PROJECT SCREENSHOTS DOCUMENTATION

Topic: Data Cleaning using Power BI & VS Code

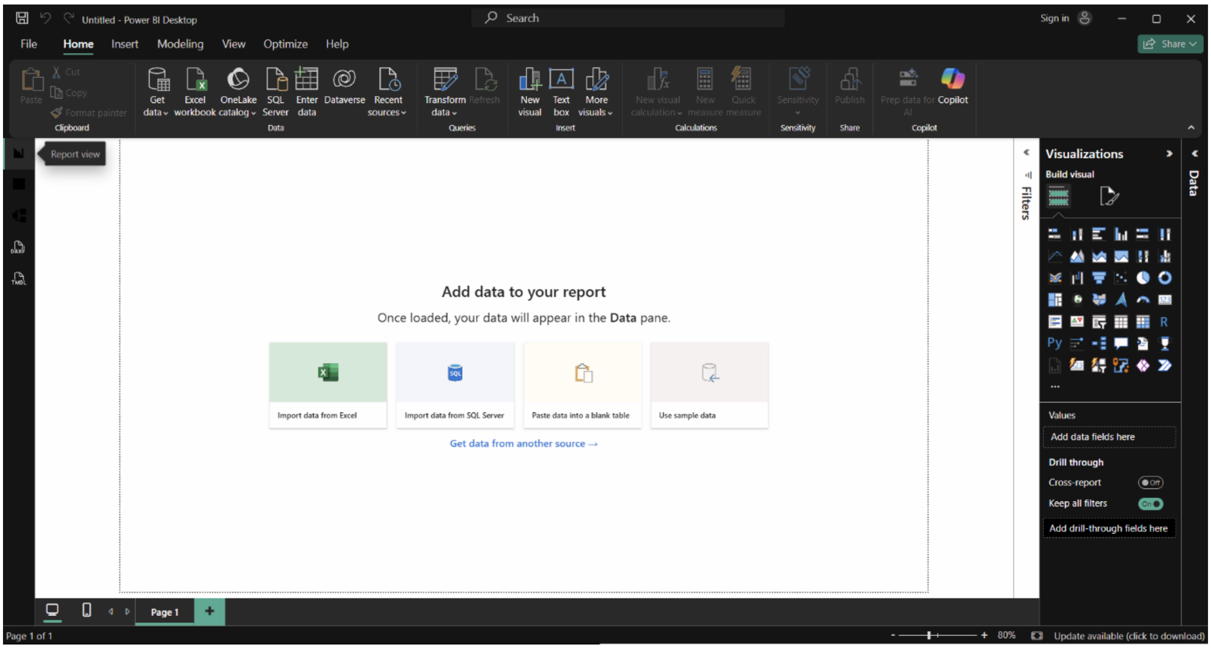
Internship: Data Analytics / Data Science Internship

Intern Name: Harika Vavilapalli

Tools Used: Microsoft Power BI (Power Query), Visual Studio Code

Description: This document contains relevant screenshots showing the step-by-step data cleaning process performed in Power BI and automated data cleaning performed using Jypter Notebook in VS Code that are covered in both the Weeks (Week 1 & Week 2).

