# data-cleaning

December 4, 2023

[2]:
```
import pandas as pd
```

[21]:
```
data = pd.read_csv("SA_FR.csv")
```

[22]:
```
data.head(10)
```

[22]:

| | product_id | product_title | retail_price | Disc__perc | disc_price | rating | summary | review | location |
|---|---|---|---|---|---|---|---|---|---|
| 0 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 4 | Nice product | Camera nicely see | Kalpi |
| 1 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Terrific | Perfect condition | Gwalior |
| 2 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Simply awesome | Justify with the price Good mobile for basic u… | Khandwa |
| 3 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Classy product | For this price, the phone is good, not for gam… | Patna |
| 4 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 1 | Waste of money! | Processor is slow | Deoria |
| 5 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Mind-blowing purchase | Super awesome  just love  it thank you so m… | Dimapur |
| 6 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Best in the market! | | |
| 7 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Worth every penny | | |
| 8 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Brilliant | | |
| 9 | MOBGZCQFCWNDK89P | POCO C51 (Royal Blue, 64 GB) | 9,999 | 35 | 6,499 | 5 | Great product | | |

```
6                    Picture quality is very good                Rengali
7  At this price range this smartphone is best..C…   Ramanagara District
8  Good but Back buttons are left , you can't cha…            Kantanagar
9                                               Good  Murshidabad District

          date  upvotes  downvotes
0  4 months ago     1198        398
1  3 months ago      223         64
2  4 months ago     1232        418
3  4 months ago      429        135
4  4 months ago      171         48
5  2 months ago      170         48
6  3 months ago      214         63
7  2 months ago      111         29
8  2 months ago      114         30
9  4 months ago      177         52
```

[23]:
```
data.isnull().sum()
```

[23]:
```
product_id        0
product_title     0
retail_price      0
Disc__perc        0
disc_price        0
rating            0
summary           0
review            9
location         12
date              0
upvotes           0
downvotes         0
dtype: int64
```

[8]:
```
import nltk
nltk.download('stopwords')
nltk.download('vader_lexicon')
import re
from nltk.corpus import stopwords
import string
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

```
[nltk_data] Error loading stopwords: <urlopen error [SSL:
[nltk_data]     CERTIFICATE_VERIFY_FAILED] certificate verify failed:
```

```
[nltk_data]     unable to get local issuer certificate (_ssl.c:1000)>
[nltk_data] Error loading vader_lexicon: <urlopen error [SSL:
[nltk_data]     CERTIFICATE_VERIFY_FAILED] certificate verify failed:
[nltk_data]     unable to get local issuer certificate (_ssl.c:1000)>
```

[ ]:
```python
# Remove unwanted emojis and special symbols
def remove_emojis(text):
    if isinstance(text, float):
        text = str(text)
    emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
                           u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                           u"\U0001F680-\U0001F6FF"  # transport & travel
                           u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                           "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)

def remove_special_symbols(text):
    if isinstance(text, float):
        text = str(text)
    special_symbols = re.compile('[^a-zA-Z0-9 ]')
    return special_symbols.sub('', text)

data['review'] = data['review'].apply(remove_emojis)
data['review'] = data['review'].apply(remove_special_symbols)

# Print the cleaned summary
print(data['review'])
```

```
0                                      Camera nicely see
1                                       Perfect condition
2        Justify with the price Good mobile for basic u…
3        For this price the phone is good not for gamin…
4                                       Processor is slow
                              …
16471              Good budget mobile for my grand mother
16472                                                Good
16473    For calling its a good choice Nokia keypad mob…
16474                                   Worst product ever
16475                                                Good
Name: review, Length: 16476, dtype: object
```

[ ]:
```python
# Save the cleaned DataFrame to a new CSV file
data.to_csv('cleaned_data.csv', index=False)
```

[ ]:
```python

```

[ ]: