Data was collected from the client for our analysis.

The following was provided by the client:

- **7 data sets** - each data set contains different columns and values

- **A data model** - this shows the relationships between all of the data sets, as well as any links that can be used to merge tables.

The following steps were followed to answer the business questions -

1. Requirements gathering

The following are 7 datasets collected from the client-

* Dataset 1: User

- ID: Unique ID of the user (automatically generated)

- Name: Full name of user

- Email: Email address of user

* Dataset 2: Profile

- User ID: Unique ID of a user that exists in the User table

- Interests: Interests of the associated user

- Age: Age of the associated user

* Dataset 3: Location

- User ID: Unique ID of a user that exists in the User table

- Address: Full address of the user

* Dataset 4: Session

- User ID: Unique ID of a user that exists in the User table

- Device: Mobile device that they used for this session on the application

- Duration: Amount of time in minutes that this user stayed active on the application during this session

* Dataset 5: Content

- ID: Unique ID of the content that was uploaded (automatically generated)

- User ID: Unique ID of a user that exists in the User table

- Type: A string detailing the type of content that was uploaded

- Category: A string detailing the category that this content is relevant to

- URL: Link to the location where this content is stored

* Dataset 6: Reaction

- Content ID: Unique ID of a piece of content that was uploaded

- User ID: Unique ID of a user that exists in the User table who reacted to this piece of content

- Type: A string detailing the type of reaction this user gave

- Datetime: The date and time of this reaction

* Dataset 7: ReactionTypes

- Type: A string detailing the type of reaction this user gave

- Sentiment: A string detailing whether this type of reaction is considered as positive, negative or neutral

- Score: This is a number calculated by Social Buzz that quantifies how “popular” each reaction is. A reaction type with a higher score should be considered as a more popular reaction

The relevant datasets selected for answering the business question were - **Reaction, Content, and Reaction Types**.

The clarification for selecting these datasets is:

- The brief carefully 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.

2. Data cleaning

The following was performed in Data Cleaning -

- removing rows that have values which are missing,

- changing the data type of some values within a column, and

- removing columns which are not relevant to this task

This data cleaning was done in excel.

3. Data modelling

The following steps were performed in Data Modelling -

3.1. Created a final data set by merging three tables together.

The reaction table was used as a base table then joining the Content data set and Reaction Types data set using VlookUp formula.

3.2. Figure out the Top 5 performing categories

Added up the total scores for each category. Using the “Sum If” formula.

The end result was a spreadsheet which contains:

- A cleaned dataset

- The top 5 categories