Statistical Outlier Detection

In statistics, if a data distribution is approximately normal, then we can use the mean and standard derivation to estimate the probability of a data point falls into a certain range:

- 68% data falls in mean +/- one standard derivation
- 95% data falls in mean +/- two standard derivations
- 99.7% data falls in mean +/- three standard derivations Thus, we can use mean +/ three standard derivations as the boundary of normal data. Any data falls out of the boundary will be considered as outliers.

Setup

```
import numpy as np
import pandas as pd
```

```
df = pd.read_csv('/content/Nov2Temp.csv')
df
```

	high	low
0	58	25
1	26	11
2	53	24
3	60	37
4	67	42
...	...	...
113	119	33
114	127	27
115	18	38
116	15	51
117	30	49

118 rows × 2 columns

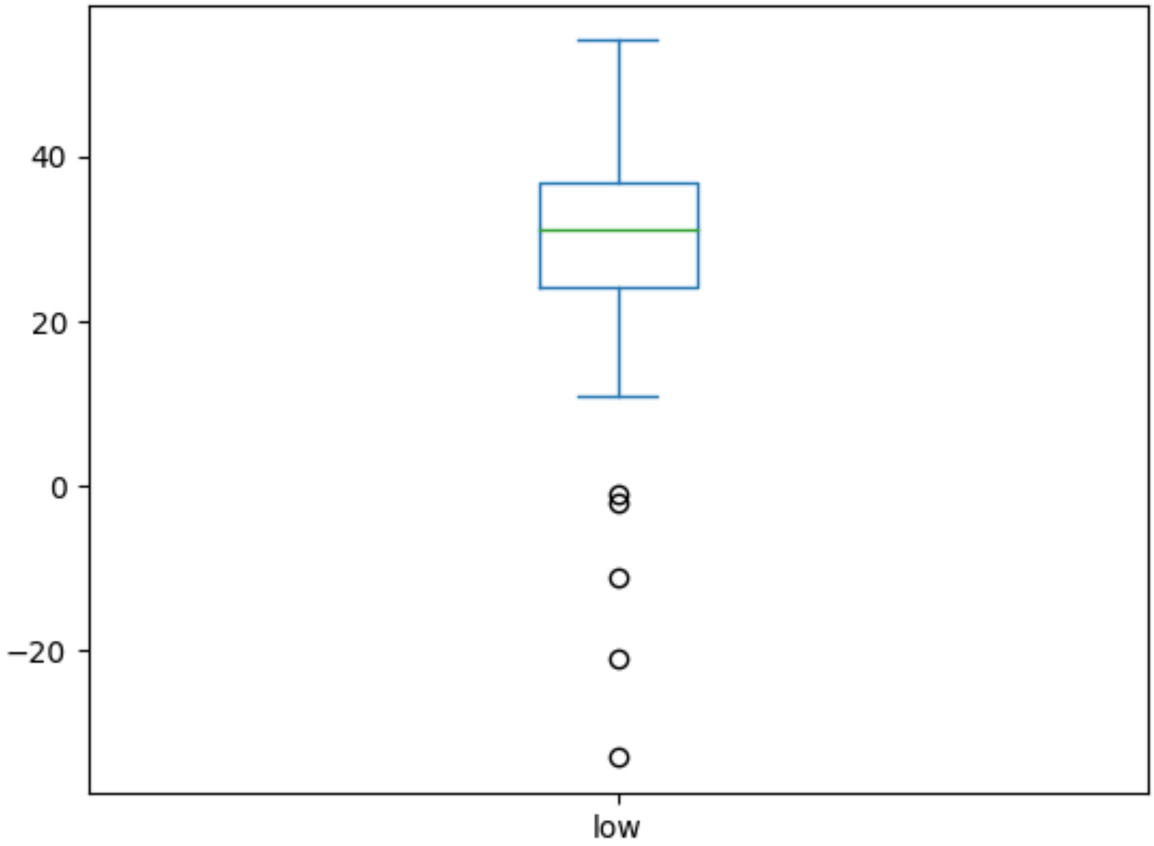
Run the detection

```
df[(df['low']< (df['low'].mean() - 3 * df['low'].std()))|
(df['low']> (df['low'].mean() + 3 * df['low'].std()))]
```

	high	low
109	48	-11
110	43	-21
111	64	-33

```
df['low'].plot(kind='box')
```

```
<Axes: >
```