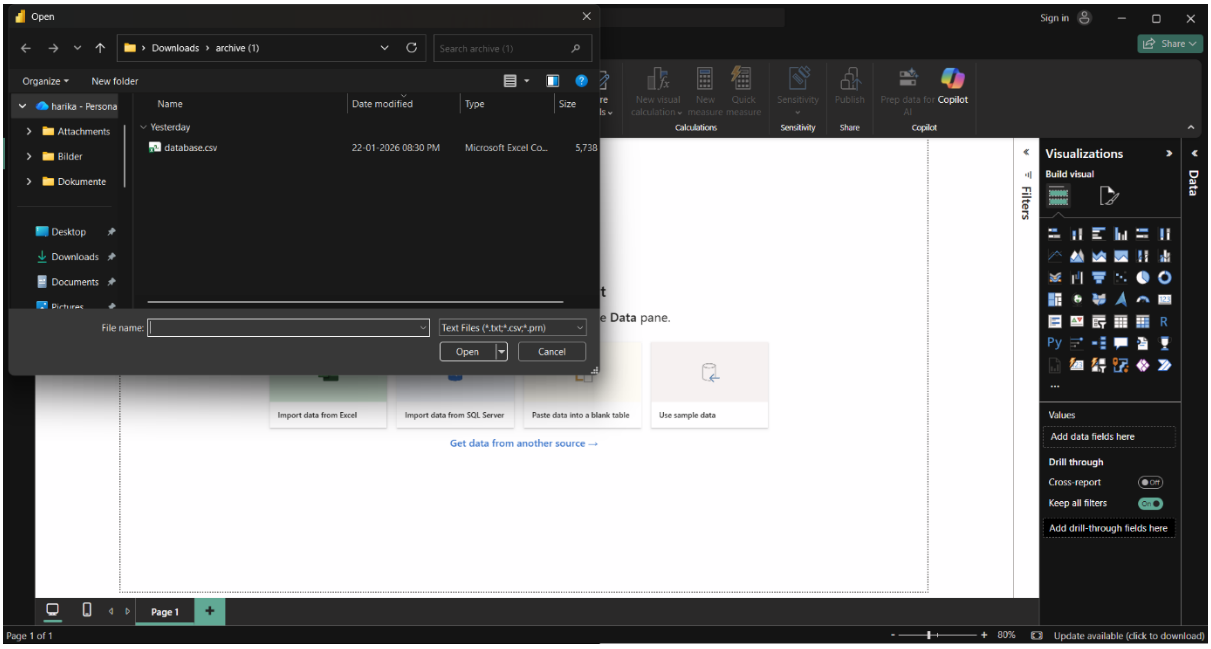
DATA CLEANING IN POWER BI



Add data to your report

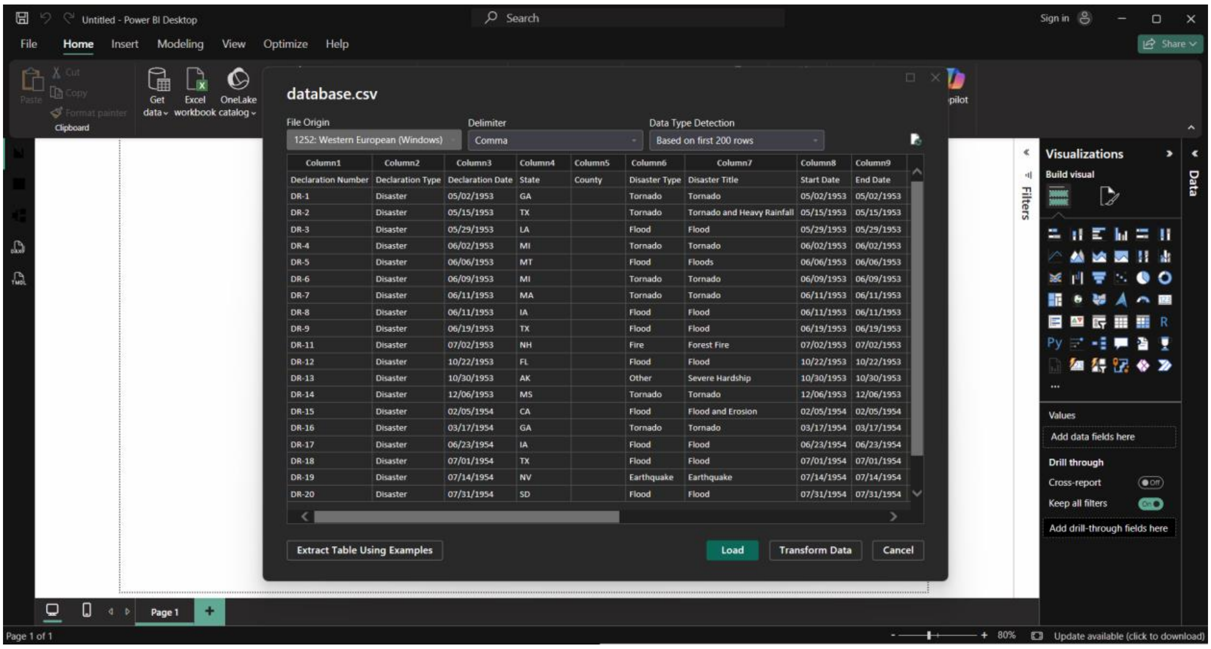
Once loaded, your data will appear in the **Data** pane.

Fig 1: Power BI Canvas.



Name	Date modified	Type	Size
database.csv	22-01-2026 08:30 PM	Microsoft Excel Co...	5,738

Fig 2: Importing Dataset Using **Get Data**.



database.csv

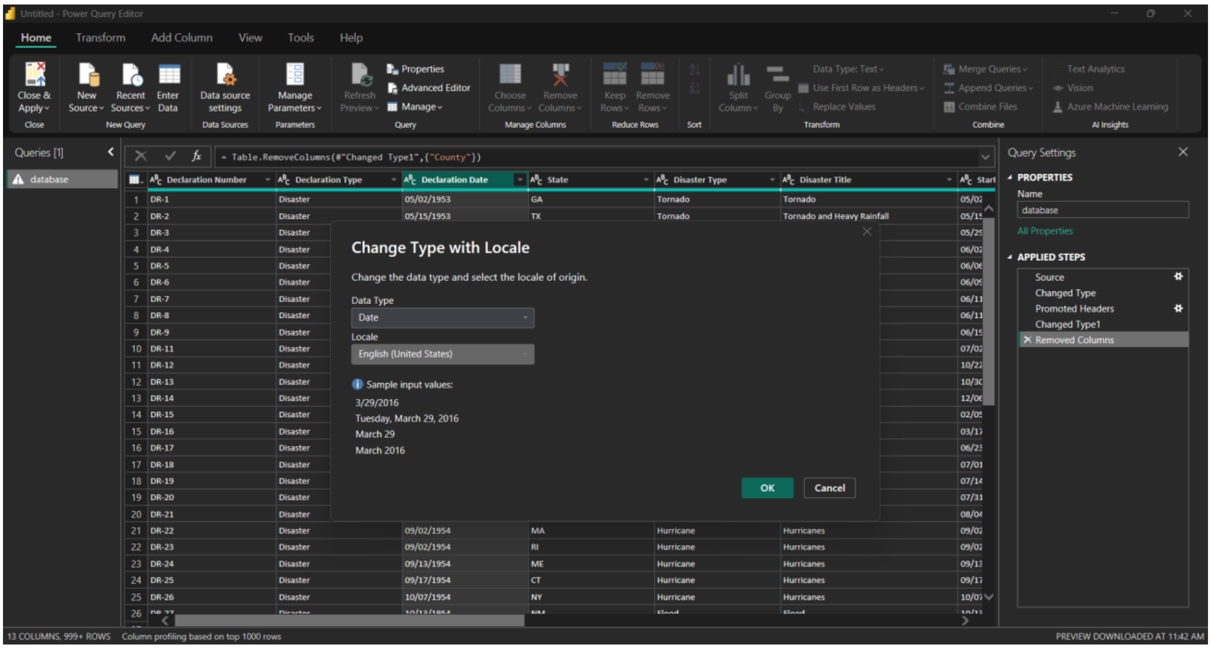
File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
Declaration Number	Declaration Type	Declaration Date	State	County	Disaster Type	Disaster Title	Start Date	End Date
DR-1	Disaster	05/02/1953	GA		Tornado	Tornado	05/02/1953	05/02/1953
DR-2	Disaster	05/15/1953	TX		Tornado	Tornado and Heavy Rainfall	05/15/1953	05/15/1953
DR-3	Disaster	05/29/1953	LA		Flood	Flood	05/29/1953	05/29/1953
DR-4	Disaster	06/02/1953	MI		Tornado	Tornado	06/02/1953	06/02/1953
DR-5	Disaster	06/06/1953	MT		Flood	Floods	06/06/1953	06/06/1953
DR-6	Disaster	06/09/1953	MI		Tornado	Tornado	06/09/1953	06/09/1953
DR-7	Disaster	06/11/1953	MA		Tornado	Tornado	06/11/1953	06/11/1953
DR-8	Disaster	06/11/1953	IA		Flood	Flood	06/11/1953	06/11/1953
DR-9	Disaster	06/19/1953	TX		Flood	Flood	06/19/1953	06/19/1953
DR-11	Disaster	07/02/1953	NH		Fire	Forest Fire	07/02/1953	07/02/1953
DR-12	Disaster	10/22/1953	FL		Flood	Flood	10/22/1953	10/22/1953
DR-13	Disaster	10/30/1953	AK		Other	Severe Hardship	10/30/1953	10/30/1953
DR-14	Disaster	12/06/1953	MS		Tornado	Tornado	12/06/1953	12/06/1953
DR-15	Disaster	02/05/1954	CA		Flood	Flood and Erosion	02/05/1954	02/05/1954
DR-16	Disaster	03/17/1954	GA		Tornado	Tornado	03/17/1954	03/17/1954
DR-17	Disaster	06/23/1954	IA		Flood	Flood	06/23/1954	06/23/1954
DR-18	Disaster	07/01/1954	TX		Flood	Flood	07/01/1954	07/01/1954
DR-19	Disaster	07/14/1954	NV		Earthquake	Earthquake	07/14/1954	07/14/1954
DR-20	Disaster	07/31/1954	SD		Flood	Flood	07/31/1954	07/31/1954

