
40
20
0
1


Boxplot of VislTedResources
100
80
60
40
20
0
1


Boxplot of AnnouncementsView
100
80
60
40
20
0
1


Boxplot of Discussion
100
80
60
40
20
0
1

In [63]:
```
df.head()
```

Out[63]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F

In [64]:
```
df['raisedhands']=df['raisedhands'].bfill()
df['raisedhands']=df['raisedhands'].bfill()
```

In [65]:
```
df
```

Out[65]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	S
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
...	...	...	...	...	...	...	...	
475	F	Jordan	Jordan	MiddleSchool	G-08	A	Chemistry	
476	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	
477	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	
478	F	Jordan	Jordan	MiddleSchool	G-08	A	History	
479	F	Jordan	Jordan	MiddleSchool	G-08	A	History	

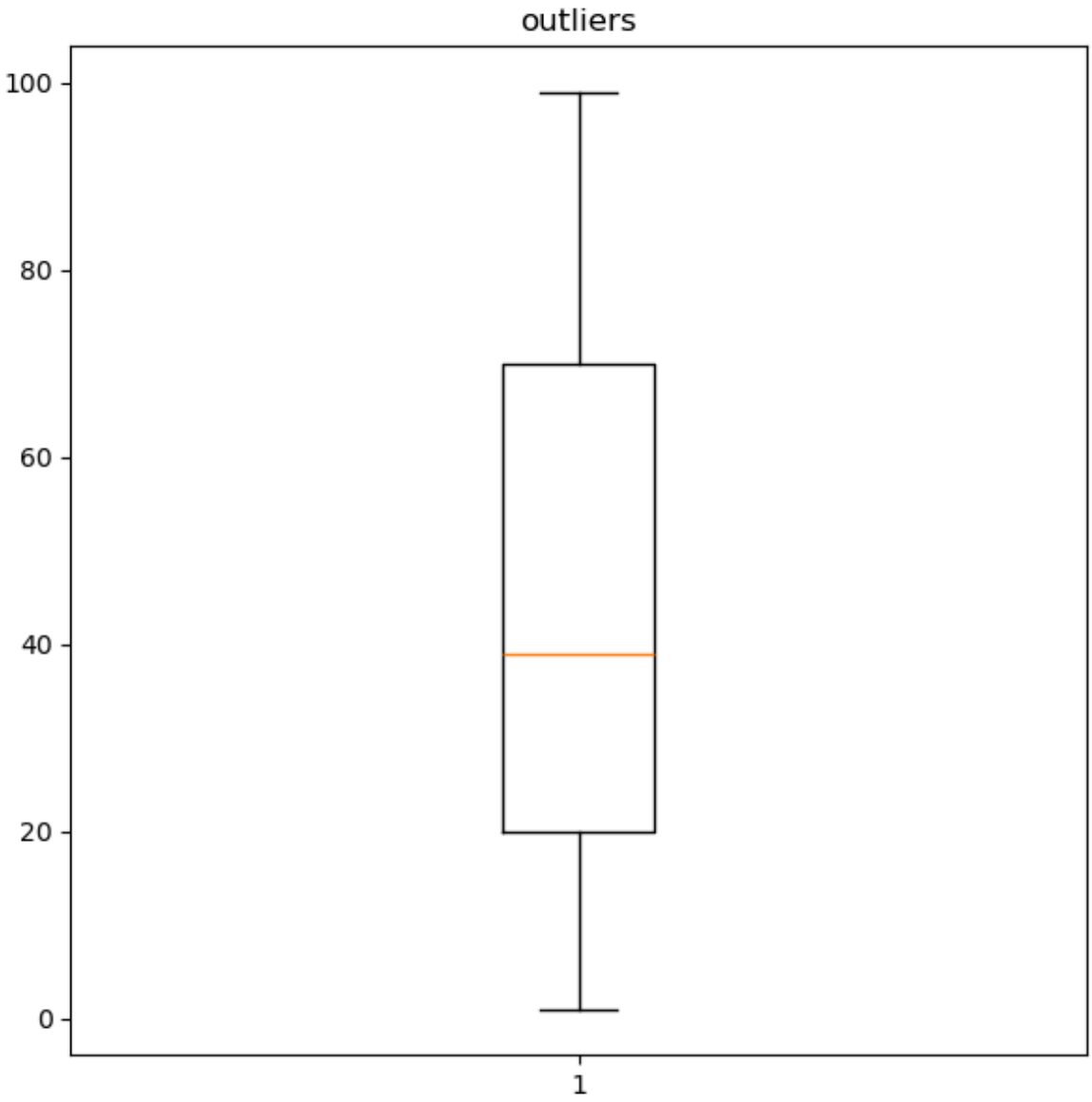
480 rows × 17 columns

In [66]:
```
#outlier
Q1=df['Discussion'].quantile(0.25)
Q3=df['Discussion'].quantile(0.75)
IQR=Q3-Q1
outliers=df[(df['Discussion']< Q1 - 1.5*IQR) | (df['Discussion']>Q3 + 1.5*IQR)]
```

In [67]:
```
print(outliers)
```

```
Empty DataFrame
Columns: [gender, NationalITy, PlaceofBirth, StageID, GradeID, SectionID, Topic,
Semester, Relation, raisedhands, VisITedResources, AnnouncementsView, Discussion,
ParentAnsweringSurvey, ParentschoolSatisfaction, StudentAbsenceDays, Class]
Index: []
```

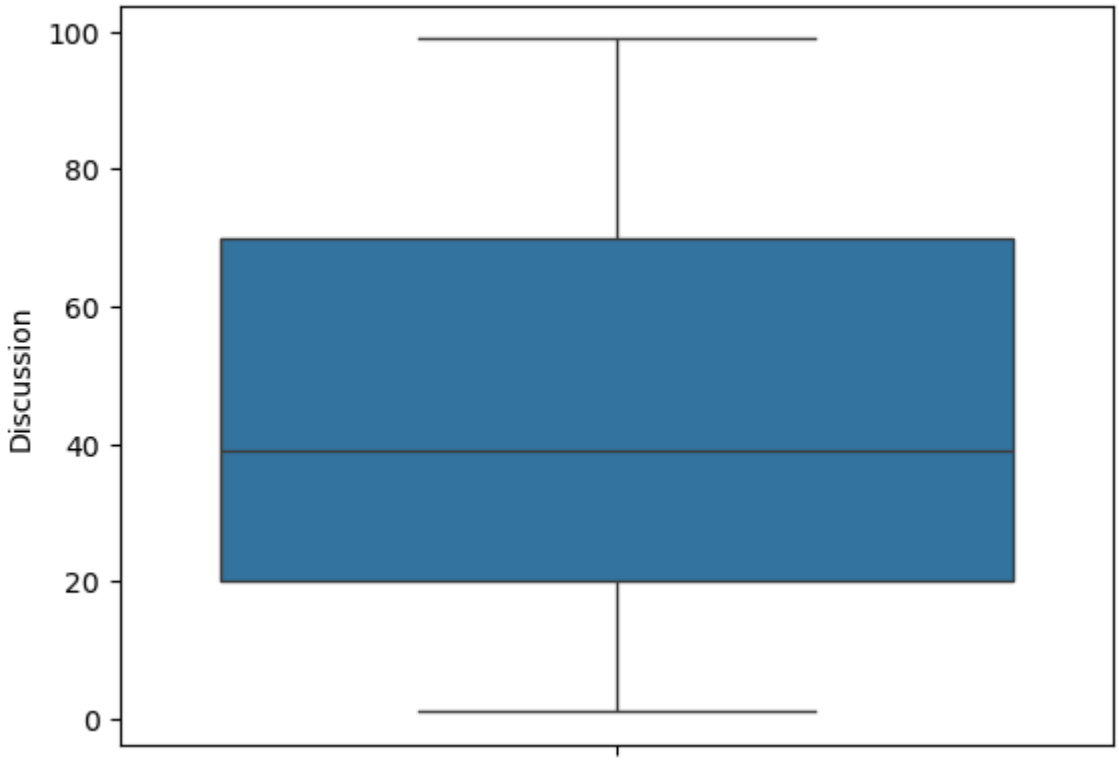
In [68]:
```
plt.figure(figsize=(7,7))
plt.boxplot(df['Discussion'])
plt.title("outliers")
plt.show()
```



In [69]:
```
import seaborn as sns
```

In [70]:
```
sns.boxplot(df['Discussion'])
```

Out[70]:
```
<Axes: ylabel='Discussion'>
```

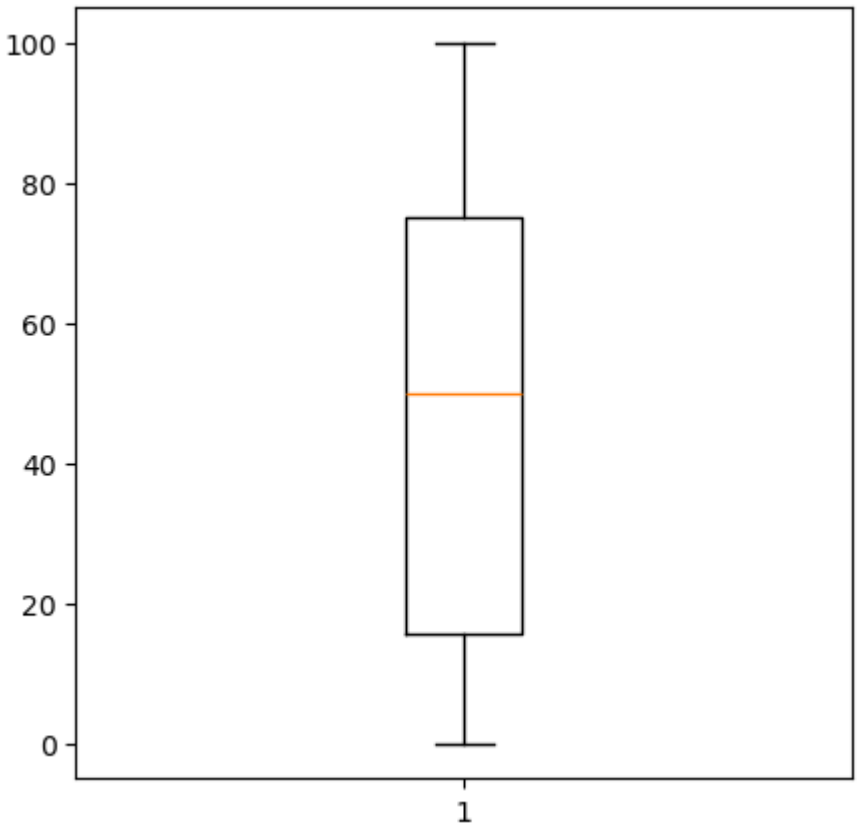


In [71]:
```
from scipy import stats
```

In [72]:
```
z_score=stats.zscore(df['raisedhands'])
z_score_outliers=df[abs(z_score)>3]
print(z_score_outliers)
```

```
Empty DataFrame
Columns: [gender, NationalITy, PlaceofBirth, StageID, GradeID, SectionID, Topic,
Semester, Relation, raisedhands, VisITedResources, AnnouncementsView, Discussion,
ParentAnsweringSurvey, ParentschoolSatisfaction, StudentAbsenceDays, Class]
Index: []
```

In [73]:
```
plt.figure(figsize=(5,5))
plt.boxplot(df['raisedhands'])
plt.show()
```



In [74]:
```
df['normalization'] = (df['VisITedResources'] - df['VisITedResources'].min()) /
```

