42047: Data Processing with Python

Assignment (Part C)

Report: Data Analysis and Visualization

Student Name and ID: Thanh Thuy An Tran

Date: 29 October 2021

Table of Contents

Abstract

This report utilizes exploratory data analysis model using Python programming language in investing the impact of social media on brand building and customer engagement. Using data of twelve posts' performance metrics extracted from the company Facebook page in a 12-month period, this report evaluates 500 posts from a renown cosmetic brand. The result indicates that social media platform has positive influence on increasing brand engagement and brand awareness.

The findings have crucial implications for the management of social media marketing in branding, maintaining customer relationship and promoting products using social media contents.

1. Introduction and Background

The growing popularity of social media platform has changed the way brands interact with customers, leading brands to seek new interactive way of engaging with customers. In 2020, over 3.6 billion people were using social media worldwide which accounted for nearly half of the world population, this number is projected to increase to almost 4.41 billion in 2025 (Statista, 2021). Therefore, it is undeniable that social media has become one of the most important channels for companies to reach and communicate with consumers.

This report will apply exploratory data analysis (EDA) method to solve a business problem. The report consists of three main parts: the first part will describe the dataset, provide background information about the company as well as identifying business problems, the second part will elaborate steps in analysing the dataset using Python and the final part will shed light on the output interpretation to answer the business problem with a conclusion at the end.

1.1 Business problem

Despite the advantages that social media provided, one of the challenges brands facing today is to measure the impact of social media marketing on brand building and customer engagement. However, along with technological advancement, the increasing availability of data information has bride the gap between companies and customers. Brands now can use data to understand about consumers and evaluate their interaction through social media. Therefore, this report focusses on examining the impact of social media factors including post category, post time, post advertisement on building brand image and drive customer engagement. The main goals of this report are as follows:

- Assessing the efficiency of Facebook posts in increasing brand awareness
- Identifying the impact of Facebook post factors such as time, frequency, types on lifetime post customers
- Defining a causal relation between the use of social media and customer relationship management by evaluating lifetime post consumers (LPC) and engagement ratio

1.2 Business Question

To solve the business problems, a listed of business question below needed to be answered:

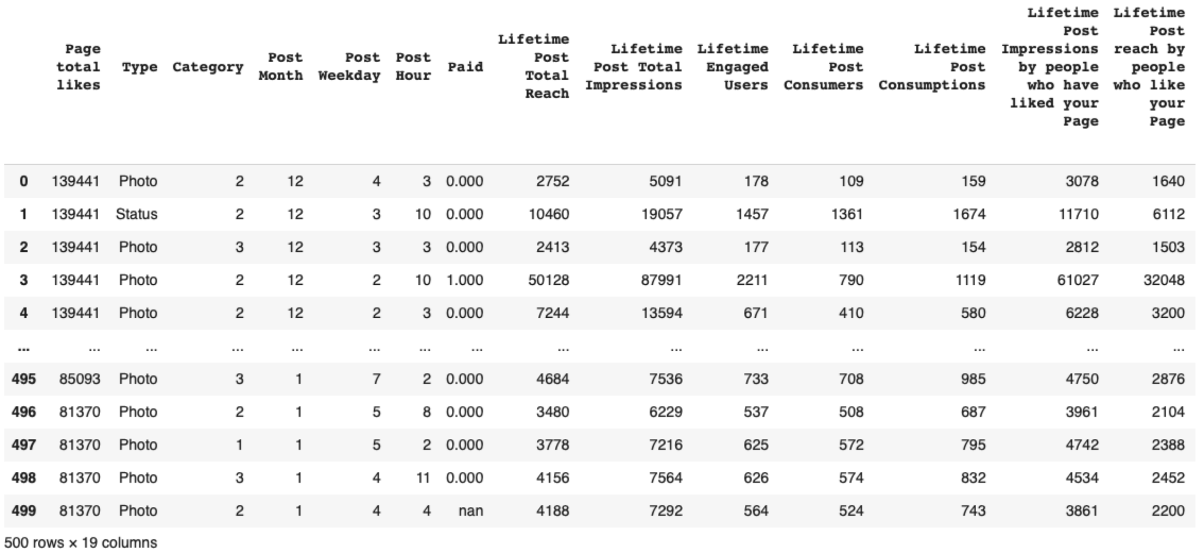
- Does the number of people like brand’s Facebook page increase when brand publish new posts? How does it affect LPC?

- What are the post features that influence more on post total interactions?

- How does social media post influence LPC and engagement ratio?

1.3 Dataset

This dataset is taken from a Facebook page of a known cosmetics brand with 500 posted published in the year of 2014. The dataset has 500 rows and 19 columns including 7 features known prior to post publication (types of post, time, promotion) and 12 features for evaluating post (reach, like, comment or share). It was collected and distributed by UCI Machine Learning Repository centre.



	Page total likes	Type	Category	Post Month	Post Weekday	Post Hour	Paid	Lifetime Post Total Reach	Lifetime Post Total Impressions	Lifetime Engaged Users	Lifetime Post Consumers	Lifetime Post Consumptions	Lifetime Post Impressions by people who have liked your Page	Lifetime Post reach by people who like your Page
0	139441	Photo	2	12	4	3	0.000	2752	5091	178	109	159	3078	1640
1	139441	Status	2	12	3	10	0.000	10460	19057	1457	1361	1674	11710	6112
2	139441	Photo	3	12	3	3	0.000	2413	4373	177	113	154	2812	1503
3	139441	Photo	2	12	2	10	1.000	50128	87991	2211	790	1119	61027	32048
4	139441	Photo	2	12	2	3	0.000	7244	13594	671	410	580	6228	3200
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	85093	Photo	3	1	7	2	0.000	4684	7536	733	708	985	4750	2876
496	81370	Photo	2	1	5	8	0.000	3480	6229	537	508	687	3961	2104
497	81370	Photo	1	1	5	2	0.000	3778	7216	625	572	795	4742	2388
498	81370	Photo	3	1	4	11	0.000	4156	7564	626	574	832	4534	2452
499	81370	Photo	2	1	4	4	nan	4188	7292	564	524	743	3861	2200

500 rows × 19 columns

Figure 1: Overview of the first 14 columns in the dataset

The dataset contains two main types of data:

- Categorization which identifies post features (type, category, time, day, month, advertisement)
- Performance which measures the impact of the post (reach, impression, engagement)

This report will explore the list of attributes in the dataset to answer the business question: page total likes, type, category, post weekday, post hour, paid, lifetime post consumers, total interactions.

2 Overview of the Data Analysis Pipeline

2.1 Workflow Diagram

Exploratory data analysis method plays a crucial role in data analysis, by summarizing and accounting statistical characteristics of data, EDA can help users to find patterns and uncover insights which can guide to further analysis (Sahoo et. al, 2019). Figure 2 below illustrates the workflow applied in this report using Python language.

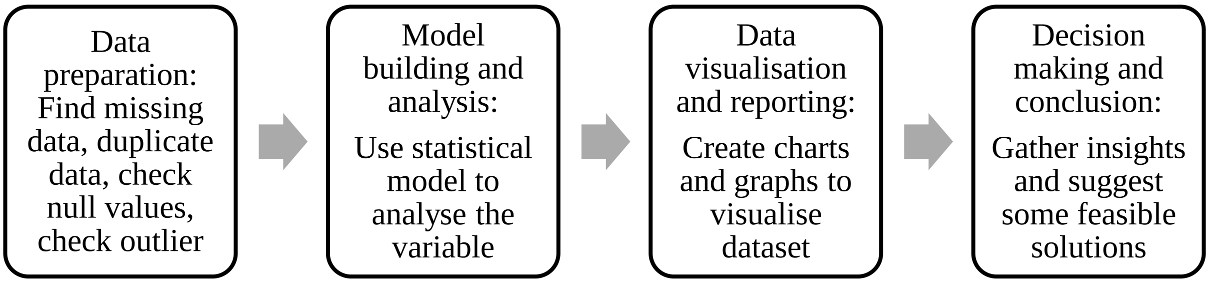
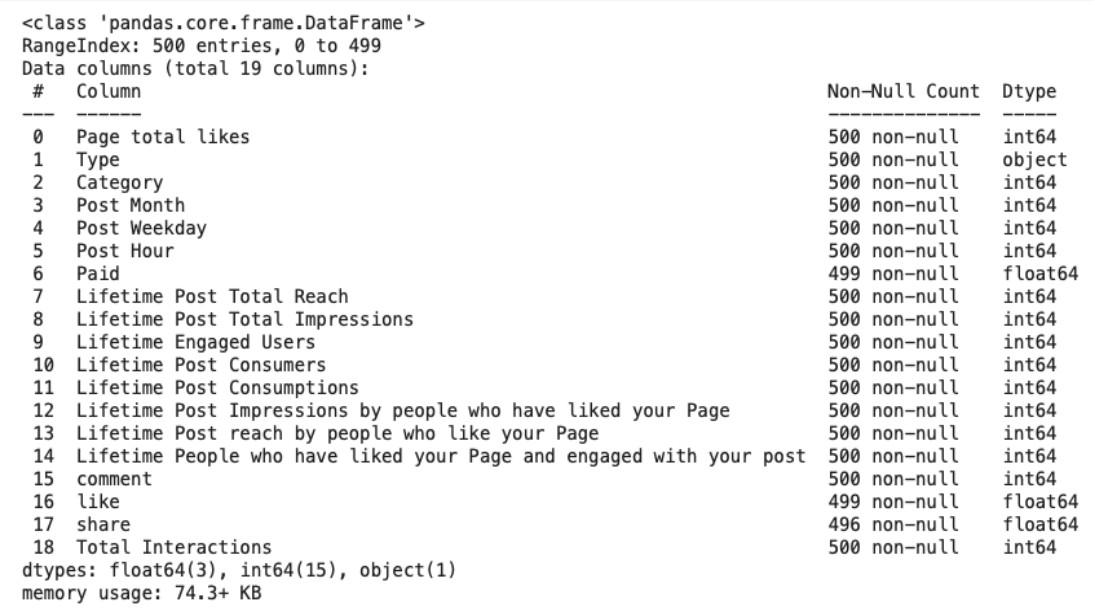

