Social media dataset casestudy

Client overview & Problem statement -

Social Buzz was founded by two former engineers from a large social media conglomerate, one from London and the other from San Francisco. They left in 2008 and both met in San Francisco to start their business. They started Social Buzz because they saw an opportunity to build on the foundation that their previous company started by creating a new platform where content took center stage. Social Buzz emphasizes content by keeping all users anonymous, only tracking user reactions on every piece of content. There are over 100 ways that users can react to content, spanning beyond the traditional reactions of likes, dislikes, and comments. This ensures that trending content, as opposed to individual users, is at the forefront of user feeds. Over the past 5 years, Social Buzz has reached over 500 million active users each month. They have scaled quicker than anticipated and need the help of an advisory firm to oversee their scaling process effectively. Due to their rapid growth and digital nature of their core product, the amount of data that they create, collect and must 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 why they are now looking at bringing in external expertise:

1) They are looking to complete an IPO 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

Objective -

1) Type of Contents posted on social media and what are top trending post contents .

2) Month that has the most post traffic.

3) Insights for further operations.

Gathering the Datasets -

The actual datasets can be downloaded from Forage's Accenture data analytics virtual internships site as this is a part of the same virtual internship ,in this case download them from my drive link given below.

drive.google.com/drive/folders/1MEINN7wsN_wn-ExIPitiBeYJZzCAAk6a?usp=sharing

Tools -

- Python (Pandas)
- MS Power BI
- Google Sheets

Cleaning the data -

Here we can see there are seven datasets named as - Content.csv ,Location.csv ,Profile.csv, Reactions.csv, Reaction types.csv , session.csv and user.csv ,here we have to analyze the matrices that these csv holds .

The whole data cleaning code is here -
https://github.com/Shubh26ham/Social-media-Dataset-Casestudy/blob/main/1.socialmedia_casestudy_wrangling_script.ipynb

1) Importing the datasets in jupyter notebook-

The data sets are imported using the commands in the Jupyter notebook.

```
contents= pd.read_csv('E:/1.DATA_SCIENCE/Datasets/social_media/Content.csv')
profile =pd.read_csv('E:/1.DATA_SCIENCE/Datasets/social_media/Profile.csv')
reactions=pd.read_csv('E:/1.DATA_SCIENCE/Datasets/social_media/Reactions.csv')
reaction_types= pd.read_csv('E:/1.DATA_SCIENCE/Datasets/social_media/ReactionTypes.csv')
```

2) Checking Data Types-

During analysis it is very important to check the data types hence getting the information in pandas using the df.dtypes command.
Types are clear in the code output.

3) Removing some columns and rows as per requirement -

# From " content " cleaning and removing data.

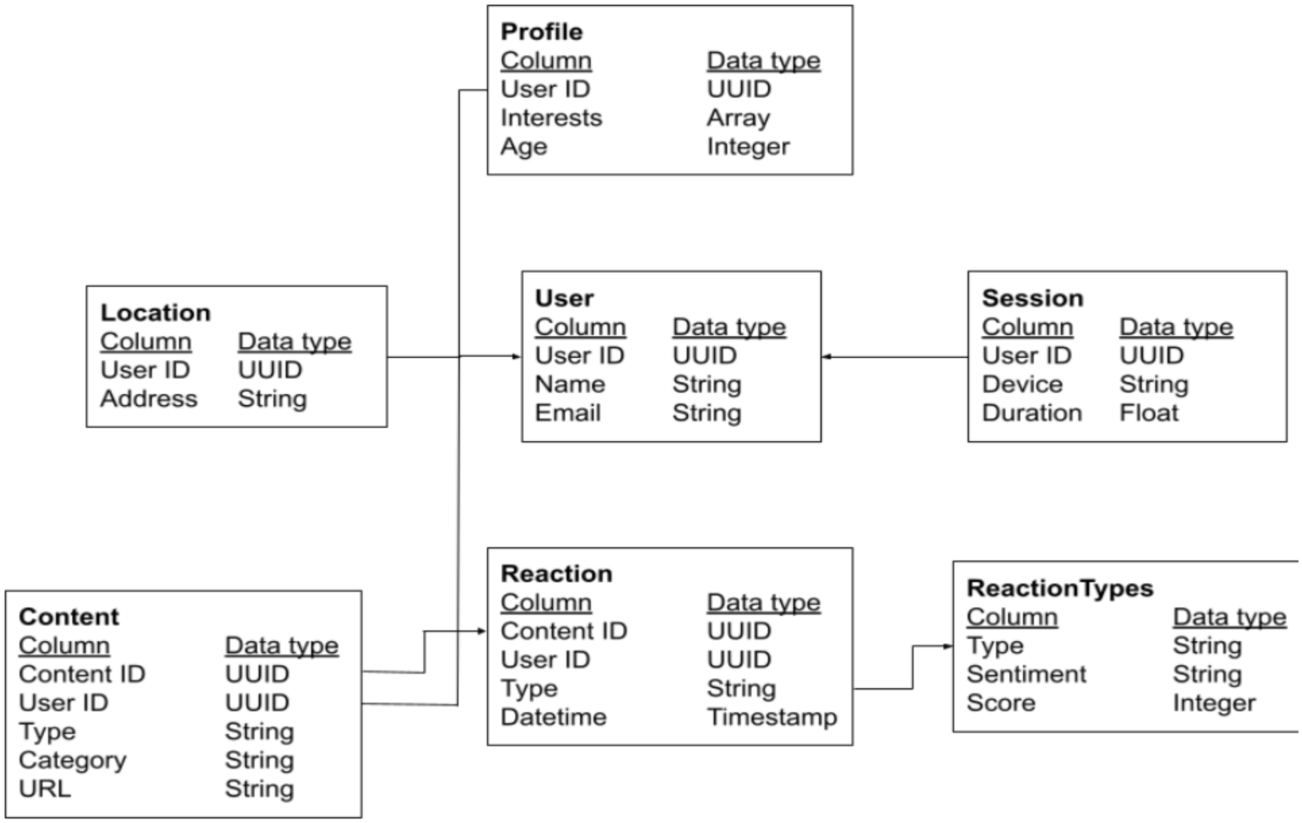
```
contents.drop(contents.columns[[0,3]], axis = 1, inplace = True)
contents.head()
```

```
# From " profile " cleaning and removing data.

profile.drop(profile.columns[0], axis = 1, inplace = True)
profile.head()
```

Removing these rows makes the data less messy and clean in view which keeps our work clean and clear for further analysis.

4) Joining the datasets as according to the diagram -



Profile

Column	Data type
User ID	UUID
Interests	Array
Age	Integer

Location

Column	Data type
User ID	UUID
Address	String

User

Column	Data type
User ID	UUID
Name	String
Email	String

Session

Column	Data type
User ID	UUID
Device	String
Duration	Float

Content

Column	Data type
Content ID	UUID
User ID	UUID
Type	String
Category	String
URL	String

Reaction

Column	Data type
Content ID	UUID
User ID	UUID
Type	String
Datetime	Timestamp

ReactionTypes

Column	Data type
Type	String
Sentiment	String
Score	Integer

According to the above diagram I connected the data using ‘merge’ in Pandas similar to outer join though here we can see some column are not required in our analysis ,means considering the files will increase size of file and will definitely waste our time hence I removed Location, User, Session csv’s.

Joining using merge statement (outer join) between Profile & Contents

```
Join_1 = pd.merge(profile, contents, how='outer', on='User ID')
Join_1.head()
```

Joining outer join between (Profile & Content) and Reactions

```
Join_2 = pd.merge(Join_1, reactions, how='outer', on='Content ID')
Join_2.head()
```

Joining outer join between (Profile & Content & Reactions) and Reaction types

```
Join_3 = pd.merge(Join_2, reaction_types , how='outer', on='Type')
Join_3.head()
```

5) Removing Null Values -

Null values ends up messing the result hence removing them is the best solution if any filling option isn’t available

```
df1 =Join_3.dropna()
```

6) Checking for duplicates -

We will remove duplicate data because it may cause unnecessary skewness in the statistics of data thus removing duplicates is necessary.

```
d = Join_3[Join_3.duplicated()]    # zero duplicates found
print(d)
```

7) Adding a month column for Month analysis -

I used the date time column for extracting the month in a separate column for convenience in analytics using strftime clause.

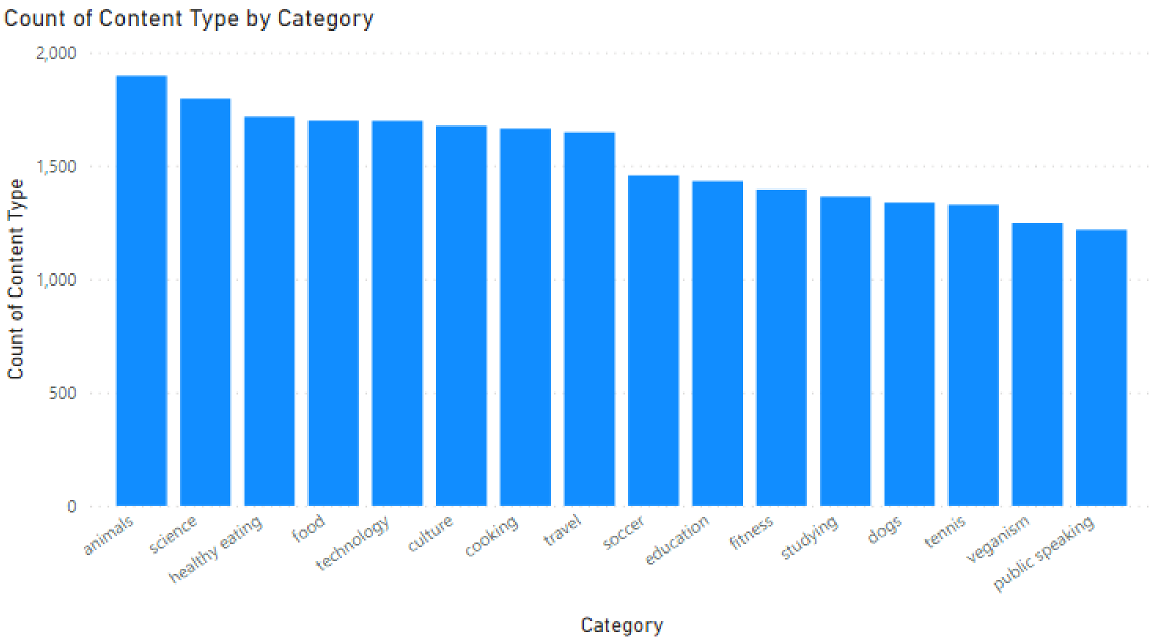
```
n1= c_df['Month'] = c_df['Datetime'].dt.strftime('%b')
```

8) **Exporting the cleaned csv file** -

After getting the cleaned data we exported it in a csv format for further visualization.

c_df.to_csv('E:/1.DATA_SCIENCE/cleaned_socialmedia_final_join.csv', index=False)

