```
[nltk_data]   Package stopwords is already
[nltk_data]   Package punkt is already
[nltk_data]   Package wordnet is already
[nltk_data]   Package averaged_perceptron_tagger is already
[nltk_data]   Package maxent_ngram_chunker is already
[nltk_data]   Package words is already
```

```
(1,048,575, 12)
```

| | product_id | category | Review | product_price | Length_of_Review |
|---|---|---|---|---|---|
| 0 | 0 | Home Improvement | 0 | 0 | 0 |
| 1 | 1 | Home Improvement | 1 | 1 | 1 |
| 2 | 2 | Home Improvement | 2 | 2 | 2 |
| 3 | 3 | Home Improvement | 3 | 3 | 3 |
| 4 | 4 | Home Improvement | 4 | 4 | 4 |
| 5 | 5 | Home Improvement | 5 | 5 | 5 |
| 6 | 6 | Home Improvement | 6 | 6 | 6 |
| 7 | 7 | Home Improvement | 7 | 7 | 7 |
| 8 | 8 | Home Improvement | 8 | 8 | 8 |
| 9 | 9 | Home Improvement | 9 | 9 | 9 |
| 10 | 10 | Home Improvement | 10 | 10 | 10 |
| 11 | 11 | Home Improvement | 11 | 11 | 11 |
| 12 | 12 | Home Improvement | 12 | 12 | 12 |
| 13 | 13 | Home Improvement | 13 | 13 | 13 |
| 14 | 14 | Home Improvement | 14 | 14 | 14 |
| 15 | 15 | Home Improvement | 15 | 15 | 15 |
| 16 | 16 | Home Improvement | 16 | 16 | 16 |
| 17 | 17 | Home Improvement | 17 | 17 | 17 |
| 18 | 18 | Home Improvement | 18 | 18 | 18 |
| 19 | 19 | Home Improvement | 19 | 19 | 19 |

```
(1,048,575, 12)
```

```
product_id        0
category          0
Review            0
product_price     0
Length_of_Review  0
dtype: int64
```

```
product_id        1048575
category          1048575
Review            1048575
product_price     1048575
Length_of_Review  1048575
```

```
In [1]: import pandas as pd
```

```
In [2]: df = pd.read_csv('Amazon_Unlocked_Mobile.csv')
```

```
In [3]: df.sample(5)
```

| | Product | Rating | Reviews | Review Length |
|---|---|---|---|---|
| 1 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 2 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 3 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 4 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 5 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |

413 rows × 4 columns

```
In [4]: df.head(5)
```

| | Product | Rating | Reviews | Review Length |
|---|---|---|---|---|
| 0 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 1 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 2 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 3 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |
| 4 | Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular) | 1 | Samsung galaxy mega 6.3 I9200 8GB U.S. Cellular | 15 |

```
In [5]: df.tail(5)
```

| | Product | Rating | Reviews | Review Length |
|---|---|---|---|---|
| 408 | Sony Ericsson C902i Unlocked Quad Band Camera 2MP Bluetooth | 5 | Great product and very fast delivery. | 50 |
| 409 | Sony Ericsson C902i Unlocked Quad Band Camera 2MP Bluetooth | 5 | Great product and very fast delivery. | 50 |
| 410 | Sony Ericsson C902i Unlocked Quad Band Camera 2MP Bluetooth | 5 | Great product and very fast delivery. | 50 |
| 411 | Sony Ericsson C902i Unlocked Quad Band Camera 2MP Bluetooth | 5 | Great product and very fast delivery. | 50 |
| 412 | Sony Ericsson C902i Unlocked Quad Band Camera 2MP Bluetooth | 5 | Great product and very fast delivery. | 50 |

```
In [6]: df.columns
```

```
Out[6]: Index(['Product', 'Rating', 'Reviews', 'Review Length'],
              dtype='object')
```

```
In [7]: df.shape[0], df.shape[1]
```

```
Out[7]: (413, 4)
```

```
In [8]: df.isnull().sum()
```

```
Out[8]: Product          0
        Rating           0
        Reviews          0
        Review Length    0
        dtype: int64
```

```
In [9]: df.describe()
```

| | Rating | Review Length |
|---|---|---|
| count | 413.000000 | 413.000000 |
| mean | 3.811138 | 19.000000 |
| std | 1.543125 | 0.000000 |
| min | 1.000000 | 19.000000 |
| 25% | 3.000000 | 19.000000 |
| 50% | 4.000000 | 19.000000 |
| 75% | 5.000000 | 19.000000 |
| max | 5.000000 | 19.000000 |

```
In [10]: df['Rating'].value_counts()
```

```
Out[10]: Rating
         5    171
         4    102
         3     71
         2     35
         1     34
         Name: count, dtype: int64
```

```
In [11]: df['Review Length'].value_counts()
```

```
Out[11]: Review Length
         19    413
         Name: count, dtype: int64
```

```
In [12]: df['Product'].value_counts()
```

```
Out[12]: Product
         Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular)                 5
         Sony Ericsson C902i Unlocked Quad Band Camera 2MP Bluetooth                5
         Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular)                 5
         Samsung Galaxy Mega 6.3 I9200 8GB Unlocked (U.S. Cellular)                 5
```

```
import pandas as pd
import sys

df = pd.read_csv('data/train.csv', encoding = 'latin1')
df.drop(['Unnamed: 0', 'Unnamed: 0.1'],
axis=1,inplace=True,errors='ignore')
df.shape, df.columns
df.head(2), df.tail(2)
df.info()
df.describe()
```

```
(8000, 7)
Index(['review', 'date', 'rating', 'title', 'product', 'helpful',
       'verified'],
      dtype='object')
```

```
(   review        date  rating  title  product  helpful  verified
0  ...  2016-08-01       5    ...  ...      ...        ...     True
1  ...  2016-07-31       5    ...  ...      ...        ...     True,
      review        date  rating  title  product  helpful  verified
7998  ...  2016-08-01       5    ...  ...      ...        ...     True
7999  ...  2016-08-01       5    ...  ...      ...        ...     True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 7 cols):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   review  8000 non-null   object
 1   date    8000 non-null   object
 2   rating  8000 non-null   int64
 3   title   8000 non-null   object
 4   product 8000 non-null   object
 5   helpful 8000 non-null   int64
 6   verified 8000 non-null  bool
dtypes: bool, int64, object
memory usage: 437.5 KB
```

| | review | date | rating | title | product | helpful | verified |
|---|---|---|---|---|---|---|---|
| 0 | ... | 2016-08-01 | 5 | ... | ... | ... | True |
| 1 | ... | 2016-07-31 | 5 | ... | ... | ... | True |

| | review | date | rating | title | product | helpful | verified |
|---|---|---|---|---|---|---|---|
| 7998 | ... | 2016-08-01 | 5 | ... | ... | ... | True |
| 7999 | ... | 2016-08-01 | 5 | ... | ... | ... | True |

```
df.info()
df.shape
df.columns
df.dtypes
df.head(), df.tail()
df.describe()
df.isnull()
df[df[df.columns[0]].isnull()].shape
df[df[df.columns[1]].isnull()].shape
df[df[df.columns[2]].isnull()].shape
df[df[df.columns[3]].isnull()].shape
df[df[df.columns[4]].isnull()].shape
df[df[df.columns[5]].isnull()].shape
df[df[df.columns[6]].isnull()].shape
df.isnull().sum(axis=1).sum()
df.isnull().sum(axis=1).sum()/df.shape[0]
df.isnull().sum(axis=1).sum()/df.shape[0]*100
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 7 cols):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   review  8000 non-null   object
 1   date    8000 non-null   object
 2   rating  8000 non-null   int64
 3   title   8000 non-null   object
 4   product 8000 non-null   object
 5   helpful 8000 non-null   int64
 6   verified 8000 non-null  bool
dtypes: bool, int64, object
memory usage: 437.5 KB
```

```
(8000, 7)
Index(['review', 'date', 'rating', 'title', 'product', 'helpful',
       'verified'],
      dtype='object')
```

| | review | date | rating | title | product | helpful | verified |
|---|---|---|---|---|---|---|---|
| 0 | ... | 2016-08-01 | 5 | ... | ... | ... | True |
| 1 | ... | 2016-07-31 | 5 | ... | ... | ... | True |

