Task: To analyse the top performing content categories

 Requirement Gathering: *Reaction*, *Content*, and *Reaction Types* were recognized as relevant data sets for the task.

To clarify why you made this selection:

* The brief carefully it states that the client wanted to see “An **analysis** of their **content categories** showing the **top 5** categories with the largest popularity”.
* As explained in the data model, popularity is quantified by the “Score” given to each reaction type.
* We therefore need data showing the content ID, category, content type, reaction type, and reaction score.
* So, to figure out popularity, we’ll have to add up which content categories have the largest score.

**Data:**

3 files:

Content.xlsx

Reactions.xlsx

Reaction Types.xlsx

**Content.xlsx**

Before cleaning:

Columns: Sl.no, Content ID, User ID, Type, Category, URL (5806,6)

After cleaning: Removed URL, User ID columns, Replaced quotation marks in category

Columns: Sl.no, Content ID, Content Type, Category (4003,4)

User ID and URL column do not add quantitative value to our analysis. Renamed Type as Content Type

**Reactions.xlsx**

Before cleaning:

Columns: Sl. No, Content ID, User ID, Type, Datetime (25554,5)

After cleaning: Removed blanks (989), and User ID column. Renamed Type as Reaction Type

Columns: Sl. No, Content ID, Reaction Type, Datetime (24573,4)

**Reaction Types.xlsx**

Before cleaning:

Columns: Sl. No, Type, Sentiment, Score (17,4)

After cleaning: Renamed Type as Reaction Type

Columns: Sl. No, Reaction Type, Sentiment, Score (17,4)