In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

load our Dataset into a Python DataFrame

In [2]:

```
Contents_path ="Data/Content.csv"
Reaction_path ="Data/Reactions.csv"
Reactiontype_path ="Data/ReactionTypes.csv"
content_db = pd.read_csv(Contents_path)
reaction_db = pd.read_csv(Reaction_path)
reactiontype_db = pd.read_csv(Reactiontype_path)
```

View our Dataset

In [3]:

```
content_db.head()
```

Out[3]:

	Unnamed: 0	Content ID	User ID	Type	Category	
0	0	97522e57-d9ab-4bd6-97bf-c24d952602d2	8d3cd87d-8a31-4935-9a4f-b319bfe05f31	photo	Studying	https://socialbuzz.cdn.com/co
1	1	9f737e0a-3cdd-4d29-9d24-753f4e3be810	beb1f34e-7870-46d6-9fc7-2e12eb83ce43	photo	healthy eating	https://socialbuzz.cdn.com/cc
2	2	230c4e4d-70c3-461d-b42c-ec09396efb3f	a5c65404-5894-4b87-82f2-d787cbee86b4	photo	healthy eating	https://socialbuzz.cdn.com/co
3	3	356fff80-da4d-4785-9f43-bc1261031dc6	9fb4ce88-fac1-406c-8544-1a899cee7aaf	photo	technology	https://socialbuzz.cdn.com/co
4	4	01ab84dd-6364-4236-abbb-3f237db77180	e206e31b-5f85-4964-b6ea-d7ee5324def1	video	food	https://socialbuzz.cdn.com/co

In [4]:

```
reaction_db.head()
```

Out[4]:

	Unnamed: 0	Content ID	User ID	Type	Datetime
0	0	97522e57-d9ab-4bd6-97bf-c24d952602d2	NaN	NaN	2021-04-22 15:17:15
1	1	97522e57-d9ab-4bd6-97bf-c24d952602d2	5d454588-283d-459d-915d-c48a2cb4c27f	disgust	2020-11-07 09:43:50
2	2	97522e57-d9ab-4bd6-97bf-c24d952602d2	92b87fa5-f271-43e0-af66-84fac21052e6	dislike	2021-06-17 12:22:51
3	3	97522e57-d9ab-4bd6-97bf-c24d952602d2	163daa38-8b77-48c9-9af6-37a6c1447ac2	scared	2021-04-18 05:13:58
4	4	97522e57-d9ab-4bd6-97bf-c24d952602d2	34e8add9-0206-47fd-a501-037b994650a2	disgust	2021-01-06 19:13:01

In [5]:

```
reactiontype_db.head()
```

Out[5]:

	Unnamed: 0	Type	Sentiment	Score
0	0	heart	positive	60
1	1	want	positive	70
2	2	disgust	negative	0
3	3	hate	negative	5
4	4	interested	positive	30

Cleaning Our Dataset

First : clean the Content Dataset

In [6]:

```
content_db.columns
```

Out[6]:

```
Index(['Unnamed: 0', 'Content ID', 'User ID', 'Type', 'Category', 'URL'], dtype='o
bject')
```

we will drop The URL column because it will not provide any insights or assist in our current analysis. and the Unnamed column and it's an index column and the python done it already and UserID it will not effect our analysis and we renamed the column type with content type to better identifying it

In [7]:

```
columns_to_drop =['Unnamed: 0','URL','User ID']
content_db = content_db.drop(columns=columns_to_drop)
content_db.rename(columns = {'Type':'content type'}, inplace = True)
content_db.head()
```

Out[7]:

	Content ID	content type	Category
0	97522e57-d9ab-4bd6-97bf-c24d952602d2	photo	Studying
1	9f737e0a-3cdd-4d29-9d24-753f4e3be810	photo	healthy eating
2	230c4e4d-70c3-461d-b42c-ec09396efb3f	photo	healthy eating
3	356fff80-da4d-4785-9f43-bc1261031dc6	photo	technology
4	01ab84dd-6364-4236-abbb-3f237db77180	video	food

In [8]:

```
print(content_db["content type"].unique())
print(content_db["Category"].unique())
```

```
['photo' 'video' 'GIF' 'audio']
['Studying' 'healthy eating' 'technology' 'food' 'cooking' 'dogs' 'soccer'
 'public speaking' 'science' 'tennis' 'travel' 'fitness' 'education'
 'studying' 'veganism' 'Animals' 'animals' 'culture' '"culture"' 'Fitness'
 '"studying"' 'Veganism' '"animals"' 'Travel' '"soccer"' 'Education'
 '"dogs"' 'Technology' 'Soccer' '"tennis"' 'Culture' '"food"' 'Food'
 '"technology"' 'Healthy Eating' '"cooking"' 'Science' '"public speaking"'
 '"veganism"' 'Public Speaking' '"science"']
```

we found that some values has written between ("")and some are not . we will replace the (double quotation mark) with nothing

In [9]:

```
content_db["Category"] = content_db["Category"].replace('"', '', regex=True)
print(content_db["Category"].unique())
```

```
['Studying' 'healthy eating' 'technology' 'food' 'cooking' 'dogs' 'soccer'
 'public speaking' 'science' 'tennis' 'travel' 'fitness' 'education'
 'studying' 'veganism' 'Animals' 'animals' 'culture' 'Fitness' 'Veganism'
 'Travel' 'Education' 'Technology' 'Soccer' 'Culture' 'Food'
 'Healthy Eating' 'Science' 'Public Speaking']
```

we also found that the same data beginning with capital letter once and in small letter once

In [10]:

```
content_db["Category"] = content_db["Category"].str.capitalize()
print(content_db["Category"].unique())
```

```
['Studying' 'Healthy eating' 'Technology' 'Food' 'Cooking' 'Dogs' 'Soccer'
 'Public speaking' 'Science' 'Tennis' 'Travel' 'Fitness' 'Education'
 'Veganism' 'Animals' 'Culture']
```

Strat Searching about the Null values in coulmns

In [11]:

```
content_db.isna().sum()
```

Out[11]:

```
Content ID      0
content type    0
Category        0
dtype: int64
```

Now it seems great we finish cleaning this Dataset

Start cleaning the Reactions Dataset

In [12]:

```
reaction_db.columns
```

Out[12]:

```
Index(['Unnamed: 0', 'Content ID', 'User ID', 'Type', 'Datetime'], dtype='object')
```

We will drop the Unnamed column becuase it's an index column and the python done it already and UserID it will not effect our analysis

In [13]:

```
columns_to_drop =['Unnamed: 0','User ID']
reaction_db = reaction_db.drop(columns=columns_to_drop)
reaction_db.head()
```

Out[13]:

	Content ID	Type	Datetime
0	97522e57-d9ab-4bd6-97bf-c24d952602d2	NaN	2021-04-22 15:17:15
1	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2020-11-07 09:43:50
2	97522e57-d9ab-4bd6-97bf-c24d952602d2	dislike	2021-06-17 12:22:51
3	97522e57-d9ab-4bd6-97bf-c24d952602d2	scared	2021-04-18 05:13:58
4	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2021-01-06 19:13:01

we change the column name to don't match the column "Type" in Content dataset and we will change the name of thr type column in the reactiontype dataset for easy joining the data on these column

In [14]:

```
reaction_db.rename(columns = {'Type':'reaction type'}, inplace = True)
```

We should know the type of each column

In [15]:

```
reaction_db.dtypes
```

```
Out[15]: Content ID       object
         reaction type    object
         Datetime         object
         dtype: object
```

we should change the dtype of Datetime to Date

```
In [16]: reaction_db["Datetime"] = pd.to_datetime(reaction_db["Datetime"])
         reaction_db.dtypes
```

```
Out[16]: Content ID               object
         reaction type            object
         Datetime         datetime64[ns]
         dtype: object
```

we sould split the Data in another column makes it's easy for analysis

```
In [17]: reaction_db["Date"] = reaction_db["Datetime"].dt.date
         reaction_db['Date'] = pd.to_datetime(reaction_db['Date'])
         reaction_db.dtypes
```

```
Out[17]: Content ID               object
         reaction type            object
         Datetime         datetime64[ns]
         Date             datetime64[ns]
         dtype: object
```

```
In [18]: reaction_db['reaction type'].unique()
```

```
Out[18]: array([nan, 'disgust', 'dislike', 'scared', 'interested', 'peeking',
                'cherish', 'hate', 'love', 'indifferent', 'super love',
                'intrigued', 'worried', 'like', 'heart', 'want', 'adore'],
               dtype=object)
```

It seems the Values it ok but we found it has a null values

We searching for the null values

```
In [19]: reaction_db.isna().sum()
```

```
Out[19]: Content ID         0
         reaction type    980
         Datetime           0
         Date               0
         dtype: int64
```

At first we calculated as the threshold. and it's 5% of the total number of rows and it's the maximum number of missing values that a column can have before it is dropped from the DataFrame

In [20]:

```
threhold = len(reaction_db) * 0.05
print("threhold : ",threhold)
cols_to_drop = reaction_db.columns[reaction_db.isna().sum() <= threhold ]
reaction_db.dropna(subset=cols_to_drop,inplace=True)
reaction_db.isna().sum()
```

```
threhold :  1277.65
```

