EDA-03 Dealing with Data Quality Problems

In [251]:
```
import pandas as pd
import numpy as np
```

In [252]:
```
data=pd.read_csv("D:\FTI\Cohort 2 EDA\Lecture 3/Customers.csv")
```

In [253]:
```
data
```

Out[253]:

	ID	Name	Gender	PO BOX	Email	CNIC	DOB	Registered_Age	
0	1001	Ali	M	A74000	Ali@gmailcom	42301-4774468-1	12/1/1980	30	300
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	14/02/1986	25	300
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	23/02/2000	15	300
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	16/05/2008	8	30(
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	17/04/2006	8	300
5	1006	Jatin	M	74500	Jatin @gmail.com	42301-4774473-1	18/03/2004	9	300
6	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	17/02/2002	12	300
7	1008	David	M	75400	David@gmail.com	42301-4774475-1	19/01/2000	14	300
8	1009	NaN	F	NaN	eetert	NaN	20/12/1997	20	300
9	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	21/11/1995	23	300
10	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	22/10/1993	20	300
11	1012	Bilal	M	74500	Bilal@hotmail.com	42301-4774479-1	23/09/1991	23	300
12	1013	Basit	M	74300	Basit@hotmail.com	42301-4774480-1	24/08/1989	25	300
13	1014	Karim	M	74500	Karimhotmail.com	42301-4774481-1	26/07/1987	25	300
14	1015	Kaleem	M	75200	Kaleem@hotmail.com	42301-4774482-1	26/06/1985	30	300
15	1016	Jafer	M	75400	Jafer@hotmail.com	42301-4774483-1	28/05/1983	31	300
16	1017	Junaid	M	74000	Junaid@hotmail.com	42301-4774484-2	28/04/1981	32	300
17	1018	Zia	M	74100	Zia@hotmail.com	42301-4774485-2	30/03/1979	30	300
18	1019	Zaki	M	74000	Zaki@fti.com	42301-474486-2	28/02/1977	35	300
19	1020	Rashid	M	74500	Rashid@fti.com	42301-4774487-2	30/01/1975	30	300
20	1021	Rana	M	74300	Rana@fti.com	42301-4774488-2	15/05/2004	14	300
21	1022	saad	M	74500	saad@fti.com	42301-4774489-2	16/05/2004	13	300
22	1023	Salim	M	75200	Salimfti.com	42301-4774490-2	17/05/2004	13	300
23	1024	Chatin	M	5400	Chatin@fti.com	42301-4774491-2	3/2/1990	23	300
24	1025	Mazhar	M	74500	Mazhar@fti.com	42301-4774492-	26/07/1987	25	
25	1026	Rafiq	M	75200	Rafiq@fti.com	42301-4774493-2	26/06/1985	26	ABI
26	1027	Iqbal	M	75400	Iqbal@fti.com	42301-4774494-2	28/05/1983	32	300
27	1028	Farooq	M	74000	Farooq@fti.com	42301-4774495-2	21/11/1995	21	300
28	1029	William	M	74100	William @fti.com	42301-4774496-2	22/10/1993	20	300
29	1030	Jaqob	M	74200	jaqob@fti.com	42301-4774476-2	23/10/1993	20	336

In [254]:
```
data.shape
```

Out[254]: (30, 10)

In [255]:
```
data.dtypes
```

Out[255]:
```
ID                 int64
Name              object
Gender            object
PO BOX            object
Email             object
CNIC              object
DOB               object
Registered_Age     int64
Mobile            object
Religion          object
dtype: object
```

1.Renaming Columns

In [256]:
```
data.columns
```

Out[256]:
```
Index(['ID ', 'Name ', 'Gender ', 'PO BOX', 'Email ', 'CNIC', 'DOB',
       'Registered_Age', 'Mobile', 'Religion'],
      dtype='object')
```

In [257]:
```
data.rename(columns={'ID ':'ID','Name ':'Name','Gender ':'Gender','PO BOX':'POBOX','Email ':'Email'},inplace=True)
```

In [258]:
```
data.columns
```

Out[258]:
```
Index(['ID', 'Name', 'Gender', 'POBOX', 'Email', 'CNIC', 'DOB',
       'Registered_Age', 'Mobile', 'Religion'],
      dtype='object')
```

