Social Media Analytics

- **Samruddhi Pawar**

Introduction

I enrolled in the Accenture Data Analytics virtual internship via Forage, serving as a data analyst to support the organization “Social Buzz” in analyzing their data. My goal was to assist them in unlocking the potential of their extensive data resources. The project encompassed three key tasks: understanding the project, cleaning and modeling the data, and creating data visualizations and storytelling.

Task 1: Project Understanding

Client Information:

- Client Name: Social Buzz
- Client Industry: Social media and Content Creation
- Establishment Year: 2010
- Employee Count: 250

Client Background:

Social Buzz was founded in 2010 by two former engineers, one from London and the other from San Francisco, who had previously worked for a major social media conglomerate. Departing from their roles in 2008, they joined forces in San Francisco to establish their venture. The motivation behind founding Social Buzz stemmed from their recognition of an opportunity to expand upon the groundwork laid by their previous employer. Their vision was to create a novel platform that places content at the forefront.

Emphasizing the significance of content, Social Buzz adopts a unique approach by maintaining user anonymity and exclusively tracking user reactions to each piece of content. The platform boasts a diverse array of over 100 reaction options, surpassing conventional responses like likes, dislikes, and comments. This distinctive approach ensures that trending content, rather than individual users, occupies the spotlight in user feeds.

Problem Statement:

In the last 5 years, Social Buzz has rapidly grown, attracting over 500 million active users monthly. The unexpected scale of their success has prompted the need for assistance from an advisory firm to oversee and streamline their ongoing scaling process effectively.

Due to their rapid growth and the digital nature of their core product, the amount of data that they create, collect, and analyze is huge. Every day over 100,000 pieces of content, ranging from text, images, videos, and GIFs are posted. All of this data is highly unstructured and requires extremely sophisticated and expensive technology to manage and maintain. Out of the 250 people working at Social Buzz, 200 of them are technical staff working on maintaining this highly complex technology.

Up until this point, they have not relied on any third-party firms to help them get to where they are. However, there are 3 main reasons they are now looking at bringing in external expertise:

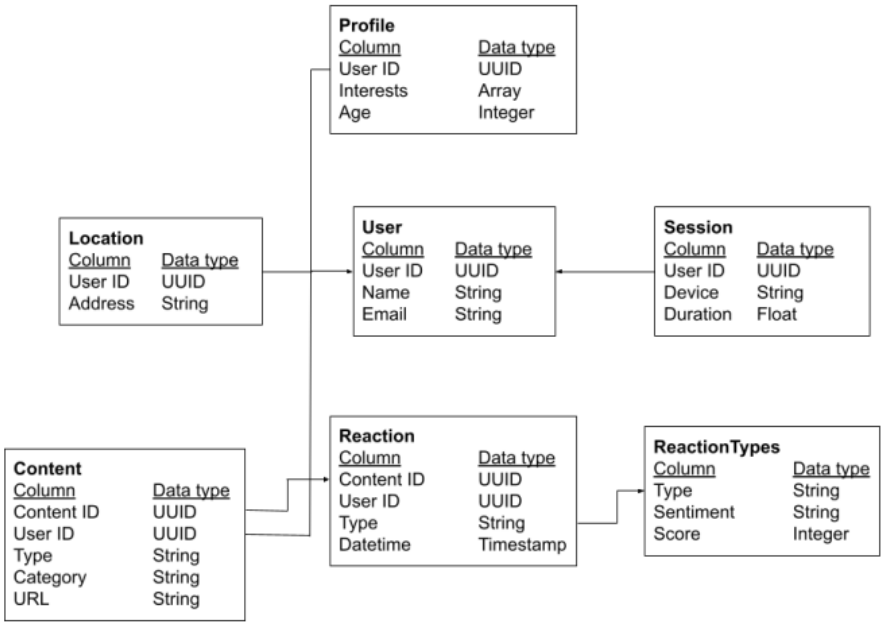
1. They are looking to complete an IPO (Initial Public Offering — this happens when a private company becomes public by selling its shares on a stock exchange) by the end of next year and need guidance to ensure that this goes smoothly.
2. They are still a small company and do not have the resources to manage the scale that they are currently at. They could hire more people, but they want an experienced practice to help instead.
3. They want to learn data best practices from a large corporation. Due to the nature of their business, they have a massive amount of data so they are keen on understanding how the world's biggest companies manage the challenges of big data.

My Task as a Data Analyst:

Analysis of sample data sets with visualizations to understand the popularity of different content categories.

Task 2: Data Cleaning and Modelling

I was provided with seven datasets and a data model. My first step is to use this data model to identify which datasets will be required to answer my business question — figuring out the top 5 categories with the largest popularity.



Data Model

Below are details about the datasets and the columns they contain.

- **User**
  The User table comprises three columns: **ID**, representing the unique user identification (automatically generated); **Name**, indicating the full name of the user; and **Email**, containing the email address of the user.

- **Profile**
  The Profile table consists of three columns: **User ID**, representing the unique identification of a user existing in the User table; **Interests**, containing the interests of the associated user; and **Age**, indicating the age of the associated user.

- **Location**
  The Location table includes two columns: **User ID**, which is the unique identification of a user existing in the User table, and **Address**, representing the full address of the user.

- **Session**
  The Session table comprises three columns: **User ID**, denoting the unique identification of a user existing in the User table; **Device**, specifying the mobile device used by the user for the session on the application; and **Duration**, indicating the amount of time in minutes that the user remained active on the application during this session.

- **Content**
  The Content table encompasses five columns: **ID**, representing the automatically generated unique identification of the uploaded content; **User ID**, denoting the unique identification of a user existing in the User table; **Type**, a string describing the type of uploaded content; **Category**, a string specifying the relevant category of the content; and **URL**, providing a link to the location where the content is stored.

- **Reaction**
  The Reaction table includes four columns: **Content ID**, representing the unique identification of an uploaded piece of content; **User ID**, denoting the unique identification of a user existing in the User table who reacted to this content; **Type**, a string describing the type of reaction given by the user; and Datetime, indicating the date and time of this reaction.

- **ReactionTypes**
  The ReactionTypes table consists of three columns: **Type**, which is a string specifying the type of reaction given by a user; **Sentiment**, a string indicating whether this reaction type is categorized as positive, negative, or neutral; and **Score**, a numerical value calculated by Social Buzz that quantifies the “popularity” of each reaction. A reaction type with a higher score is deemed more popular.

For my analysis, I'll be focusing solely on the **Reaction, Content, and Reaction Types tables**, disregarding the other four tables in the dataset.

The reasons below state why I chose these 3 datasets among the 7 available.

- The business task clearly stated 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.

However, before delving into the analysis of the datasets, it is essential to ensure that the data is clean and prepared for analysis.

Data Cleaning

In the Reaction table,

- I checked for duplicate rows
- I deleted the User ID column, it is not relevant to this task
- I deleted the blanks in the **type** column
- I formatted the **DateTime** column as 'date'.

In the Content table,

- I checked for duplicate rows
- In the **category** column, I noticed that some cells contained double quotes, such as "animals" and animals. To standardize the entries, I utilized the Find and Replace option in Excel, removing every instance of a quote.
- I removed columns that are not relevant to this task, the URL and User ID columns.

In the Reation Type table,

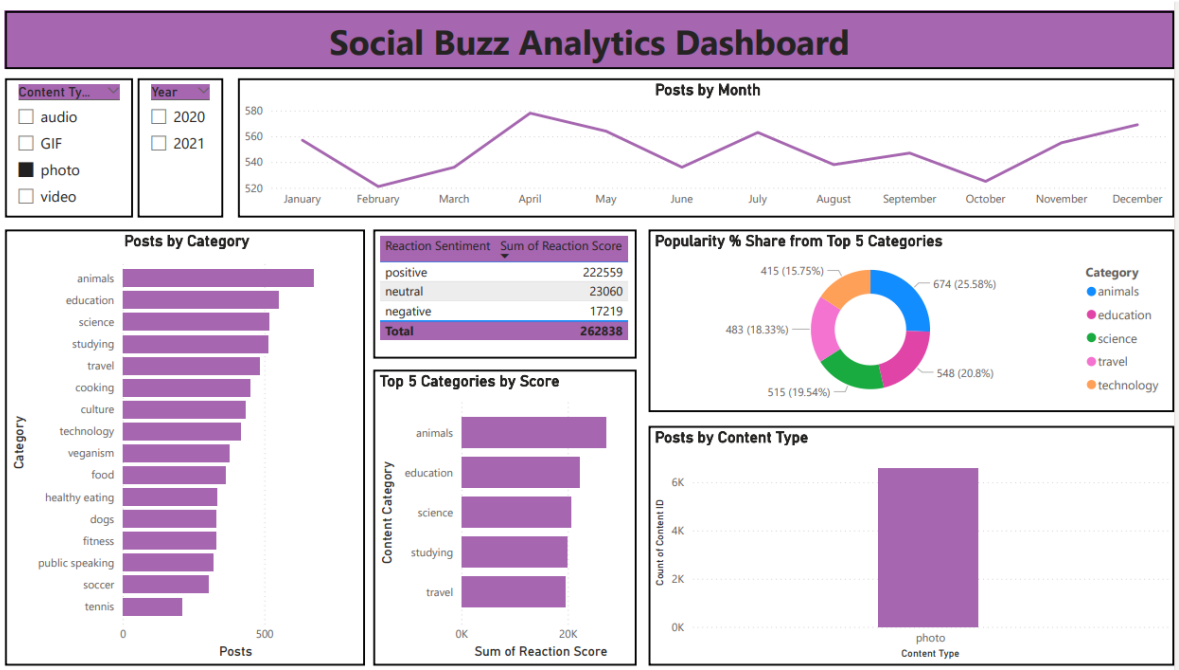
- The table didn't need cleaning.

I consolidated the three tables into a final dataset through a process of merging using MS Excel. Below is the final dataset.

Content ID	Datetime	Reaction Ty	Reaction Sentiment	Reaction Score	Content Ty	Content Categ
97522e57-d9ab-4bd6-97bf-c24d952602d2	07-11-2020 09:43	disgust	negative	0	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	17-06-2021 12:22	dislike	negative	10	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	18-04-2021 05:13	scared	negative	15	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	06-01-2021 19:13	disgust	negative	0	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	23-08-2020 12:25	interested	positive	30	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	07-12-2020 06:27	peeking	neutral	35	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	11-04-2021 17:35	cherish	positive	70	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	27-01-2021 08:32	hate	negative	5	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	01-04-2021 22:54	peeking	neutral	35	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	04-08-2020 05:05	love	positive	65	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	07-11-2020 08:36	indifferent	neutral	20	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	02-11-2020 06:28	scared	negative	15	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	01-11-2020 01:16	interested	positive	30	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	07-10-2020 18:39	hate	negative	5	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	03-09-2020 18:51	scared	negative	15	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	24-02-2021 05:09	super love	positive	75	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	23-09-2020 06:24	peeking	neutral	35	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	24-02-2021 11:37	indifferent	neutral	20	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	22-05-2021 19:44	interested	positive	30	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	31-01-2021 16:03	intrigued	positive	45	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	20-11-2020 17:26	peeking	neutral	35	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	11-04-2021 20:47	worried	negative	12	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	13-06-2021 16:46	like	positive	50	photo	studying
97522e57-d9ab-4bd6-97bf-c24d952602d2	11-04-2021 14:29	heart	positive	60	photo	studying

Task 3: Data Visualization and Storytelling

The dataset comprises 16 distinct content categories and 4 content types.



Insights and Recommendations:

1. Given the popularity of “animals” and “science” as the top content categories, it is recommended to prioritize and strategically incorporate more content related to these themes. Leveraging real-life and factual content within these categories aligns well with audience preferences. Consider diversifying the content within these categories to maintain engagement, potentially exploring subtopics, trending

subjects, or interactive formats to enhance user experience and further capitalize on the audience's interest in authentic and informative content.

2. The most posts were made in May and January.
3. Healthy eating and food both appeared in the top 5 categories by score. This may indicate the audience within your user base. This insight could be used to create a campaign and work with healthy eating brands to boost user engagement.

Conclusions

In conclusion, delving into the world of social buzz's social media analytics opens up a realm of possibilities for understanding user behavior, content preferences, and overall engagement. The insights gained from the data analysis serve as a compass, guiding decisions toward strategies that resonate with their audience.

So, let's embrace the insights, adapt to the changing dynamics, and continue to elevate our social media presence through the lens of analytics. The data tells a story, and it's up to us to craft a narrative that resonates, inspires, and evolves with the ever-shifting currents of the digital landscape.