Data preparation: Find missing data, duplicate data, check null values, check outlier
Model building and analysis: Use statistical model to analyse the variable
Data visualisation and reporting: Create charts and graphs to visualise dataset
Decision making and conclusion: Gather insights and suggest some feasible solutions

Figure 2: Exploratory data analysis workflow

2.2 Data Preparation

Data preparation is the first stage of data analysis to help us understand the content and characteristics of the dataset. Function head(), tail() and info() were applied to get the brief information about the dataset. Figure 3 shows the initial overview of the dataset characteristics:



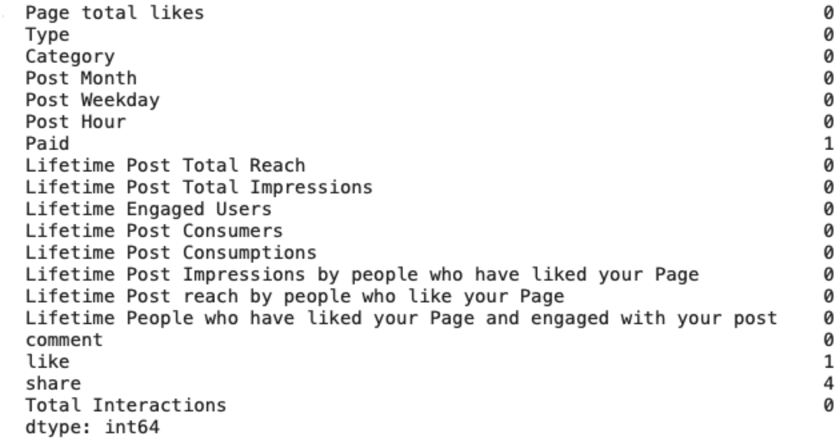
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 19 columns):
 #   Column                                                               Non-Null Count  Dtype
---  ------                                                               --------------  -----
 0   Page total likes                                                     500 non-null    int64
 1   Type                                                                 500 non-null    object
 2   Category                                                             500 non-null    int64
 3   Post Month                                                           500 non-null    int64
 4   Post Weekday                                                         500 non-null    int64
 5   Post Hour                                                            500 non-null    int64
 6   Paid                                                                 499 non-null    float64
 7   Lifetime Post Total Reach                                            500 non-null    int64
 8   Lifetime Post Total Impressions                                      500 non-null    int64
 9   Lifetime Engaged Users                                               500 non-null    int64
 10  Lifetime Post Consumers                                              500 non-null    int64
 11  Lifetime Post Consumptions                                           500 non-null    int64
 12  Lifetime Post Impressions by people who have liked your Page         500 non-null    int64
 13  Lifetime Post reach by people who like your Page                     500 non-null    int64
 14  Lifetime People who have liked your Page and engaged with your post  500 non-null    int64
 15  comment                                                              500 non-null    int64
 16  like                                                                 499 non-null    float64
 17  share                                                                496 non-null    float64
 18  Total Interactions                                                   500 non-null    int64
dtypes: float64(3), int64(15), object(1)
memory usage: 74.3+ KB
```

Figure 3: Overview of data characteristics

Apart from checking the data, new variables which are engagement and click-through rate (CTR) ratio were created for the purpose of data analysis. Statistical method was also applied to calculate the mean value of total interactions, engagement ratio and CTR per Facebook post, results are shown in the code notebook.

2.3 Missing value exploration

Finding missing/ null/ duplicate values is the important part of preparing data, as it can cause inconsistency that impact the process of data analysis. Figure 4 shows the overview of null values in the dataset. As the number of null values are quite small compared to overall dataset, remove them will not affect the analysis.