2.Checking for Null Values

In [259]:
```
data.isnull().sum()
```

Out[259]:
```
ID                 0
Name               1
Gender             0
POBOX              1
Email              0
CNIC               1
DOB                0
Registered_Age     0
Mobile             0
Religion          24
dtype: int64
```

a) Drop all coulmns where there are more than 30% missing values / where Non-NaN values are 70% of the entire dataset

In [260]:
```
data.dropna(axis=1, thresh=0.70*len(data),inplace=True)
```

In [261]:
```
data.head()
```

Out[261]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	Mo
0	1001	Ali	M	A74000	Ali@gmailcom	42301-4774468-1	12/1/1980	30	300209(
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	14/02/1986	25	300209(
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	23/02/2000	15	300209(
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	16/05/2008	8	300205
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	17/04/2006	8	300209(

In [262]:
```
data.shape
```

Out[262]: (30, 9)

b) lets define a subset of columns and drop records if all the values are NaN

In [263]:
```
data.dropna(how='all', subset=['Email','CNIC','POBOX','Mobile'],inplace=True)
```

In [264]:
```
data.shape
```

Out[264]: (30, 9)

c) lets define a subset of columns and drop records if there is any NaN value

In [265]:
```
data.dropna(how='any', subset=['Email','CNIC','POBOX','Mobile'], inplace=True)
```

In [266]:
```
data.shape
```

Out[266]: (29, 9)

In [267]:
```
data.head(10)
```

Out[267]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	
0	1001	Ali	M	A74000	Ali@gmailcom	42301-4774468-1	12/1/1980	30	3002(
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	14/02/1986	25	3002(
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	23/02/2000	15	3002(
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	16/05/2008	8	300:
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	17/04/2006	8	3002(
5	1006	Jatin	M	74500	Jatin @gmail.com	42301-4774473-1	18/03/2004	9	3002(
6	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	17/02/2002	12	3002(
7	1008	David	M	75400	David@gmail.com	42301-4774475-1	19/01/2000	14	3002(
9	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	21/11/1995	23	3002(
10	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	22/10/1993	20	3002(

In [268]:
```
data.reset_index(drop=True, inplace=True )
```

In [269]:
```
data
```

Out[269]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	
0	1001	Ali	M	A74000	Ali@gmailcom	42301-4774468-1	12/1/1980	30	300
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	14/02/1986	25	300
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	23/02/2000	15	300
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	16/05/2008	8	30
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	17/04/2006	8	300
5	1006	Jatin	M	74500	Jatin @gmail.com	42301-4774473-1	18/03/2004	9	300
6	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	17/02/2002	12	300
7	1008	David	M	75400	David@gmail.com	42301-4774475-1	19/01/2000	14	300
8	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	21/11/1995	23	300
9	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	22/10/1993	20	300
10	1012	Bilal	M	74500	Bilal@hotmail.com	42301-4774479-1	23/09/1991	23	300
11	1013	Basit	M	74300	Basit@hotmail.com	42301-4774480-1	24/08/1989	25	300
12	1014	Karim	M	74500	Karimhotmail.com	42301-4774481-1	26/07/1987	25	300
13	1015	Kaleem	M	75200	Kaleem@hotmail.com	42301-4774482-1	26/06/1985	30	300
14	1016	Jafer	M	75400	Jafer@hotmail.com	42301-4774483-1	28/05/1983	31	300
15	1017	Junaid	M	74000	Junaid@hotmail.com	42301-4774484-2	28/04/1981	32	300
16	1018	Zia	M	74100	Zia@hotmail.com	42301-4774485-2	30/03/1979	30	300
17	1019	Zaki	M	74000	Zaki@fti.com	42301-474486-2	28/02/1977	35	300
18	1020	Rashid	M	74500	Rashid@fti.com	42301-4774487-2	30/01/1975	30	300
19	1021	Rana	M	74300	Rana@fti.com	42301-4774488-2	15/05/2004	14	300
20	1022	saad	M	74500	saad@fti.com	42301-4774489-2	16/05/2004	13	300
21	1023	Salim	M	75200	Salimfti.com	42301-4774490-2	17/05/2004	13	300
22	1024	Chatin	M	5400	Chatin@fti.com	42301-4774491-2	3/2/1990	23	300
23	1025	Mazhar	M	74500	Mazhar@fti.com	42301-4774492-	26/07/1987	25	
24	1026	Rafiq	M	75200	Rafiq@fti.com	42301-4774493-2	26/06/1985	26	AB
25	1027	Iqbal	M	75400	Iqbal@fti.com	42301-4774494-2	28/05/1983	32	300
26	1028	Farooq	M	74000	Farooq@fti.com	42301-4774495-2	21/11/1995	21	300
27	1029	William	M	74100	William @fti.com	42301-4774496-2	22/10/1993	20	300
28	1030	Jaqob	M	74200	jaqob@fti.com	42301-4774476-2	23/10/1993	20	336

In [270]:
```
data.isnull().sum()
```

Out[270]:
```
ID                0
Name              0
Gender            0
POBOX             0
Email             0
CNIC              0
DOB               0
Registered_Age    0
Mobile            0
dtype: int64
```

3. Changing Data Types

In [271]:
```
data.dtypes
```

Out[271]:
```
ID                int64
Name             object
Gender           object
POBOX            object
Email            object
CNIC             object
DOB              object
Registered_Age    int64
Mobile           object
dtype: object
```

Changing Data Types

In [272]:
```
data.head()
```

Out[272]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	Mo
0	1001	Ali	M	A74000	Ali@gmailcom	42301-4774468-1	12/1/1980	30	300209(
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	14/02/1986	25	300209(
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	23/02/2000	15	300209(
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	16/05/2008	8	30020
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	17/04/2006	8	300209(

In [273]:
```
data['ID']=data['ID'].astype(object)
data['DOB']=pd.to_datetime (data['DOB'])
```

In [274]:
```
data.dtypes
```

Out[274]:
```
ID                        object
Name                      object
Gender                    object
POBOX                     object
Email                     object
CNIC                      object
DOB               datetime64[ns]
Registered_Age             int64
Mobile                    object
dtype: object
```

4.Dealing with Incorrect Data

a) Validating POBOX (Length of POBOX should be 5 and it should not contain any alphabet)

In [275]:
```
cnt=0
for row in data['POBOX']:
    if(len(str(row))!=5)or (any(c.isalpha()for c in row)):
        data.loc[cnt,'POBOX']=np.nan
    cnt+=1
```

In [276]:
```
c='?'
c.isalpha()
```

Out[276]:
```
False
```

In [277]:
```
data.head()
```

Out[277]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	M
0	1001	Ali	M	NaN	Ali@gmailcom	42301-4774468-1	1980-12-01	30	300209
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300209
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300209
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	2008-05-16	8	30020
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300209

b) Validating Mobile

In [278]:
```
cnt=0
for row in data['Mobile']:
    if(len(str(row))>15)or (len(str(row))<10) or (any(c.isalpha()for c in ro
w)):
        data.loc[cnt,'Mobile']=np.nan
        #print (cnt, row , len(str(row)))
    else:
        pass
    cnt+=1
```

In [279]:
```
data
```

Out[279]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	
0	1001	Ali	M	NaN	Ali@gmailcom	42301-4774468-1	1980-12-01	30	300
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	2008-05-16	8	
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300
5	1006	Jatin	M	74500	Jatin @gmail.com	42301-4774473-1	2004-03-18	9	300
6	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	2002-02-17	12	300
7	1008	David	M	75400	David@gmail.com	42301-4774475-1	2000-01-19	14	300
8	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	1995-11-21	23	300
9	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	1993-10-22	20	300
10	1012	Bilal	M	74500	Bilal@hotmail.com	42301-4774479-1	1991-09-23	23	300
11	1013	Basit	M	74300	Basit@hotmail.com	42301-4774480-1	1989-08-24	25	300
12	1014	Karim	M	74500	Karimhotmail.com	42301-4774481-1	1987-07-26	25	300
13	1015	Kaleem	M	75200	Kaleem@hotmail.com	42301-4774482-1	1985-06-26	30	300
14	1016	Jafer	M	75400	Jafer@hotmail.com	42301-4774483-1	1983-05-28	31	300
15	1017	Junaid	M	74000	Junaid@hotmail.com	42301-4774484-2	1981-04-28	32	300
16	1018	Zia	M	74100	Zia@hotmail.com	42301-4774485-2	1979-03-30	30	300
17	1019	Zaki	M	74000	Zaki@fti.com	42301-474486-2	1977-02-28	35	300
18	1020	Rashid	M	74500	Rashid@fti.com	42301-4774487-2	1975-01-30	30	300
19	1021	Rana	M	74300	Rana@fti.com	42301-4774488-2	2004-05-15	14	300
20	1022	saad	M	74500	saad@fti.com	42301-4774489-2	2004-05-16	13	300
21	1023	Salim	M	75200	Salimfti.com	42301-4774490-2	2004-05-17	13	300
22	1024	Chatin	M	NaN	Chatin@fti.com	42301-4774491-2	1990-03-02	23	300
23	1025	Mazhar	M	74500	Mazhar@fti.com	42301-4774492-	1987-07-26	25	
24	1026	Rafiq	M	75200	Rafiq@fti.com	42301-4774493-2	1985-06-26	26	
25	1027	Iqbal	M	75400	Iqbal@fti.com	42301-4774494-2	1983-05-28	32	300
26	1028	Farooq	M	74000	Farooq@fti.com	42301-4774495-2	1995-11-21	21	300
27	1029	William	M	74100	William @fti.com	42301-4774496-2	1993-10-22	20	300
28	1030	Jaqob	M	74200	jaqob@fti.com	42301-4774476-2	1993-10-23	20	336