9) **Visualization** -



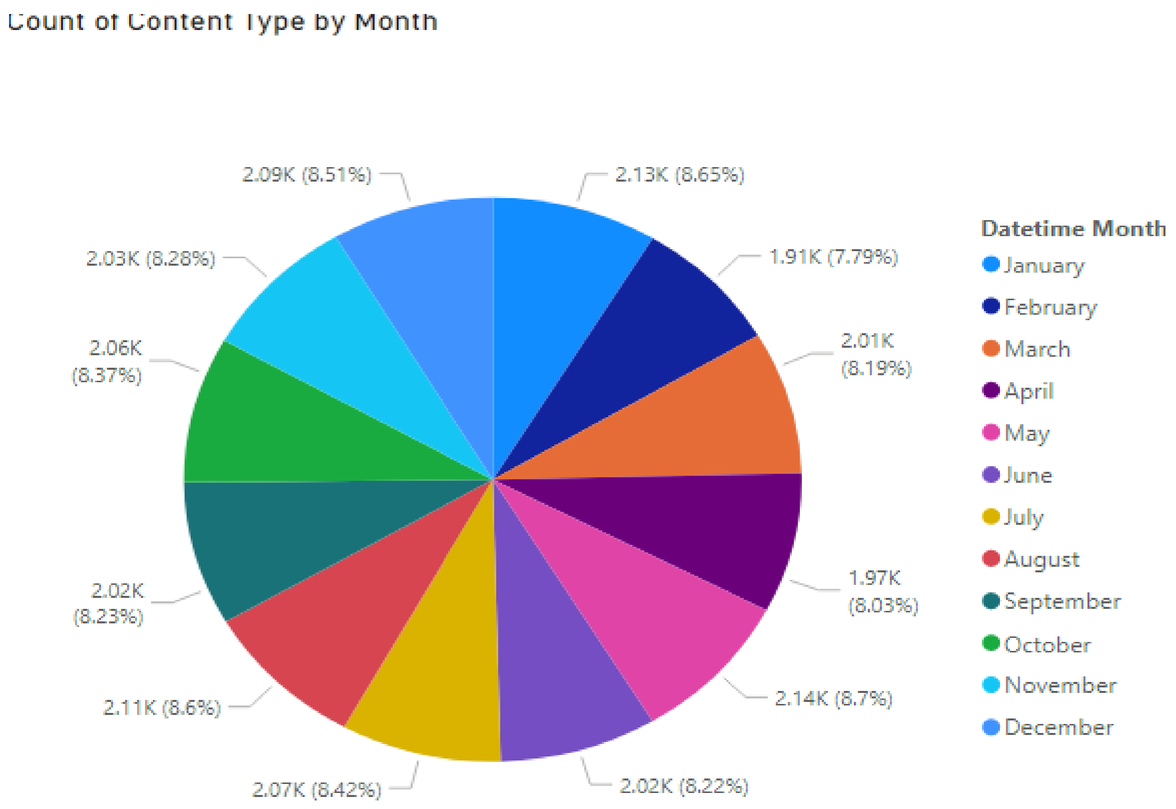
Let's analyze the results from data

By looking at the data we can see the there are 16 unique Categories following as- Animals ,Science , Healthy eating, food ,Technology , Culture ,Cooking , Travel ,Soccer ,Education , Fitness , Studying, Dogs, Tennis ,Veganism and Public speaking. Here we can see the graph shows the relations between Counts of content type and category ,the count function is used because we want the number of contents that falls under certain categories.

We can determine that Animals ,Science , Healthy Eating, Food , Technology are few top categories.

“Type of Contents posted on social media and what are top trending post contents”

We can conclude that Animals ,Science , Healthy Eating, Food , Technology are the top 5 categories that can be used for advertisement purposes.



This pie-chart shows the distribution of Months by Content type. This is done because count of content type shows the number of contents that have been posted and when plotted along a month it will show the distribution of posts posted in a particular month. From this pie chart we can conclude May has most number of posts with 8.7% of share in the pie chart and followed by January with 8.65% of share in the pie chart

“ Month that has the most post traffic.“

We can easily see May and January have the most posts with 2.14 k and 2.13 k posts respectively .

10) Insights -

According to the statement the main problem is to solve the issue of using the large public reach and capitalization .The best solution according to the experts Advertisements are the great way to generate revenue and according to the above analysis we can easily say.

1) Use the Advertisement in the segment in categories like Animals ,Science , Healthy Eating, Food , Technology because they are liked by the users therefore using ads related to these categories would be extremely beneficial.