Fig 3: Transforming the Data Using **Transform Data**.



```
= Table.RemoveColumns(#"Changed Type1",{"County"})
```

Name: database

APPLIED STEPS
- Source
- Changed Type
- Promoted Headers
- Changed Type1
- Removed Columns

	Declaration Number	Declaration Type	Declaration Date	State	Disaster Type	Disaster Title	Start
1	DR-1	Disaster	05/02/1953	GA	Tornado	Tornado	05/02
2	DR-2	Disaster	05/15/1953	TX	Tornado	Tornado and Heavy Rainfall	05/15
3	DR-3	Disaster					05/29
4	DR-4	Disaster					06/02
5	DR-5	Disaster					06/06
6	DR-6	Disaster					06/09
7	DR-7	Disaster					06/11
8	DR-8	Disaster					06/11
9	DR-9	Disaster					06/19
10	DR-11	Disaster					07/02
11	DR-12	Disaster					10/22
12	DR-13	Disaster					10/30
13	DR-14	Disaster					12/06
14	DR-15	Disaster					02/05
15	DR-16	Disaster					03/17
16	DR-17	Disaster					06/23
17	DR-18	Disaster					07/01
18	DR-19	Disaster					07/14
19	DR-20	Disaster					07/31
20	DR-21	Disaster					08/04
21	DR-22	Disaster	09/02/1954	MA	Hurricane	Hurricanes	09/02
22	DR-23	Disaster	09/02/1954	RI	Hurricane	Hurricanes	09/02
23	DR-24	Disaster	09/13/1954	ME	Hurricane	Hurricanes	09/13
24	DR-25	Disaster	09/17/1954	CT	Hurricane	Hurricanes	09/17
25	DR-26	Disaster	10/07/1954	NY	Hurricane	Hurricanes	10/07
26	DR-27	Disaster	10/13/1954	NM	Flood	Flood	10/13

Change Type with Locale

Change the data type and select the locale of origin.

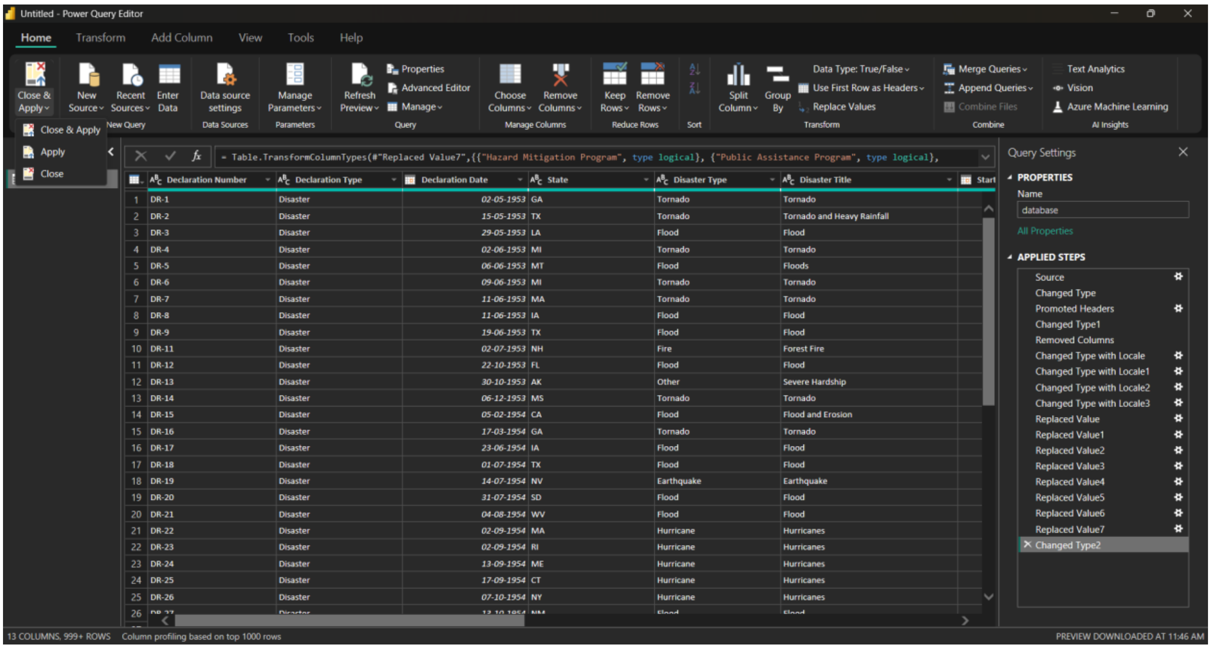
Data Type: Date

Locale: English (United States)

Sample input values:
3/29/2016
Tuesday, March 29, 2016
March 29
March 2016

13 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 11:42 AM

Fig 4: Changing The Data Type of the column Using **Locale**.



```
= Table.TransformColumnTypes(#"Replaced Value7",{{"Hazard Mitigation Program", type logical}, {"Public Assistance Program", type logical},
```

