Data Alchemy

From Listening to Informed Decision Making

Guest Lecture | Universiti Sains Malaysia

Presenter

Lalitha Shamugam

- Google Scholar
- LinkedIn

Materials

- Jupyter Notebook, Dataset, Slides --> GitHub
- Visualisation, Dashboard, Story --> Tableau Dashboard

Overview

1. Social Listening
2. Hypothesis Development
3. Data Collection & Pre-processing
4. Case Study
5. Dashboard

1. Social Listening

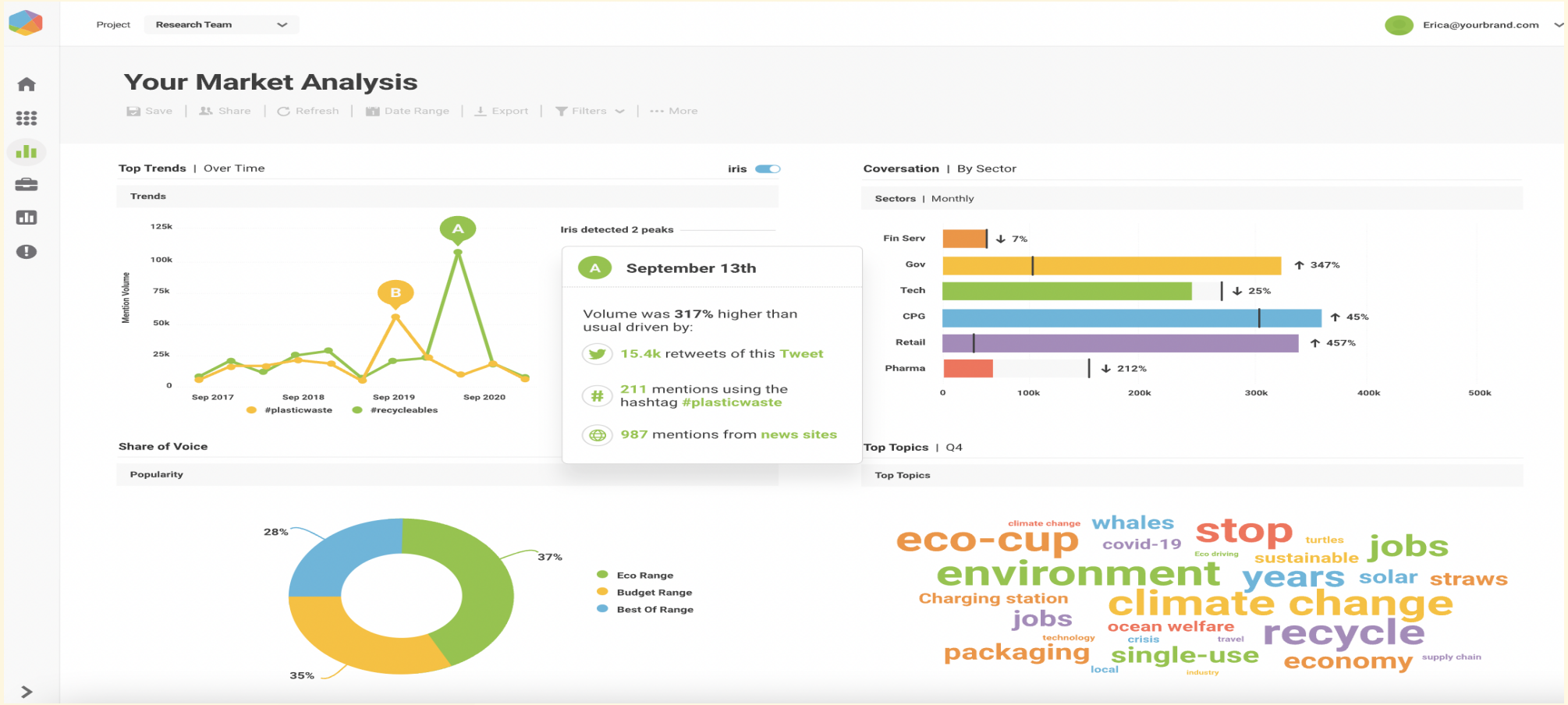
Introduction

- **Social listening** - process of identifying and assessing what is being said about a company, individual, product, or brand on various social media platforms.
- Importance of social listening includes:

