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

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](https://cdn.theforage.com/vinternships/companyassets/T6kdcdKSTfg2aotxT/MsAqi7SNLKw3C6LAr/1664298399720/ReactionTypes.csv" \t "_blank)**

**[Click to download file →](https://cdn.theforage.com/vinternships/companyassets/T6kdcdKSTfg2aotxT/MsAqi7SNLKw3C6LAr/1664298399720/ReactionTypes.csv" \t "_blank)**

**Reactions**

**Click to download file →**

**[Content](https://cdn.theforage.com/vinternships/companyassets/T6kdcdKSTfg2aotxT/MsAqi7SNLKw3C6LAr/1664298350004/Content.csv" \t "_blank)**

**[Click to download file →](https://cdn.theforage.com/vinternships/companyassets/T6kdcdKSTfg2aotxT/MsAqi7SNLKw3C6LAr/1664298350004/Content.csv" \t "_blank)**

**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!**

1. In the content table ‘category’ column have duplicate category values means animal category has written in many different form like ‘Animal’ , ‘animal’ , animal , etc.

**Solution** : Click transform data – Home tab(transform section) – Replace value .

1. In the “reaction” table ‘User ID’ and ‘Type’ column have some null values .

* I have noticed one thing that If I remove null values from the user id then many of values which are not null in the ‘Type’ column are also getting removed . But If I removed the null values from ‘Type’ column only then only null values are getting removed from ‘User ID’ column .

**Solution:** So, I am thinking to apply second step by filtering null values from ‘Type’ column.

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.

* I am using power bi so, it automatically make the modelling correctly . The content table and rection both table have content id column he connected relation with them with ‘many to 1 ‘ relationship . Then reaction table and reaction type table had the reaction type column common so, he find out the relationship with it and joined.
* Now , I have to merge this three tables together so, I like to merge this table with the help of merge queries (present in transform data).

Build your presentation Structure

Importantly, before you start creating your data charts, you should plan your presentation structure. This will ensure your presentation answers the right questions.

Here is a template that we would use at Accenture to create a presentation. You need to **download this template and populate slides 2-6.**

For each slide, think about:

* Agenda - What will your presentation cover?
* Project Recap - What are the key points from the brief?
* Problem - What is the problem that you answer in this presentation?
* The Analytics team - Who is on your team?
  + As a reminder from the earlier task - this includes: Andrew Fleming (Chief Technical Architect), Marcus Rompton (Senior Principle), and yourself!
* Process - How did you complete your analysis?

**Once you’ve populated slides 2-6 - complete the quick knowledge check to move onto the next step.**

In the next step, we’ll review what you’ve included before we start charting.

If you can't open the powerpoint template, here is a PDF version for you to use:

Create your charts

The final step of this task is**to populate slides 7 - 10 with your data insights.**

If you’ve lost your spreadsheet, we’ve attached one here again for you to use.

**[Cleaned data set for analysis](https://cdn.theforage.com/vinternships/companyassets/T6kdcdKSTfg2aotxT/MsAqi7SNLKw3C6LAr/1664564187354/Task%203_Final%20Content%20Data%20set.csv" \t "_blank)**

**[Click to download file →](https://cdn.theforage.com/vinternships/companyassets/T6kdcdKSTfg2aotxT/MsAqi7SNLKw3C6LAr/1664564187354/Task%203_Final%20Content%20Data%20set.csv" \t "_blank)**

We know that the client wants to understand the top 5 content categories. Have a think about what are the best ways to present this? In a pie chart? In a bar chart?

There are also some other interesting insights that we might want to share with them. For example:

* How many unique categories are there?
* How many reactions are there to the most popular category?
* What was the month with the most posts?

If you get stuck, [here](https://www.excel-easy.com/data-analysis/charts.html) is a resource on how to make charts in excel.

**Once your slides are complete with your data insights and charts, upload your work to see the example answer in the next step.**