	Declaration Number	Declaration Type	Declaration Date	State	Disaster Type	Disaster Title
1	DR-1	Disaster	02-05-1953	GA	Tornado	Tornado
2	DR-2	Disaster	15-05-1953	TX	Tornado	Tornado and Heavy Rainfall
3	DR-3	Disaster	29-05-1953	LA	Flood	Flood
4	DR-4	Disaster	02-06-1953	MI	Tornado	Tornado
5	DR-5	Disaster	06-06-1953	MT	Flood	Floods
6	DR-6	Disaster	09-06-1953	MI	Tornado	Tornado
7	DR-7	Disaster	11-06-1953	MA	Tornado	Tornado
8	DR-8	Disaster	11-06-1953	IA	Flood	Flood
9	DR-9	Disaster	19-06-1953	TX	Flood	Flood
10	DR-11	Disaster	02-07-1953	NH	Fire	Forest Fire
11	DR-12	Disaster	22-10-1953	FL	Flood	Flood
12	DR-13	Disaster	30-10-1953	AK	Other	Severe Hardship
13	DR-14	Disaster	06-12-1953	MS	Tornado	Tornado
14	DR-15	Disaster	05-02-1954	CA	Flood	Flood and Erosion
15	DR-16	Disaster	17-03-1954	GA	Tornado	Tornado
16	DR-17	Disaster	23-06-1954	IA	Flood	Flood
17	DR-18	Disaster	01-07-1954	TX	Flood	Flood
18	DR-19	Disaster	14-07-1954	NV	Earthquake	Earthquake
19	DR-20	Disaster	31-07-1954	SD	Flood	Flood
20	DR-21	Disaster	04-08-1954	WV	Flood	Flood
21	DR-22	Disaster	02-09-1954	MA	Hurricane	Hurricanes
22	DR-23	Disaster	02-09-1954	RI	Hurricane	Hurricanes
23	DR-24	Disaster	13-09-1954	ME	Hurricane	Hurricanes
24	DR-25	Disaster	17-09-1954	CT	Hurricane	Hurricanes
25	DR-26	Disaster	07-10-1954	NY	Hurricane	Hurricanes

Query Settings

PROPERTIES

Name: database

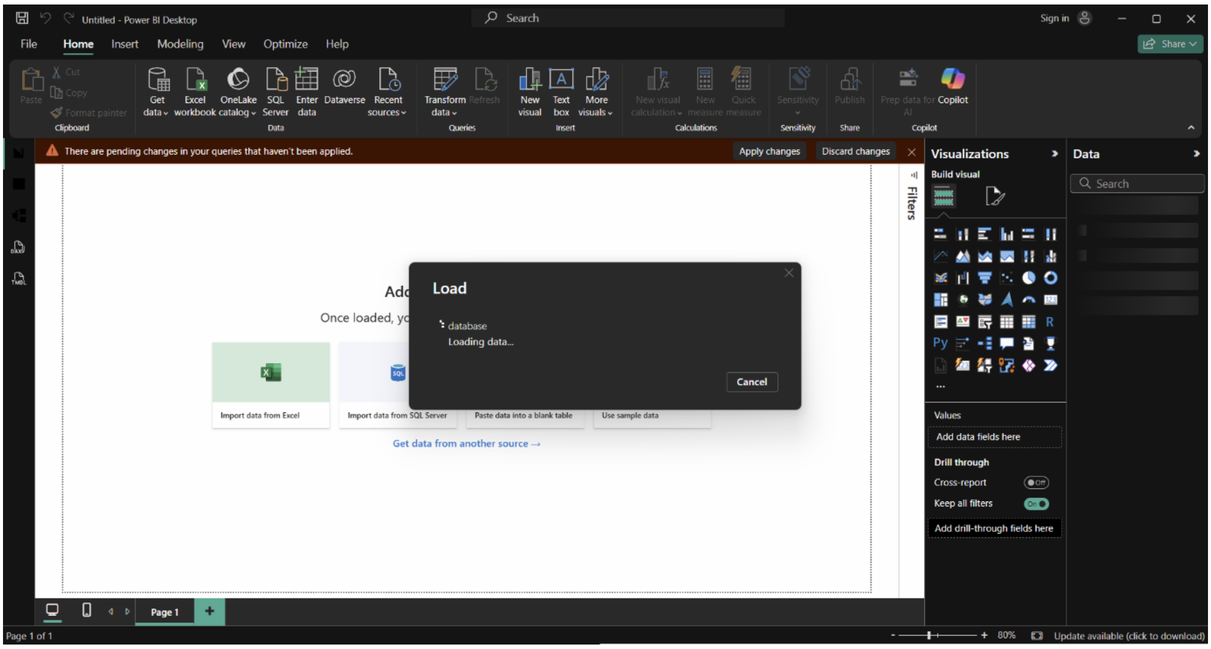
APPLIED STEPS

- Source
- Changed Type
- Promoted Headers
- Changed Type1
- Removed Columns
- Changed Type with Locale
- Changed Type with Locale1
- Changed Type with Locale2
- Changed Type with Locale3
- Replaced Value
- Replaced Value1
- Replaced Value2
- Replaced Value3
- Replaced Value4
- Replaced Value5
- Replaced Value6
- Replaced Value7
- Changed Type2

13 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 11:46 AM

Fig 5: Selecting **Close & Apply** after data cleaning.



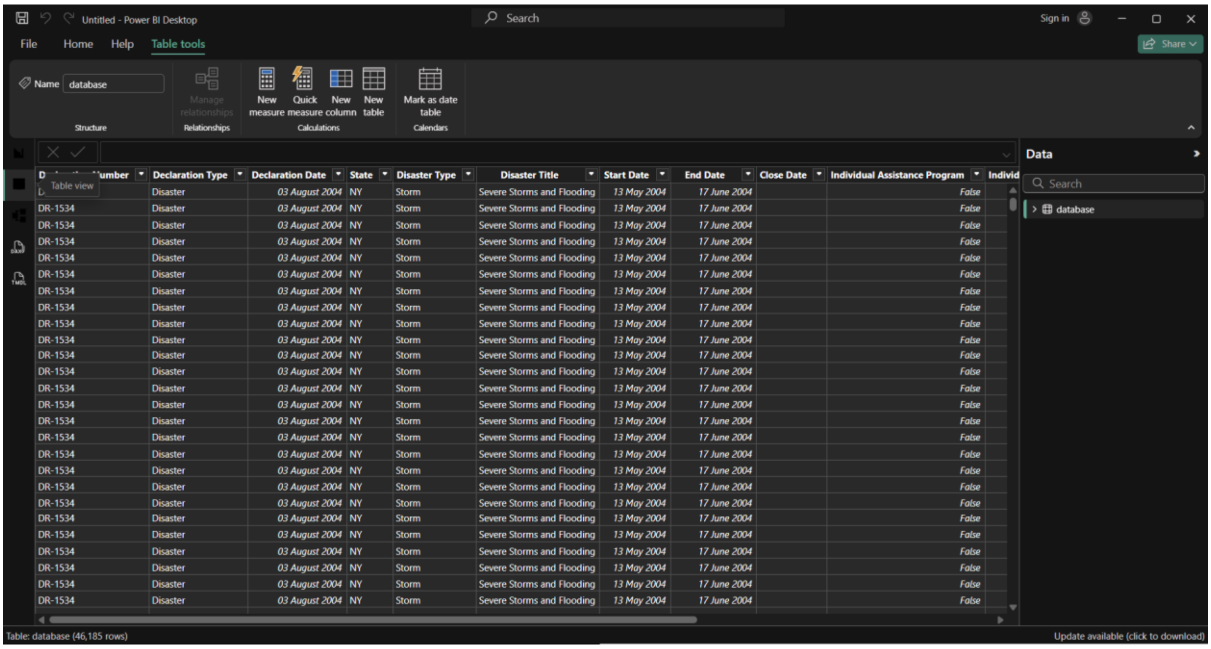
There are pending changes in your queries that haven't been applied.

