Accenture Data Analytics and Visualization Virtual Experience

Task 1: Project Understanding

This task involves getting acquainted with the company, their situation and what they want to do.

Nothing complicated but you must exercise some active listening because it will help you in task 2.

Task 2: Data Cleaning and Modeling

For this task, you have to identify what is your objective with the datasets the client provided and then build an analytical file to obtain the insights.

The cleaning processes I performed on the data were:

From Content (1) file:

1) Removed the index column
2) Capitalized the first letter of each word keeping GIF all capitals.
3) Removed the quotation marks / Capitalized the first letter of each word / removed duplicates
4) Twenty percent of records in the column URL are missing but losing the other columns is worse than keeping the empty URL's.

From Location (1) file:

1) Removed the index column
2) Split the address in street name and number, city, state, and Zip code.

The extra thought: As the client's brief explained, we are using the events' buzz to attract interactions with the post, having information on the user's location would give us an edge to monetize it.

Think travel agencies, transportation companies, among others.

From Profile (1) file:

1) Removed the index column
   From the interest column
2) Transformed the interest column containing lists into individual colums.
3) Eliminated the numerical lists
4) Eliminated the repeated interest (e.g. ['science', 'studying', 'culture', 'culture'])
5) AS 28% of the records contain an invalid age (Nearly all social networking sites only allow users aged 13 and over because Children's Online Privacy Protection Act (COPPA), the age values were removed and replaced with the average age and rounded to the closest integer.

The extra thought:

a) The site's nature is showing events and sometimes those events involve the use and advertising of smoking cigarettes, drinking alcohol and in some jurisdictions, cannabis use. Therefore, the validation process should be restrictive for individuals under the legal age set by the local jurisdiction.
b) Even if records with less than a registered age of 10 years old are very unlikely to be actually done by kids, it's necessary to review the sign-up process to avoid falling into legal sanctions.

From Reactions (1) file:

1) Removed the index column
2) Almost 12% (3.018) of the records on the USER ID column are empty, so it's possible that were posted by someone or something not registered.
   The register process should be scrutinized to avoid fake reactions. Almost 1 out of 3 reactions are posted by unregistered users (980) and that affect the popularity of the posts.

From ReactionsType (1) file:

1) Removed the index column
2) Type, sentiment and score adjusted accordingly. (e.g. *interest* is a positive sentiment but the score ranks it below a neutral one)

Originally

Type	Sentiment	Score
Super Love	Positive	75
Adore	Positive	72
Want	Positive	70
Cherish	Positive	70
Love	Positive	65
Heart	Positive	60
Like	Positive	50
Intrigued	Positive	45
Peeking	Neutral	35
Interested	Positive	30
Indifferent	Neutral	20
Scared	Negative	15
Worried	Negative	12
Dislike	Negative	10
Hate	Negative	5
Disgust	Negative	0

Adjusted

Type	Sentiment	Score
Super Love	Positive	75
Adore	Positive	72
Want	Positive	70
Cherish	Positive	70
Love	Positive	65
Heart	Positive	60
Like	Positive	50
Intrigued	Positive	45
Interested	Positive	35
Peeking	Neutral	30
Indifferent	Neutral	20
Worried	Negative	15
Dislike	Negative	12
Scared	Negative	10
Hate	Negative	5
Disgust	Negative	0

From Session (1) file:

1) Removed the index column
2) Removed the devices' brands that does not exist.
3) Removed the sessions where the Duration = 0

The extra thought:

There should exist a table with records of each user interaction. Bases on the information available for the exercise, we are assuming the average time that each user spend on the platform is the same every time they log in and vote on the posted events.

From User (1) file:

1) Removed the index column
2) Removed the invalid email address (does not contain the @ symbol) 2.6% (13 records)
3) The registered emails are incorrect because one provider can't have two identical email addresses for different people and there are 257 records (51.4%) duplicated.

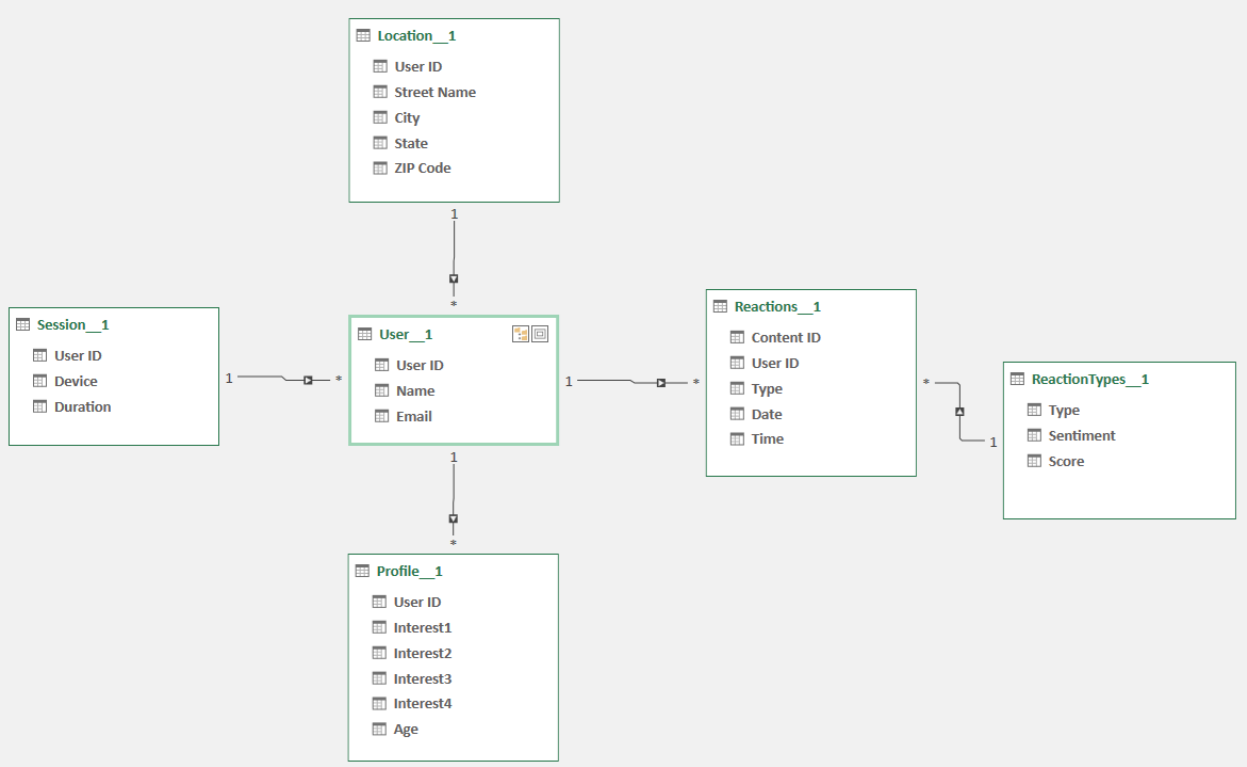

Figure 1 Data Model – Power Pivot

Task 3: Data Visualization & Storytelling

Task 4: Present to the client: YouTube Video