  1. Identifies brand mentions and customer feedback
  2. Tracks market trends and consumer sentiments

Tools

Social Listening Dashboard



Key Aspect and Technique

1. Monitoring mention

- Track real-time mentions of your brand across platforms
- Use sentiment analysis to gauge customer feelings

2. Tracking Hashtags and Keywords

- Identify trending topics related to your brand or industry
- Monitor specific campaigns or events through targeted hashtags

Key Aspect and Technique

3. Audience Insights

- Understand demographics, behaviors, and preferences of your audience
- Adapt marketing strategies based on audience engagement patterns

2. Hypothesis Development

Definition

A hypothesis is a tentative explanation that accounts for a set of facts and can be tested by further investigation.

Characteristics

Well-formed business hypotheses have three key characteristics:

1. Testable
2. Precise
3. Discrete

Example

"Engaging daily on Twitter for one month will increase our brand's follower count by 10%"

- **Testable**: This hypothesis can be easily tested by tracking the number of followers on Twitter over a month
- **Precise**: It specifies the platform (Twitter), the frequency of engagement (daily), and the duration (one month)
- **Discrete**: Focuses on a single, clear outcome – increasing the follower count by 10%

3. Data Collection & Pre-processing

First Step...

Data Collection:
Employ Boolean search techniques to target specific brand-related content and use Isentia tools for data crawling

Data Cleaning:
Researchers clean the data to ensure accuracy and relevance

Categorization:
Classify data based on sentiments and topics

4. Case Study

Topic

Malaysian Twitter User Engagement for Lalitha Sdn Bhd

Description:

This project entails a data analysis of Twitter engagement for Lalitha Sdn. Bhd., utilizing a synthetic dataset generated via Python. The objective is to formulate and test four hypotheses through detailed data visualizations, aiming to uncover insights and patterns that drive engagement on the platform.

Objective

1. To determine if social media posts made during evening hours receive higher engagement than posts made at other times.
2. To assess whether including images in social media posts increases engagement metrics such as likes, comments, and retweets.
3. To explore the relationship between the sentiment of social media posts (negative, positive, neutral) and the number of retweets they receive.
4. To investigate the correlation between the number of followers a user has and the frequency of their social media posts.

Dataset Info

- This is a synthetic dataset created for today's class
- It is available in the following link.

social_media_data

User ID	Frequency of Posts	Post Type	Likes	Comments	Retweets	Sentiment of Post	Sentiment of Comments	Date of Post	Number of Followers	Engagement
8270	29	Image	34	26	32	Neutral	Negative	2023-11-13 13:32:18	3480	92
1860	23	Image	339	168	178	Negative	Neutral	2023-11-21 08:30:00	1380	685
6390	7	Image	230	42	14	Negative	Neutral	2023-11-01 19:23:32	595	369
6191	3	Image	438	114	117	Positive	Positive	2023-11-06 03:57:52	303	669
6734	10	No Image	229	58	26	Neutral	Neutral	2023-11-05 22:41:26	800	406
7265	10	Image	702	261	24	Negative	Positive	2023-11-05 16:02:25	820	987
1466	19	Image	464	188	108	Positive	Neutral	2023-11-17 15:03:24	1178	759

Dataset Description

1. This synthetic dataset captures a month's worth of Twitter interactions among followers of Lalitha Sdn Bhd.
2. It consists of **1,000** entries, each representing a distinct tweet from users who follow the company.
3. Ideal for social media analytics, this dataset can help the company in strategizing its Twitter presence.
4. It offers insights for enhancing user engagement, determining effective content strategies, and identifying optimal posting times.

Dataset Attributes

- User ID
- Frequency of Tweets
- Tweet Type
- Likes, Comments, Retweets
- Sentiment of Tweet and Comments
- Date and Time of Tweet
- Number of Followers (of Users)
- Engagement Score

Read File

- Import the CSV file as a DataFrame using the pandas library.

```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# read csv file
df = pd.read_csv('social_media_data.csv')

# view file
df.head()
```

Basic Info

- The info() method in pandas was used to inspect the details.

```
# find data info
df.info()
```

```
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   User ID                 1000 non-null   int64
 1   Frequency of Posts      1000 non-null   int64
 2   Post Type               1000 non-null   object
 3   Likes                   1000 non-null   int64
 4   Comments                1000 non-null   int64
 5   Retweets                1000 non-null   int64
 6   Sentiment of Post       1000 non-null   object
 7   Sentiment of Comments   1000 non-null   object
 8   Date of Post            1000 non-null   datetime64[ns]
 9   Number of Followers     1000 non-null   int64
 10  Engagement              1000 non-null   int64
 11  Hour of Post            1000 non-null   int32
dtypes: datetime64[ns](1), int32(1), int64(7), object(3)
memory usage: 90.0+ KB
```

Statistical Description

- The describe() method was used to obtain a statistical summary.

```
# statistical description
df.describe(include='all')
```

	User ID	Frequency of Posts	Post Type	Likes	Comments	Retweets	Sentiment of Post	Sentiment of Comments	Date of Post	Number of Followers	Engagement	Hour of Post
count	1000.00000	1000.00000	1000	1000.00000	1000.000000	1000.000000	1000	1000	1000	1000.00000	1000.000000	1000.000000
unique	NaN	NaN	2	NaN	NaN	NaN	3	3	NaN	NaN	NaN	NaN
top	NaN	NaN	No Image	NaN	NaN	NaN	Neutral	Negative	NaN	NaN	NaN	NaN
freq	NaN	NaN	513	NaN	NaN	NaN	349	337	NaN	NaN	NaN	NaN
mean	5594.25600	15.09700	NaN	312.47300	122.982000	78.394000	NaN	NaN	2023-11-15 22:28:33.809000192	1494.98400	551.614000	11.725000
min	1004.00000	1.00000	NaN	0.00000	0.000000	0.000000	NaN	NaN	2023-11-01 06:41:49	64.00000	48.000000	0.000000
25%	3467.50000	8.00000	NaN	154.75000	61.000000	37.000000	NaN	NaN	2023-11-08 12:01:35	701.50000	368.750000	6.000000
50%	5750.00000	15.00000	NaN	301.00000	117.000000	70.000000	NaN	NaN	2023-11-16 01:30:57	1362.00000	522.000000	12.000000
75%	7806.75000	23.00000	NaN	443.25000	178.000000	108.000000	NaN	NaN	2023-11-23 06:52:18.500000	2142.75000	708.250000	18.000000
max	9996.00000	29.00000	NaN	748.00000	298.000000	265.000000	NaN	NaN	2023-11-30 17:38:59	4263.00000	1408.000000	23.000000
std	2533.55383	8.44683	NaN	190.64476	75.646018	54.183361	NaN	NaN	NaN	970.69703	245.557157	6.883077

Hypothesis 1

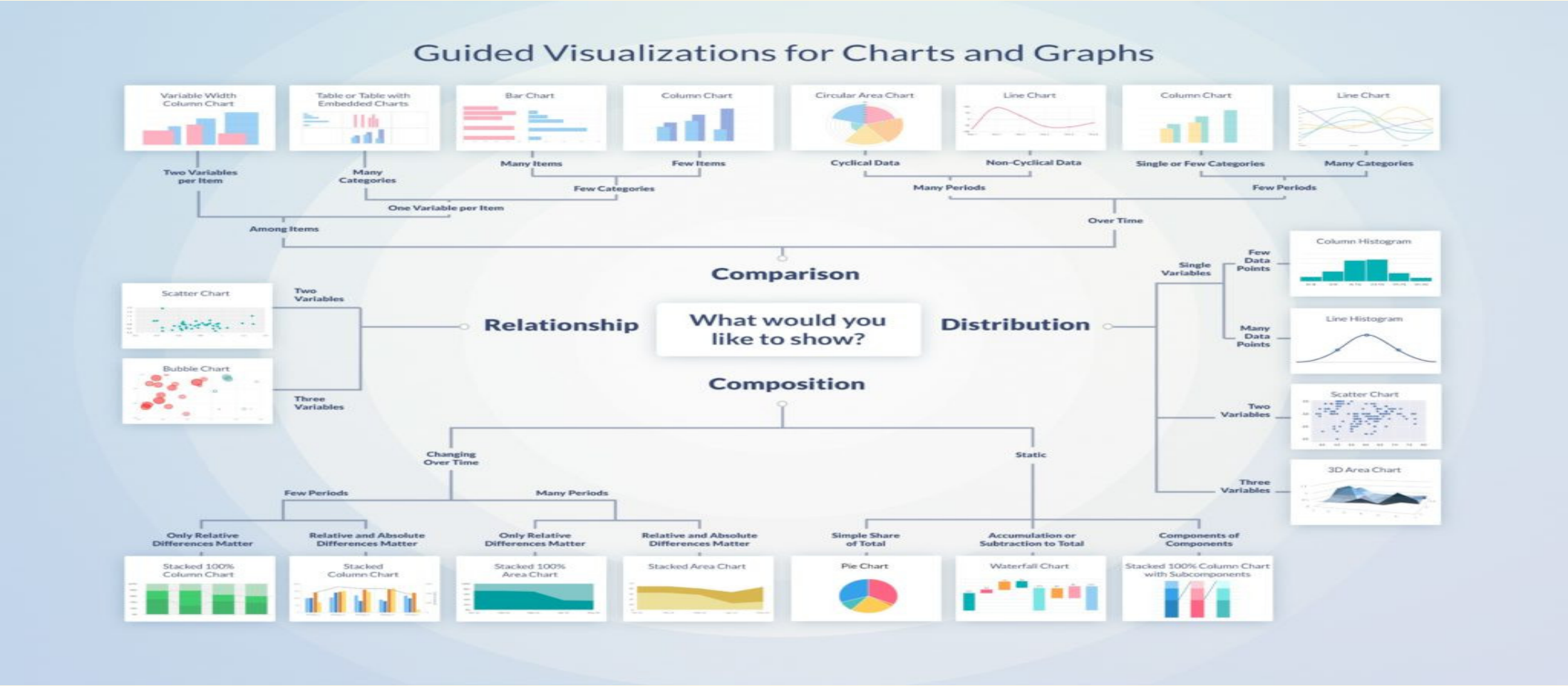
Date of Post and *Engagement* Relationship:

- **Hypothesis**: *Posts made during evening hours receive higher engagement compared to posts made at other times of the day*