Load

database

Loading data...

Fig 6: Power BI automatically Loads the Cleaned Data



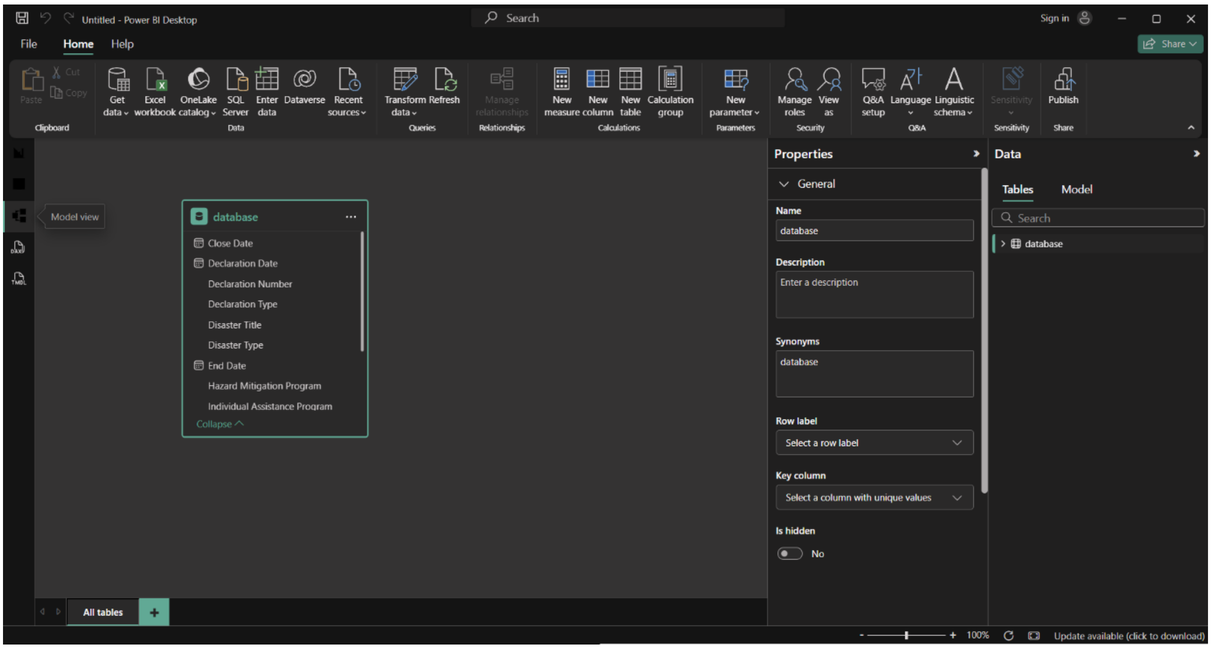
Name: database

D[illegible]umber	Declaration Type	Declaration Date	State	Disaster Type	Disaster Title	Start Date	End Date	Close Date	Individual Assistance Program	Individ
D	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	
DR-1534	Disaster	03 August 2004	NY	Storm	Severe Storms and Flooding	13 May 2004	17 June 2004		False	

Table: database (46,185 rows)

database

Fig 7: Cleaned Data in **Table View**.



database
- Close Date
- Declaration Date
- Declaration Number
- Declaration Type
- Disaster Title
- Disaster Type
- End Date
- Hazard Mitigation Program
- Individual Assistance Program

General

Name: database

Description: Enter a description

Synonyms: database

Row label: Select a row label

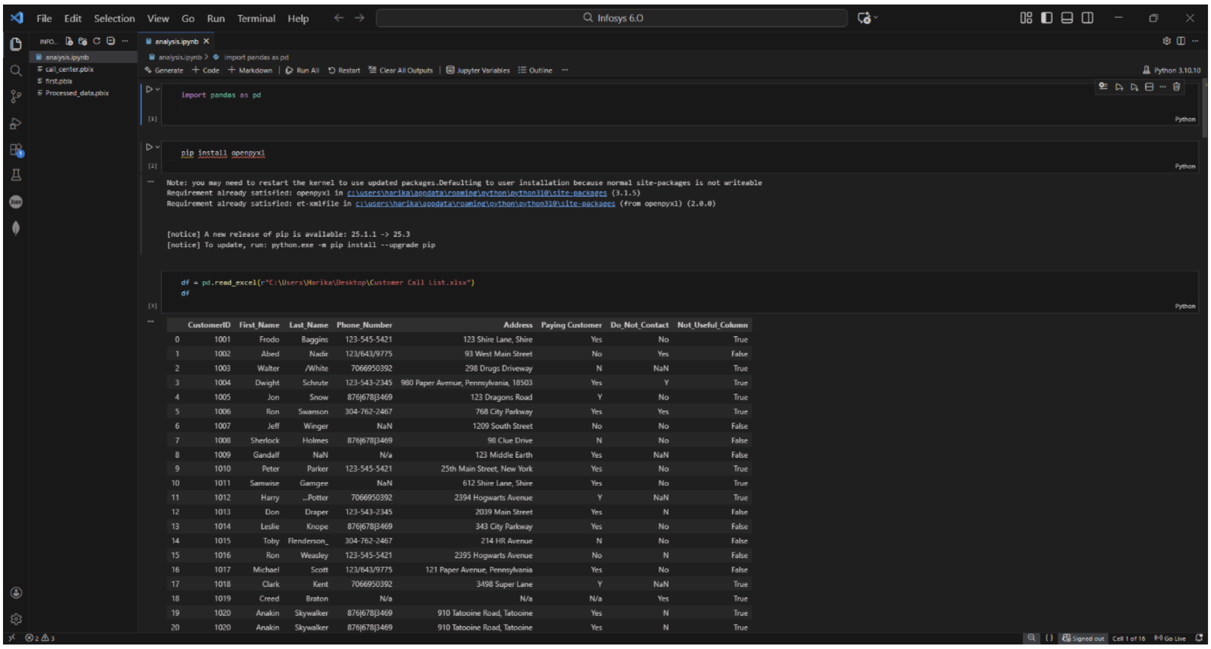
Key column: Select a column with unique values

Is hidden: No

database

Fig 8: Cleaned Data In **Model View**.

