Accenture Data Analytics Internship: Social Buzz

May 8th, 2023

Social media and content creation is the industry, and they offer an opportunity for users to anonymously react to content that is posted. 500 million users and over 100000 pieces of content per day.

Why do they need an analysis?

- They need to complete an IPO and need guidance.
- They are still a small-scale company and need help analyzing their large dataset.
- They also want to learn standard data practices from industry giants so they can apply it to their company.

What do they need from us:

- An audit of their data practice
- Recommendations for a successful IPO (initial public opening, they want to go public with their shares)
- An analysis that shows their top 5 most popular categories.

Delegated tasks

- Creation of an up-to-date big data best practices presentation
- Extraction of sample data sets using SQL
- On-site audit of their data-center - Merging of sample data set tables
- Virtual session with Social Buzz team to present previous client success stories relevant to them
- Preparation of best practice document for IPO
- Loading of sample data sets into Accenture sandbox database
- Technology architecture workshop with Social Buzz Data Team to understand their technology landscape
- Stress testing of their technology to identify weak spots
- Communication with previous IPO companies within our client base for reference stories
- Analysis of sample data sets with visualizations
- Full documentation of the process that we can guide them through for IPO

Team:

Industry experts: people with experience in the social media sector

IPO experts: they will provide the IPO requirement

Data experts: they will provide big data insights and content analysis

STEP 1

There are 7 different data sets and a data model. The goal is to create a final dataset that has the right combination of data to meet the requirements of the client.

The processes of this step are as follows:

- Requirement gathering
- Data cleaning

- Data modeling

Data Model

We have been sent numerous datasets that have contained the following data.

User

- User ID: Unique ID of the user (automatically generated)
- Name: Full name of user
- Email: Email address of user

Profile

- User ID: Unique ID of a user that exists in the User table
- Interests: Interests of the associated user
- Age: Age of the associated user

Location

- User ID: Unique ID of a user that exists in the User table
- Address: Full address of the user

Session

- User ID: Unique ID of a user that exists in the User table
- Device: Mobile device that they used for this session on the application
- Duration: Amount of time in minutes that this user stayed active on the application during this session

Content

- ID: Unique ID of the content that was uploaded (automatically generated)
- User ID: Unique ID of a user that exists in the User table
- Type: A string detailing the type of content that was uploaded
- Category: A string detailing the category that this content is relevant to
- URL: Link to the location where this content is stored

Reaction

- Content ID: Unique ID of a piece of content that was uploaded
- User ID: Unique ID of a user that exists in the User table who reacted to this piece of content
- Type: A string detailing the type of reaction this user gave
- Datetime: The date and time of this reaction

Reaction Types

- Type: A string detailing the type of reaction this user gave
- Sentiment: A string detailing whether this type of reaction is considered as positive, negative, or neutral
- Score: This is a number calculated by Social Buzz that quantifies how “popular” each reaction is. A reaction type with a higher score should be considered as a more popular reaction.

Needed datasets for our requirements: The only datasets we need are Content, Reaction, Reaction Types.

Substep 2: Data Cleaning

We must clean the collected datasets.

- **Content dataset:** Delete the initial serial number column, URL column and User ID column. Eliminate the quotation marks in the Content Type column. Convert data in the type and category columns to text datatype.
- **Reaction dataset:** Delete the initial serial number column and User ID column. There are 980 blanks in reaction type column (so filter and delete those missing values). Split the datetime column into date and time (To get date use: **=INT(DateTime column)**, to get time into a separate column, use: **DateTime column – INT(DateTime column)**)
- **ReactionType dataset:** Delete the initial serial number column.

Substep 3: Merge the columns from each cleaned data set into one major dataset for analysis.

We are going to use the reaction dataset as a base.

To merge the content type column from content dataset, use: =VLOOKUP(A2, [Content_cleaned.xlsx]Sheet1!$1:$1048576, 2, FALSE)

To merge the category column from content dataset, use: =VLOOKUP(A2, [Content_cleaned.xlsx]Sheet1!$1:$1048576, 3, FALSE)

To merge the score from the reactiontype dataset, use: =VLOOKUP(B2,[ReactionType_cleaned.xlsx]Sheet1!$1:$1048576, 3, FALSE)

STEP 2: Answer the following questions

What are the top 5 categories by scores?

We can use a pivot table, or we can use the SUMIF function.

	F	G	H	I	J	K	L
1	Type	Category	Score			Categories	Sum_of_scores
2		Studying	0			Animals	74965
3		Studying	10			cooking	64756
4		Studying	15			culture	66579
5		Studying	0			dogs	52511
6		Studying	30			education	57436
7		Studying	35			fitness	55323
8		Studying	70			food	66676
9		Studying	5			healthy eating	69339
10		Studying	35			public speaking	49264
11		Studying	65			science	71168
12		Studying	20			soccer	57783
13		Studying	15			Studying	53984
14		Studying	30			technology	68738
15		Studying	5			tennis	50339
16		Studying	15			travel	64880
17		Studying	75			veganism	49619
18		Studying	35				
19		Studying	20				
20		Studying	30				
21		Studying	45				
22		Studying	35				

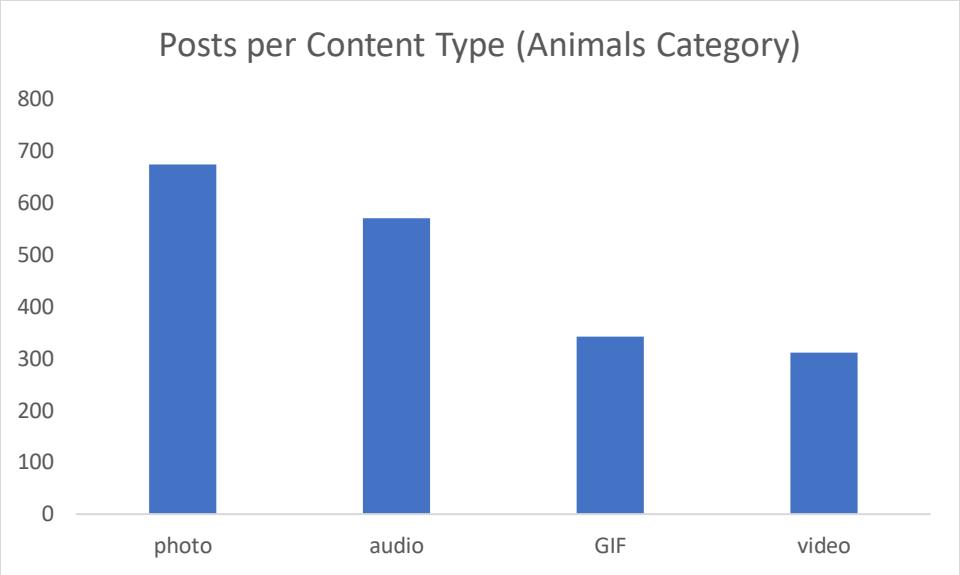
To find the top 5 categories as shown above, use the following formula: =SUMIF(G2:G24574, K2, H2:H24574) and repeat for the other items in column K.

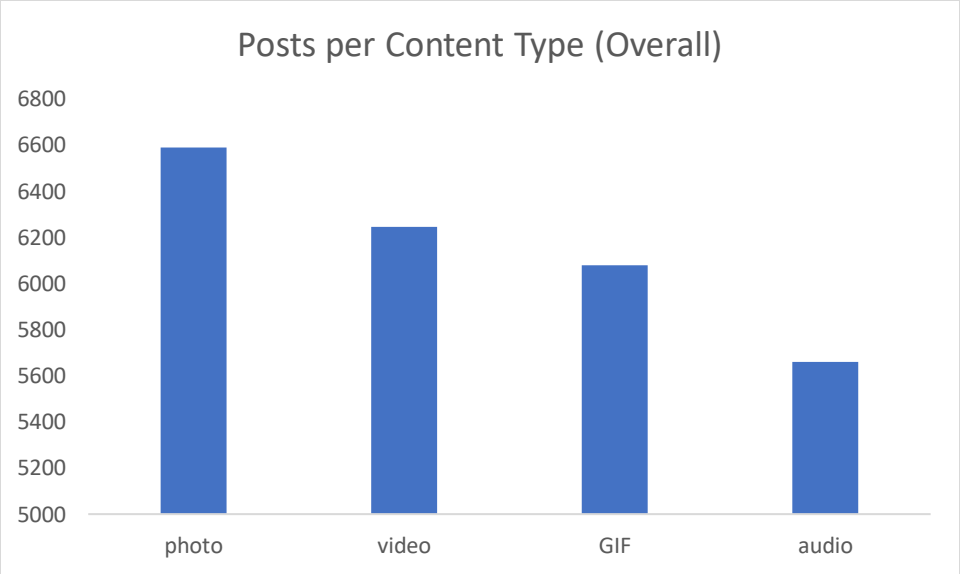
The top 5 categories are: Animals, science, healthy eating, technology, food

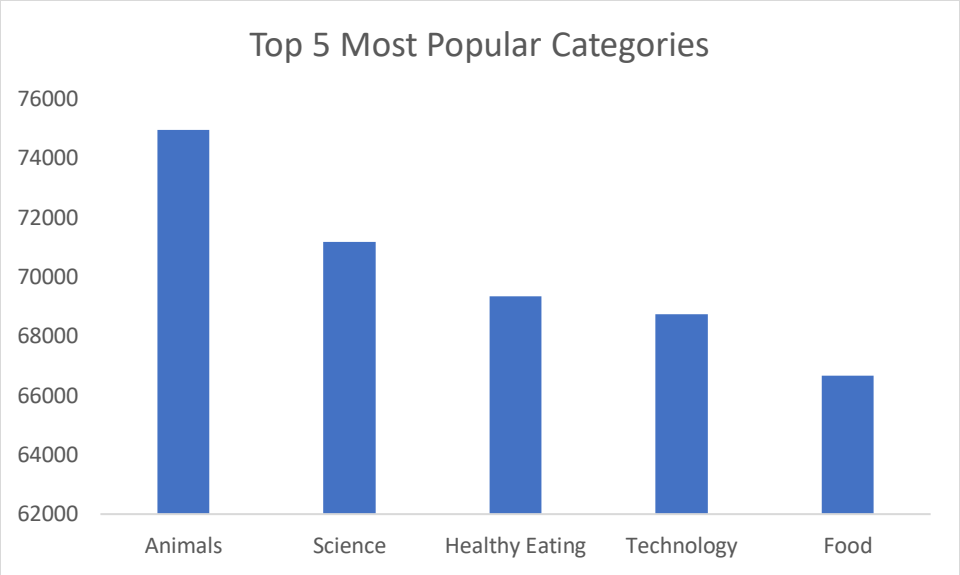
STEP 3: VISUALIZE THE DATA

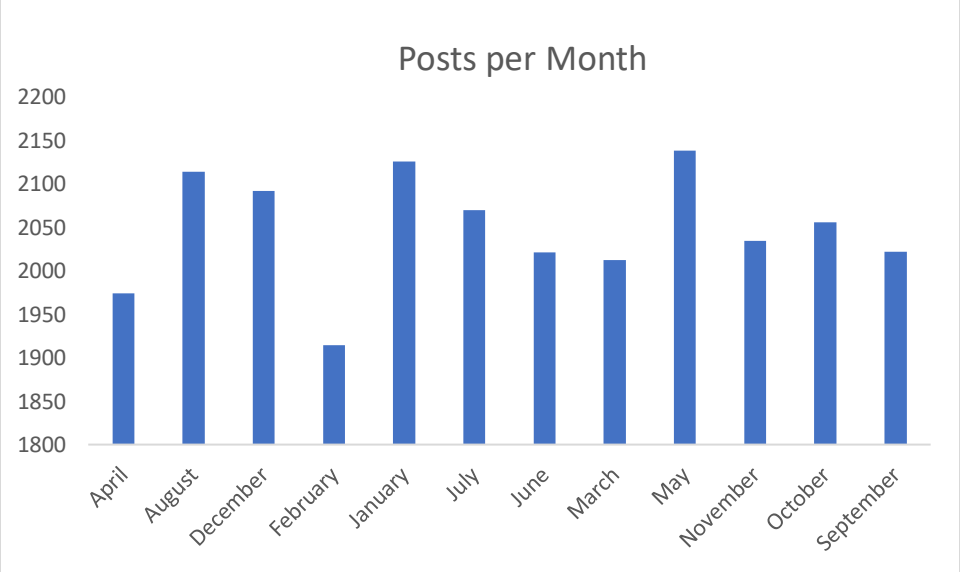
The next step is to create a PowerPoint presentation to present to our stakeholders. The presentation will include:

- 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?


Posts per Content Type (Animals Category)
800
700
600
500
400
300
200
100
0
photo
audio
GIF
video


Posts per Content Type (Overall)
6800
6600
6400
6200
6000
5800
5600
5400
5200
5000
photo
video
GIF
audio





Project Summary

This is an end-to-end project for a social media company with a limited data analytics team. The scope of this project included:

- Gathering the required data from the social media company

- Modelling data based on relationships between datasets and columns

- Cleaning and merging different datasets into one for convenient data analysis

- Using advanced Excel functions such as VLOOKUPs for preliminary analysis before loading the data set onto PowerBI for further transformation and visualization.

- Creating a visually appealing Power BI dashboard using common design elements and dashboard design principles.