Project Work:

Reading the brief to: -

* Understand the client and business problem at hand.
* Identify the requirements that need to be delivered for this project.
* Identify which tasks you should focus on as a Data Analyst.

let’s have a look at what data I 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.

To make sure I am using the right data to answer the business questions I’ll follow these steps:

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

First up, requirements gathering: -

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

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)

The required data sets needed to complete the analysis are

Ans- Reaction, Content, Reaction Types

To clarify why I 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…

**Second:**Cleaning 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.

Imported the data in Excel Power Query and performed data exploration, data cleaning and data transformation operations.

Solution: -

* Loaded all Required files to Excel power query editor.
* Merged query as (Content + Reaction) as Left Outer join and then this with ReactionTypes as Left Outer join because we need all rows from Contents table to analyze its reaction scores.
* Cleaning data for blank, duplicate, mismatching categorical type data and see if the data types are correctly assigned.
* Removing Content\_UserID, Conternt\_URL, Reaction\_UserID because I am not going to use this for any kind of analysis.
* Content\_Type is of 4 types (audio, video, image and GIF).
* Content\_Category contains some value with double quotation and some without so changing all values with no double quotation also some values are in proper case and some are in lower case so changing everything in Proper case.
* Other columns are with correct data type and with correct values.
* Applying this logic to divide daytime into different sections to analyze wrt different day sessions.
* Morning: 6 AM–12 PM
* Afternoon: 12 PM–6 PM
* Evening: 6 PM–9 PM
* Late evening/night: 9 PM–12 AM
* Late night: 12 AM–6 AM

Building presentation Structure

For each slide, thought about:

* Agenda - What will my presentation cover?
* Project Recap - What are the key points from the brief?
* Problem - What is the problem that needed to be answered in this presentation?
* The Analytics team - Who is on my team?
* Process - How did I completed my analysis?