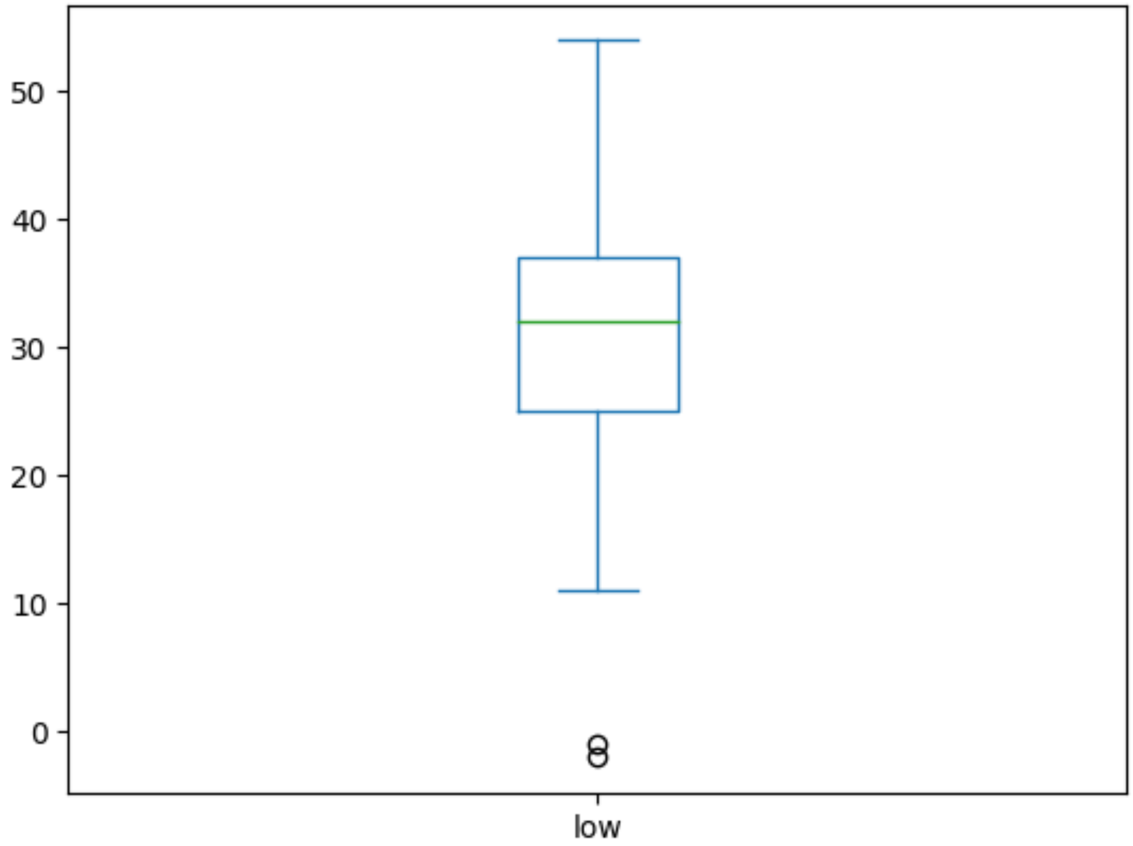


Remove the outliers

```
df.drop((df[(df['low']< (df['low'].mean() - 3 * df['low'].std()))|
(df['low']> (df['low'].mean() + 3 * df['low'].std()))]).index, inplace = True)
```

```
df['low'].plot(kind = 'box')
```

```
<Axes: >
```



Practice

Play with df['high']

Colab paid products - Cancel contracts here