# Sumit Sharma

Data cleaning of data

In [181]:

```
import pandas as pd
```

In [156]:

```
data=pd.read_csv("D:/assienment file unmessenger/Data Cleaning with Excel.csv")
```

In [183]:

```
data.head()
```

Out[183]:

| | Date | Ticker | Name | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|---|---|
| **0** | 21-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **1** | 21-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **2** | 22-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **3** | 23-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **4** | 24-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |

In [179]:

```
data.shape
```

Out[179]:

```
(34, 8)
```

In [180]:

```
data.dtypes
```

Out[180]:

```
Date       object
Ticker     object
Name       object
Open      float64
High      float64
Low       float64
Close     float64
Volume    float64
dtype: object
```

In [161]:

```
data.columns
```

Out[161]:

```
Index(['Date', 'Ticker', 'Name', 'Open', 'High', 'Low', 'Close', 'Volum
e'], dtype='object')
```

In [162]:

```
# shorting the missssing values in rows in decending order
data.isnull().sum().sort_values(ascending=False)
```

Out[162]:

```
Date      2
Ticker    2
Name      2
Open      2
High      2
Low       2
Close     2
Volume    2
dtype: int64
```

In [163]:

```
# checking if there are any missing values in rows
data.isnull().any(axis=1).sum()
```

Out[163]:

```
2
```

In [184]:

```
# checking for the rows which having missing values greter than 50
data[data.isnull().sum(axis=1)<10].head()
```

Out[184]:

| | Date | Ticker | Name | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|---|---|
| **0** | 21-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **1** | 21-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **2** | 22-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **3** | 23-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **4** | 24-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |

In [164]:

```
print("Before deleting the row ",data.shape[0])
```

```
Before deleting the row  34
```

In [165]:

```
# cheecking any row is miss value

print("Before deleting the row ",data.shape[0])
data=data[data.isnull().sum(axis=1)<10]
print("After removing the rows having more the 50 missing values ",data.shape[0])
```

```
Before deleting the row  34
After removing the rows having more the 50 missing values  34
```

In [166]:

```
data.shape
```

Out[166]:

```
(34, 8)
```

In [167]:

```
# checking values in columns
data.isnull().sum()
```

Out[167]:

```
Date      2
Ticker    2
Name      2
Open      2
High      2
Low       2
Close     2
Volume    2
dtype: int64
```

In [117]:

```
x=data.isnull().sum()
y=(data.isnull().sum()/data.shape[0])*100
z=({'number of missing values':x,'percentage of missing values':y})
df=pd.DataFrame(z,columns=['number of missing values','percentage of missing values'])
df.sort_values(by="percentage of missing values",ascending=False)
```

Out[117]:

| | number of missing values | percentage of missing values |
|---|---|---|
| **Date** | 2 | 5.882353 |
| **Ticker** | 2 | 5.882353 |
| **Name** | 2 | 5.882353 |
| **Open** | 2 | 5.882353 |
| **High** | 2 | 5.882353 |
| **Low** | 2 | 5.882353 |
| **Close** | 2 | 5.882353 |
| **Volume** | 2 | 5.882353 |

In [168]:

```
data.describe()
```

Out[168]:

| | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|
| **count** | 32.000000 | 32.000000 | 32.000000 | 32.000000 | 3.200000e+01 |
| **mean** | 1514.407815 | 1530.415639 | 1502.440939 | 1517.100616 | 1.449311e+07 |
| **std** | 1436.315838 | 1447.150306 | 1426.281440 | 1437.324118 | 1.099436e+07 |
| **min** | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 2.037100e+06 |
| **25%** | 350.832512 | 355.997505 | 347.002510 | 353.765015 | 3.345100e+06 |
| **50%** | 676.984985 | 690.399994 | 669.510010 | 678.410004 | 1.470610e+07 |
| **75%** | 3439.380066 | 3458.955017 | 3415.890015 | 3441.092407 | 1.972350e+07 |
| **max** | 3507.639893 | 3524.860107 | 3483.199951 | 3510.979980 | 4.598240e+07 |

In [178]:

```
print("lets check the columns after removing loaned from coolumn",data.columns)
```

```
lets check the columns after removing loaned from coolumn Index(['Date',
'Ticker', 'Name', 'Open', 'High', 'Low', 'Close', 'Volume'], dtype='objec
t')
```

In [169]:

```
data.dtypes
```

Out[169]:

```
Date       object
Ticker     object
Name       object
Open      float64
High      float64
Low       float64
Close     float64
Volume    float64
dtype: object
```

In [170]:

```
data['Date'].fillna('NA',inplace = True)
data['Ticker'].fillna('NA',inplace = True)
data['Name'].fillna("we don't have data",inplace = True)
data['Open']=data['Open'].fillna(data['Open']).mode()[0]
data['High']=data['High'].fillna(data['High']).mode()[0]
data['Low']=data['Low'].fillna(data['Low']).mode()[0]
data['Close']=data['Close'].fillna(data['Close']).mode()[0]
data['Volume']=data['Volume'].fillna(data['Volume']).mode()[0]
```

In [172]:

```
data.isnull().sum().sum()
```

Out[172]:

```
0
```

In [173]:

```
data['Ticker'] = data['Ticker'].str.title()
data['Name'] = data['Name'].str.title()
```

In [185]:

```
data
```

Out[185]:

| | Date | Ticker | Name | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|---|---|
| **0** | 21-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **1** | 21-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **2** | 22-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **3** | 23-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |
| **4** | 24-06-2021 | Fb | Facebook, Inc. | 331.089996 | 332.920013 | 327.649994 | 332.290009 | 3345100.0 |

In [187]:

```
data.to_excel
```

```
Out[187]:
```

```
<bound method NDFrame.to_excel of           Date Ticker
Name         Open  \
0   21-06-2021     Fb                       Facebook, Inc.  331.089996
1   21-06-2021     Fb              Facebook, Inc.           331.089996
2   22-06-2021     Fb                       Facebook, Inc.  331.089996
3   23-06-2021     Fb                       Facebook, Inc.  331.089996
4   24-06-2021     Fb                       Facebook, Inc.  331.089996
5   25-06-2021     Fb         Facebook,     Inc.            331.089996
6   28-06-2021     Fb  Facebook,                      Inc.  331.089996
7   29-06-2021     Fb                       Facebook, Inc.  331.089996
8   30-06-2021     Fb                       Facebook, Inc.  331.089996
9   01-07-2021     Fb                       Facebook, Inc.  331.089996
10  02-07-2021     Fb                       Facebook, Inc.  331.089996
11          NA     Na                   We Don'T Have Data  331.089996
12  21-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
13  22-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
14  22-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
15  23-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
16  24-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
17  25-06-2021   Amzn                  Amazon.Com, Inc.     331.089996
18  28-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
19  29-06-2021   Amzn               Amazon.Com,       Inc.  331.089996
20  30-06-2021   Amzn                     Amazon.Com, Inc.  331.089996
21  01-07-2021   Amzn                     Amazon.Com, Inc.  331.089996
```

```
In [ ]:
```