Out[20]:

```
Content ID       0
reaction type    0
Datetime         0
Date             0
dtype: int64
```

it seems we now handle all the null values

Start cleaning the Reactionstype Dataset

In [21]:

```
reactiontype_db.columns
```

Out[21]:

```
Index(['Unnamed: 0', 'Type', 'Sentiment', 'Score'], dtype='object')
```

We will drop the "Unnamed" column becuase it's an index column and the python done it already and changing the column "type" name

In [22]:

```
columns_to_drop =['Unnamed: 0']
reactiontype_db = reactiontype_db.drop(columns=columns_to_drop)
reactiontype_db.rename(columns = {'Type':'reaction type'}, inplace = True)
reactiontype_db.columns
```

Out[22]:

```
Index(['reaction type', 'Sentiment', 'Score'], dtype='object')
```

In [23]:

```
reactiontype_db.dtypes
```

Out[23]:

```
reaction type    object
Sentiment        object
Score             int64
dtype: object
```

In [24]:

```
print("Types :",reactiontype_db['reaction type'].unique())
print("Sentiments :",reactiontype_db['Sentiment'].unique())
print("Scores :",reactiontype_db['Score'].unique())
```

```
Types : ['heart' 'want' 'disgust' 'hate' 'interested' 'indifferent' 'love'
 'super love' 'cherish' 'adore' 'like' 'dislike' 'intrigued' 'peeking'
 'scared' 'worried']
Sentiments : ['positive' 'negative' 'neutral']
Scores : [60 70  0  5 30 20 65 75 72 50 10 45 35 15 12]
```

It seems everything is ok we finish cleaning our

Datasets

Joining our Datasets

Making a new dataset result joining the Dataset reaction_db with content_db on the Content ID as the an unique identifier in these dataset

In [25]:

```
joining_first = pd.merge(reaction_db, content_db, on='Content ID')
joining_first.head()
```

Out[25]:

	Content ID	reaction type	Datetime	Date	content type	Category
0	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2020-11-07 09:43:50	2020-11-07	photo	Studying
1	97522e57-d9ab-4bd6-97bf-c24d952602d2	dislike	2021-06-17 12:22:51	2021-06-17	photo	Studying
2	97522e57-d9ab-4bd6-97bf-c24d952602d2	scared	2021-04-18 05:13:58	2021-04-18	photo	Studying
3	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2021-01-06 19:13:01	2021-01-06	photo	Studying
4	97522e57-d9ab-4bd6-97bf-c24d952602d2	interested	2020-08-23 12:25:58	2020-08-23	photo	Studying

joining our new dataset with the Dataset reactiontype_db on the reaction type as an unique identifier

In [26]:

```
analysis_dataset = pd.merge(joining_first ,reactiontype_db,on='reaction type')
```

In [27]:

```
analysis_dataset.head()
```

Out[27]:

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

Cleaning our new Dataset

In [28]:

```
analysis_dataset.isna().sum()
```

Out[28]:

```
Content ID        0
reaction type     0
Datetime          0
Date              0
content type      0
Category          0
Sentiment         0
Score             0
dtype: int64
```

We should drop the column Datetime becuase the Date column is enough for analysis our data

In [29]:

```
analysis_dataset = analysis_dataset.drop(columns=['Datetime'])
analysis_dataset.head()
```

Out[29]:

	Content ID	reaction type	Date	content type	Category	Sentiment	Score
0	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2020-11-07	photo	Studying	negative	0
1	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2021-01-06	photo	Studying	negative	0
2	97522e57-d9ab-4bd6-97bf-c24d952602d2	disgust	2021-04-09	photo	Studying	negative	0
3	9f737e0a-3cdd-4d29-9d24-753f4e3be810	disgust	2021-03-28	photo	Healthy eating	negative	0
4	230c4e4d-70c3-461d-b42c-ec09396efb3f	disgust	2020-08-04	photo	Healthy eating	negative	0

Statistical analysis

In [30]:

```
analysis_dataset.describe()
```

Out[30]:

	Date	Score
count	24573	24573.000000
mean	2020-12-16 18:35:49.188133376	39.622553
min	2020-06-18 00:00:00	0.000000
25%	2020-09-16 00:00:00	15.000000
50%	2020-12-17 00:00:00	35.000000
75%	2021-03-17 00:00:00	65.000000
max	2021-06-18 00:00:00	75.000000
std	NaN	26.043011

In [31]:

```
analysis_dataset.to_csv("Data After cleaning.csv",index=False)
```

In [ ]: