In [1]:

```
import pandas as pd
data = {'cars': ["BMW", "Volvo", "Ford"],
       'passings': [3, 7, 2]}

df = pd.DataFrame(data)
print(df)
```

```
    cars  passings
0    BMW         3
1  Volvo         7
2   Ford         2
```

In [3]:

```
print(pd.__version__)
```

```
1.2.4
```

Series

In [5]:

```
import pandas as pd
a = [1,7,2]
df1 = pd.Series(a)
print(df1)
```

```
0    1
1    7
2    2
dtype: int64
```

In [14]:

```
print(df1[0])
```

```
1
```

In [8]:

```
import pandas as pd
a = [1,7,2]
df1 = pd.Series(a, index = ["x", "y", "z"])
print(df1)
```

```
x    1
y    7
z    2
dtype: int64
```

In [9]:

```
print(df1["y"])
```

```
7
```

In [12]:

```
import pandas as pd
students ={"hemant": 500, "hitesh": 800, "aniket": 900}
df2 = pd.Series(students)
print(df2)
```

```
hemant    500
hitesh    800
aniket    900
dtype: int64
```

In [13]:

```
import pandas as pd
students ={"hemant": 500, "hitesh": 800, "aniket": 900}
df2 = pd.Series(students, index = ["hemant", "hitesh"])
print(df2)
```

```
hemant    500
hitesh    800
dtype: int64
```

DataFrame

In [64]:

```
import pandas as pd
data = {"hemant":[500,600,7000],
        "shawn":[700,800,900]}
df = pd.DataFrame(data)
print(df)
```

```
   hemant  shawn
0     500    700
1     600    800
2    7000    900
```

In [65]:

```
print(df.loc[0])
```

```
hemant    500
shawn     700
Name: 0, dtype: int64
```

In [66]:

```
print(df.loc[[0,1]])
```

```
   hemant  shawn
0     500    700
1     600    800
```

```
In [46]:
```

```python
import pandas as pd
data = {"hemant":[500,600,7000],
        "shawn":[700,800,900]}
df = pd.DataFrame(data, index = ["maths","history","scince"])
print(df)
```

```
         hemant  shawn
maths       500    700
history     600    800
scince     7000    900
```

```
In [48]:
```

```python
print(df.loc["history"])
```

```
hemant    600
shawn     800
Name: history, dtype: int64
```

Read CSV

```
In [49]:
```

```python
import pandas as pd
df = pd.read_csv('data.csv')
print(df.to_string())
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112       NaN
```

Tip: use to_string() to print the entire DataFrame.

In [50]:

```
import pandas as pd
df = pd.read_csv('data.csv')
print(df)
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
..        ...    ...       ...       ...
164        60    105       140     290.8
165        60    110       145     300.0
166        60    115       145     310.2
167        75    120       150     320.4
168        75    125       150     330.4

[169 rows x 4 columns]
```

In [51]:

```
print(pd.options.display.max_rows)
```

```
60
```

In [53]:

```
import pandas as pd
pd.options.display.max_rows = 9999
df = pd.read_csv('data.csv')
print(df)
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112       NaN
```

Read JSON

In [2]:

```
import pandas as pd
df = pd.read_json('data.json')
print(df.to_string())
```

```
    Duration  Pulse  Maxpulse  Calories
0         60    110       130     409.1
1         60    117       145     479.0
2         60    103       135     340.0
3         45    109       175     282.4
4         45    117       148     406.0
5         60    102       127     300.5
6         60    110       136     374.0
7         45    104       134     253.3
8         30    109       133     195.1
9         60     98       124     269.0
10        60    103       147     329.3
11        60    100       120     250.7
12        60    106       128     345.3
13        60    104       132     379.3
14        60     98       123     275.0
15        60     98       120     215.2
16        60    100       120     300.0
17        45     90       112       NaN
```

In [4]:

```
import pandas as pd

data = {
  "Duration":{
    "0":60,
    "1":60,
    "2":60,
    "3":45,
    "4":45,
    "5":60
  },
  "Pulse":{
    "0":110,
    "1":117,
    "2":103,
    "3":109,
    "4":117,
    "5":102
  },
  "Maxpulse":{
    "0":130,
    "1":145,
    "2":135,
    "3":175,
    "4":148,
    "5":127
  },
  "Calories":{
    "0":409,
    "1":479,
    "2":340,
    "3":282,
    "4":406,
    "5":300
  }
}

df = pd.DataFrame(data)

print(df)
```

```
   Duration  Pulse  Maxpulse  Calories
0        60    110       130       409
1        60    117       145       479
2        60    103       135       340
3        45    109       175       282
4        45    117       148       406
5        60    102       127       300
```

Analyze Data

In [5]:

```
import pandas as pd
df = pd.read_csv('data.csv')
print(df.head(10))
```

```
   Duration  Pulse  Maxpulse  Calories
0        60    110       130     409.1
1        60    117       145     479.0
2        60    103       135     340.0
3        45    109       175     282.4
4        45    117       148     406.0
5        60    102       127     300.0
6        60    110       136     374.0
7        45    104       134     253.3
8        30    109       133     195.1
9        60     98       124     269.0
```

In [6]:

```
print(df.head())
```

```
   Duration  Pulse  Maxpulse  Calories
0        60    110       130     409.1
1        60    117       145     479.0
2        60    103       135     340.0
3        45    109       175     282.4
4        45    117       148     406.0
```

In [9]:

```
print(df.tail())
```

```
     Duration  Pulse  Maxpulse  Calories
164        60    105       140     290.8
165        60    110       145     300.0
166        60    115       145     310.2
167        75    120       150     320.4
168        75    125       150     330.4
```

In [10]:

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 169 entries, 0 to 168
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Duration  169 non-null    int64  
 1   Pulse     169 non-null    int64  
 2   Maxpulse  169 non-null    int64  
 3   Calories  164 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 5.4 KB
None
```

Cleaning Empty cells

In [18]:

```
import pandas as pd
df = pd.read_csv('data.csv')
new_df = df.dropna()
print(df.info())
print(new_df.info())
#print(new_df.to_string())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 169 entries, 0 to 168
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Duration  169 non-null    int64  
 1   Pulse     169 non-null    int64  
 2   Maxpulse  169 non-null    int64  
 3   Calories  164 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 5.4 KB
None
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 168
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Duration  164 non-null    int64  
 1   Pulse     164 non-null    int64  
 2   Maxpulse  164 non-null    int64  
 3   Calories  164 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 6.4 KB
None
```

In [19]:

```
import pandas as pd
df = pd.read_csv('data.csv')
df.dropna(inplace = True)
print(df.info())
#print(df.to_string())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 168
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Duration  164 non-null    int64  
 1   Pulse     164 non-null    int64  
 2   Maxpulse  164 non-null    int64  
 3   Calories  164 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 6.4 KB
None
```

In [25]:

```
import pandas as pd
df = pd.read_csv('data.csv')
df.fillna("hemant", inplace =True)
print(df.to_string())
```

```
    Duration  Pulse  Maxpulse Calories
0         60    110       130    409.1
1         60    117       145    479.0
2         60    103       135    340.0
3         45    109       175    282.4
4         45    117       148    406.0
5         60    102       127    300.0
6         60    110       136    374.0
7         45    104       134    253.3
8         30    109       133    195.1
9         60     98       124    269.0
10        60    103       147    329.3
11        60    100       120    250.7
12        60    106       128    345.3
13        60    104       132    379.3
14        60     98       123    275.0
15        60     98       120    215.2
16        60    100       120    300.0
17        45     90       112   hemant
```

In [31]:

```
import pandas as pd
df = pd.read_csv('data.csv')
x = df["Calories"].mean()
df["Calories"].fillna(x, inplace = True)
print('mean = ', x)
print(df.to_string())
```

```
mean =  375.79024390243916
    Duration  Pulse  Maxpulse    Calories
0         60    110       130  409.100000
1         60    117       145  479.000000
2         60    103       135  340.000000
3         45    109       175  282.400000
4         45    117       148  406.000000
5         60    102       127  300.000000
6         60    110       136  374.000000
7         45    104       134  253.300000
8         30    109       133  195.100000
9         60     98       124  269.000000
10        60    103       147  329.300000
11        60    100       120  250.700000
12        60    106       128  345.300000
13        60    104       132  379.300000
14        60     98       123  275.000000
15        60     98       120  215.200000
16        60    100       120  300.000000
```

In [36]:

```
import pandas as pd
df = pd.read_csv('data.csv')
x = df["Calories"].median()
df["Calories"].fillna(x, inplace = True)
print('median = ', x)
print(df.to_string())
```

```
median =  318.6
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112     318.6
```

In [38]:

```
import pandas as pd
df = pd.read_csv('data.csv')
x = df["Calories"].mode()[0]
df["Calories"].fillna(x, inplace = True)
print('mode = ', x)
print(df.to_string())
```

```
mode =  300.0
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112     300.0
```

Clean wrong format

In [44]:

```python
import pandas as pd
df = pd.read_csv('data.csv')
df['Calories'] = pd.to_numeric(df['Calories']) #pd.to_datetime()
print(df.to_string())
```

```
    Duration  Pulse  Maxpulse  Calories
0         60    110       130     409.1
1         60    117       145     479.0
2         60    103       135     340.0
3         45    109       175     282.4
4         45    117       148     406.0
5         60    102       127     300.0
6         60    110       136     374.0
7         45    104       134     253.3
8         30    109       133     195.1
9         60     98       124     269.0
10        60    103       147     329.3
11        60    100       120     250.7
12        60    106       128     345.3
13        60    104       132     379.3
14        60     98       123     275.0
15        60     98       120     215.2
16        60    100       120     300.0
17        45     90       112       NaN
```

In [46]:

```python
df.dropna(subset=["Calories"], inplace = True)
print(df.to_string())
```

```
    Duration  Pulse  Maxpulse  Calories
0         60    110       130     409.1
1         60    117       145     479.0
2         60    103       135     340.0
3         45    109       175     282.4
4         45    117       148     406.0
5         60    102       127     300.0
6         60    110       136     374.0
7         45    104       134     253.3
8         30    109       133     195.1
9         60     98       124     269.0
10        60    103       147     329.3
11        60    100       120     250.7
12        60    106       128     345.3
13        60    104       132     379.3
14        60     98       123     275.0
15        60     98       120     215.2
16        60    100       120     300.0
18        60    103       123     323.0
```

Clean Wrong Data

In [57]:

```
import pandas as pd
df = pd.read_csv('data.csv')
for i in range(0,len(df["Duration"])):
    if(df.loc[i, "Duration"] > 60):
        df.loc[i, "Duration"] = 59
    else:
        continue
print(df.to_string())
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112       NaN
```

In [58]:

```
import pandas as pd
df = pd.read_csv('data.csv')
for i in df.index:
    if(df.loc[i, "Duration"] > 60):
        df.drop(i, inplace = True)
print(df.to_string())
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112       NaN
```

Removing Duplicates

In [77]:

```
import pandas as pd
df = pd.read_csv('data.csv')
print(df.duplicated())
```

```
0      False
1      False
2      False
3      False
4      False
       ...  
164    False
165    False
166    False
167    False
168    False
Length: 169, dtype: bool
```

In [80]:

```
import pandas as pd
df = pd.read_csv('data.csv')
df.drop_duplicates(inplace = True)
print(df.to_string())
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
5          60    102       127     300.0
6          60    110       136     374.0
7          45    104       134     253.3
8          30    109       133     195.1
9          60     98       124     269.0
10         60    103       147     329.3
11         60    100       120     250.7
12         60    106       128     345.3
13         60    104       132     379.3
14         60     98       123     275.0
15         60     98       120     215.2
16         60    100       120     300.0
17         45     90       112       NaN
```

Correlation

Result Explained The Result of the corr() method is a table with a lot of numbers that represents how well the relationship is between two columns.

The number varies from -1 to 1.

1 means that there is a 1 to 1 relationship (a perfect correlation), and for this data set, each time a value went up in the first column, the other one went up as well.

0.9 is also a good relationship, and if you increase one value, the other will probably increase as well.

-0.9 would be just as good relationship as 0.9, but if you increase one value, the other will probably go down.

0.2 means NOT a good relationship, meaning that if one value goes up does not mean that the other will.

What is a good correlation? It depends on the use, but I think it is safe to say you have to have at least 0.6 (or -0.6) to call it a good correlation.

Perfect Correlation: We can see that "Duration" and "Duration" got the number 1.000000, which makes sense, each column always has a perfect relationship with itself.

Good Correlation: "Duration" and "Calories" got a 0.922721 correlation, which is a very good correlation, and we can predict that the longer you work out, the more calories you burn, and the other way around: if you burned a lot of calories, you probably had a long work out.

Bad Correlation: "Duration" and "Maxpulse" got a 0.009403 correlation, which is a very bad correlation, meaning that we can not predict the max pulse by just looking at the duration of the work out, and vice versa.

In [82]:

```
import pandas as pd
df = pd.read_csv('data.csv')
df.corr()
```

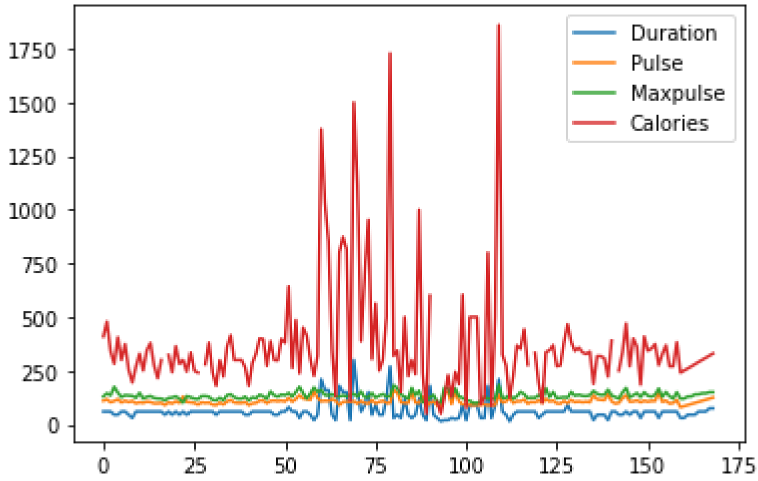
Out[82]:

	Duration	Pulse	Maxpulse	Calories
Duration	1.000000	-0.155408	0.009403	0.922717
Pulse	-0.155408	1.000000	0.786535	0.025121
Maxpulse	0.009403	0.786535	1.000000	0.203813
Calories	0.922717	0.025121	0.203813	1.000000

Plotting

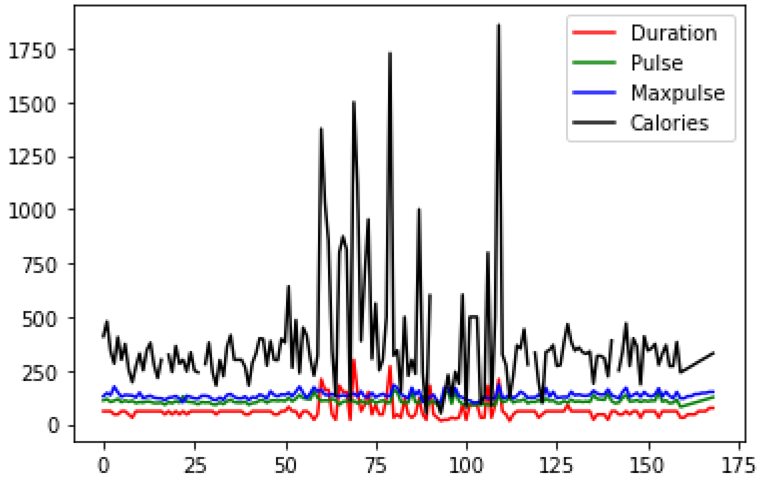
In [99]:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('data.csv')
df.plot()
plt.legend()
plt.show()
```



