In [42]:

```
import pandas as pd
```

In [43]:

```
import numpy as np
```

In [44]:

```
df=pd.read_csv("Datasets/diabetes_unclean.csv")
```

In [45]:

```
df.head(7)
```

Out[45]:

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	502	17975	F	50.0	4.7	46.0	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
1	735	34221	M	26.0	4.5	62.0	4.9	3.7	1.4	1.1	2.1	0.6	23.0	N
2	420	47975	F	50.0	4.7	46.0	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
3	680	87656	F	50.0	4.7	46.0	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
4	504	34223	M	33.0	7.1	46.0	4.9	4.9	1.0	0.8	2.0	0.4	21.0	N
5	634	34224	F	45.0	2.3	24.0	4.0	2.9	1.0	1.0	1.5	0.4	21.0	N
6	721	34225	F	50.0	2.0	50.0	4.0	3.6	1.3	0.9	2.1	0.6	24.0	N

In [46]:

```
df.tail(3)
```

Out[46]:

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
1006	193	454316	M	62.0	6.3	82.0	6.7	5.3	2.0	1.0	3.5	NaN	30.1	Y
1007	194	454316	F	57.0	4.1	70.0	9.3	5.3	3.3	1.0	1.4	1.3	29.0	Y
1008	195	4543	f	55.0	4.1	34.0	13.9	5.4	1.6	1.6	3.1	0.7	33.0	Y

In [47]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1009 entries, 0 to 1008
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   ID         1009 non-null   int64  
 1   No_Pation  1009 non-null   int64  
 2   Gender     1009 non-null   object 
 3   AGE        1008 non-null   float64
 4   Urea       1008 non-null   float64
 5   Cr         1007 non-null   float64
 6   HbA1c      1006 non-null   float64
 7   Chol       1007 non-null   float64
 8   TG         1007 non-null   float64
 9   HDL        1008 non-null   float64
 10  LDL        1007 non-null   float64
 11  VLDL       1008 non-null   float64
 12  BMI        1009 non-null   float64
 13  CLASS      1009 non-null   object 
dtypes: float64(10), int64(2), object(2)
memory usage: 110.5+ KB
```

In [48]:

```
df.describe(exclude=np.number)
```

Out[48]:

	Gender	CLASS
count	1009	1009
unique	3	5
top	M	Y
freq	570	840

In [49]:

```python
df.describe(include="all")
```

Out[49]:

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL
count	1009.000000	1.009000e+03	1009	1008.000000	1008.000000	1007.000000	1006.000000	1007.000000	1007.000000	1008.000000	1007.000000
unique	NaN	NaN	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	M	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	570	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	339.161546	2.717448e+05	NaN	53.620040	5.131094	68.973188	8.284155	4.863873	2.348769	1.204216	2.610119
std	239.738169	3.365681e+06	NaN	8.740975	2.931136	59.813297	2.533576	1.297326	1.397487	0.658158	1.116095
min	1.000000	1.230000e+02	NaN	25.000000	0.500000	6.000000	0.900000	0.000000	0.300000	0.200000	0.300000
25%	127.000000	2.406500e+04	NaN	51.000000	3.700000	48.000000	6.500000	4.000000	1.500000	0.900000	1.800000
50%	296.000000	3.439900e+04	NaN	55.000000	4.600000	60.000000	8.000000	4.800000	2.000000	1.100000	2.500000
75%	548.000000	4.539000e+04	NaN	59.000000	5.700000	73.000000	10.200000	5.600000	2.900000	1.300000	3.300000
max	800.000000	7.543566e+07	NaN	79.000000	38.900000	800.000000	16.000000	10.300000	13.800000	9.900000	9.900000

In [50]:

```python
df.isna().sum()
```

Out[50]:

```
ID           0
No_Pation    0
Gender       0
AGE          1
Urea         1
Cr           2
HbA1c        3
Chol         2
TG           2
HDL          1
LDL          2
VLDL         1
BMI          0
CLASS        0
dtype: int64
```

In [51]:

```
df['Cr']=df.Cr.fillna(df.Cr.median())
```

In [52]:

```
df['HDL']=df.Cr.fillna(df.HDL.mode()[0])
```

In [68]:

```
df['AGE']=df.Cr.fillna(df.AGE.mean())
```

In [54]:

```
df.isna().sum()
```

Out[54]:

```
ID            0
No_Pation     0
Gender        0
AGE           0
Urea          1
Cr            0
HbA1c         3
Chol          2
TG            2
HDL           0
LDL           2
VLDL          1
BMI           0
CLASS         0
dtype: int64
```

In [55]:

```
df=df.dropna()
```

In [56]:

```
df.isna().sum()
```

Out[56]:

```
ID            0
No_Pation     0
Gender        0
AGE           0
Urea          0
Cr            0
HbA1c         0
Chol          0
TG            0
HDL           0
LDL           0
VLDL          0
BMI           0
CLASS         0
dtype: int64
```

In [57]:

```
df.duplicated().sum()
```

Out[57]:

```
0
```

In [58]:

```
df.Gender.unique()
```

Out[58]:

