Data Cleansing

February 17, 2025

1 Data Cleansing

[76]:
```
import pandas as pd
```

1.0.1 Read File

[77]:
```
df = pd.read_excel(r'D:\Analyst materials\Python\Customer Call List.xlsx')
df
```

[77]:

	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	7066950392	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			
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			

```
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
```

[78]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21 entries, 0 to 20
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
```

```
 0   CustomerID         21 non-null     int64
 1   First_Name         21 non-null     object
 2   Last_Name          20 non-null     object
 3   Phone_Number       19 non-null     object
 4   Address            21 non-null     object
 5   Paying Customer    21 non-null     object
 6   Do_Not_Contact     17 non-null     object
 7   Not_Useful_Column  21 non-null     bool
dtypes: bool(1), int64(1), object(6)
memory usage: 1.3+ KB
```

1.0.2 Check and Remove Duplicates

[79]:
```
df[df.duplicated()]
```

[79]:
```
    CustomerID First_Name  Last_Name  Phone_Number  \
20        1020     Anakin  Skywalker  876|678|3469

                       Address Paying Customer Do_Not_Contact  \
20  910 Tatooine Road, Tatooine             Yes              N

    Not_Useful_Column
20               True
```

[80]:
```
df = df.drop_duplicates()
```

1.0.3 Drop not useful column

[81]:
```
df = df.drop(columns = 'Not_Useful_Column')
```

[81]:
```
   CustomerID First_Name Last_Name  Phone_Number                Address  \
0        1001      Frodo   Baggins  123-545-5421  123 Shire Lane, Shire

  Paying Customer Do_Not_Contact
0             Yes             No
```

1.0.4 Cleaning the Last Name column by removing extra characters

[82]:
```
df['Last_Name'] = df['Last_Name'].str.strip("123._/")
df[['Last_Name']]
```

[82]:
```
     Last_Name
0      Baggins
1        Nadir
2        White
3      Schrute
4         Snow
5      Swanson
```

```
6       Winger
7       Holmes
8          NaN
9       Parker
10      Gamgee
11      Potter
12      Draper
13       Knope
14  Flenderson
15     Weasley
16       Scott
17        Kent
18      Braton
19   Skywalker
```

1.0.5 Cleaning Phone Number column by first stripping all special characters using regex

[83]:
```
df['Phone_Number'] = df['Phone_Number'].str.replace('[^A-Za-z0-9]', '',␣
↪regex=True)
df[['Phone_Number']]
```

[83]:
```
   Phone_Number
0    1235455421
1    1236439775
2           NaN
3    1235432345
4    8766783469
5    3047622467
6           NaN
7    8766783469
8            Na
9    1235455421
10          NaN
11          NaN
12   1235432345
13   8766783469
14   3047622467
15   1235455421
16   1236439775
17          NaN
18           Na
19   8766783469
```

1.0.6 and then insert the new format back onto Phone Number

[84]:
```
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']]
```

[84]:

	Phone_Number
0	123-545-5421
1	123-643-9775
2	nan--
3	123-543-2345
4	876-678-3469
5	304-762-2467
6	nan--
7	876-678-3469
8	Na--
9	123-545-5421
10	nan--
11	nan--
12	123-543-2345
13	876-678-3469
14	304-762-2467
15	123-545-5421
16	123-643-9775
17	nan--
18	Na--
19	876-678-3469

1.0.7 ...and then strip off extra characters on null values for later cleaning

[85]:
```
df['Phone_Number'] = df['Phone_Number'].str.strip('--')
df[['Phone_Number']]
```

[85]:

	Phone_Number
0	123-545-5421
1	123-643-9775
2	nan
3	123-543-2345
4	876-678-3469
5	304-762-2467
6	nan
7	876-678-3469
8	Na
9	123-545-5421
10	nan
11	nan
12	123-543-2345