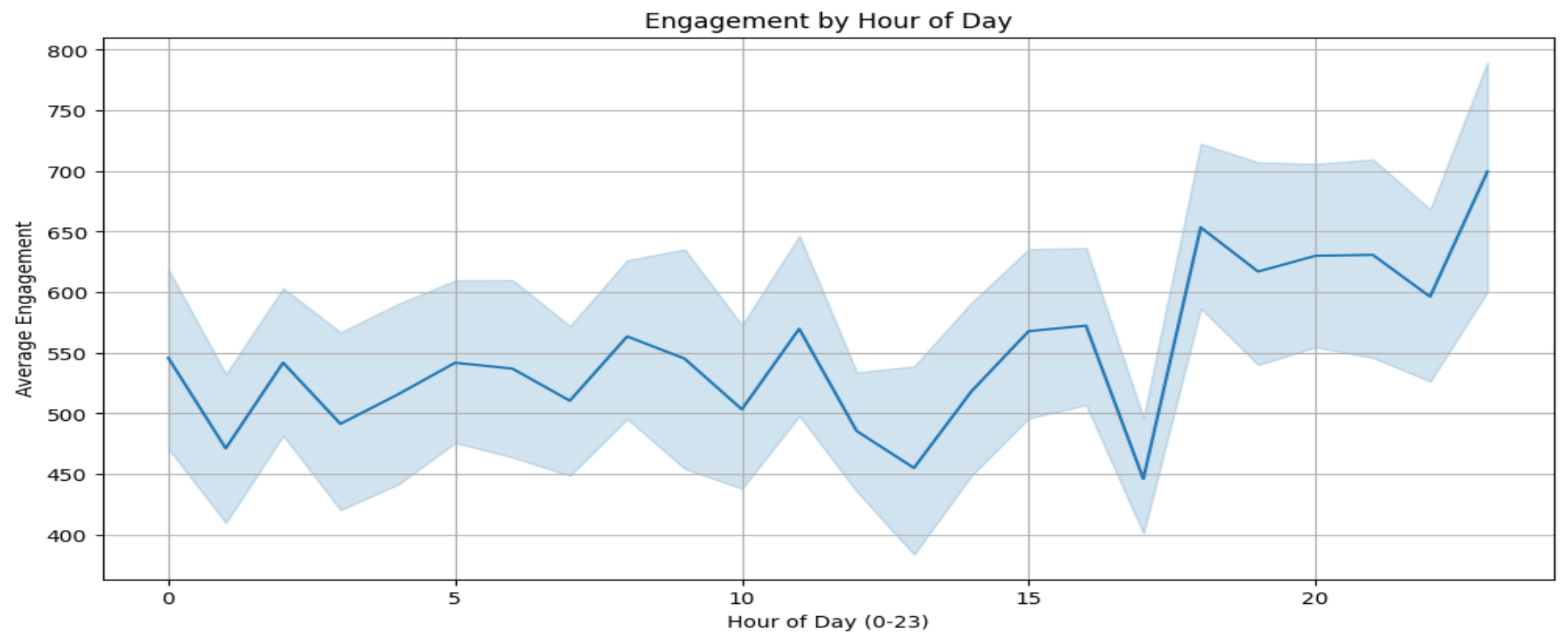
Choice of Visualisation

A line chart is ideal for this hypothesis because it will clearly illustrates trends and comparisons over time, effectively showing how engagement varies across different times of the day.

Data Visualisation Guide



Visualisation 1



Hypothesis 2

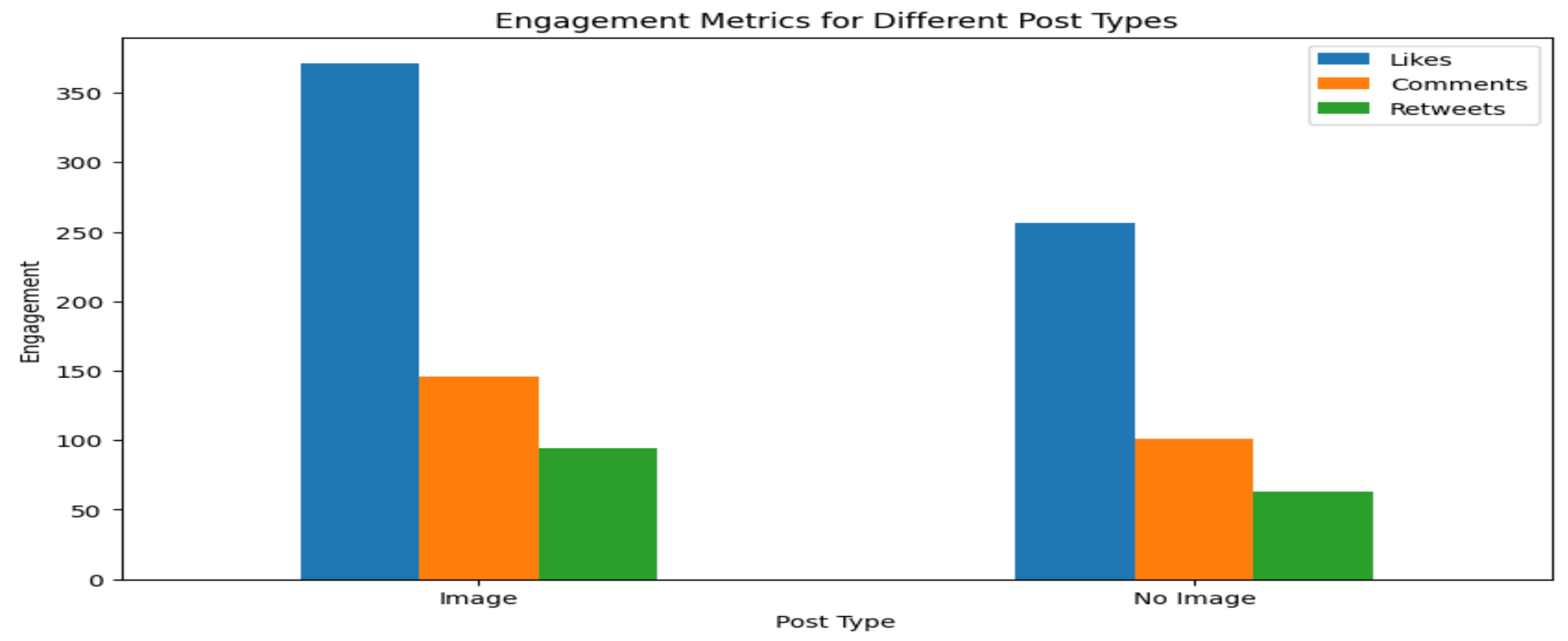
Social Media Post and *Engagement* Relationship:

- **Hypothesis**: *Social media posts that include images will receive higher engagement metrics (likes, comments, and retweets) compared to posts without images*

Choice of Visualisation

A bar chart is suitable for this hypothesis as it clearly compares discrete categories—posts with and without images—showing the difference in engagement metrics in a visually distinct manner.

Visualisation 2



Hypothesis 3

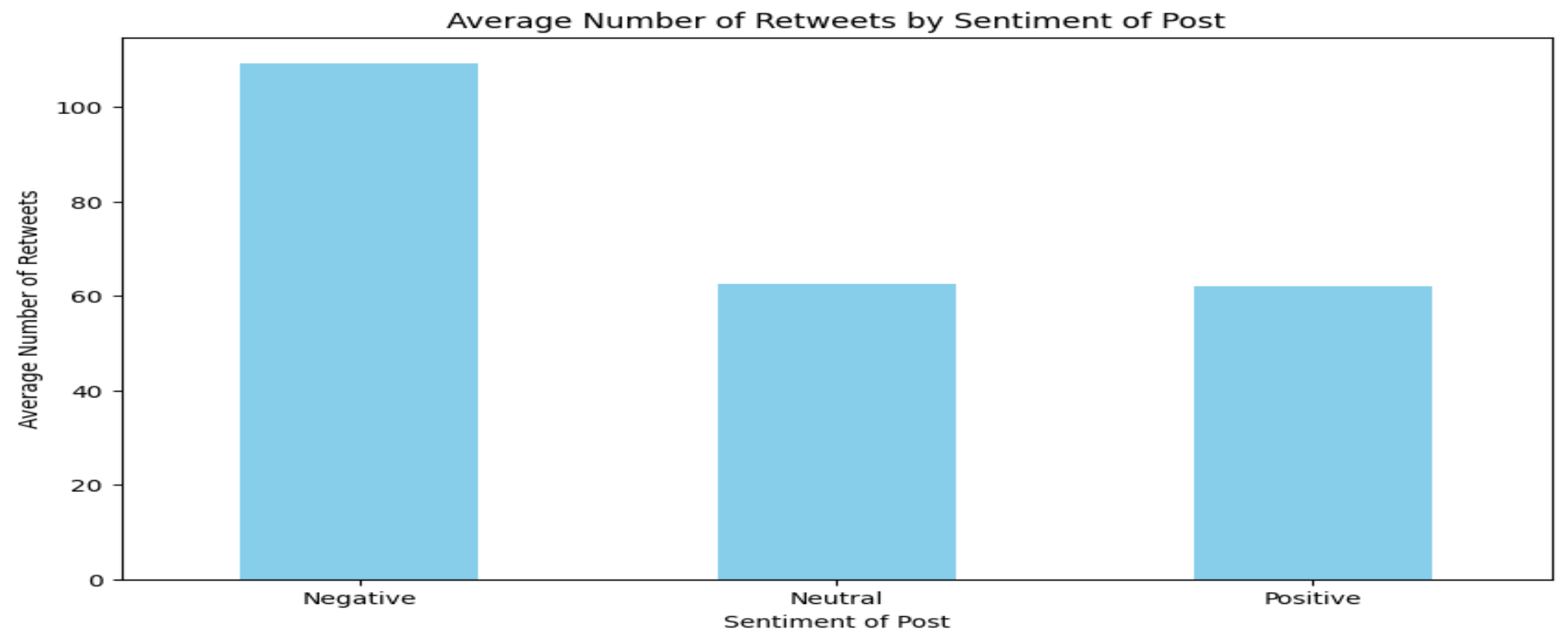
Post with Negative Sentiment and *Retweets* Relationship:

- **Hypothesis**: *Posts with negative sentiment are associated with a higher number of retweets than posts with positive or neutral sentiments*

Choice of Visualisation

A bar chart is ideal for this hypothesis because it effectively compares engagement across distinct categories—negative, positive, and neutral sentiments—allowing for a clear visual representation of differences in user interaction based on the sentiment of the posts.

Visualisation 3



Hypothesis 4

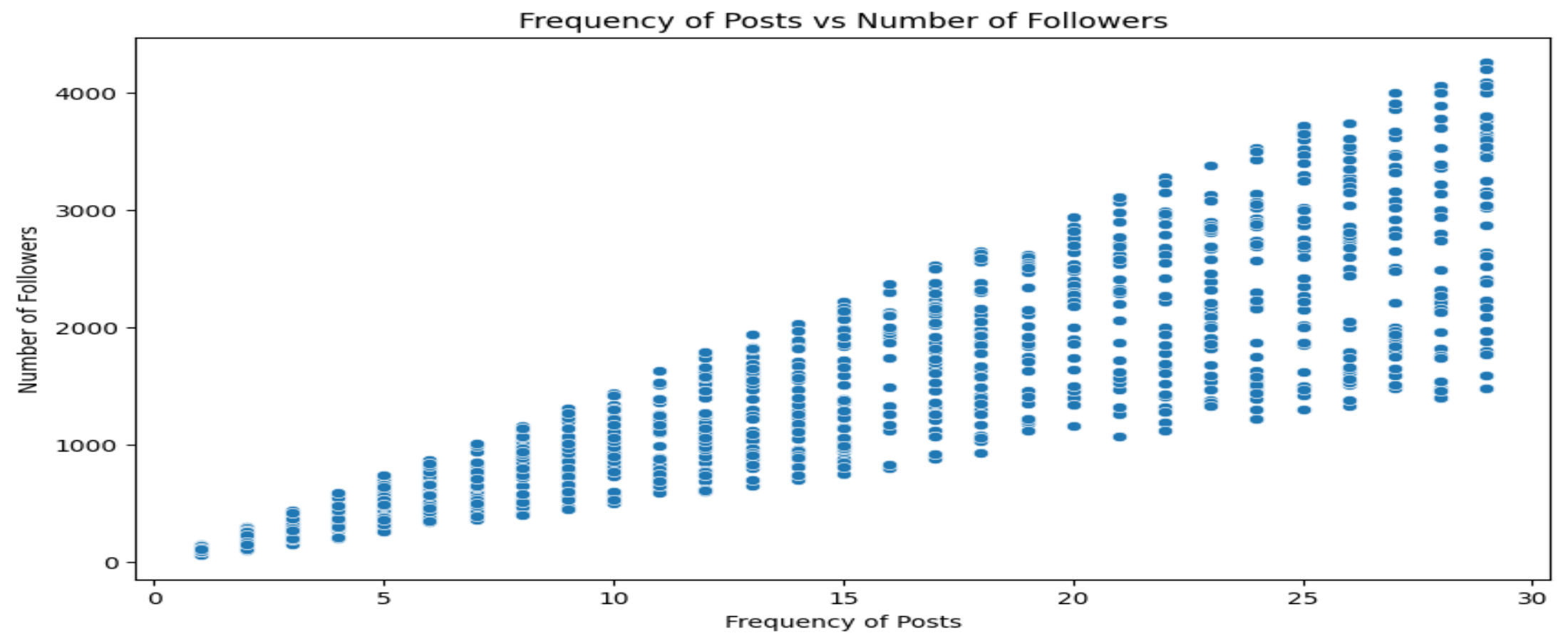
Frequency of posts and *Number of Followers* Relationship:

- **Hypothesis**: *There is a positive correlation between the frequency of posts by social media users and their number of followers*

Choice of Visualisation

A scatter plot is suitable for this hypothesis as it visually represents the correlation between two continuous variables—frequency of posts and number of followers—allowing for the observation of patterns or trends in the data.