c) Validating Email

Validating Email Using Regular Expressions

In [280]:
```
import re
regexp = re.compile ("[a-z0-9._%+-]+@[a-z0-9.-]+\.[a-z]{2,}")
cnt=0
for row in data['Email']:
    if re.search(regexp,row):
        print(cnt, " " , row, "  ",  "valid Email Address")
    else:
        print(cnt, row, "Invalid Email Address")
    cnt+=1
```

```
0 Ali@gmailcom Invalid Email Address
1   Ahmed@gmail.com    valid Email Address
2   Sara@gmail.com    valid Email Address
3   Janifer@gmail.com    valid Email Address
4   Aslam@gmail.com    valid Email Address
5 Jatin @gmail.com Invalid Email Address
6   Arham@gmail.com    valid Email Address
7   David@gmail.com    valid Email Address
8   Somro@hotmail.com    valid Email Address
9   Khan@hotmail.com    valid Email Address
10   Bilal@hotmail.com    valid Email Address
11   Basit@hotmail.com    valid Email Address
12 Karimhotmail.com Invalid Email Address
13   Kaleem@hotmail.com    valid Email Address
14   Jafer@hotmail.com    valid Email Address
15   Junaid@hotmail.com    valid Email Address
16   Zia@hotmail.com    valid Email Address
17   Zaki@fti.com    valid Email Address
18   Rashid@fti.com    valid Email Address
19   Rana@fti.com    valid Email Address
20   saad@fti.com    valid Email Address
21 Salimfti.com Invalid Email Address
22   Chatin@fti.com    valid Email Address
23   Mazhar@fti.com    valid Email Address
24   Rafiq@fti.com    valid Email Address
25   Iqbal@fti.com    valid Email Address
26   Farooq@fti.com    valid Email Address
27 William @fti.com Invalid Email Address
28   jaqob@fti.com    valid Email Address
```

In [281]:
```
import re
regexp = re.compile ("[a-z0-9._%+-]+@[a-z0-9.-]+\.[a-z]{2,}")
cnt=0
for row in data2['Email']:
    if re.search(regexp,row):
        pass
    else:
        data.loc[cnt, 'Email']=np.nan
    cnt+=1
```

In [282]:
```
data
```

Out[282]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	
0	1001	Ali	M	NaN	NaN	42301-4774468-1	1980-12-01	30	300
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	2008-05-16	8	
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300
5	1006	Jatin	M	74500	NaN	42301-4774473-1	2004-03-18	9	300
6	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	2002-02-17	12	300
7	1008	David	M	75400	David@gmail.com	42301-4774475-1	2000-01-19	14	300
8	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	1995-11-21	23	300
9	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	1993-10-22	20	300
10	1012	Bilal	M	74500	Bilal@hotmail.com	42301-4774479-1	1991-09-23	23	300
11	1013	Basit	M	74300	Basit@hotmail.com	42301-4774480-1	1989-08-24	25	300
12	1014	Karim	M	74500	NaN	42301-4774481-1	1987-07-26	25	300
13	1015	Kaleem	M	75200	Kaleem@hotmail.com	42301-4774482-1	1985-06-26	30	300
14	1016	Jafer	M	75400	Jafer@hotmail.com	42301-4774483-1	1983-05-28	31	300
15	1017	Junaid	M	74000	Junaid@hotmail.com	42301-4774484-2	1981-04-28	32	300
16	1018	Zia	M	74100	Zia@hotmail.com	42301-4774485-2	1979-03-30	30	300
17	1019	Zaki	M	74000	Zaki@fti.com	42301-474486-2	1977-02-28	35	300
18	1020	Rashid	M	74500	Rashid@fti.com	42301-4774487-2	1975-01-30	30	300
19	1021	Rana	M	74300	Rana@fti.com	42301-4774488-2	2004-05-15	14	300
20	1022	saad	M	74500	saad@fti.com	42301-4774489-2	2004-05-16	13	300
21	1023	Salim	M	75200	NaN	42301-4774490-2	2004-05-17	13	300
22	1024	Chatin	M	NaN	Chatin@fti.com	42301-4774491-2	1990-03-02	23	300
23	1025	Mazhar	M	74500	Mazhar@fti.com	42301-4774492-	1987-07-26	25	
24	1026	Rafiq	M	75200	Rafiq@fti.com	42301-4774493-2	1985-06-26	26	
25	1027	Iqbal	M	75400	Iqbal@fti.com	42301-4774494-2	1983-05-28	32	300
26	1028	Farooq	M	74000	Farooq@fti.com	42301-4774495-2	1995-11-21	21	300
27	1029	William	M	74100	NaN	42301-4774496-2	1993-10-22	20	300
28	1030	Jaqob	M	74200	jaqob@fti.com	42301-4774476-2	1993-10-23	20	336

Validating CNIC

In [283]:
```
regexp = re.compile ("^[0-9+]{5}-[0-9+]{7}-[0-9]{1}$")
cnt=0
for row in data['CNIC']:
    if re.search(regexp, row):
        pass
    else:
        data.loc[cnt, 'CNIC']=np.nan
    cnt+=1
```

In [284]:
```
data
```

Out[284]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	
0	1001	Ali	M	NaN	NaN	42301-4774468-1	1980-12-01	30	300
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	2008-05-16	8	
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300
5	1006	Jatin	M	74500	NaN	42301-4774473-1	2004-03-18	9	300
6	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	2002-02-17	12	300
7	1008	David	M	75400	David@gmail.com	42301-4774475-1	2000-01-19	14	300
8	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	1995-11-21	23	300
9	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	1993-10-22	20	300
10	1012	Bilal	M	74500	Bilal@hotmail.com	42301-4774479-1	1991-09-23	23	300
11	1013	Basit	M	74300	Basit@hotmail.com	42301-4774480-1	1989-08-24	25	300
12	1014	Karim	M	74500	NaN	42301-4774481-1	1987-07-26	25	300
13	1015	Kaleem	M	75200	Kaleem@hotmail.com	42301-4774482-1	1985-06-26	30	300
14	1016	Jafer	M	75400	Jafer@hotmail.com	42301-4774483-1	1983-05-28	31	300
15	1017	Junaid	M	74000	Junaid@hotmail.com	42301-4774484-2	1981-04-28	32	300
16	1018	Zia	M	74100	Zia@hotmail.com	42301-4774485-2	1979-03-30	30	300
17	1019	Zaki	M	74000	Zaki@fti.com	NaN	1977-02-28	35	300
18	1020	Rashid	M	74500	Rashid@fti.com	42301-4774487-2	1975-01-30	30	300
19	1021	Rana	M	74300	Rana@fti.com	42301-4774488-2	2004-05-15	14	300
20	1022	saad	M	74500	saad@fti.com	42301-4774489-2	2004-05-16	13	300
21	1023	Salim	M	75200	NaN	42301-4774490-2	2004-05-17	13	300
22	1024	Chatin	M	NaN	Chatin@fti.com	42301-4774491-2	1990-03-02	23	300
23	1025	Mazhar	M	74500	Mazhar@fti.com	NaN	1987-07-26	25	
24	1026	Rafiq	M	75200	Rafiq@fti.com	42301-4774493-2	1985-06-26	26	
25	1027	Iqbal	M	75400	Iqbal@fti.com	42301-4774494-2	1983-05-28	32	300
26	1028	Farooq	M	74000	Farooq@fti.com	42301-4774495-2	1995-11-21	21	300
27	1029	William	M	74100	NaN	42301-4774496-2	1993-10-22	20	300
28	1030	Jaqob	M	74200	jaqob@fti.com	42301-4774476-2	1993-10-23	20	336

5.Handling Outdated Data

In [285]:
```
data.head()
```

Out[285]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	M
0	1001	Ali	M	NaN	NaN	42301-4774468-1	1980-12-01	30	300209
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300209
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300209
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	2008-05-16	8	
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300209

a) Create a new Feature

In [286]:
```
data ['DOB_Year'] = data['DOB'].dt.year
```

In [287]:
```
data ['DOB_Year']
```

Out[287]:
```
0     1980
1     1986
2     2000
3     2008
4     2006
5     2004
6     2002
7     2000
8     1995
9     1993
10    1991
11    1989
12    1987
13    1985
14    1983
15    1981
16    1979
17    1977
18    1975
19    2004
20    2004
21    2004
22    1990
23    1987
24    1985
25    1983
26    1995
27    1993
28    1993
Name: DOB_Year, dtype: int64
```

2. Create another feature as Age_Updated and compute current age

In [288]:
```
now = pd.to_datetime('today')
data['Age_Updated']=now.year-data['DOB'].dt.year
```

In [289]:
```
data.head()
```

Out[289]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	M
0	1001	Ali	M	NaN	NaN	42301-4774468-1	1980-12-01	30	300209
1	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300209
2	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300209
3	1004	Janifer	F	74500	Janifer@gmail.com	42301-4774471-1	2008-05-16	8	
4	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300209

3.Delete Unnecessary Columns

In [290]:
```
del data['DOB_Year']
```

4. Delete Missing values

In [291]:
```
data.isnull().sum()
```

Out[291]:
```
ID                0
Name              0
Gender            0
POBOX             2
Email             5
CNIC              2
DOB               0
Registered_Age    0
Mobile            3
Age_Updated       0
dtype: int64
```

In [292]:
```
data.dropna(inplace=True)
data.reset_index(drop=True, inplace=True)
```

In [293]:
```
data.shape
```

Out[293]:
```
(19, 10)
```

In [294]:
```
data.isnull().sum()
```

Out[294]:
```
ID                0
Name              0
Gender            0
POBOX             0
Email             0
CNIC              0
DOB               0
Registered_Age    0
Mobile            0
Age_Updated       0
dtype: int64
```

In [295]:
```
data
```

Out[295]:

	ID	Name	Gender	POBOX	Email	CNIC	DOB	Registered_Age	
0	1002	Ahmed	M	74100	Ahmed@gmail.com	42301-4774469-1	1986-02-14	25	300
1	1003	Sara	F	74000	Sara@gmail.com	42301-4774470-1	2000-02-23	15	300
2	1005	Aslam	M	74300	Aslam@gmail.com	42301-4774472-1	2006-04-17	8	300
3	1007	Arham	M	75200	Arham@gmail.com	42301-4774474-1	2002-02-17	12	300
4	1008	David	M	75400	David@gmail.com	42301-4774475-1	2000-01-19	14	300
5	1010	Somro	M	74100	Somro@hotmail.com	42301-4774477-1	1995-11-21	23	300
6	1011	Khan	M	74000	Khan@hotmail.com	42301-4774478-1	1993-10-22	20	300
7	1012	Bilal	M	74500	Bilal@hotmail.com	42301-4774479-1	1991-09-23	23	300
8	1013	Basit	M	74300	Basit@hotmail.com	42301-4774480-1	1989-08-24	25	300
9	1015	Kaleem	M	75200	Kaleem@hotmail.com	42301-4774482-1	1985-06-26	30	300
10	1016	Jafer	M	75400	Jafer@hotmail.com	42301-4774483-1	1983-05-28	31	300
11	1017	Junaid	M	74000	Junaid@hotmail.com	42301-4774484-2	1981-04-28	32	300
12	1018	Zia	M	74100	Zia@hotmail.com	42301-4774485-2	1979-03-30	30	300
13	1020	Rashid	M	74500	Rashid@fti.com	42301-4774487-2	1975-01-30	30	300
14	1021	Rana	M	74300	Rana@fti.com	42301-4774488-2	2004-05-15	14	300
15	1022	saad	M	74500	saad@fti.com	42301-4774489-2	2004-05-16	13	300
16	1027	Iqbal	M	75400	Iqbal@fti.com	42301-4774494-2	1983-05-28	32	300
17	1028	Farooq	M	74000	Farooq@fti.com	42301-4774495-2	1995-11-21	21	300
18	1030	Jaqob	M	74200	jaqob@fti.com	42301-4774476-2	1993-10-23	20	336

6. Saving the dataframe as a CSV

In [296]:
```
data.to_csv("D:/FTI/cleaned.csv", index=False)
```