```
import pandas as pd
```

```
content = pd.read_csv("/content/Content.csv")
reactions = pd.read_csv("/content/Reactions.csv")
reactions_type = pd.read_csv("/content/ReactionTypes.csv")
```

```
content.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  1000 non-null   int64 
 1   Content ID  1000 non-null   object
 2   User ID     1000 non-null   object
 3   Type        1000 non-null   object
 4   Category    1000 non-null   object
 5   URL         801 non-null    object
dtypes: int64(1), object(5)
memory usage: 47.0+ KB
```

```
reactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25553 entries, 0 to 25552
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  25553 non-null  int64 
 1   Content ID  25553 non-null  object
 2   User ID     22534 non-null  object
 3   Type        24573 non-null  object
 4   Datetime    25553 non-null  object
dtypes: int64(1), object(4)
memory usage: 998.3+ KB
```

Clean the dataset "Content"

```
content.drop(columns=["Unnamed: 0","URL","User ID"], inplace=True)
```

```
content['Category'].value_counts()
```

```
technology           71
animals              67
travel               67
culture              63
science              63
fitness              61
food                 61
healthy eating       61
cooking              60
soccer               58
tennis               58
education            57
dogs                 56
studying             55
veganism             48
public speaking      48
Fitness               5
Animals               4
Science               4
"soccer"              3
"culture"             3
Soccer                3
"dogs"                2
Education             2
Studying              2
Travel                2
Food                  2
"veganism"            1
"public speaking"     1
Public Speaking       1
"technology"          1
"cooking"             1
Healthy Eating        1
"studying"            1
"food"                1
Culture               1
"tennis"              1
Technology            1
"animals"             1
Veganism              1
"science"             1
Name: Category, dtype: int64
```

```
content['Category']= content['Category'].replace('"', '', regex=True)
content['Category']=content['Category'].str.lower()
```

```
content = content.rename(columns={"Type": "Content Type"})
```

```
content.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Content ID  1000 non-null   object
 1   Type        1000 non-null   object
 2   Category    1000 non-null   object
dtypes: object(3)
memory usage: 23.6+ KB
```

Clean the dataset "Reactions"

```
reactions.drop(columns=["Unnamed: 0","User ID"],inplace=True)
```

```
reactions = reactions.dropna()
```

```
# Rename the "Type" column to "Reaction Type".
reactions = reactions.rename(columns={"Type": "Reaction Type"})
```

```
reactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 24573 entries, 1 to 25552
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Content ID     24573 non-null  object
 1   Reaction Type  24573 non-null  object
 2   Datetime       24573 non-null  object
dtypes: object(3)
memory usage: 767.9+ KB
```

Clean the dataset "Reaction Types"

```
reactions_type.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Reaction Type  16 non-null     object
 1   Sentiment      16 non-null     object
 2   Score          16 non-null     int64
dtypes: int64(1), object(2)
memory usage: 512.0+ bytes
```

```
reactions_type.drop(columns = ["Unnamed: 0"],inplace = True)
```

```
reactions_type = reactions_type.rename(columns={"Type": "Reaction Type"})
```

We join the 3 files

```
# We join the DataFrames "content" and "reactions" by the column "Content ID".
df = pd.merge(content, reactions, on="Content ID")

# We join the resulting DataFrame with the DataFrame "reactions_type" by the column "Reaction Type".
df = pd.merge(df, reactions_type, on="Reaction Type")
```

```
df.head()
```

	Content ID	Content Type	Category	Reaction Type	Datetime	Sentiment	Score
0	97522e57-d9ab-4bd6-97bf-c24d952602d2	photo	studying	disgust	2020-11-07 09:43:50	negative	0
1	97522e57-d9ab-4bd6-97bf-c24d952602d2	photo	studying	disgust	2021-01-06 19:13:01	negative	0
2	97522e57-d9ab-4bd6-97bf-c24d952602d2	photo	studying	disgust	2021-04-09 02:46:20	negative	0
3	9f737e0a-3cdd-4d29-9d24-753f4e3be810	photo	healthy eating	disgust	2021-03-28 21:15:26	negative	0
4	230c4e4d-70c3-461d-b42c-ec09396efb3f	photo	healthy eating	disgust	2020-08-04 05:40:33	negative	0

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 24573 entries, 0 to 24572
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Content ID     24573 non-null  object
 1   Content Type   24573 non-null  object
 2   Category       24573 non-null  object
 3   Reaction Type  24573 non-null  object
 4   Datetime       24573 non-null  object
 5   Sentiment      24573 non-null  object
 6   Score          24573 non-null  int64
dtypes: int64(1), object(6)
memory usage: 1.5+ MB
```

```
# Group the DataFrame "df" by the column "Category" and count the number of reactions per category
category_counts = df.groupby("Category")["Reaction Type"].count()

# We obtained the top 5 categories with most reactions
top_categories = category_counts.nlargest(5)

# We created a new DataFrame with the top 5 categories with the most reactions.
top_categories_df = pd.DataFrame(top_categories).reset_index()
```

```
top_categories_df
```

	Category	Reaction Type
0	animals	1897
1	science	1796
2	healthy eating	1717
3	food	1699
4	technology	1698

```
# Group the DataFrame "df" by the column "Category" and add the Score by category.
category_scores = df.groupby("Category").agg({"Score": "sum"}).reset_index()

# We sort the categories by Score in descending order and obtain the top 5
top_categories_by_score = category_scores.sort_values(by="Score", ascending=False).head()

# We create a new DataFrame with the top 5 categories with the highest Score.
top_categories_df_score = pd.DataFrame(top_categories_by_score)
```

```
top_categories_df_score
```

	Category	Score
0	animals	74965
9	science	71168
7	healthy eating	69339
12	technology	68738
6	food	66676

```
# We create an ExcelWriter object
writer = pd.ExcelWriter("output.xlsx")

# Save the DataFrame "df" in the first sheet of the Excel file
df.to_excel(writer, sheet_name="Cleaned_Table", index=False)

# We save the DataFrame "top_categories_df" in the second sheet of the Excel file
top_categories_df_score.to_excel(writer, sheet_name="Top_5_Categories", index=False)

# We save the Excel file
writer.save()
```

```
<ipython-input-79-68cd930a59ee>:11: FutureWarning: save is not part of the public API, usage can give unexpected results and will be removed in a future version
  writer.save()
```