Visualisation 4



5. Dashboard

Dashboard Design [Website][Journal article]

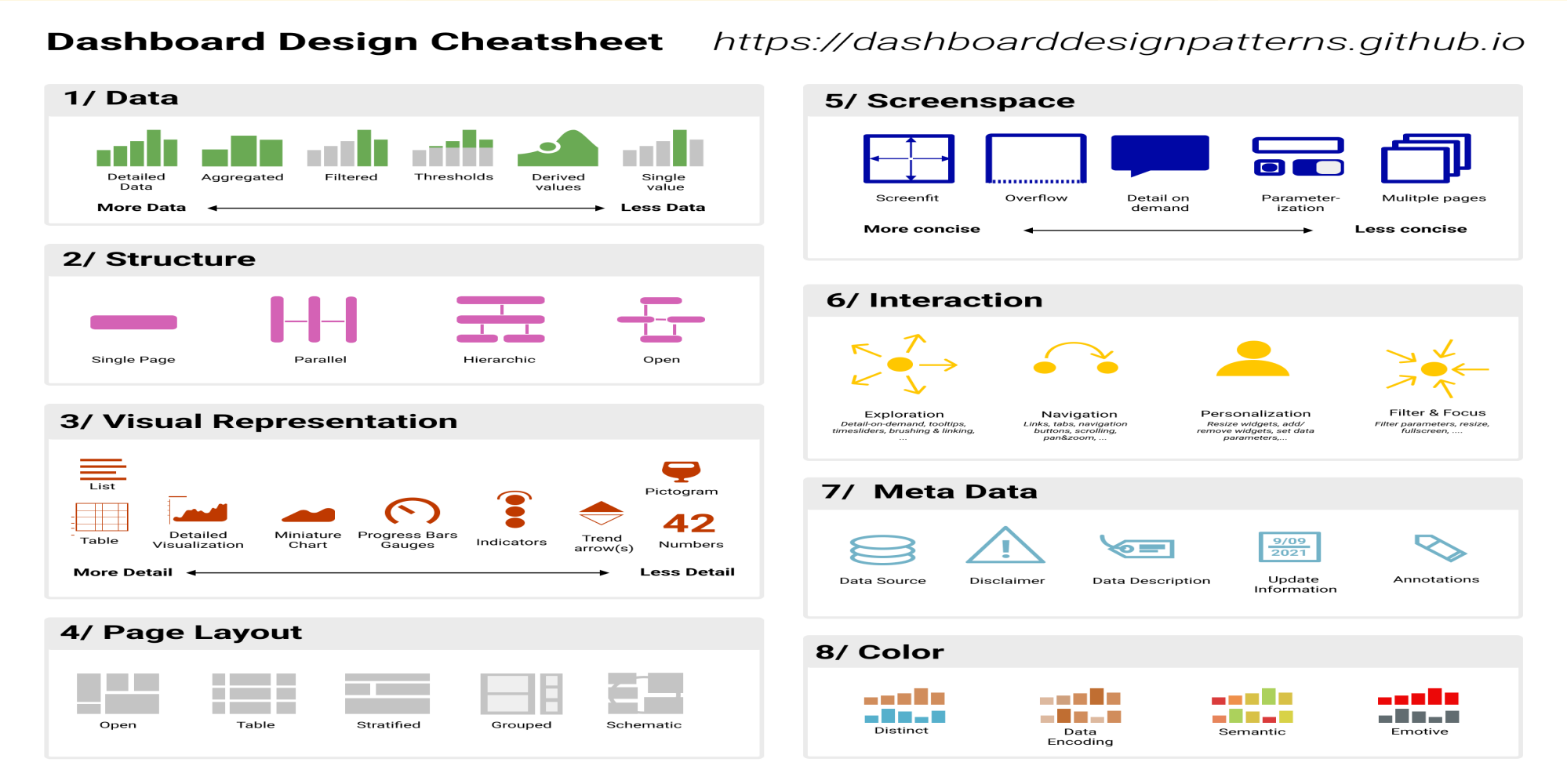


Tableau Dashboard

Reference

- The Data Visualisation Catalogue
- Dashboard Design Pattern [Website][Journal article]

Activity

- Go to GitHub to download the dataset
- Go to Tableau Public and sign up an account