## What is Feature engineering ?

Feature engineering is a process of extracting useful feature from data raw data using math,statisics and domain knowledge

## Handling outliers

## What is outliers?

An outlier is an extremely high or extremely low data point relative to the nearest data point.

## Outlier detection and removal using percentile

## What is percentile?

In [60]:
```
import pandas as pd
import matplotlib.pylab as plt
import seaborn as sns
```

In [2]:
```
df=pd.read_csv('height.csv')
df.head()
```

Out[2]:

| | name | height |
|---|---|---|
| 0 | mohan | 5.9 |
| 1 | maria | 5.2 |
| 2 | sakib | 5.1 |
| 3 | tao | 5.5 |
| 4 | virat | 4.9 |

In [4]:
```
df.describe()
```

Out[4]:

| | height |
|---|---|
| count | 14.000000 |
| mean | 6.050000 |
| std | 2.779804 |
| min | 1.200000 |
| 25% | 5.250000 |
| 50% | 5.550000 |
| 75% | 6.175000 |
| max | 14.500000 |

In [17]:
```
df['height'].quantile(0.7)
```

Out[17]:
```
6.109999999999999
```

In [11]:
```
max_thresold=df['height'].quantile(0.95)
max_thresold
```

Out[11]:
```
9.689999999999998
```

In [18]:
```
df[df['height']>max_thresold]
```

Out[18]:

| | name | height |
|---|---|---|
| 9 | imran | 14.5 |

In [20]:
```
min_thresold=df['height'].quantile(0.05)
min_thresold
```

Out[20]:
```
3.6050000000000004
```

In [21]:
```
df[df['height']<min_thresold]
```

Out[21]:

| | name | height |
|---|---|---|
| 12 | yoseph | 1.2 |

In [58]:
```
df[(df['height']<max_thresold) & (df['height']>min_thresold)]
```

Out[58]:

| | name | height |
|---|---|---|
| 0 | mohan | 5.9 |
| 1 | maria | 5.2 |
| 2 | sakib | 5.1 |
| 3 | tao | 5.5 |
| 4 | virat | 4.9 |
| 5 | khusbu | 5.4 |
| 6 | dmitry | 6.2 |
| 7 | selena | 6.5 |
| 8 | john | 7.1 |
| 10 | jose | 6.1 |
| 11 | deepika | 5.6 |
| 13 | binod | 5.5 |

In [24]:
```
#Compulsory Task
# Take a data set and remove the outliers using a percentile
```

## Outlier detection and removal using IQR

In [26]:
```
df.head(10)
```

Out[26]:

| | name | height |
|---|---|---|
| 0 | mohan | 5.9 |
| 1 | maria | 5.2 |
| 2 | sakib | 5.1 |
| 3 | tao | 5.5 |
| 4 | virat | 4.9 |
| 5 | khusbu | 5.4 |
| 6 | dmitry | 6.2 |
| 7 | selena | 6.5 |
| 8 | john | 7.1 |
| 9 | imran | 14.5 |

In [27]:
```
df.describe()
```

Out[27]:

| | height |
|---|---|
| count | 14.000000 |
| mean | 6.050000 |
| std | 2.779804 |
| min | 1.200000 |
| 25% | 5.250000 |
| 50% | 5.550000 |
| 75% | 6.175000 |
| max | 14.500000 |

In [29]:
```
Q1=df.height.quantile(0.25)
Q3=df.height.quantile(0.75)
Q1,Q3
```

Out[29]:
```
(5.25, 6.175)
```

In [31]:
```
IQR=Q3-Q1
IQR
```

Out[31]:
```
0.9249999999999998
```

In [33]:
```
lower_limit=Q1-1.5*IQR
upper_limit=Q3+1.5*IQR
lower_limit,upper_limit
```

Out[33]:
```
(3.8625000000000003, 7.5625)
```

In [40]:
```
df[(df['height']<lower_limit) | (df['height']>upper_limit)]
```

Out[40]:

| | name | height |
|---|---|---|
| 9 | imran | 14.5 |
| 12 | yoseph | 1.2 |

In [36]:
```
df[df['height']>upper_limit]
```

Out[36]:

| | name | height |
|---|---|---|
| 9 | imran | 14.5 |

In [37]:
```
df[(df['height']<lower_limit)]
```

Out[37]:

| | name | height |
|---|---|---|
| 12 | yoseph | 1.2 |

In [49]:
```
#chnage in orignial data set
df1=df[(df['height']>lower_limit) & (df['height']<upper_limit)]
```

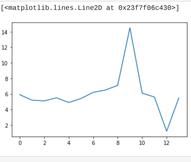
In [50]:
```
df1
```

Out[50]:

| | name | height |
|---|---|---|
| 0 | mohan | 5.9 |
| 1 | maria | 5.2 |
| 2 | sakib | 5.1 |
| 3 | tao | 5.5 |
| 4 | virat | 4.9 |
| 5 | khusbu | 5.4 |
| 6 | dmitry | 6.2 |
| 7 | selena | 6.5 |
| 8 | john | 7.1 |
| 10 | jose | 6.1 |
| 11 | deepika | 5.6 |
| 13 | binod | 5.5 |

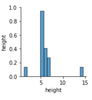
In [59]:
```
plt.plot(df.height)
```

Out[59]:
```
[<matplotlib.lines.Line2D at 0x23f7f06c430>]
```




In [61]:
```
sns.pairplot(df)
```

Out[61]:
```
<seaborn.axisgrid.PairGrid at 0x23f0094c0d0>
```




In [ ]:
```
```