```
array(['F', 'M', 'f'], dtype=object)
```

In [59]:

```
df['Gender']=df['Gender'].replace('f','F')
```

In [60]:

```
df.Gender.unique()
```

Out[60]:

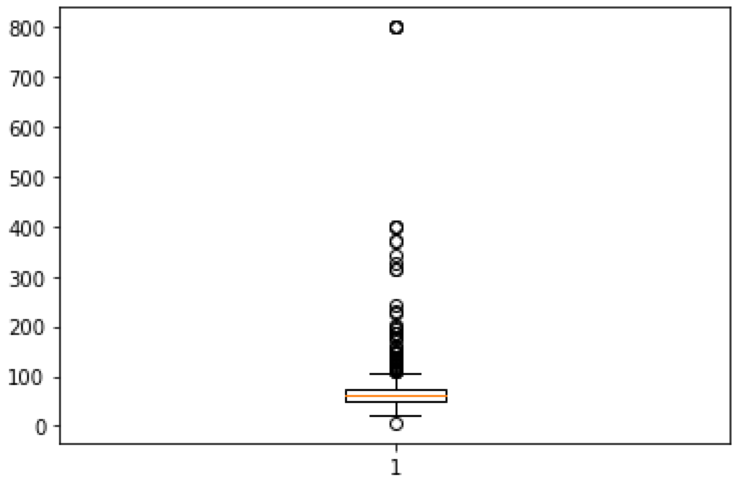
```
array(['F', 'M'], dtype=object)
```

In [61]:

```
import matplotlib.pyplot as plt
```

In [62]:

```
plt.boxplot(df.Cr)
plt.show()
```

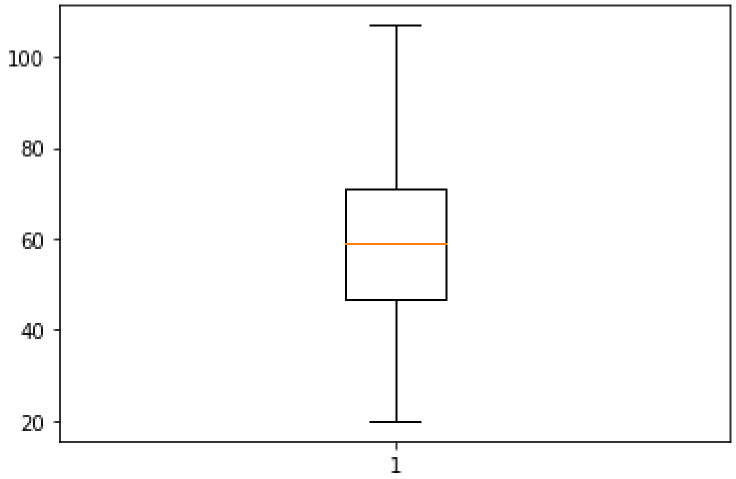


In [63]:

```
#remove outlier in Cr column
q1=df['Cr'].quantile(0.25)
q3=df['Cr'].quantile(0.75)
iqr=q3-q1
ul=q3+(1.5)*iqr
ll=q1-(1.5)*iqr
df=df[(df['Cr']>=ll) & (df['Cr']<=ul)]
```

In [64]:

```
plt.boxplot(df.Cr)
plt.show()
```



In [65]:

```
# find corrilation in given dataframe
df.corr()
```

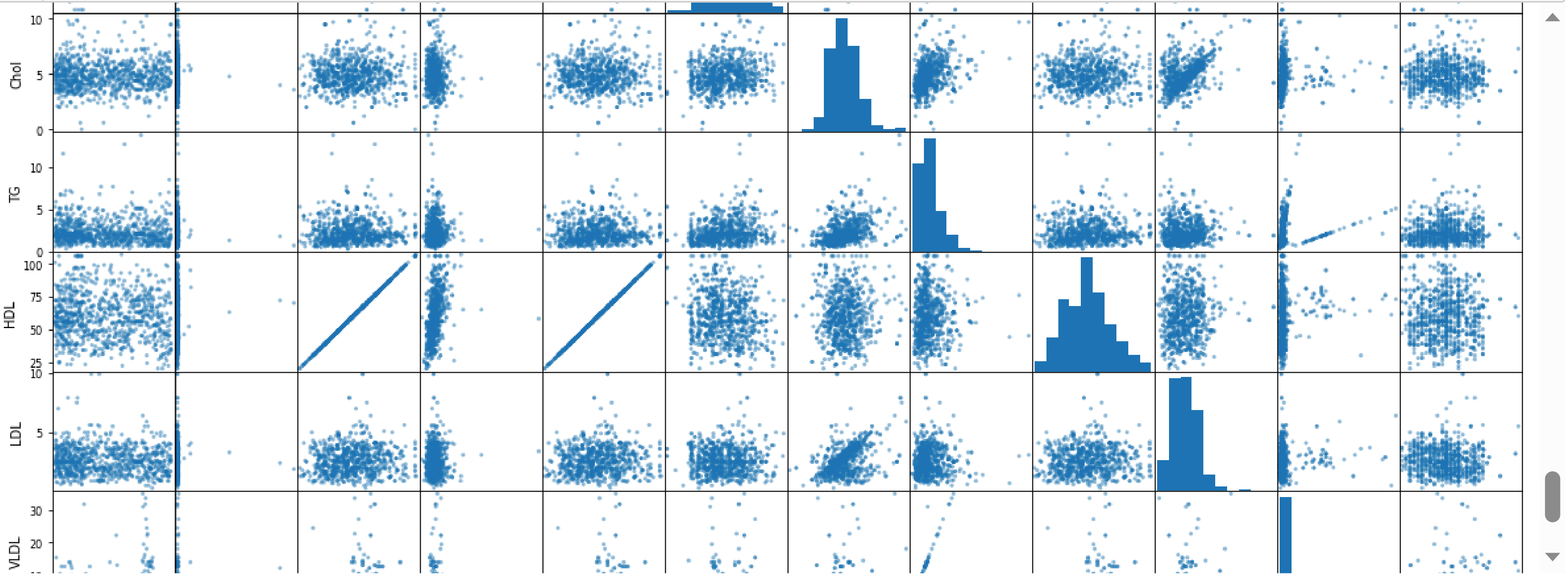
Out[65]:

	ID	No_Pation	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
ID	1.000000	0.065089	-0.069225	-0.038072	-0.069225	-0.016719	0.043226	-0.040850	-0.069225	-0.055461	0.142097	0.041155
No_Pation	0.065089	1.000000	0.040006	-0.014675	0.040006	-0.032952	-0.030948	-0.040056	0.040006	-0.003560	0.113511	0.018335
AGE	-0.069225	0.040006	1.000000	0.394575	1.000000	-0.132797	-0.018307	0.018130	1.000000	0.076335	0.121891	-0.011111
Urea	-0.038072	-0.014675	0.394575	1.000000	0.394575	-0.020306	0.022223	0.018001	0.394575	-0.003328	0.017614	0.034809
Cr	-0.069225	0.040006	1.000000	0.394575	1.000000	-0.132797	-0.018307	0.018130	1.000000	0.076335	0.121891	-0.011111
HbA1c	-0.016719	-0.032952	-0.132797	-0.020306	-0.132797	1.000000	0.168250	0.225676	-0.132797	0.014643	0.069974	0.414565
Chol	0.043226	-0.030948	-0.018307	0.022223	-0.018307	0.168250	1.000000	0.328054	-0.018307	0.423856	0.072181	0.018462
TG	-0.040850	-0.040056	0.018130	0.018001	0.018130	0.225676	0.328054	1.000000	0.018130	0.002472	0.150595	0.100708
HDL	-0.069225	0.040006	1.000000	0.394575	1.000000	-0.132797	-0.018307	0.018130	1.000000	0.076335	0.121891	-0.011111
LDL	-0.055461	-0.003560	0.076335	-0.003328	0.076335	0.014643	0.423856	0.002472	0.076335	1.000000	0.064721	-0.058008
VLDL	0.142097	0.113511	0.121891	0.017614	0.121891	0.069974	0.072181	0.150595	0.121891	0.064721	1.000000	0.203209
BMI	0.041155	0.018335	-0.011111	0.034809	-0.011111	0.414565	0.018462	0.100708	-0.011111	-0.058008	0.203209	1.000000

In [66]:

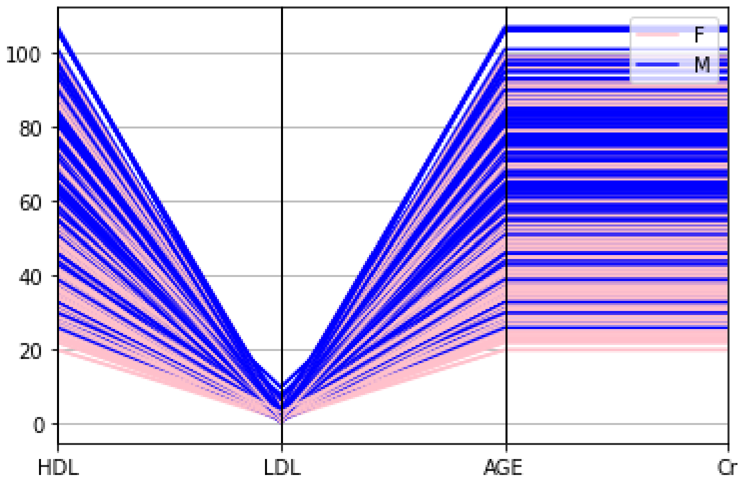
```
pd.plotting.scatter_matrix(df,figsize=(20,20))

```



In [69]:

```
pd.plotting.parallel_coordinates(df,'Gender',cols=['HDL','LDL','AGE','Cr'],color=['pink','blue'])
plt.show()
```



In [70]:

```
pd.crosstab(df.Gender,df.CLASS)
```

Out[70]:

CLASS	N	N	P	Y	Y
Gender					
F	60	0	17	334	3
M	36	1	35	454	6

In [ ]:

```

```