Text

Description automatically generated

Logo

Description automatically generated

Assignment 1 – Initial Analysis

By,

Sharon Appoline Rosary  
 ALY6070-21789 SEC 09 Communication and Visualization for Data Analytics  
Northeastern University  
Instructor: Yoni Dvorkis  
Date of submission: 30 January 2022

**Introduction**

*Data Visualization and its importance:*

In a world full of smartphones in every pocket, it’s interesting to perceive the transformation from bits to exabytes and this data is collected and analyzed to gain meaningful information that can drive major innovations. With so much information being collected through data analysis in today’s business world, we require a way to paint a picture of that data so we can communicate our findings. After drawing meaningful insights and interpretation of results, the final step in the data analysis process, Data visualization translates information into a visual context, such as graphs, charts, or maps. Then, it is presented to stakeholders and decision-makers in an easily comprehensible way. The main goal of data visualization is to make it easier to identify patterns, trends, and outliers in large data sets.

The data set used here for analysis/visualization is a social media dataset containing information about twitter data. Using this data, a twitter dashboard with visualizations and KPIs will be created which will help us to investigate in-depth of data and enable data driven insights for decision making. Understanding how to analyze Twitter data thoroughly helps keep a brand out of the dark. Twitter Analytics allows users to collect data about how a campaign was performed and the trends to follow in the future.

**Data Collection**

A social media dataset containing web-scrapped twitter data can be used to create our dashboard. The dataset contains 181 observations and 21 variables. It contains numerical variables mostly such as impressions, engagement, retweets, replies, clicks, follows, etc. The dataset is attached in the Appendix section. The fields in the dataset are given below.

*Description of the Dataset:*

The dataset contains the following variables –

* Tweet - Tweets posted on twitter.
* Id - id of each tweet.
* Time - time at which the tweet was posted.
* Impressions - The no. of places where it was shown.
* Engagement - The no. of engagements for the tweet (replies, retweets, likes).
* Engagement ratio - Engagement ratio
* Retweets - The no. of retweets for the post.
* Replies - The no. of replies for the post.
* Likes - The no. of likes for the post.
* User profile - The no. of visits of user profile who posted the tweet.
* URL clicks - The no. of clicks made on the tweet’s URL.
* Hashtag clicks - The no. of clicks on the tweet’s hashtag.
* Detail expands - The no. of clicks on the details in case of a huge tweet.
* Permalink clicks - The no. of clicks on permalinks.
* App opens - The no. of app opens from that tweet.
* App installs - The no. of app installs from that tweet.
* Follows - The no. of follows after the tweet.
* Email tweet - The no. of email tweets for the tweet.
* Dial phone - The no. of dialed phones after the tweet.
* Media views - The media views that has happened based on the tweet.
* Media engagements - The media engagements happened based on the tweet.

**Initial Analysis**

* ***Variables in the data:***

There are 21 variables present in the dataset out of which 18 are numerical variables.   
The variables are Tweet, id, time, impressions, engagement, engagement ratio, retweets, replies, likes, user\_ profile, url\_clicks, hashtag\_clicks, detail\_expands, permalink\_clicks, app\_opens, app\_installs, follows, email\_tweet, dial\_phone, media\_views and media\_engagements.

* ***Correlations, patterns, and trends in the data:***

The following correlations, patterns and trends are observed in the data:

* A pattern where the engagement rate increases as the number of engagements increase was observed.

Graphical user interface, chart

Description automatically generated

* The above scatter plot depicts the correlation between engagements and follows, likes, replies, retweets. There is no correlation between engagements and follows whereas there is a slight correlation between engagement & likes, engagement & replies , engagement & retweets.
* ***Audience of the data and questions they would ask about the data:***

There is a wide array of audience right from individual users to enterprises.

* Normal users can use this data to keep track of individual account activities.
* Companies can use this data to do a market research and check reach of their products. Also, to understand the current trends.
* Consulting firms working for a political cause use twitter analytics to strategize their campaigns and follow up on other campaigns, etc.

Some generic questions include:

* Give an overview of daily activity stats.
* What is the reach of the newly launched product?
* How are people feeling about the new product?
* How do people feel about politicians in an area?
* What is the success or reach of the campaign?
* ***Questions that I propose to answer to my audience through data visualizations:***

Questions that I propose to answer to my audience related to the twitter data set include:

* What are the total tweets and media engagement per tweet?
* Compare average impression rate and engagement rate.
* Compare tweets vs details expand.
* What are the top 10 tweets trending this week?
* What is the tweet volume on a weekday basis?
* ***Dashboards being used to show data clearly:***

Dashboards are information management tools that track, analyze, and display key performance indicators, metrics, and data points. They can be used to monitor the overall health of a company, department, or team. Dashboards serve different purposes for different roles. There is no such thing as a one-size-fits-all dashboard. Well-designed dashboards answer business questions and provide insights before one must ask for them.

Benefits of using a dashboard for business:

* Monitor multiple metrics and KPIs at once.
* Easy to read.
* Save time by reducing or eliminating manual reporting tasks.
* Consistent and timely cross-functional communication.

A dashboard can be used to show data clearly by:

* Making sure the data is relevant.
* Using strategic metrics.
* Choosing measurable metrics.
* Keeping dashboard clean.
* Take the ‘at-a-glance’ test.
* Categorize metrics in a nice grid.
* Using dashboard filters.
* ***Types of graphs and charts that can be used to clearly explain the data and answer the business questions:***

The business questions can be answered by visualizations like:

* The total tweets and media engagement per tweet will be visualized using a dual axes chart.
* The comparison between average impression rate and engagement rate can be shown using a dual axes chart. All the four are key metrics(KPIs).
* A bar chart with filter options and dual axes is used to show the top 10 tweets trending rank wise.
* A lollipop chart to shoe the tweet volume on a weekday basis.

**Summary**

Thus, the initial analysis of twitter data set was performed where the key metrics KPIs were identified and the graphs and charts that will be used to answer the business questions were chosen. The next step is to setup the dashboard with all the information gathered during initial analysis.

**References**

[1] Why Data Visualization Is Important. Stark, M. (2020, June 10).  Analytiks. <https://analytiks.co/importance-of-data-visualization/>

[2] Data visualization: What it is and why it matters. SAS. (n.d.). Retrieved January 12, 2022, from <https://www.sas.com/en_us/insights/big-data/data-visualization.html>

[3] Hayward, E. (2021, December 7). *What is a data dashboard?* Klipfolio.Com. <https://www.klipfolio.com/blog/what-is-a-data-dashboard#:%7E:text=A%20data%20dashboard%20is%20an,department%2C%20or%20a%20specific%20process.&text=A%20dashboard%20transforms%20the%20raw%20data%20into%20something%20human%2Dreadable>.

**Appendix**

The twitter dataset is attached below:

