**Task 2: Data Cleaning & Modeling**

The client has sent through:

* **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 you can use to merge tables.

There is a lot of information here and it’s easy to get lost in the data. So, to make sure we are using the right data to answer the business questions we’ll follow these steps:

* Requirements gathering
* Data cleaning
* Data modelling

The first step is to **use** **this** **data model to** **identify which datasets will be required to answer your business question -**which is to figure out the**top 5 categories with the largest popularity.**

Definitions of different data types:

* **String** - Sequence of characters, digits, or symbols—always treated as text
* **UUID -** Universally Unique Identifiers
* **Array -** List with a number of elements in a specific order—typically of the same type
* **Integer -** Numeric data type for numbers without fractions
* **Timestamp** - Number of seconds that have elapsed since midnight (00:00:00 UTC), 1st January 1970 (Unix time)

****

We’ve identified *Reaction*, *Content*, and *Reaction Types* as our relevant data sets.

To clarify why we 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.

But! Before we begin to work with the data sets, we’ll need to ensure that the data is clean and ready for analysis…

Data cleaning is a common and very important task when working with data.

****

**Second:**Clean the data by:

* 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.
  + *Think about how each column might be relevant to the business question you’re investigating. If you can’t think of why a column may be useful, it may not be worth including it.*

Now we want to figure out the top 5 categories. To complete your data modelling, follow these steps:

**1. Create a final data set by merging your three tables together**

* We recommend using the Reaction table as your base table, then first join the relevant columns from your Content data set, and then the Reaction Types data set.
* Hint: You can use a “VLookUp” formula

**2. Figure out the Top 5 performing categories**

* Add up the total scores for each category.
* Hint: You can use the “Sum If” formula

The**end result**should be one spreadsheet which contains:

1. A cleaned dataset
2. The top 5 categories