datacleaning

September 19, 2024

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_excel(r"D:\python tut\csv files\Customer Call List.xlsx")
df
```

[1]:

	CustomerID	First_Name	Last_Name	Phone_Number	\
0	1001	Frodo	Baggins	123-545-5421	
1	1002	Abed	Nadir	123/643/9775	
2	1003	Walter	/White	7066950392	
3	1004	Dwight	Schrute	123-543-2345	
4	1005	Jon	Snow	876|678|3469	
5	1006	Ron	Swanson	304-762-2467	
6	1007	Jeff	Winger	NaN	
7	1008	Sherlock	Holmes	876|678|3469	
8	1009	Gandalf	NaN	N/a	
9	1010	Peter	Parker	123-545-5421	
10	1011	Samwise	Gamgee	NaN	
11	1012	Harry	…Potter	7066950392	
12	1013	Don	Draper	123-543-2345	
13	1014	Leslie	Knope	876|678|3469	
14	1015	Toby	Flenderson_	304-762-2467	
15	1016	Ron	Weasley	123-545-5421	
16	1017	Michael	Scott	123/643/9775	
17	1018	Clark	Kent	7066950392	
18	1019	Creed	Braton	N/a	
19	1020	Anakin	Skywalker	876|678|3469	
20	1020	Anakin	Skywalker	876|678|3469	

	Address	Paying Customer	Do_Not_Contact	\
0	123 Shire Lane, Shire	Yes	No	
1	93 West Main Street	No	Yes	
2	298 Drugs Driveway	N	NaN	
3	980 Paper Avenue, Pennsylvania, 18503	Yes	Y	
4	123 Dragons Road	Y	No	
5	768 City Parkway	Yes	Yes	
6	1209 South Street	No	No	
7	98 Clue Drive	N	No	
8	123 Middle Earth	Yes	NaN	

```
9              25th Main Street, New York            Yes             No
10                  612 Shire Lane, Shire            Yes             No
11                   2394 Hogwarts Avenue              Y            NaN
12                       2039 Main Street            Yes              N
13                       343 City Parkway            Yes             No
14                          214 HR Avenue              N             No
15                   2395 Hogwarts Avenue             No              N
16         121 Paper Avenue, Pennsylvania            Yes             No
17                         3498 Super Lane              Y            NaN
18                                     N/a            N/a            Yes
19            910 Tatooine Road, Tatooine            Yes              N
20            910 Tatooine Road, Tatooine            Yes              N

    Not_Useful_Column
0                True
1               False
2                True
3                True
4                True
5                True
6               False
7               False
8               False
9                True
10               True
11               True
12              False
13              False
14              False
15              False
16              False
17               True
18               True
19               True
20               True
```

[3]:
```
df["Last_Name"] = df["Last_Name"].str.strip("123/_.")
df["First_Name"]
df["Phone_Number"] = df["Phone_Number"].str.replace('[^a-zA-Z0-9]', '',␣
 ↪regex=True)
df["Phone_Number"] = df["Phone_Number"].apply(lambda x: str(x))
df["Phone_Number"]  = df["Phone_Number"].apply(lambda x: x[0:3]+ "-" + x[3:6] +␣
 ↪"-"+ x[6:10])
df["Phone_Number"] = df["Phone_Number"].str.replace('nan--', '')
```

[8]:
```
df["Phone_Number"] = df["Phone_Number"].str.replace('--', '')
```

[10]:
```
df
```

[10]:

	CustomerID	First_Name	Last_Name	Phone_Number	Address	Paying Customer	Do_Not_Contact
0	1001	Frodo	Baggins	123-545-5421	123 Shire Lane, Shire	Yes	No
1	1002	Abed	Nadir	123-643-9775	93 West Main Street	No	Yes
2	1003	Walter	White		298 Drugs Driveway	N	NaN
3	1004	Dwight	Schrute	123-543-2345	980 Paper Avenue, Pennsylvania, 18503	Yes	Y
4	1005	Jon	Snow	876-678-3469	123 Dragons Road	Y	No
5	1006	Ron	Swanson	304-762-2467	768 City Parkway	Yes	Yes
6	1007	Jeff	Winger		1209 South Street	No	No
7	1008	Sherlock	Holmes	876-678-3469	98 Clue Drive	N	No
8	1009	Gandalf	NaN		123 Middle Earth	Yes	NaN
9	1010	Peter	Parker	123-545-5421	25th Main Street, New York	Yes	No
10	1011	Samwise	Gamgee		612 Shire Lane, Shire	Yes	No
11	1012	Harry	Potter		2394 Hogwarts Avenue	Y	NaN
12	1013	Don	Draper	123-543-2345	2039 Main Street	Yes	N
13	1014	Leslie	Knope	876-678-3469	343 City Parkway	Yes	No
14	1015	Toby	Flenderson	304-762-2467	214 HR Avenue	N	No
15	1016	Ron	Weasley	123-545-5421	2395 Hogwarts Avenue	No	N
16	1017	Michael	Scott	123-643-9775	121 Paper Avenue, Pennsylvania	Yes	No
17	1018	Clark	Kent		3498 Super Lane	Y	NaN
18	1019	Creed	Braton		N/a	N/a	Yes
19	1020	Anakin	Skywalker	876-678-3469	910 Tatooine Road, Tatooine	Yes	N
20	1020	Anakin	Skywalker	876-678-3469	910 Tatooine Road, Tatooine	Yes	N

```
    Not_Useful_Column
0                True
1               False
2                True
3                True
4                True
5                True
6               False
7               False
8               False
9                True
10               True
11               True
12              False
13              False
14              False
15              False
16              False
17               True
18               True
19               True
20               True
```

[14]:
```
address_split = df["Address"].str.split(',', expand = True)
df = df.assign(Street_Address = address_split[0], State = address_split[1],␣
 ↪Zip_Code = address_split[2])
```

[30]:
```
df["Paying Customer        "] = df["Paying Customer"].str.replace("Yes","Y").
 ↪str.replace("No","N")
df["Do_Not_Contact"] = df["Do_Not_Contact"].str.replace("Yes","Y").str.
 ↪replace("No","N")
```

[30]:
```
    CustomerID First_Name   Last_Name  Phone_Number Paying Customer  \
0         1001      Frodo     Baggins  123-545-5421                Y
1         1002       Abed       Nadir  123-643-9775                N
2         1003     Walter       White                              N
3         1004     Dwight     Schrute  123-543-2345                Y
4         1005        Jon        Snow  876-678-3469                Y
5         1006        Ron     Swanson  304-762-2467                Y
6         1007       Jeff      Winger                              N
7         1008   Sherlock      Holmes  876-678-3469                N
8         1009    Gandalf         NaN                              Y
9         1010      Peter      Parker  123-545-5421                Y
10        1011    Samwise      Gamgee                              Y
11        1012      Harry      Potter                              Y
12        1013        Don      Draper  123-543-2345                Y
```

```
13        1014     Leslie       Knope  876-678-3469              Y
14        1015       Toby  Flenderson  304-762-2467              N
15        1016        Ron     Weasley  123-545-5421              N
16        1017   Michael        Scott  123-643-9775              Y
17        1018      Clark         Kent                            Y
18        1019      Creed      Braton                          N/a
19        1020     Anakin   Skywalker  876-678-3469              Y
20        1020     Anakin   Skywalker  876-678-3469              Y

   Do_Not_Contact  Not_Useful_Column        Street_Address          State  \
0               N               True        123 Shire Lane          Shire
1               Y              False   93 West Main Street           None
2             NaN               True    298 Drugs Driveway           None
3               Y               True      980 Paper Avenue   Pennsylvania
4               N               True      123 Dragons Road           None
5               Y               True      768 City Parkway           None
6               N              False     1209 South Street           None
7               N              False         98 Clue Drive           None
8             NaN              False      123 Middle Earth           None
9               N               True      25th Main Street       New York
10              N               True        612 Shire Lane          Shire
11            NaN               True  2394 Hogwarts Avenue           None
12              N              False      2039 Main Street           None
13              N              False      343 City Parkway           None
14              N              False         214 HR Avenue           None
15              N              False  2395 Hogwarts Avenue           None
16              N              False      121 Paper Avenue   Pennsylvania
17            NaN               True       3498 Super Lane           None
18              Y               True                   N/a           None
19              N               True     910 Tatooine Road       Tatooine
20              N               True     910 Tatooine Road       Tatooine