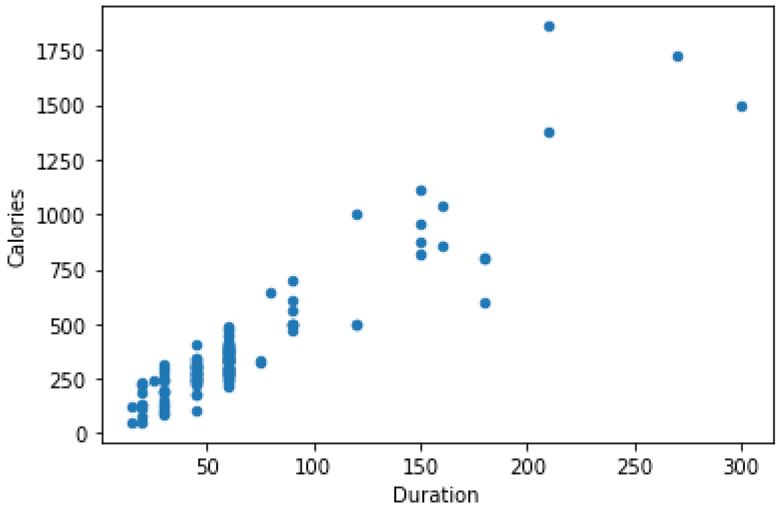
In [100]:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('data.csv')
plt.plot(df["Duration"], c = "red", label = "Duration")
plt.plot(df["Pulse"], c = "green", label = "Pulse")
plt.plot(df["Maxpulse"], c = "blue", label = "Maxpulse")
plt.plot(df["Calories"], c = "black", label = "Calories")
plt.legend()
plt.show()
```



In [102]:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('data.csv')
df.plot(kind = 'scatter', x = "Duration", y = "Calories")
plt.show()
```

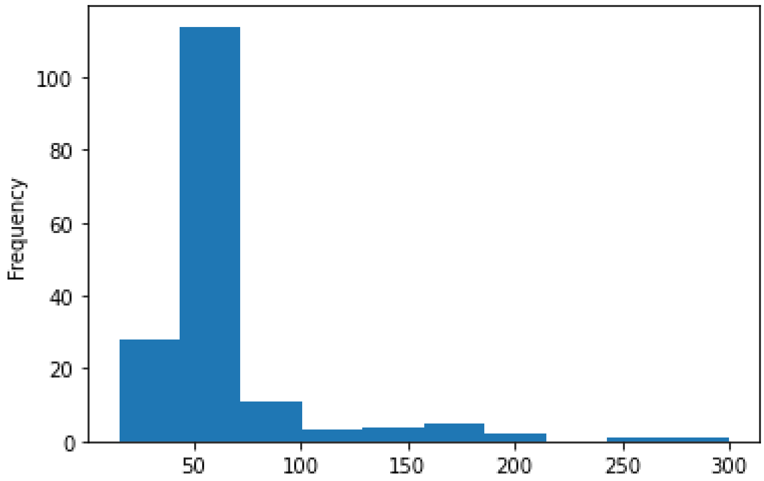


In [103]:

```
df["Duration"].plot(kind = 'hist')
```

Out[103]:

```
<AxesSubplot:ylabel='Frequency'>
```

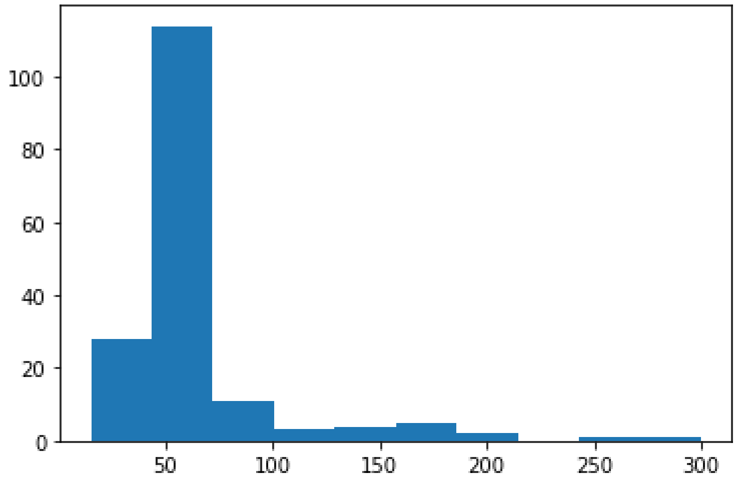


In [107]:

```
plt.hist(df["Duration"])
```

Out[107]:

```
(array([ 28., 114.,  11.,   3.,   4.,   5.,   2.,   0.,   1.,   1.]),
 array([ 15. ,  43.5,  72. , 100.5, 129. , 157.5, 186. , 214.5, 243. ,
        271.5, 300. ]),
 <BarContainer object of 10 artists>)
```



In [ ]:

```

```