In [1]:

```python
import pandas as pd
df=pd.read_csv("uber.csv")
df
```

Out[1]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pick
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	
...	...	...	...	...	...	
199995	42598914	2012-10-28 10:49:00.00000053	3.0	2012-10-28 10:49:00 UTC	-73.987042	
199996	16382965	2014-03-14 01:09:00.0000008	7.5	2014-03-14 01:09:00 UTC	-73.984722	
199997	27804658	2009-06-29 00:42:00.00000078	30.9	2009-06-29 00:42:00 UTC	-73.986017	
199998	20259894	2015-05-20 14:56:25.0000004	14.5	2015-05-20 14:56:25 UTC	-73.997124	
199999	11951496	2010-05-15 04:08:00.00000076	14.1	2010-05-15 04:08:00 UTC	-73.984395	

200000 rows × 9 columns

In [2]:

```python
df.head()
```

Out[2]:

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_lat
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.73
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.72
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.74
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.79
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.74

In [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 9 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   Unnamed: 0         200000 non-null  int64  
 1   key                200000 non-null  object 
 2   fare_amount        200000 non-null  float64
 3   pickup_datetime    200000 non-null  object 
 4   pickup_longitude   200000 non-null  float64
 5   pickup_latitude    200000 non-null  float64
 6   dropoff_longitude  199999 non-null  float64
 7   dropoff_latitude   199999 non-null  float64
 8   passenger_count    200000 non-null  int64  
dtypes: float64(5), int64(2), object(2)
memory usage: 13.7+ MB
```

In [4]:

```
#preprocessing
df.isna().sum()
```

Out[4]:

```
Unnamed: 0           0
key                  0
fare_amount          0
pickup_datetime      0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    1
dropoff_latitude     1
passenger_count      0
dtype: int64
```

In [5]:

```
df.shape
```

Out[5]:

```
(200000, 9)
```

In [6]:

```
df1=df.drop(["Unnamed: 0","key","pickup_datetime"], axis=1)
```

In [7]:

```
df1.dropna(inplace=True)
```

In [8]:

```
df1.isna().sum()
```

Out[8]:

```
fare_amount          0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    0
dropoff_latitude     0
passenger_count      0
dtype: int64
```

In [9]:

```
#correlation
df1.corr()
```

Out[9]:

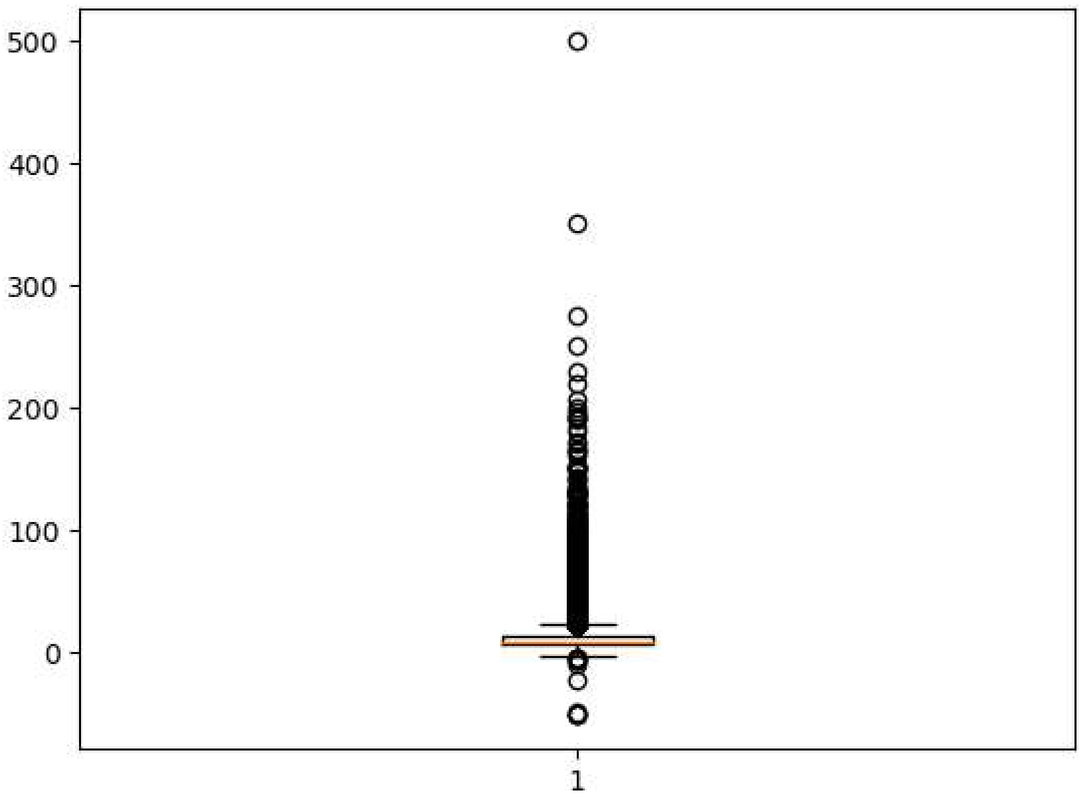
	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_
fare_amount	1.000000	0.010458	-0.008482	0.008986	-C
pickup_longitude	0.010458	1.000000	-0.816461	0.833026	-0
pickup_latitude	-0.008482	-0.816461	1.000000	-0.774787	0
dropoff_longitude	0.008986	0.833026	-0.774787	1.000000	-0
dropoff_latitude	-0.011014	-0.846324	0.702367	-0.917010	1
passenger_count	0.010158	-0.000415	-0.001559	0.000033	-0

In [10]:

```
import matplotlib.pyplot as plt
plt.boxplot(df1['fare_amount'])
```

Out[10]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x213633ad0d0>,
  <matplotlib.lines.Line2D at 0x213633ad3a0>],
 'caps': [<matplotlib.lines.Line2D at 0x213633ad670>,
  <matplotlib.lines.Line2D at 0x213633ad940>],
 'boxes': [<matplotlib.lines.Line2D at 0x2136338fdc0>],
 'medians': [<matplotlib.lines.Line2D at 0x213633adc10>],
 'fliers': [<matplotlib.lines.Line2D at 0x213633adee0>],
 'means': []}
```



In [11]:
```
import numpy as np
def removeoutlier(data):
    Q1=np.percentile(data,25)
    Q2=np.percentile(data,50)
    Q3=np.percentile(data,75)
    IQR=Q3-Q1
    lb=Q1-1.5*IQR
    ub=Q3+1.5*IQR
    return(lb,ub)
```

In [12]:
```
lower_bound,upper_boud=removeoutlier(df1["fare_amount"])
```

In [13]:
```
lower_bound,upper_boud
```

Out[13]:
```
(-3.75, 22.25)
```

In [14]:
```
df1=df1[(df1.fare_amount>=lower_bound) & (df1.fare_amount<=upper_boud)]
```

In [15]:
```
plt.boxplot(df1["fare_amount"])
```

Out[15]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x21361759cd0>,
  <matplotlib.lines.Line2D at 0x21361759f40>],
 'caps': [<matplotlib.lines.Line2D at 0x2136176a250>,
  <matplotlib.lines.Line2D at 0x2136176a550>],
 'boxes': [<matplotlib.lines.Line2D at 0x21361759a00>],
 'medians': [<matplotlib.lines.Line2D at 0x2136176a820>],
 'fliers': [<matplotlib.lines.Line2D at 0x2136176aaf0>],
 'means': []}
```