DATA CLEANING USING VS CODE (JYPTER NOTEBOOK)



```
import pandas as pd
```

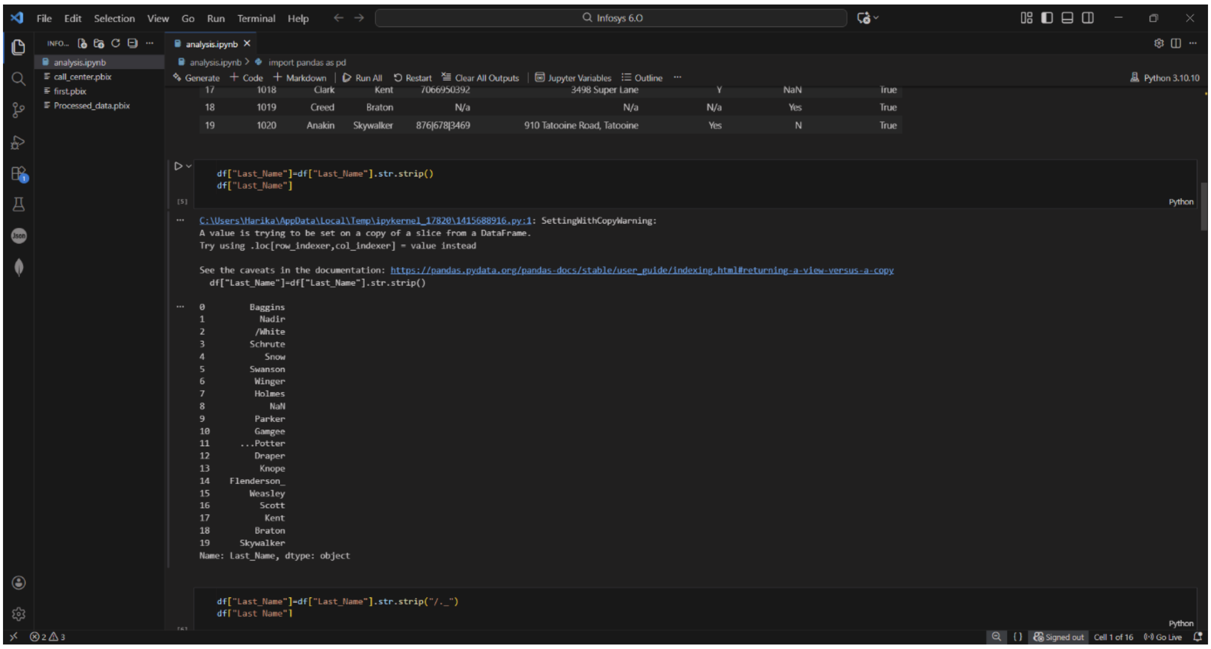
```
pip install openpyxl
```

```
Note: you may need to restart the kernel to use updated packages.Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: openpyxl in c:\users\harika\appdata\roaming\python\python310\site-packages (3.1.5)
Requirement already satisfied: et-xmlfile in c:\users\harika\appdata\roaming\python\python310\site-packages (from openpyxl) (2.0.0)


[notice] A new release of pip is available: 25.1.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
df = pd.read_excel(r"C:\Users\Harika\Desktop\Customer Call List.xlsx")
df
```

	CustomerID	First_Name	Last_Name	Phone_Number	Address	Paying Customer	Do_Not_Contact	Not_Useful_Column
0	1001	Frodo	Baggins	123-545-5421	123 Shire Lane, Shire	Yes	No	True
1	1002	Abed	Nadir	123/643/9775	93 West Main Street	No	Yes	False
2	1003	Walter	/White	7066950392	298 Drugs Driveway	N	NaN	True
3	1004	Dwight	Schrute	123-543-2345	980 Paper Avenue, Pennsylvania, 18503	Yes	Y	True
4	1005	Jon	Snow	876|678|3469	123 Dragons Road	Y	No	True
5	1006	Ron	Swanson	304-762-2467	768 City Parkway	Yes	Yes	True
6	1007	Jeff	Winger	NaN	1209 South Street	No	No	False
7	1008	Sherlock	Holmes	876|678|3469	98 Clue Drive	N	No	False
8	1009	Gandalf	NaN	N/a	123 Middle Earth	Yes	NaN	False
9	1010	Peter	Parker	123-545-5421	25th Main Street, New York	Yes	No	True
10	1011	Samwise	Gamgee	NaN	612 Shire Lane, Shire	Yes	No	True
11	1012	Harry	...Potter	7066950392	2394 Hogwarts Avenue	Y	NaN	True
12	1013	Don	Draper	123-543-2345	2039 Main Street	Yes	N	False
13	1014	Leslie	Knope	876|678|3469	343 City Parkway	Yes	No	False
14	1015	Toby	Flenderson_	304-762-2467	214 HR Avenue	N	No	False
15	1016	Ron	Weasley	123-545-5421	2395 Hogwarts Avenue	No	N	False
16	1017	Michael	Scott	123/643/9775	121 Paper Avenue, Pennsylvania	Yes	No	False
17	1018	Clark	Kent	7066950392	3498 Super Lane	Y	NaN	True
18	1019	Creed	Braton	N/a	N/a	N/a	Yes	True
19	1020	Anakin	Skywalker	876|678|3469	910 Tatooine Road, Tatooine	Yes	N	True
20	1020	Anakin	Skywalker	876|678|3469	910 Tatooine Road, Tatooine	Yes	N	True

Fig 1: Creating a Jypter file using ipynb extension and Loading the Data.



17	1018	Clark	Kent	7066950392	3498 Super Lane	Y	NaN	True
18	1019	Creed	Braton	N/a	N/a	N/a	Yes	True
19	1020	Anakin	Skywalker	876|678|3469	910 Tatooine Road, Tatooine	Yes	N	True

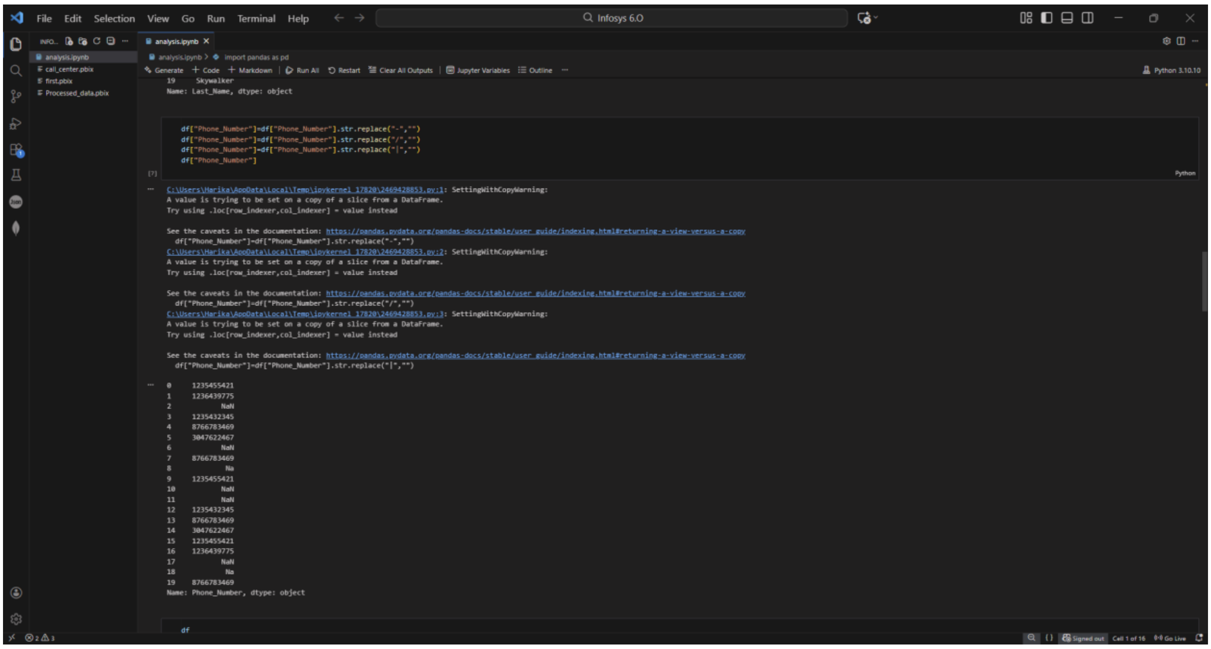
```
df["Last_Name"]=df["Last_Name"].str.strip()
df["Last_Name"]
```

```
C:\Users\Harika\AppData\Local\Temp\ipykernel_17820\1415688916.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Last_Name"]=df["Last_Name"].str.strip()
```

```
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
Name: Last_Name, dtype: object
```

```
df["Last_Name"]=df["Last_Name"].str.strip("/._")
df["Last_Name"]
```

Fig 2: Removing the Unwanted Elements in the column.



```
19      Skywalker
Name: Last_Name, dtype: object
```

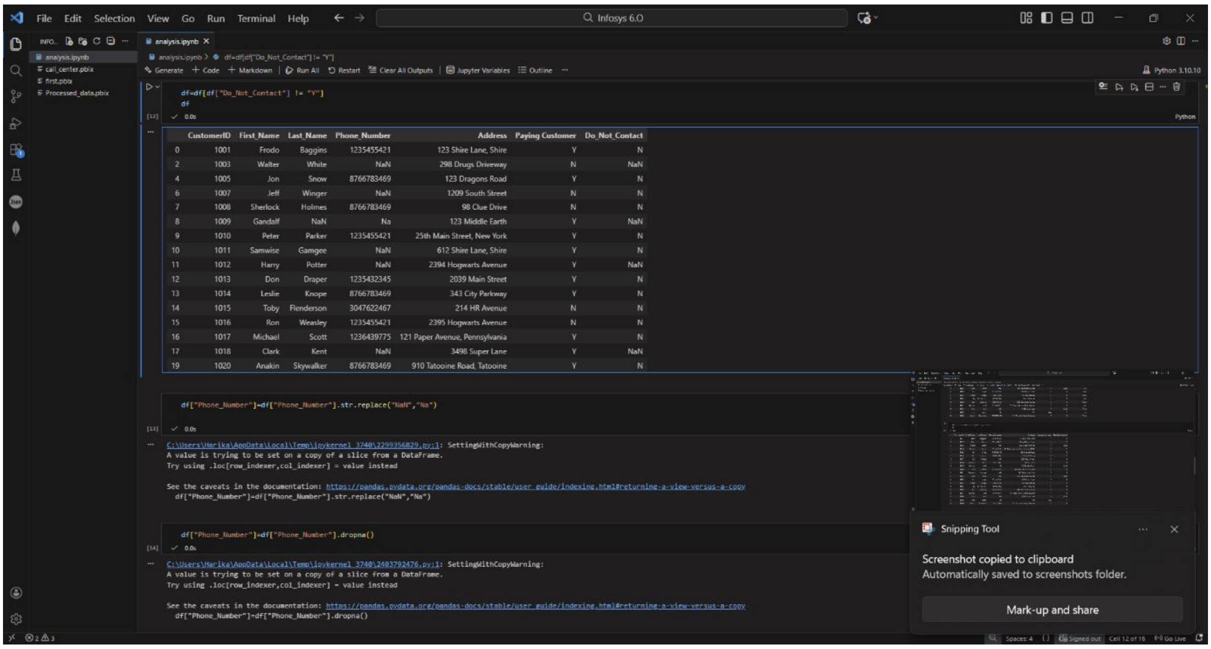
```
df["Phone_Number"]=df["Phone_Number"].str.replace("-","")
df["Phone_Number"]=df["Phone_Number"].str.replace("/","")
df["Phone_Number"]=df["Phone_Number"].str.replace("|","")
df["Phone_Number"]
```

```
C:\Users\Harika\AppData\Local\Temp\ipykernel_17820\2469428853.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Phone_Number"]=df["Phone_Number"].str.replace("-","")
C:\Users\Harika\AppData\Local\Temp\ipykernel_17820\2469428853.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Phone_Number"]=df["Phone_Number"].str.replace("/","")
C:\Users\Harika\AppData\Local\Temp\ipykernel_17820\2469428853.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Phone_Number"]=df["Phone_Number"].str.replace("|","")
```

```
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
Name: Phone_Number, dtype: object
```

```
df
```

Fig 3: Replacing the values with suitable data types.



```
df=df[df["Do_Not_Contact"] != "Y"]
df
```

	CustomerID	First_Name	Last_Name	Phone_Number	Address	Paying Customer	Do_Not_Contact
0	1001	Frodo	Baggins	1235455421	123 Shire Lane, Shire	Y	N
2	1003	Walter	White	NaN	298 Drugs Driveway	N	NaN
4	1005	Jon	Snow	8766783469	123 Dragons Road	Y	N
6	1007	Jeff	Winger	NaN	1209 South Street	N	N
7	1008	Sherlock	Holmes	8766783469	98 Clue Drive	N	N
8	1009	Gandalf	NaN	Na	123 Middle Earth	Y	NaN
9	1010	Peter	Parker	1235455421	25th Main Street, New York	Y	N
10	1011	Samwise	Gamgee	NaN	612 Shire Lane, Shire	Y	N
11	1012	Harry	Potter	NaN	2394 Hogwarts Avenue	Y	NaN
12	1013	Don	Draper	1235432345	2039 Main Street	Y	N
13	1014	Leslie	Knope	8766783469	343 City Parkway	Y	N
14	1015	Toby	Flenderson	3047622467	214 HR Avenue	N	N
15	1016	Ron	Weasley	1235455421	2395 Hogwarts Avenue	N	N
16	1017	Michael	Scott	1236439775	121 Paper Avenue, Pennsylvania	Y	N
17	1018	Clark	Kent	NaN	3498 Super Lane	Y	NaN
19	1020	Anakin	Skywalker	8766783469	910 Tatooine Road, Tatooine	Y	N

```
df["Phone_Number"]=df["Phone_Number"].str.replace("NaN","Na")
```

```
C:\Users\Harika\AppData\Local\Temp\ipykernel_3740\2299356829.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Phone_Number"]=df["Phone_Number"].str.replace("NaN","Na")
```

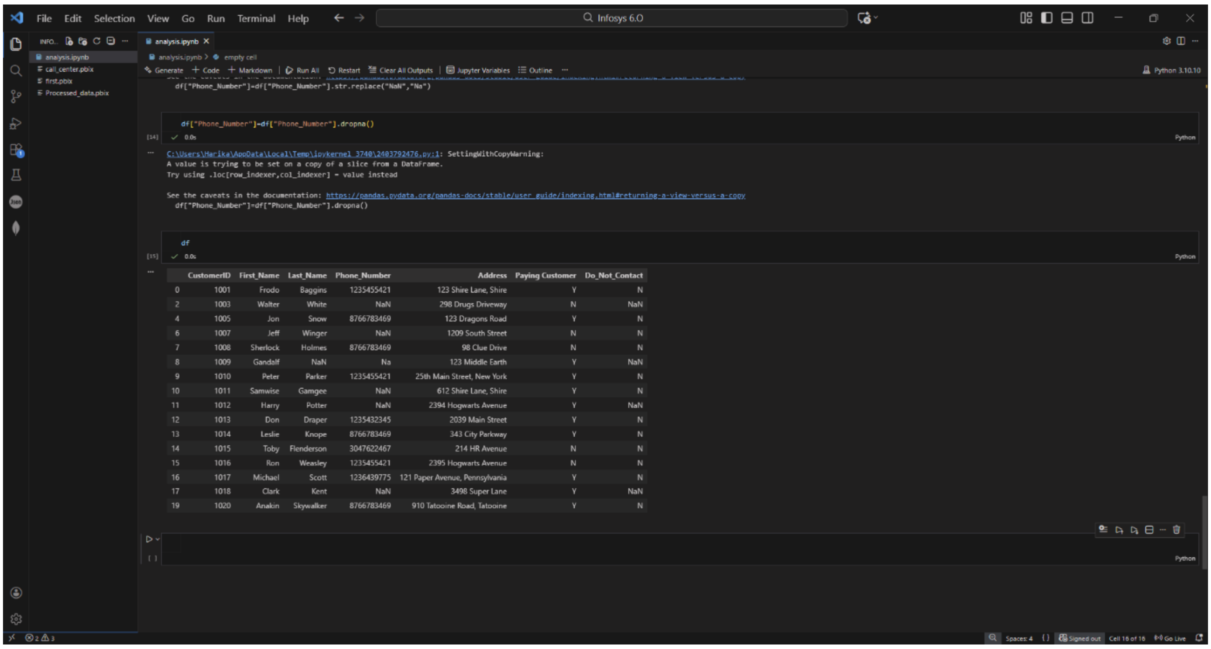
```
df["Phone_Number"]=df["Phone_Number"].dropna()
```

```
C:\Users\Harika\AppData\Local\Temp\ipykernel_3740\2403792476.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Phone_Number"]=df["Phone_Number"].dropna()
```

Snipping Tool

Screenshot copied to clipboard
Automatically saved to screenshots folder.

Fig 4 : Dropping the irrelevant or unused elements in the Data.



```
  df["Phone_Number"]=df["Phone_Number"].str.replace("NaN","Na")
```

```
df["Phone_Number"]=df["Phone_Number"].dropna()
```

[14] ✓ 0.0s

```
C:\Users\Harika\AppData\Local\Temp\ipykernel_3740\2403792476.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["Phone_Number"]=df["Phone_Number"].dropna()
```

```
df
```

[15] ✓ 0.0s

	CustomerID	First_Name	Last_Name	Phone_Number	Address	Paying Customer	Do_Not_Contact
0	1001	Frodo	Baggins	1235455421	123 Shire Lane, Shire	Y	N
2	1003	Walter	White	NaN	298 Drugs Driveway	N	NaN
4	1005	Jon	Snow	8766783469	123 Dragons Road	Y	N
6	1007	Jeff	Winger	NaN	1209 South Street	N	N
7	1008	Sherlock	Holmes	8766783469	98 Clue Drive	N	N
8	1009	Gandalf	NaN	Na	123 Middle Earth	Y	NaN
9	1010	Peter	Parker	1235455421	25th Main Street, New York	Y	N
10	1011	Samwise	Gamgee	NaN	612 Shire Lane, Shire	Y	N
11	1012	Harry	Potter	NaN	2394 Hogwarts Avenue	Y	NaN
12	1013	Don	Draper	1235432345	2039 Main Street	Y	N
13	1014	Leslie	Knope	8766783469	343 City Parkway	Y	N
14	1015	Toby	Flenderson	3047622467	214 HR Avenue	N	N
15	1016	Ron	Weasley	1235455421	2395 Hogwarts Avenue	N	N
16	1017	Michael	Scott	1236439775	121 Paper Avenue, Pennsylvania	Y	N
17	1018	Clark	Kent	NaN	3498 Super Lane	Y	NaN
19	1020	Anakin	Skywalker	8766783469	910 Tatooine Road, Tatooine	Y	N

Fig 5: Loading the Cleaned Dataset.