   Zip_Code Paying Customer\t
0      None                 Y
1      None                 N
2      None                 N
3     18503                 Y
4      None                 Y
5      None                 Y
6      None                 N
7      None                 N
8      None                 Y
9      None                 Y
10     None                 Y
11     None                 Y
12     None                 Y
13     None                 Y
```

```
14     None                 N
15     None                 N
16     None                 Y
17     None                 Y
18     None               N/a
19     None                 Y
20     None                 Y
```

[32]:
```
df = df.drop(columns="Paying Customer\t")
```

[35]:
```
df.fillna('', inplace=True)
df
```

[35]:

	CustomerID	First_Name	Last_Name	Phone_Number	Paying Customer	\
0	1001	Frodo	Baggins	123-545-5421	Y	
1	1002	Abed	Nadir	123-643-9775	N	
2	1003	Walter	White		N	
3	1004	Dwight	Schrute	123-543-2345	Y	
4	1005	Jon	Snow	876-678-3469	Y	
5	1006	Ron	Swanson	304-762-2467	Y	
6	1007	Jeff	Winger		N	
7	1008	Sherlock	Holmes	876-678-3469	N	
8	1009	Gandalf			Y	
9	1010	Peter	Parker	123-545-5421	Y	
10	1011	Samwise	Gamgee		Y	
11	1012	Harry	Potter		Y	
12	1013	Don	Draper	123-543-2345	Y	
13	1014	Leslie	Knope	876-678-3469	Y	
14	1015	Toby	Flenderson	304-762-2467	N	
15	1016	Ron	Weasley	123-545-5421	N	
16	1017	Michael	Scott	123-643-9775	Y	
17	1018	Clark	Kent		Y	
18	1019	Creed	Braton		N/a	
19	1020	Anakin	Skywalker	876-678-3469	Y	
20	1020	Anakin	Skywalker	876-678-3469	Y	

	Do_Not_Contact	Not_Useful_Column	Street_Address	State	\
0	N	True	123 Shire Lane	Shire	
1	Y	False	93 West Main Street		
2		True	298 Drugs Driveway		
3	Y	True	980 Paper Avenue	Pennsylvania	
4	N	True	123 Dragons Road		
5	Y	True	768 City Parkway		
6	N	False	1209 South Street		
7	N	False	98 Clue Drive		
8		False	123 Middle Earth		
9	N	True	25th Main Street	New York	

```
10              N               True        612 Shire Lane          Shire
11                              True  2394 Hogwarts Avenue
12              N              False      2039 Main Street
13              N              False      343 City Parkway
14              N              False         214 HR Avenue
15              N              False  2395 Hogwarts Avenue
16              N              False      121 Paper Avenue   Pennsylvania
17                              True       3498 Super Lane
18              Y               True                  N/a
19              N               True     910 Tatooine Road       Tatooine
20              N               True     910 Tatooine Road       Tatooine

   Zip_Code
0
1
2
3     18503
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

[39]:
```
df
```

[39]:
```
    CustomerID First_Name   Last_Name  Phone_Number Paying Customer  \
0         1001      Frodo     Baggins  123-545-5421                Y
1         1002       Abed       Nadir  123-643-9775                N
2         1003     Walter       White                              N
3         1004     Dwight     Schrute  123-543-2345                Y
4         1005        Jon        Snow  876-678-3469                Y
5         1006        Ron     Swanson  304-762-2467                Y
6         1007       Jeff      Winger                              N
7         1008   Sherlock      Holmes  876-678-3469                N
8         1009    Gandalf                                          Y
```

```
9         1010      Peter      Parker  123-545-5421              Y
10        1011    Samwise      Gamgee                            Y
11        1012      Harry      Potter                            Y
12        1013        Don      Draper  123-543-2345              Y
13        1014     Leslie       Knope  876-678-3469              Y
14        1015       Toby  Flenderson  304-762-2467              N
15        1016        Ron     Weasley  123-545-5421              N
16        1017   Michael        Scott  123-643-9775              Y
17        1018      Clark         Kent                            Y
18        1019      Creed      Braton                          N/a
19        1020     Anakin   Skywalker  876-678-3469              Y
20        1020     Anakin   Skywalker  876-678-3469              Y

   Do_Not_Contact  Not_Useful_Column        Street_Address          State  \
0               N               True        123 Shire Lane          Shire
1               Y              False   93 West Main Street
2                              True    298 Drugs Driveway
3               Y              True      980 Paper Avenue   Pennsylvania
4               N              True      123 Dragons Road
5               Y              True      768 City Parkway
6               N              False     1209 South Street
7               N              False         98 Clue Drive
8                              False      123 Middle Earth
9               N              True      25th Main Street       New York
10              N              True        612 Shire Lane          Shire
11                             True  2394 Hogwarts Avenue
12              N              False      2039 Main Street
13              N              False      343 City Parkway
14              N              False         214 HR Avenue
15              N              False  2395 Hogwarts Avenue
16              N              False      121 Paper Avenue   Pennsylvania
17                             True       3498 Super Lane
18              Y              True                   N/a
19              N              True     910 Tatooine Road       Tatooine
20              N              True     910 Tatooine Road       Tatooine

   Zip_Code
0
1
2
3     18503
4
5
6
7
8
9
```

```
10
11
12
13
14
15
16
17
18
19
20
```

[45]:
```
for x in df.index:
    if df.loc[x,"Do_Not_Contact"] == "Y":
        df.drop(x, inplace=True)
```

[47]:
```
for x in df.index:
    if df.loc[x,"Phone_Number"] == " ":
        df.drop(x, inplace= True)
```

[48]:
```
df
```

[48]:

	CustomerID	First_Name	Last_Name	Phone_Number	Paying Customer	\
0	1001	Frodo	Baggins	123-545-5421	Y	
2	1003	Walter	White		N	
4	1005	Jon	Snow	876-678-3469	Y	
6	1007	Jeff	Winger		N	
7	1008	Sherlock	Holmes	876-678-3469	N	
8	1009	Gandalf			Y	
9	1010	Peter	Parker	123-545-5421	Y	
10	1011	Samwise	Gamgee		Y	
11	1012	Harry	Potter		Y	
12	1013	Don	Draper	123-543-2345	Y	
13	1014	Leslie	Knope	876-678-3469	Y	
14	1015	Toby	Flenderson	304-762-2467	N	
15	1016	Ron	Weasley	123-545-5421	N	
16	1017	Michael	Scott	123-643-9775	Y	
17	1018	Clark	Kent		Y	
19	1020	Anakin	Skywalker	876-678-3469	Y	
20	1020	Anakin	Skywalker	876-678-3469	Y	

	Do_Not_Contact	Not_Useful_Column	Street_Address	State	\
0	N	True	123 Shire Lane	Shire	
2		True	298 Drugs Driveway		

```
4                N               True      123 Dragons Road                
6                N              False     1209 South Street                
7                N              False         98 Clue Drive                
8                               False      123 Middle Earth                
9                N               True      25th Main Street       New York
10               N               True        612 Shire Lane          Shire
11                               True  2394 Hogwarts Avenue                
12               N              False      2039 Main Street                
13               N              False      343 City Parkway                
14               N              False         214 HR Avenue                
15               N              False  2395 Hogwarts Avenue                
16               N              False      121 Paper Avenue   Pennsylvania
17                              True       3498 Super Lane                
19               N               True     910 Tatooine Road       Tatooine
20               N               True     910 Tatooine Road       Tatooine

   Zip_Code
0
2
4
6
7
8
9
10
11
12
13
14
15
16
17
19
20
```

[49]:
```
df.to_csv(r"D:\Dataset\new_data.csv")
```

[ ]:
```

```