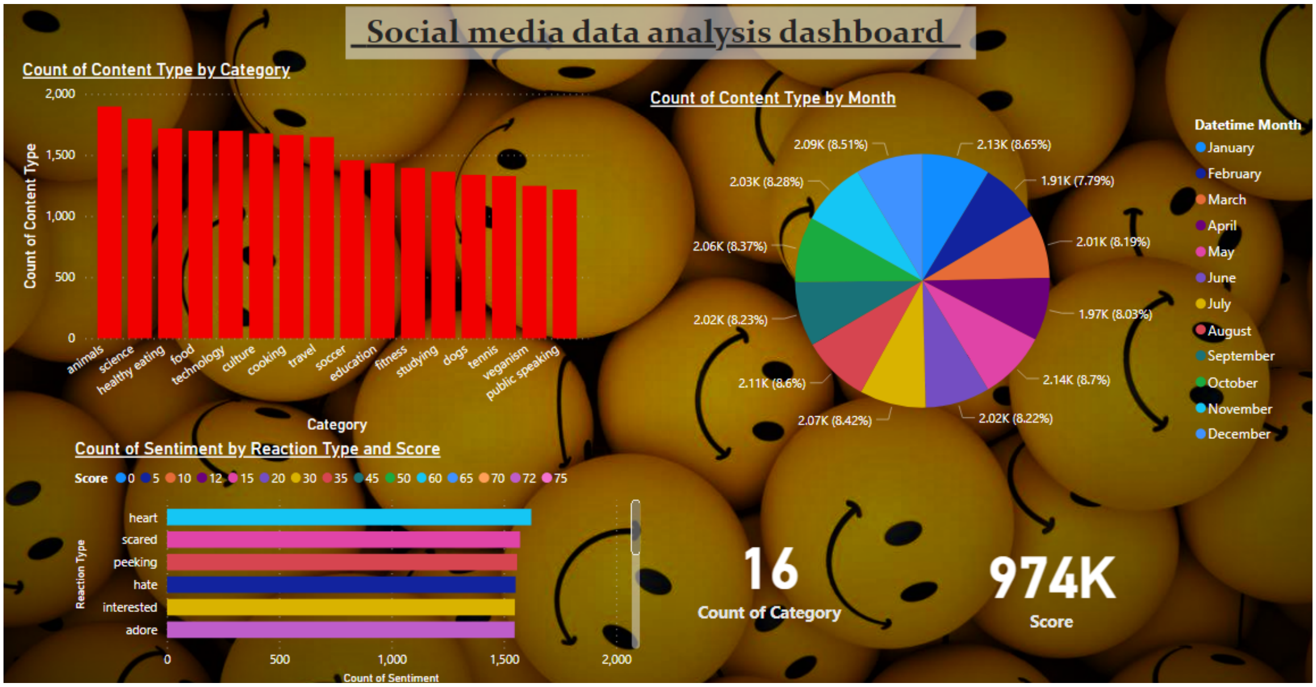
2) As the months of May and January have max posts ,therefore the time of Christmas and New year and summer vacations would be great time for advertise sale and discount offers on merchandise products if you want ,these may help you generate revenue.

3) Increasing and promoting the topics of Animals and Science awareness can help you run successful environmental campaigns with help of NGOs working in the field.

11) Dashboard -

Using Power-BI we have created the dashboard .

Link -
https://github.com/Shubh26ham/Social-media-Dataset-Casestudy/blob/main/2.socialmedia_dashboard.pbix



Social media data analysis dashboard
Count of Content Type by Category
2,000
1,500
1,000
500
0
Count of Content Type
animals
science
healthy eating
food
technology
culture
cooking
travel
soccer
education
fitness
studying
dogs
tennis
veganism
public speaking
Category
Count of Content Type by Month
2.09K (8.51%)
2.13K (8.65%)
2.03K (8.28%)
1.91K (7.79%)
2.06K (8.37%)
2.01K (8.19%)
2.02K (8.23%)
1.97K (8.03%)
2.11K (8.6%)
2.14K (8.7%)
2.07K (8.42%)
2.02K (8.22%)
Datetime Month
January
February
March
April
May
June
July
August
September
October
November
December
Count of Sentiment by Reaction Type and Score
Score 0 5 10 12 15 20 30 35 45 50 60 65 70 72 75
heart
scared
peeking
hate
interested
adore
Reaction Type
0
500
1,000
1,500
2,000
Count of Sentiment
16
Count of Category
974K
Score