Task 2: Data Cleaning & Modeling

Let's dive into the data

Now you have a good understanding of the project and your role - it's time to get to work!

Don't worry if you haven't done data analysis before, we'll take you through each step and provide support along the way.

So, let's have a look at what data you have to work with. 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 you are using the right data to answer the business questions you'll follow these steps:

1. Requirements gathering
2. Data cleaning
3. Data modelling

First up, requirements gathering

As we mentioned, you have been sent 7 datasets and a data model.

Often you won't need all these datasets to find what you're looking for.

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

When you think you've identified the right data sets to include, complete the multi choice quiz to move onto the next step.

Data Model

Click to download file →

Data sets - Quick Explanation

Great work! You've identified *Reaction*, *Content*, and *Reaction Types* as our relevant data sets.

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.

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

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

What you need to do:

First: Open the three data sets below

Reaction Types

Click to download file →

Reactions

Click to download file →

Content

Click to download file →

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.*

Your end result should be three cleaned data sets.

If you get stuck, we'll provide some guidance in the next step. But we encourage you to give it a go first!

Data Modelling

Okay, we're nearly there! You're doing a great job.

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

Once you have a final data file, upload it to complete this task! We'll provide you with some explanation videos in the next step - but first give it a go to see if you can figure it out.

You can use Excel or any other tool of your choice to create your final data set.