```
Page total likes                                                       0
Type                                                                   0
Category                                                               0
Post Month                                                             0
Post Weekday                                                           0
Post Hour                                                              0
Paid                                                                   1
Lifetime Post Total Reach                                              0
Lifetime Post Total Impressions                                        0
Lifetime Engaged Users                                                 0
Lifetime Post Consumers                                                0
Lifetime Post Consumptions                                             0
Lifetime Post Impressions by people who have liked your Page          0
Lifetime Post reach by people who like your Page                      0
Lifetime People who have liked your Page and engaged with your post   0
comment                                                                0
like                                                                   1
share                                                                  4
Total Interactions                                                     0
dtype: int64
```

Figure 4: Overview of null values in the dataset

After dropping the null values, a heatmap was created to check the overall dataset, it is clear from the heatmap that there is no missing value in the data (Figure 5).

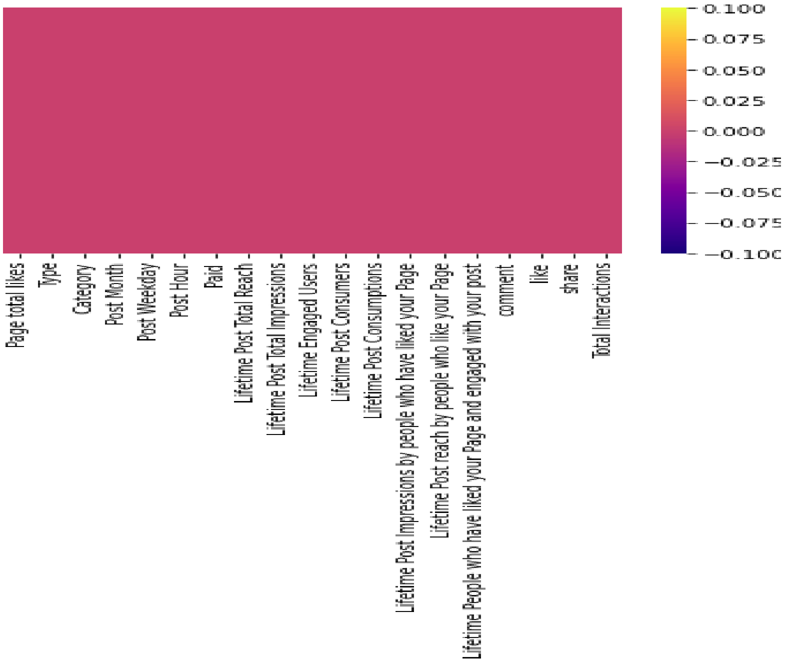


Figure 5: Heatmap of all values in the dataset

2.4 Outlier identification

Outlier detection is the process of identifying data points that have extreme values compared to the rest of the distribution. Box plot was chosen to be the method of identifying outliers in the dataset. A box plot was used on certain variables in the dataset such as: “Page total likes”, “Post Month”, “Post Weekday”, “Post Hour”, “Total interactions”, “LPC”. Among those, only “Total interactions” and “LPC” have outliers with 40 and 37 outliers respectively (Figure 6).

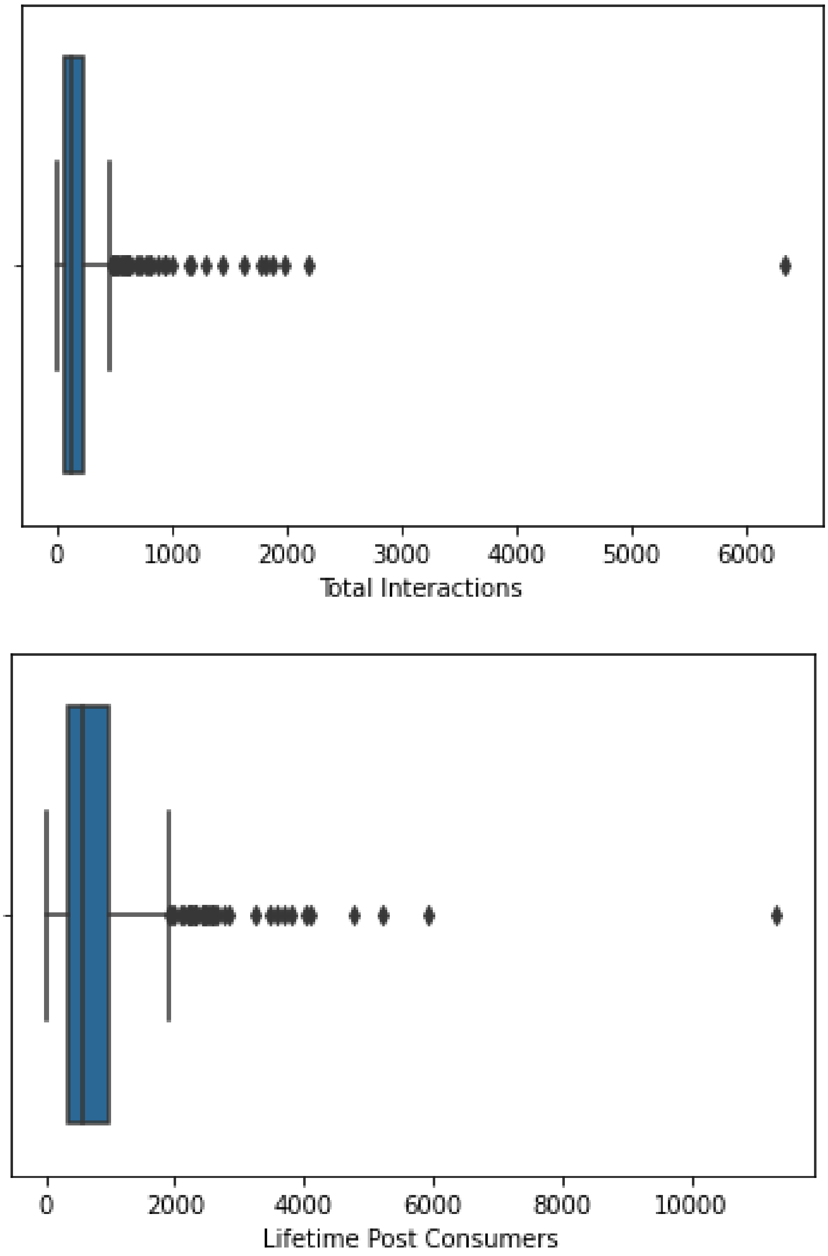


Figure 6: Box Plot for outlier

Given the size of the dataset with 500 rows and 19 columns, dropping all the outliers off the data would potentially affect the accuracy of overall dataset. Therefore, the outliers will not be eliminated.

2.5 Data Visualization

Data visualisation is a crucial process to display a massive amount of data in a way that is easily accessible and understandable (Sadiku et. al. 2016). Choosing the right visualisation method is the key when displaying the data in graphical form. For this I have imported Matplot lib, Seaborn library packages in Python notebook. First, a line chart was made to display the total page likes of Facebook over the studied period (see Figure 7) and a scatter plot was created to represent the relationship of the number of lifetime post consumers and total page like (see Figure 8):

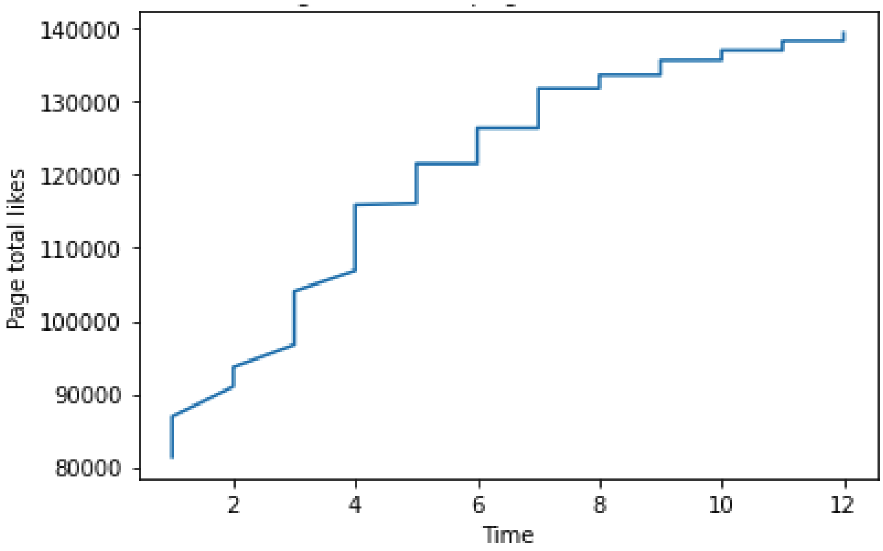


Figure 7: Total page like overtime

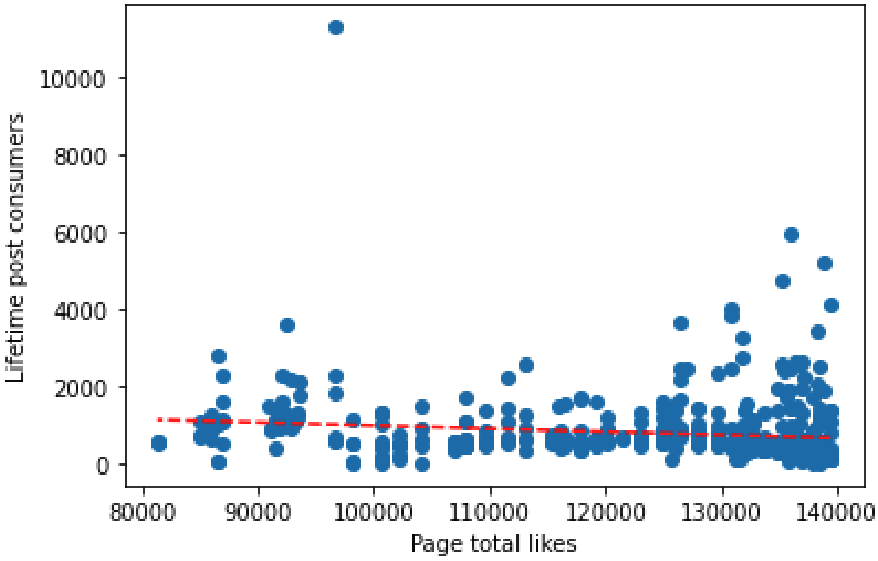
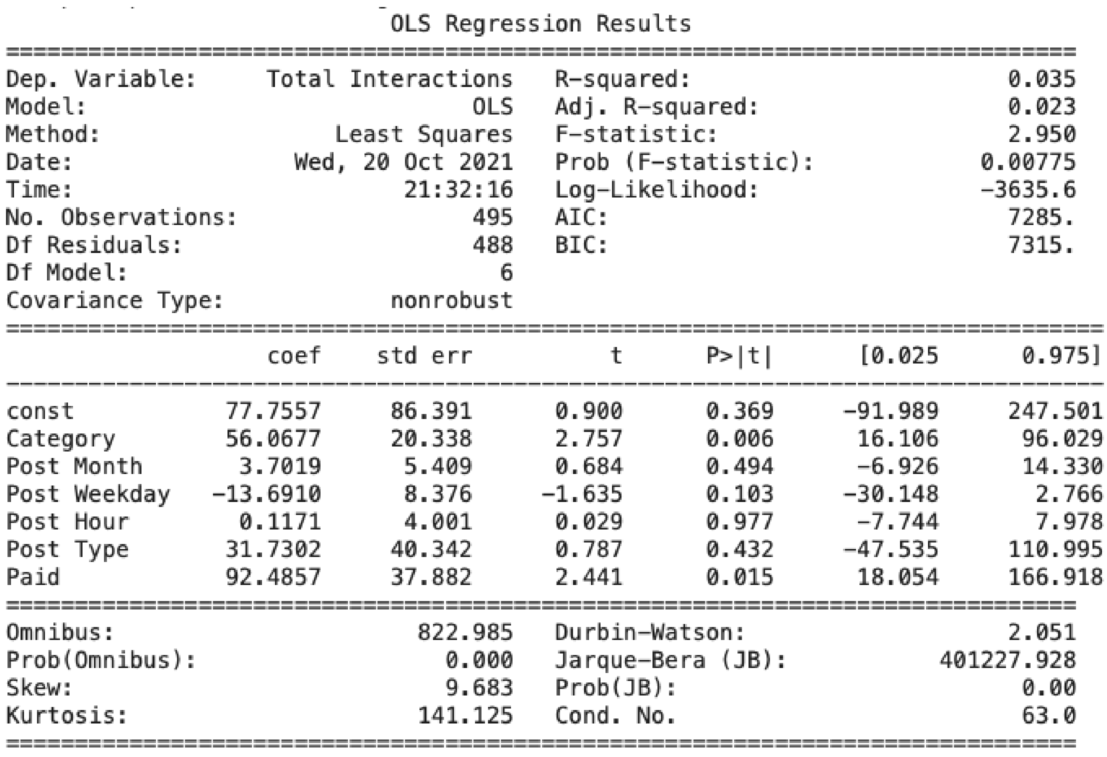


Figure 8: Influence of total page likes on lifetime post consumers

From the two charts created above, it is clear from Figure 7 that Page Likes increase overtime as brand satisfaction is translated into social media interaction with the company's page. However, we can see that in Figure 8, LPC decreased which means users are not so keen to engage with posts being published. This report also used statistical regression model to evaluate the relationship between dependent variable “Total interactions” and other independent variables as describe in Figure 9:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Total Interactions   R-squared:                       0.035
Model:                            OLS   Adj. R-squared:                  0.023
Method:                 Least Squares   F-statistic:                     2.950
Date:                Wed, 20 Oct 2021   Prob (F-statistic):            0.00775
Time:                        21:32:16   Log-Likelihood:                -3635.6
No. Observations:                 495   AIC:                             7285.
Df Residuals:                     488   BIC:                             7315.
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           77.7557     86.391      0.900      0.369     -91.989     247.501
Category        56.0677     20.338      2.757      0.006      16.106      96.029
Post Month       3.7019      5.409      0.684      0.494      -6.926      14.330
Post Weekday   -13.6910      8.376     -1.635      0.103     -30.148       2.766
Post Hour        0.1171      4.001      0.029      0.977      -7.744       7.978
Post Type       31.7302     40.342      0.787      0.432     -47.535     110.995
Paid            92.4857     37.882      2.441      0.015      18.054     166.918
==============================================================================
Omnibus:                      822.985   Durbin-Watson:                   2.051
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           401227.928
Skew:                           9.683   Prob(JB):                         0.00
Kurtosis:                     141.125   Cond. No.                         63.0
==============================================================================
```

Figure 9: Regression model

From the regression model, there are two factors that have a significantly positive impact on Total interactions: Category (P-value = 0.006) and Paid (P-value = 0.015). A density plot and bar plot were created on factors “Category” and “Paid” to get further insights into these variables, more interpretation will be at Python Notebook.

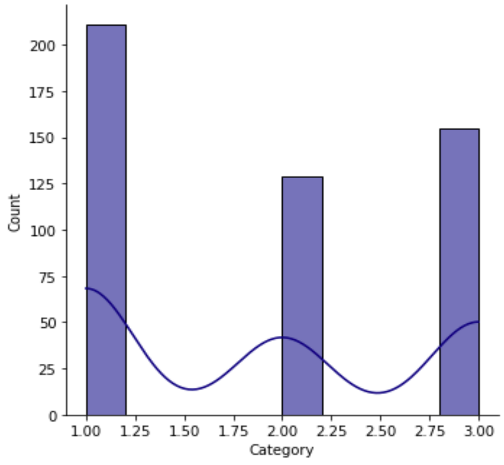


Figure 10: Density plot of Category

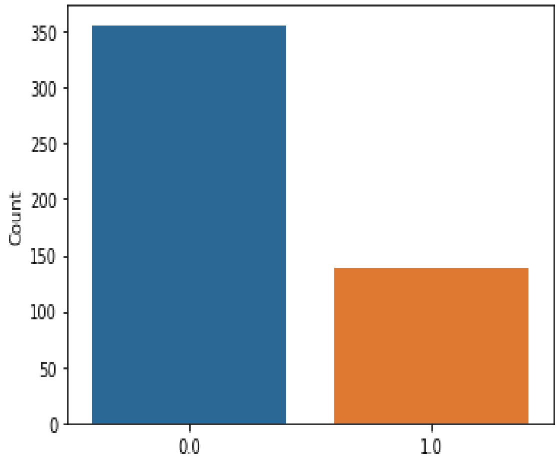
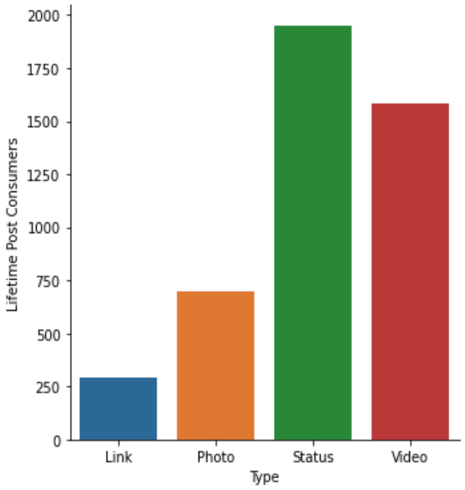


Figure 11: Facebook Ad bar chart

To understand more about the impact of Facebook on improving brand engagement and maintaining customer relationship, further visual analysis had been conducted. Figure 12 displays the impact of Post Type on LPC and Figure 13 shows the distribution of different content types throughout the studied period. As we can see, despite accounting for only 9.09%

of overall post type, status has the most significant impact on LPC, followed by video and photo. Link is the least attractive post type as it generates the lowest number of LPC.



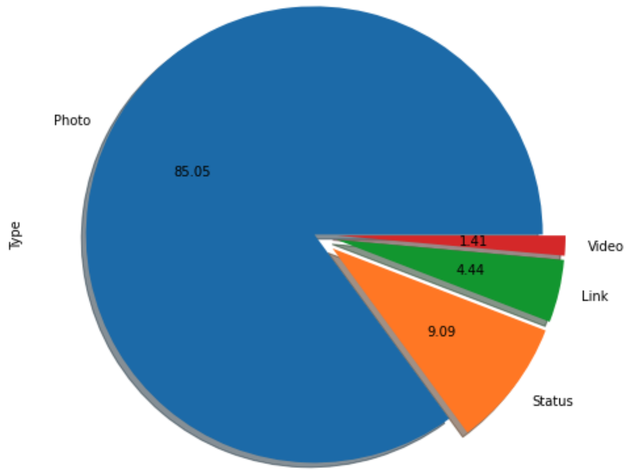
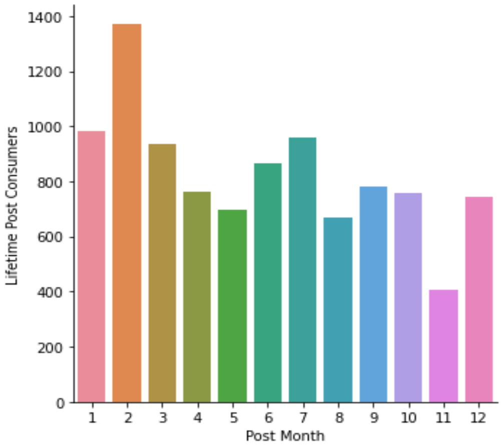


Figure 12: Impact of Post type on LPC

Figure 13: Post type distribution

The following three graphs shows the impact of different factors on LPC in terms of Post month, Post hour, Post weekday. Figure 17 displays the engagement ratio and CTR distribution over the year. Overall, a seasonality pattern is clearly witnessed in all the charts with more detailed interpretation written in Python Notebook.



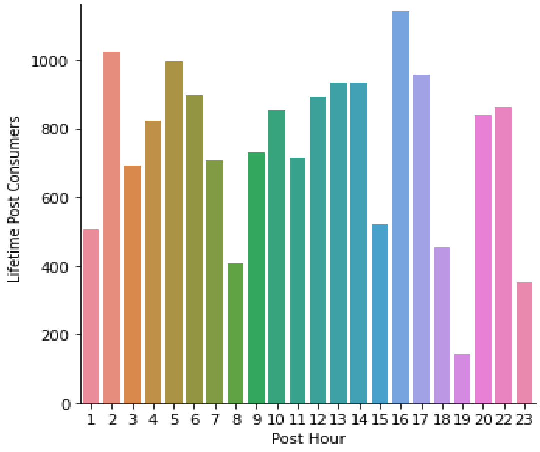
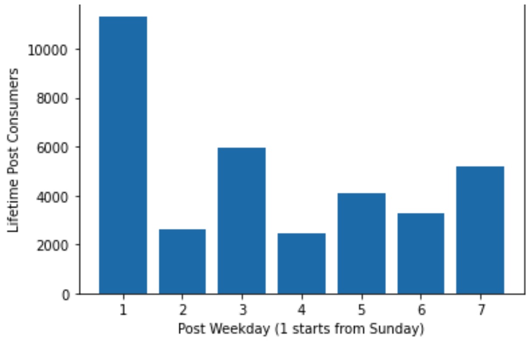


Figure 14: Impact of Post month on LPC

Figure 15: Impact of Post hour on LPC



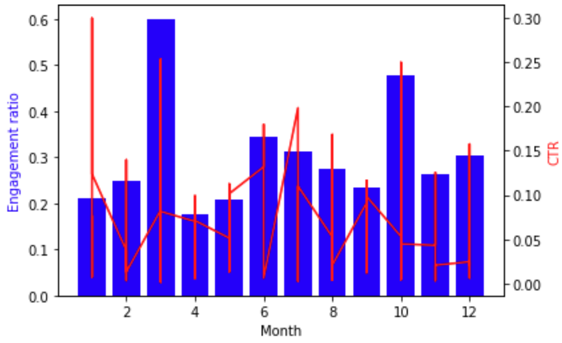


Figure 16: Impact of Post weekday on LPC

Figure 17: Engagement ratio and CTR distribution

3 Discussion and Conclusions

From the data analysis, it can be concluded that the use of social media platform such as Facebook has positive impact on brand's awareness, as the number of people who likes the page has increased steadily over time. However, the rate of Lifetime Post Consumers decreased when page likes reach a certain number which means users are not so keen to engage with posts being published. Such issue may disclose some erosion of the company's Facebook page, since users are seeing the contents published, but are not interacting with it. Categorization and Facebook Ad are the two factors that influence post total interactions. However, how post category and Facebook Ad is used depends on the company's strategy and goal. Furthermore, there is a causal relationship between the use of social media and customer relationship management as both engagement ratio and CTR witnessed an upward trend over the year. The "Type" content is considered the most relevant input that affect lifetime post consumers. Posts use the "Status" type are likely to result in twice the impact of the remaining types. Also, seasonality pattern was found regarding the "Month", "Hour" and "Weekday" that the post was published.

4 References

[1] Statista. 2021. *Number of social media users 2025 | Statista*. [online] Available at: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/> [Accessed 13 October 2021].

[2] Archive.ics.uci.edu. 2021. *UCI Machine Learning Repository: Facebook metrics Data Set*. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Facebook+metrics> [Accessed 14 October 2021].

[3] N. O. Sadiku, M., E. Shadare, A., M. Musa, S. and M. Akujuobi, C., 2016. DATA VISUALIZATION. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, Volume. 02(Issue.12).

[4] Sahoo, K., Pramanik, J., Kumar Samal, A. and Kumar Pani, S., 2019. Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), pp.4727-4735.