```
13  876-678-3469
14  304-762-2467
15  123-545-5421
16  123-643-9775
17           nan
18            Na
19  876-678-3469
```

1.0.8 Clearing all null values

[86]:
```
df = df.fillna('')
df = df.replace('N/a', '')
df = df.replace('nan', '')
df = df.replace('Na', '')
df
```

[86]:

	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		
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		
7	1008	Sherlock	Holmes	876-678-3469	
8	1009	Gandalf			
9	1010	Peter	Parker	123-545-5421	
10	1011	Samwise	Gamgee		
11	1012	Harry	Potter		
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		
18	1019	Creed	Braton		
19	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	
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
9              25th Main Street, New York           Yes             No
10                  612 Shire Lane, Shire           Yes             No
11                   2394 Hogwarts Avenue             Y
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
18                                                                 Yes
19            910 Tatooine Road, Tatooine           Yes              N
```

1.0.9 Splitting the Address into 3 columns for more detailed view

[87]:
```
df[['Street_Address', 'State', 'Zip_Code']] = df['Address'].str.split(',', n=2,␣
↪expand=True)
df[['Address', 'Street_Address', 'State', 'Zip_Code']]
```

[87]:
```
                                  Address        Street_Address  \
0                   123 Shire Lane, Shire        123 Shire Lane
1                     93 West Main Street   93 West Main Street
2                      298 Drugs Driveway    298 Drugs Driveway
3   980 Paper Avenue, Pennsylvania, 18503      980 Paper Avenue
4                         123 Dragons Road      123 Dragons Road
5                         768 City Parkway      768 City Parkway
6                        1209 South Street     1209 South Street
7                            98 Clue Drive         98 Clue Drive
8                         123 Middle Earth      123 Middle Earth
9                25th Main Street, New York      25th Main Street
10                  612 Shire Lane, Shire        612 Shire Lane
11                   2394 Hogwarts Avenue  2394 Hogwarts Avenue
12                       2039 Main Street      2039 Main Street
13                       343 City Parkway      343 City Parkway
14                          214 HR Avenue         214 HR Avenue
15                   2395 Hogwarts Avenue  2395 Hogwarts Avenue
16         121 Paper Avenue, Pennsylvania      121 Paper Avenue
17                         3498 Super Lane       3498 Super Lane
18
19            910 Tatooine Road, Tatooine     910 Tatooine Road

              State Zip_Code
0           Shire     None
1            None     None
2            None     None
3    Pennsylvania    18503
4            None     None
```

```
5            None     None
6            None     None
7            None     None
8            None     None
9        New York     None
10          Shire     None
11           None     None
12           None     None
13           None     None
14           None     None
15           None     None
16   Pennsylvania     None
17           None     None
18           None     None
19       Tatooine     None
```

1.0.10 Formatting all 'Yes' and 'No' into 'Y' and 'N' for coherence

[88]:
```
df['Paying Customer'] = df['Paying Customer'].str.replace('Yes', 'Y')
df['Paying Customer'] = df['Paying Customer'].str.replace('No', 'N')
df['Do_Not_Contact'] = df['Do_Not_Contact'].str.replace('Yes', 'Y')
df['Do_Not_Contact'] = df['Do_Not_Contact'].str.replace('No', 'N')
df[['Paying Customer', 'Do_Not_Contact']]
```

[88]:
```
   Paying Customer Do_Not_Contact
0                Y              N
1                N              Y
2                N
3                Y              Y
4                Y              N
5                Y              Y
6                N              N
7                N              N
8                Y
9                Y              N
10               Y              N
11               Y
12               Y              N
13               Y              N
14               N              N
15               N              N
16               Y              N
17               Y
18                              Y
19               Y              N
```

1.0.11 Changing the response of customers who did not specify they don't want to be contacted

[89]:
```
df['Do_Not_Contact'] = df['Do_Not_Contact'].replace('', 'N')
df[['Do_Not_Contact']]
```

[89]:

	Do_Not_Contact
0	N
1	Y
2	N
3	Y
4	N
5	Y
6	N
7	N
8	N
9	N
10	N
11	N
12	N
13	N
14	N
15	N
16	N
17	N
18	Y
19	N

1.0.12 Drop rows where customers specified they don't want to be contacted

[90]:
```
for x in df.index:
    if df.loc[x, 'Do_Not_Contact'] == 'Y':
        df.drop(x, inplace=True)
```

[90]:

	CustomerID	First_Name	Last_Name	Phone_Number	\
0	1001	Frodo	Baggins	123-545-5421	
2	1003	Walter	White		
4	1005	Jon	Snow	876-678-3469	
6	1007	Jeff	Winger		
7	1008	Sherlock	Holmes	876-678-3469	
8	1009	Gandalf			
9	1010	Peter	Parker	123-545-5421	
10	1011	Samwise	Gamgee		
11	1012	Harry	Potter		
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
19        1020     Anakin   Skywalker  876-678-3469

                          Address Paying Customer Do_Not_Contact  \
0            123 Shire Lane, Shire               Y              N
2               298 Drugs Driveway               N              N
4                 123 Dragons Road               Y              N
6                1209 South Street               N              N
7                    98 Clue Drive               N              N
8                 123 Middle Earth               Y              N
9       25th Main Street, New York               Y              N
10           612 Shire Lane, Shire               Y              N
11            2394 Hogwarts Avenue               Y              N
12                2039 Main Street               Y              N
13                343 City Parkway               Y              N
14                   214 HR Avenue               N              N
15            2395 Hogwarts Avenue               N              N
16  121 Paper Avenue, Pennsylvania               Y              N
17                3498 Super Lane               Y              N
19    910 Tatooine Road, Tatooine               Y              N

          Street_Address          State Zip_Code
0          123 Shire Lane          Shire     None
2     298 Drugs Driveway           None     None
4       123 Dragons Road           None     None
6      1209 South Street           None     None
7          98 Clue Drive           None     None
8       123 Middle Earth           None     None
9       25th Main Street       New York     None
10        612 Shire Lane          Shire     None
11  2394 Hogwarts Avenue           None     None
12      2039 Main Street           None     None
13      343 City Parkway           None     None
14         214 HR Avenue           None     None
15  2395 Hogwarts Avenue           None     None
16      121 Paper Avenue   Pennsylvania     None
17       3498 Super Lane           None     None
19     910 Tatooine Road       Tatooine     None
```

1.0.13 Drop rows where customers' phone numbers don't exist

[91]:
```
for x in df.index:
    if df.loc[x, 'Phone_Number'] == '':
        df.drop(x, inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 10 entries, 0 to 19
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CustomerID       10 non-null     int64
 1   First_Name       10 non-null     object
 2   Last_Name        10 non-null     object
 3   Phone_Number     10 non-null     object
 4   Address          10 non-null     object
 5   Paying Customer  10 non-null     object
 6   Do_Not_Contact   10 non-null     object
 7   Street_Address   10 non-null     object
 8   State            4 non-null      object
 9   Zip_Code         0 non-null      object
dtypes: int64(1), object(9)
memory usage: 1.2+ KB
```

1.0.14 Reset index

[92]:
```
df = df.reset_index(drop=True)
df
```

[92]:

	CustomerID	First_Name	Last_Name	Phone_Number	Address	Paying Customer	Do_Not_Contact	Street_Address	State	Zip_Code
0	1001	Frodo	Baggins	123-545-5421	123 Shire Lane, Shire	Y	N	123 Shire Lane	Shire	None
1	1005	Jon	Snow	876-678-3469	123 Dragons Road	Y	N			
2	1008	Sherlock	Holmes	876-678-3469	98 Clue Drive	N	N			
3	1010	Peter	Parker	123-545-5421	25th Main Street, New York	Y	N			
4	1013	Don	Draper	123-543-2345	2039 Main Street	Y	N			
5	1014	Leslie	Knope	876-678-3469	343 City Parkway	Y	N			
6	1015	Toby	Flenderson	304-762-2467	214 HR Avenue	N	N			
7	1016	Ron	Weasley	123-545-5421	2395 Hogwarts Avenue	N	N			
8	1017	Michael	Scott	123-643-9775	121 Paper Avenue, Pennsylvania	Y	N			
9	1020	Anakin	Skywalker	876-678-3469	910 Tatooine Road, Tatooine	Y	N			

1	123 Dragons Road	None	None
2	98 Clue Drive	None	None
3	25th Main Street	New York	None
4	2039 Main Street	None	None
5	343 City Parkway	None	None
6	214 HR Avenue	None	None
7	2395 Hogwarts Avenue	None	None
8	121 Paper Avenue	Pennsylvania	None
9	910 Tatooine Road	Tatooine	None