"Data is messy" We will be performing the following operation on our Onion price to refine it.

1. Remove e.g. remove redundant data from the data frame
2. Parse e.g. extract date from year and month column

1 Importing essential libraries

In [20]:

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

2 Reading the DataFrame

In [21]:

```
df = pd.read_csv('MonthWiseMarketArrivals_Chennai.csv')
```

In [22]:

```
df.head(10)
```

Out[22]:

	market	month	year	quantity	priceMin	priceMax	priceMod
0	CHENNAI	January	2004	103400	798	1019	910
1	CHENNAI	January	2005	120500	430	638	533
2	CHENNAI	January	2006	111900	428	621	524
3	CHENNAI	January	2007	84800	900	1370	1129
4	CHENNAI	January	2008	127400	588	1000	797
5	CHENNAI	January	2009	111320	1428	2028	1762
6	CHENNAI	January	2010	110000	1639	2259	1980
7	CHENNAI	January	2011	102000	3583	4583	4083
8	CHENNAI	January	2012	126000	771	1013	892
9	CHENNAI	January	2013	116700	1786	2132	1964

In [23]:

```
df.shape
```

Out[23]:

```
(196, 7)
```

3 Checking for null values and removing it

In [24]:

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

Out[24]:

```
market      1
month       1
year        0
quantity    0
priceMin    0
priceMax    0
priceMod    0
dtype: int64
```

In [25]:

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

In [26]:

```
df.dtypes
```

Out[26]:

```
market      object
month       object
year        object
quantity     int64
priceMin    object
priceMax    object
priceMod    object
dtype: object
```

4 Changing the datatypes for integer values

In [27]:

```
df.iloc[:,2:7] = df.iloc[:,2:7].astype(int)
```

In [28]:

```
df.dtypes
```

Out[28]:

```
market      object
month       object
year         int32
quantity     int32
priceMin     int32
priceMax     int32
priceMod     int32
dtype: object
```

In [29]:

```
df.describe()
```

Out[29]:

	year	quantity	priceMin	priceMax	priceMod
count	195.000000	195.000000	195.000000	195.000000	195.000000
mean	2011.630769	111527.435897	1435.497436	1778.266667	1611.712821
std	4.704360	14863.354493	1165.613388	1328.164341	1244.411557
min	2004.000000	63900.000000	304.000000	456.000000	384.000000
25%	2008.000000	103300.000000	741.000000	1000.000000	874.000000
50%	2012.000000	111200.000000	1092.000000	1457.000000	1263.000000
75%	2016.000000	121950.000000	1764.500000	2073.000000	1935.000000
max	2020.000000	150400.000000	8696.000000	11130.000000	9876.000000

5 Finding the dates

In [30]:

```
df["date"] = df["month"] + "-" + df["year"].map(str)
df.head()
```

Out[30]:

	market	month	year	quantity	priceMin	priceMax	priceMod	date
0	CHENNAI	January	2004	103400	798	1019	910	January-2004
1	CHENNAI	January	2005	120500	430	638	533	January-2005
2	CHENNAI	January	2006	111900	428	621	524	January-2006
3	CHENNAI	January	2007	84800	900	1370	1129	January-2007
4	CHENNAI	January	2008	127400	588	1000	797	January-2008

In [31]:

```
index = pd.to_datetime(df.date)
df.date = pd.DatetimeIndex(df.date)
df.index
```

Out[31]:

```
Int64Index([  0,   1,   2,   3,   4,   5,   6,   7,   8,   9,
            ...
            185, 186, 187, 188, 189, 190, 191, 192, 193, 194],
           dtype='int64', length=195)
```

In [32]:

```
df.head(10)
```

Out[32]:

	market	month	year	quantity	priceMin	priceMax	priceMod	date
0	CHENNAI	January	2004	103400	798	1019	910	2004-01-01
1	CHENNAI	January	2005	120500	430	638	533	2005-01-01
2	CHENNAI	January	2006	111900	428	621	524	2006-01-01
3	CHENNAI	January	2007	84800	900	1370	1129	2007-01-01
4	CHENNAI	January	2008	127400	588	1000	797	2008-01-01
5	CHENNAI	January	2009	111320	1428	2028	1762	2009-01-01
6	CHENNAI	January	2010	110000	1639	2259	1980	2010-01-01
7	CHENNAI	January	2011	102000	3583	4583	4083	2011-01-01
8	CHENNAI	January	2012	126000	771	1013	892	2012-01-01
9	CHENNAI	January	2013	116700	1786	2132	1964	2013-01-01

6 Saving the cleaned Dataframe

In [33]:

```
df.to_csv('MonthWiseMarketArrivals_ChennaiCleaned.csv', index = False)
```