| | review | date | rating | title | product | helpful | verified |
|---|---|---|---|---|---|---|---|
| 7998 | ... | 2016-08-01 | 5 | ... | ... | ... | True |
| 7999 | ... | 2016-08-01 | 5 | ... | ... | ... | True |

```
df.shape
df.info()
df.columns
df.dtypes
df.head(), df.tail()
df.describe()
df.isnull()
df[df[df.columns[0]].isnull()].shape
df[df[df.columns[1]].isnull()].shape
df[df[df.columns[2]].isnull()].shape
df[df[df.columns[3]].isnull()].shape
df[df[df.columns[4]].isnull()].shape
df[df[df.columns[5]].isnull()].shape
df[df[df.columns[6]].isnull()].shape
df.isnull().sum(axis=1).sum()
df.isnull().sum(axis=1).sum()/df.shape[0]
df.isnull().sum(axis=1).sum()/df.shape[0]*100
```

```
(8000, 7)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 7 cols):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   review  8000 non-null   object
 1   date    8000 non-null   object
 2   rating  8000 non-null   int64
 3   title   8000 non-null   object
 4   product 8000 non-null   object
 5   helpful 8000 non-null   int64
 6   verified 8000 non-null  bool
dtypes: bool, int64, object
memory usage: 437.5 KB
```

```
```

```
import pandas as pd

df = pd.read_csv("Flipkart.csv")

df.head()
```

| | Unnamed: 0 | Product_name | Review_Rating | Review_Length |
|---|---|---|---|---|
| 0 | 0 | Lenovo Ideapad Gaming 3 Ryzen 5 Hexa Core 5600H... | Nice under 50k Great performance gud in this ... | 70 |
| 1 | 1 | Lenovo Ideapad Gaming 3 Ryzen 5 Hexa Core 5600H... | Great performance no issues in the last few ... | 51 |
| 2 | 2 | Lenovo Ideapad Gaming 3 Ryzen 5 Hexa Core 5600H... | Great performance but usually if you take flip... | 54 |
| 3 | 3 | DELL Inspiron Ultra Dual Core 3400 - (4 GB/1 ... | Ms ofc is happy and happ product. 👍 👍 | 5 |
| 4 | 4 | DELL Inspiron Ultra Dual Core 3400 - (4 GB/1 ... | Light weight laptop with free amazing features... | 5 |

```
df.tail()
```

| | Unnamed: 0 | Product_name | Review_Rating | Review_Length |
|---|---|---|---|---|
| 1195 | 1195 | MSI GF Series Intel 10th Gen Core i5 Processor... | Great display, accurate colours and the price ... | 5 |
| 1196 | 1196 | MSI GF Series Intel 10th Gen Core i5 Processor... | Superb product Not bought it yet but it looks ... | 5 |
| 1197 | 1197 | MSI GF Series Intel 10th Gen Core i5 Processor... | Awesome | 5 |
| 1198 | 1198 | MSI GF Series Intel 10th Gen Core i5 Processor... | Only one issue with adapter | 5 |
| 1199 | 1199 | MSI GF Series Intel 10th Gen Core i5 Processor... | Worth the money u spend for this monster beast | 5 |

```
df.columns
```

```
Index(['Unnamed: 0', 'Product_name', 'Review_Rating', 'Review_Length'],
      dtype='object')
```

```
df.duplicated().sum()
```

```
0
```

```
df.isnull()
```

```
Unnamed: 0  Product_name  Review_Rating  Review_Length
0       False       False          False          False
1       False       False          False          False
2       False       False          False          False
3       False       False          False          False
4       False       False          False          False
...       ...         ...            ...            ...
1195    False       False          False          False
1196    False       False          False          False
1197    False       False          False          False
1198    False       False          False          False
1199    False       False          False          False

[1200 rows x 4 columns]
```

```
df.shape
```

```
(1200, 4)
```

```
df.info()
```

```
<class pandas.core.frame.DataFrame>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 4 columns):
 #   Unnamed: 0   1200 non-null   int64
 #   Product_name 1200 non-null   object
 #   Review_Rating 1200 non-null  int64
 #   Review_Length 1200 non-null  int64
dtype: int64
```