SOCIAL BUZZ – Data Analysis

Dashboard Link

I Used Excel (Basic Data Cleaning), Power BI (ETL process) and Chat GPT (to gather efficient suggestions for data visualization) to complete this project.

Project Understanding:

Social Buzz is a fast-growing tech unicorn that need to adapt quickly to their global scaling process efficiently. Accenture will be running 3 months POC, and the Data Team will be focusing on an analysis of their content categories that highlights the top 5 categories with the largest aggregate popularity. Social Buzz has over 100,000 posts per day, that can be assumed to have 36,500,000 posts per year. We will analyse the sample of this dataset from them.

Data understanding:

- Client Social Buzz's data is provided in three raw CSV files. The "Content" file includes 1000 rows and five columns: 'Content ID' (1000 distinct values), 'User ID' (446 distinct values), 'Type' (4 distinct values), 'Category' (16 distinct values), and 'URL' for the posted content.
- The "Reactions Type" file has 16 rows and three columns: 'Type' (16 distinct values), 'Score' for each reaction type, and 'Sentiment' categorizing reactions into positive (9 types), neutral (2 types), and negative (5 types).
- The "Reaction" file, with 25553 rows and four columns, captures user interactions: 'User ID' reacting to a post, 'Date Time' of the reaction, 'Content ID,' and the reaction 'Type.'

Data Cleaning & Modelling:

In this step I first used MS excel to do the Data cleaning like adjusting column width, removing duplicate values, Data formatting, removing unwanted columns, and deleting rows with blank spaces. Since Power BI has more advanced data transformation options, I uploaded the cleaned individual excel files into the Power BI as tables.

Processes done in Power BI:

Data Transformation:

- Combined multiple tables into one single cleaned named ‘**Cleaned table’** by using merge queries option in the power query editor.
- Created multiple columns of Days (in text), hours (in numbers), month (in text) by using duplicate column option, date option, time option in transform ribbon in the power query editor.
- Extracted the first 3 letters of the day to use it in the visualization using the extract option the transform ribbon in the power query. Added an index column which acts as a primary column for other columns in the cleaned table.
- Assigned the data types to each of the columns and modify the ones that are wrong using the data type option.
- Created calculated columns using DAX formulas to find the total score and total count of reactions each category got.
  - DAX formula for total score accumulated by category,

    ```
    Total Score = CALCULATE(SUM('Cleaned Table'[Reaction Score]), ALLEXCEPT('Cleaned Table', 'Cleaned Table'[Content Category]))
    ```
  - DAX formula for total count of reactions each category got,

    ```
    Total_count = CALCULATE(count('Cleaned Table'[Content Category]), ALLEXCEPT('Cleaned Table','Cleaned Table'[TotalScore]))
    ```

- Created new table by duplicating a table and deleting the unwanted columns in it. Renamed it as '**Fact table'** which contains all unique content categories.
- Added 'Count of reactions', 'Total Reaction score' for each category and, 'Average reaction score' for each category using left outer in the merge queries option and using aggregate option in the window while expanding the merge table to get the required columns wanted.
- Checked and assigned the relationships between the tables, if it's not assigned to avoid any problems while creating visualizations.

Chat GPT Prompts:

Since it's a project I gave Chat GPT custom instructions to assist me in this project using ctrl + Shift + I shortcut key. I used Chain of thought prompt technique and role prompting technique to get efficient suggestions.

- Custom instructions:
  - My requirements: I am working in a data analysis project for a social media company. I will need some suggestions and assistance in this project during the process.
  - My instructions to Chat GPT: Generate responses more effective and clearer.
- Prompt 1: Imagine yourself as an expert in data analysis with 5+ years of experience in power bi.
- Prompt 2: you need to assist me in a project that i been working on. The client is a social media company who grew rapidly in the last 5 years, but they are facing scaling problem due to the unmanageable data that they are getting. So, I need to provide some insights and make the visualizations that will be presented to the stake holders. I'll provide the details of the data that needs to be analysed.
- Prompt 3: I have a data set of a social media company called social buzz. The "Content" table includes 1000 rows and five columns: 'Content ID' (962 distinct values), 'User ID' (438 distinct values), 'Type' (4 distinct values) and 'Category' (16 distinct values). The "Reactions Type" table has 16 rows and three columns: 'Type' (16 distinct values), 'Score' for each reaction type, and 'Sentiment' categorizing reactions into positive (9 types), neutral (2 types), and negative (5 types). The "Reaction" table, with 24573 rows and four columns, captures user interactions: 'User ID' reacting to a post, 'Date Time' of the reaction, 'Content ID,' and the reaction 'Type.'
- Prompt 4: Suggest me all possible and meaningful visualization ideas.

Data Visualization:

The main requirement of the POC is to find the top 5 performing categories and find the key insights to improve the user engagement. I created 6 visualizations and 3 text boxes to present the data and insights found to the stakeholders.

- Top 5 Performing Content Categories (Clustered bar chart), visualization of aggregate score per category.
- Reaction count by days (Stacked column chart), shows the no. of reactions from the users during different days.
- Reaction count by content type and content category (Clustered column chart with small multiples), no. of reactions for each content type by content categories.
- Percentage of reaction type by sentiments (Pie chart), shows the percentage distribution of sentiments by reaction types.
- Reaction count by year and month (Line chart), shows the number of reactions from June 2020 to June 2021.
- Average reaction score during a day (Stacked column chart), displays the average reaction score during hours in a day for top 5 performing content categories.

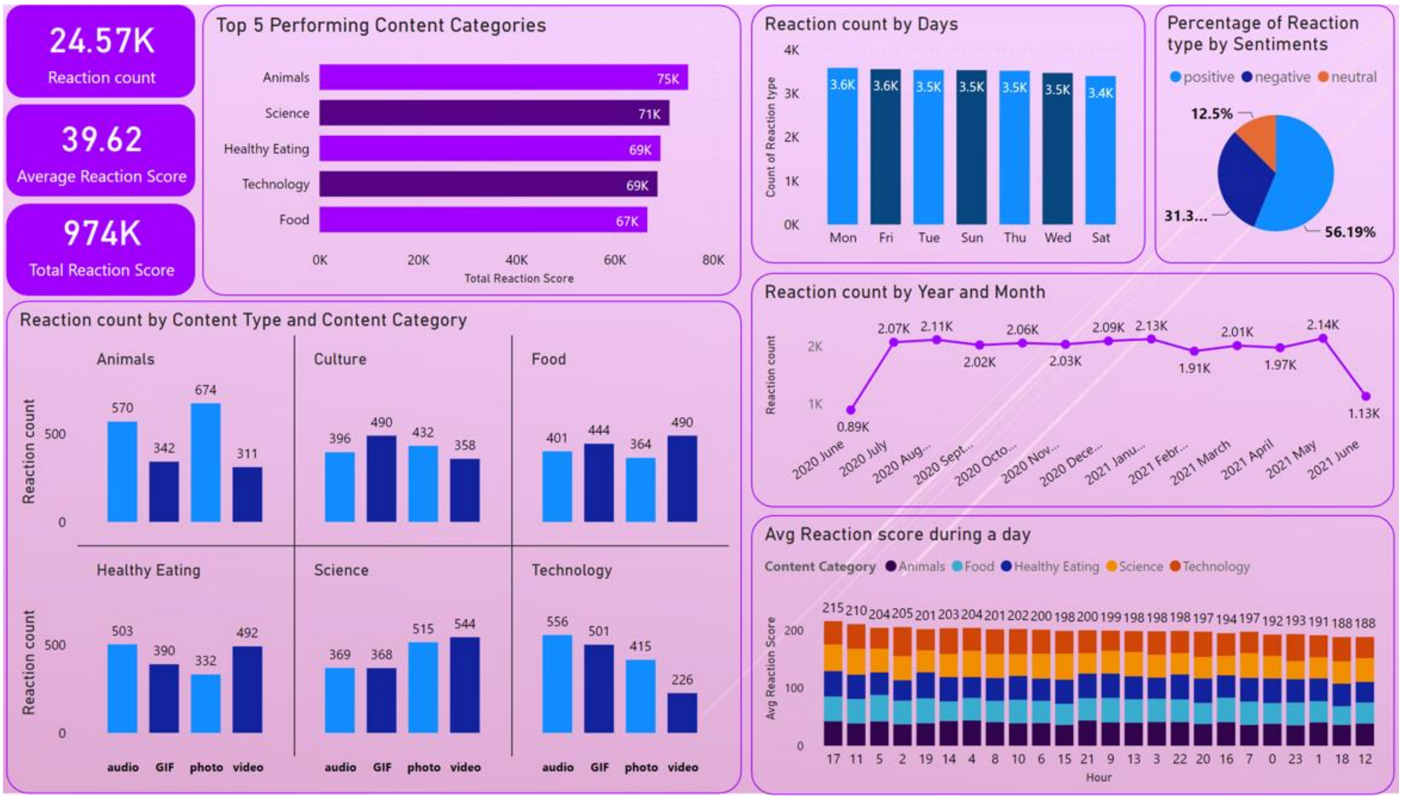
Key Insights:

- The top 5 performing content categories are Animals, science, Health Eating, Technology and Food.
- Health eating is the top-ranking category with common theme of food, suggesting that users are interested in content related to health, wellness, and food.
- May 2021 had the most posts, with the most active days being Saturday, Sunday, and Monday.
- The hours with the highest number of posts in May 2021 were 6, 7, and 8 in the morning, and 17, 23, and 0 in the evening/night.
- The most popular content types for each category,
    - Animals - Photo and audio formats,
    - Science - Photo and video formats,
    - Healthy Eating - Audio and video formats,
    - Technology - Audio and gif formats,
    - Food - Video and gif formats.

Possible impacts using insights:

- Create more content on Healthy Eating, Animals, and Science to engage users. Create specific types of contents for each category to maximize user engagement.
- Using the insights Social Buzz can optimize the timing of the content release to increase engagement, leading to platform growth.

Social Buzz Dashboard:



24.57K
Reaction count
39.62
Average Reaction Score
974K
Total Reaction Score
Top 5 Performing Content Categories
Animals 75K
Science 71K
Healthy Eating 69K
Technology 69K
Food 67K
0K 20K 40K 60K 80K
Total Reaction Score
Reaction count by Days
Count of Reaction type
4K 3K 2K 1K 0K
3.6K 3.6K 3.5K 3.5K 3.5K 3.5K 3.4K
Mon Fri Tue Sun Thu Wed Sat
Percentage of Reaction type by Sentiments
● positive ● negative ● neutral
12.5%
31.3...
56.19%
Reaction count by Year and Month
Reaction count
2K 1K
0.89K 2.07K 2.11K 2.02K 2.06K 2.03K 2.09K 2.13K 1.91K 2.01K 1.97K 2.14K 1.13K
2020 June 2020 July 2020 Aug... 2020 Sept... 2020 Octo... 2020 Nov... 2020 Dece... 2021 Janu... 2021 Febr... 2021 March 2021 April 2021 May 2021 June
Reaction count by Content Type and Content Category
Animals
570 342 674 311
Culture
396 490 432 358
Food
401 444 364 490
Healthy Eating
503 390 332 492
Science
369 368 515 544
Technology
556 501 415 226
Reaction count
500 0
audio GIF photo video
Avg Reaction score during a day
Content Category ● Animals ● Food ● Healthy Eating ● Science ● Technology
Avg Reaction Score
200 100 0
215 210 204 205 201 203 204 201 202 200 198 200 199 198 198 198 197 194 197 192 193 191 188 188
17 11 5 2 19 14 4 8 10 6 15 21 9 13 3 22 20 16 7 0 23 1 18 12
Hour