In [75]:
```
df
```

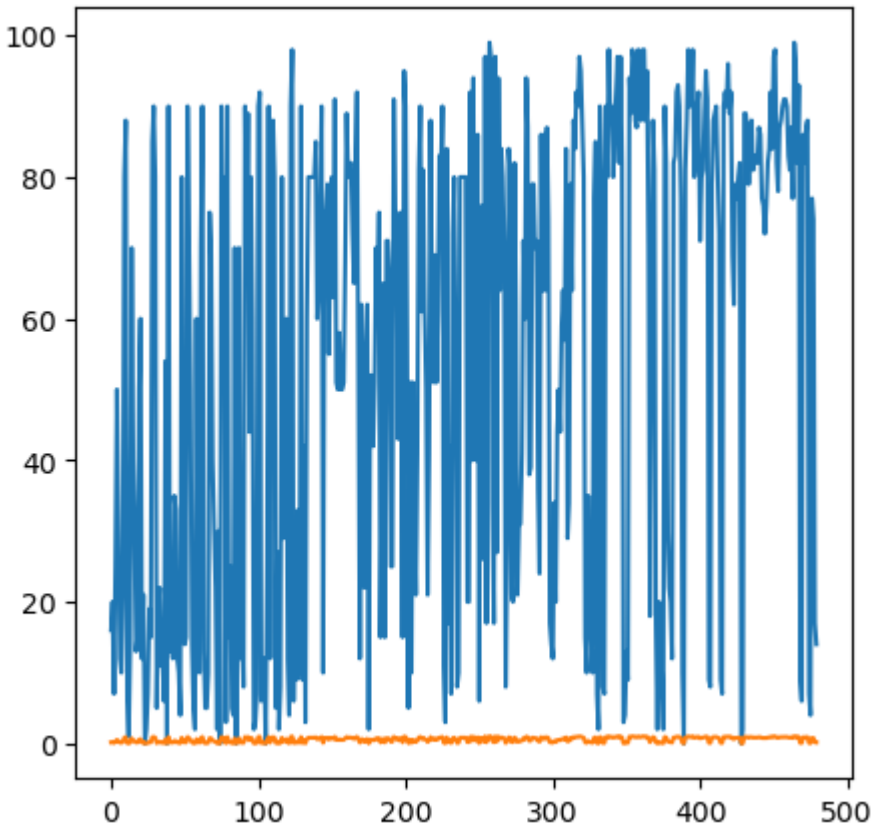
Out[75]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	S
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	
...	...	...	...	...	...	...	...	
475	F	Jordan	Jordan	MiddleSchool	G-08	A	Chemistry	
476	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	
477	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	
478	F	Jordan	Jordan	MiddleSchool	G-08	A	History	
479	F	Jordan	Jordan	MiddleSchool	G-08	A	History	

480 rows × 18 columns

In [76]:
```
plt.figure(figsize=(5,5))
plt.plot(df['VisITedResources'])
```

```
plt.plot(df['normalization'])
plt.show()
```



In [77]:
```
from sklearn.preprocessing import MaxAbsScaler
scaler=MaxAbsScaler()
df['VisITedResources']=scaler.fit_transform(df[['VisITedResources']])
print(df)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04         A
2        M          KW       KuwaIT    lowerlevel    G-04         A
3        M          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father           15          0.161616
1           IT        F   Father           20          0.202020
2           IT        F   Father           10          0.070707
3           IT        F   Father           30          0.252525
4           IT        F   Father           40          0.505051
..         ...      ...      ...          ...               ...
475  Chemistry        S   Father            5          0.040404
476    Geology        F   Father           50          0.777778
477    Geology        S   Father           55          0.747475
478    History        F   Father           30          0.171717
479    History        S   Father           35          0.141414

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 ...         ...                   ...
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class  normalization
0                       Good            Under-7     M       0.161616
1                       Good            Under-7     M       0.202020
2                        Bad            Above-7     L       0.070707
3                        Bad            Above-7     L       0.252525
4                        Bad            Above-7     M       0.505051
..                       ...                ...   ...            ...
475                      Bad            Above-7     L       0.040404
476                      Bad            Under-7     M       0.777778
477                      Bad            Under-7     M       0.747475
478                      Bad            Above-7     L       0.171717
479                      Bad            Above-7     L       0.141414

[480 rows x 18 columns]
```