

In [1]: `import pandas as pd`

In [2]: `data = {
 'date': ['2021-12-01', '01-12-2022', '2022/12/01', '12-01-2021'],
 'country': ['USA', 'U.S.A.', 'America', 'United States'],
 'name': ['anwaar', 'anwar', 'Hamza', 'Hazma'],
 'sales_2020': [100, 200, None, 200],
 'sales_2021': [None, 150, 300, 150]
}`
make pandas dataframe
`df = pd.DataFrame(data)`

In [3]: `df.head()`

Out[3]:

	date	country	name	sales_2020	sales_2021
0	2021-12-01	USA	anwaar	100.0	NaN
1	01-12-2022	U.S.A.	anwar	200.0	150.0
2	2022/12/01	America	Hamza	NaN	300.0
3	12-01-2021	United States	Hazma	200.0	150.0

In [4]: *# standardizing the date format*
`df['date'] = pd.to_datetime(df['date'], errors='coerce')
df['date'] = df['date'].dt.strftime('%Y-%m-%d')
df.head()`

Out[4]:

	date	country	name	sales_2020	sales_2021
0	2021-12-01	USA	anwaar	100.0	NaN
1	NaN	U.S.A.	anwar	200.0	150.0
2	NaN	America	Hamza	NaN	300.0
3	NaN	United States	Hazma	200.0	150.0

```
In [5]: # Harmonize the name of the country
country_mapping = {'USA': 'United States', 'U.S.A.': 'United States', 'America': 'United States'}
df['country'] = df['country'].replace(country_mapping)
df.head()
```

```
Out[5]:
```

	date	country	name	sales_2020	sales_2021
0	2021-12-01	United States	anwaar	100.0	NaN
1	NaN	United States	anwar	200.0	150.0
2	NaN	United States	Hamza	NaN	300.0
3	NaN	United States	Hazma	200.0	150.0

```
In [6]: # Correct the typographical Mistakes in name
# Let's assume we want to correct 'Jonh Doe' to 'John Doe'
df['name'] = df['name'].replace({'anwaar': 'anwar', 'Hazma': 'Hamza'})
df.head()
```

```
Out[6]:
```

	date	country	name	sales_2020	sales_2021
0	2021-12-01	United States	anwar	100.0	NaN
1	NaN	United States	anwar	200.0	150.0
2	NaN	United States	Hamza	NaN	300.0
3	NaN	United States	Hamza	200.0	150.0

```
In [7]: # remove duplicates
df = df.drop_duplicates(subset="name")
df.head()
```

```
Out[7]:
```

	date	country	name	sales_2020	sales_2021
0	2021-12-01	United States	anwar	100.0	NaN
2	NaN	United States	Hamza	NaN	300.0

In [8]: `df.head()`

Out[8]:

	date	country	name	sales_2020	sales_2021
0	2021-12-01	United States	anwar	100.0	NaN
2	NaN	United States	Hamza	NaN	300.0

In [9]: *# 5. Resolving Contradictory Data*
For demonstration, Let's assume sales_2021 should always be higher than sales_2020
We'll remove rows where this condition is not met
`df = df.drop(df[df['sales_2021'] <= df['sales_2020']].index)`
`df.head()`

Out[9]:

	date	country	name	sales_2020	sales_2021
0	2021-12-01	United States	anwar	100.0	NaN
2	NaN	United States	Hamza	NaN	300.0

In []: