Quarterly and Monthly Analysis for Facebook Live Fashion Sellers in Thailand

Bienvenido Hiyas Jr. Student ID: 13824819

***Abstract*— Online selling is common nowadays especially in social media like Instagram, twitter and Facebook. However, there is no guarantee that every time you post or sell your products every day, people will be interested in it. This report will provide a statistical analysis for quarterly and monthly of a data set from Facebook live sellers in Thailand from 2012 to 2018. This will also provide a clear comparison of the successful posts that gathered many reactions, comments and shares within the quarter and months of the year. The data set is from 10 Thai fashion and retail online post of different nature like links, videos, pictures and statuses. The data set is then manipulated to provide a visualization of total reactions, comments and shares in each quarter and months. It is found that the number of likes in a post is directly proportional to the number of engagements metrics like share, comments and reaction. In addition, most of the posts with many reactions, shares and comments are highest in the 1st and 2nd quarters of the year. With this, online sellers when to focus their online selling activity.**

I. Introduction

Selling on the Thai ecommerce market is a great option for growth-minded online sellers according to the Web Interpret website (www.webinterpret.com). In addition, there are 29,078,158 internet users in Thailand and Internet penetration amounts to 42.70%. One of the most common internet platform used in Thailand is Facebook. Thailand has 46 million registered Facebook users. Therefore, of all the Facebook accounts in the world, 2% of them are logging on from Thailand and this number represents a huge percentage of the Thai population and clearly Facebook is an essential part of daily life for most Thai people [1].

With this, most Thai fashion and cosmetics retail sellers are actively posting their products in Facebook via videos, pictures, statuses and links and gathering responses like comments, sharing and reactions. However, there seems to be a significant difference of Facebook users engagements in every quarter and months within the year with regards to their post. This report will provide a quarterly and monthly statistical analysis of the Facebook posts from 10 Thai fashion and cosmetic retail sellers as well as to provide a visual presentation of the comments, shares and reactions as well as the type of posts that gathered the highest numbers of reactions, comments and shares.

II. Data

The data set used in this investigative analysis is the Facebook Live Sellers in Thailand Data Set, from Nassim Dehouche, Mahidol University International College, nassim.deh '@' mahidol.edu.

The data file is a comma separated values file named, Live.csv [2] and the Data Set Description can be found in UCI Machine Learning Repository [3]. The Live.csv file has 7051 rows and 16 Columns. However, the last 4 columns named Column1, Column2, Column3 and Column4 don't have values data on each column. The first row contains the variables indicated on table 1 below.

Variables	Type
status_id	Qualitative Nominal
status_type	Qualitative Nominal
status_published	Qualitative Ordinal
num_reactions	Quantitative Discrete
num_comments	Quantitative Discrete
num_shares	Quantitative Discrete
num_likes	Quantitative Discrete
num_loves	Quantitative Discrete
num_wows	Quantitative Discrete
num_hahas	Quantitative Discrete
num_sads	Quantitative Discrete
num_angrys	Quantitative Discrete
Column1	N/A
Column2	N/A
Column3	N/A
Column4	N/A

Table1 – List of Variables and Types.

The data set was taken from Facebook pages of 10 Thai fashion and cosmetics retail sellers posts of a different nature **status_type** (video, photos, statuses, and links). The engagement metrics consist of comments, shares, and reactions. Moreover, the number of reactions is equivaent to the total number of likes, loves, wows, hahas, sads and angrys.

Prior to these investigation, the variability of consumer engagement is analysed through a Principal Component Analysis, highlighting the changes induced by the use of Facebook Live. The seasonal component is analysed through a study of the averages of the different engagement metrics for different time-frames (hourly, daily and monthly). Finally, the statistical outlier posts were identified, that are qualitatively analysed further, in terms of their selling approach and activities.

III. Methods

A range of Data Science methods with R studio commands and libraries are used for the pre-processing of the Live.csv data. These were conducted in the R-Studio software[4]. The libraries used are dplyr, ggplot2 and tidyr.

A. Data Presentation

The R function **read.table()** was used to import the Live.csv data to R studio. Then the argument **na.strings** = **“”** was used to import the missing values. Finally, the argument **header** was set to **T**(True) to instruct R that the first line of the file are the variable names. The new data is saved in the new data named, **Data**. **Data** has now a total of 7050 observations with 16 variables. **Data** was then inspected with the **summary()** command and found that there are no missing values in every observation except for the last 4 variables; **Column1, Column2, Column3 and Column4** where they don’t have data or values. Therefore in order to eliminate them, the **select()** command was used excluding the unnecessary four variables mentioned above. In addition, since the purpose of this investigation is to investigate the selling activities of the online sellers per quarters and months, three new variables **Quarters**, **Months** and **Years** were added to the data using the **mutate()** function of the dplyr package. **Quarters** variable was transformed from the **status_published** variable using the command **quarters.POSIXt(). Years** variable was also taken from the **status_published** variable using the command **as.numeric()** with **format()** argument. Furthermore, to have an ordered visualization for the **Months** variable, the **factor()** command with arguments **level()** was used. To have a fair data, the Data was manipulated to excluded the 2012 rows since this year has only Q3 and Q4 data as well as the 2018 rows since this year only has Q1 and Q4 data. The function used was **filter()** with argument **Years>2012** and **Years<2018** along with **Pipe** function.

Finally, the data was then manipulated to have new data called **Data_Years, Data_Quarters and Data_Months** that contains specific data for years, quarters and months to satisfy the objective of this investigative analysis. The new variables used the **Pipe** function to have a sequential operation. For the **Data_Years** variable, the data from the Data was group by Years, Quarters and Months using the function **group_by()** then using **Pipe** summarizes the means of the number of reactions, comments and shares using the **summarise()** function. This code also filtered out the other variables since they are unnecessary for the objective. The **Data_Quarters and Data_Months** data also follows the method **for Data_Years**, however, it groups the **Data_Years** data instead of the **Data** data.

B. Type Conversion

When the data is imported using the function **read.table()**, the variable type is automatically assigned. To perform some types of numerical analysis, the data types of these variables were manually assigned. The variables **status_id** and **status_type** factors were already assigned correctly but the other variables were not. The status_published variable was assigned to **as.POSIXct()** and the rest were assigned as **as.numeric()** variables**.**

In result to above data presentation and type conversion methods, the data is now structured and tidy data set.

C. Data Exploration

In order to explore the data, correlation matrices between numerical data types was used. A new data **Data2** was created that includes only the numerical variables. Then the function **cor()** was used with the argument **method="pearson"** to measure the linear correlation between the numeric variables using the Pearson Correlation method.

D. Data Visualization

Data visualization was done using the **ggplot()** function added with scatter plot **geom_point()** and additional **geom_smooth()** with arguments "**method=lm**" in order to see patterns on the plot. Furthermore, Bar Graph **geom_col()** with **fill** arguments for colors was also used along with **ggplot** for proper and tidy visualization. For Boxplot, **geom_boxplot()** was used in the status_type visualization with arguments **fill** for colors and **coord_cartesian()** arguments to have a fair grid in the plot

IV. Results and Discussions

The original data from the Live.csv file was spreadsheet of 7051 rows and 16 columns. After doing the data presentation, the result is a data set of 7050 observation and 12 variables where the first row of the data as the variables and the unnecessary columns were omitted. Furthermore, to meet the objective of this investigation, thre new variables were added; **Quarters**, **Months** and **Years**, which were taken from the **status_published** variable. With this our data set was ready for further statistical analysis.

Data exploration was then implemented where the correlation matrices was used. The result can be found in the figure 1 below.

```
> cor(Data2, method ="pearson")
               num_reactions num_comments num_shares  num_likes num_loves   num_wows num_hahas   num_sads num_angrys
num_reactions    1.00000000    0.1508429  0.2507225 0.99492337 0.3050028 0.26775182 0.1760280 0.07513801 0.12432606
num_comments     0.15084290    1.0000000  0.6406374 0.10168704 0.5212235 0.16239399 0.3250480 0.23645268 0.22518437
num_shares       0.25072252    0.6406374  1.0000000 0.17249193 0.8200002 0.40762786 0.3998256 0.19996997 0.31251256
num_likes        0.99492337    0.1016870  0.1724919 1.00000000 0.2093081 0.20779976 0.1207838 0.05216884 0.08743065
num_loves        0.30500278    0.5212235  0.8200002 0.20930809 1.0000000 0.50879804 0.5078296 0.20760019 0.37100127
num_wows         0.26775182    0.1623940  0.4076279 0.20779976 0.5087980 1.00000000 0.2877562 0.08650272 0.18308651
num_hahas        0.17602804    0.3250480  0.3998256 0.12078383 0.5078296 0.28775619 1.0000000 0.14142086 0.21191016
num_sads         0.07513801    0.2364527  0.1999700 0.05216884 0.2076002 0.08650272 0.1414209 1.00000000 0.14207247
num_angrys       0.12432606    0.2251844  0.3125126 0.08743065 0.3710013 0.18308651 0.2119102 0.14207247 1.00000000
```

Figure 1 – Pearson Correlation Result for all Numeric variables.

In addition, a data visualization on the relationship in figure 1 above was also done to have a clear visualization as shown below Figure 2.

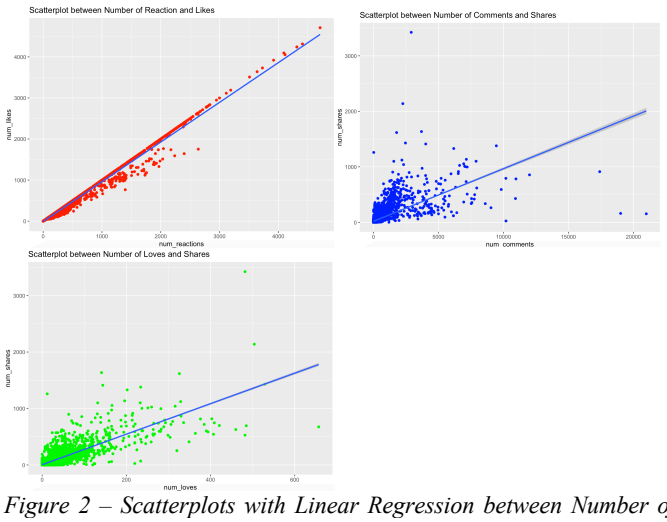

Figure 2 – Scatterplots with Linear Regression between Number of Reaction and Likes, Number of Comments and Shares and Number of Loves and Shares.

As shown in the figure 1, there is an almost perfect correlation of 0.995 between the number of reactions and number of likes. We can visualize this as shown in the figure 2 where almost all that follows the straight blue diagonal line. This means that as the number of reaction increases, the number of likes also increases. Moreover, a strong correlation of 0.82 between number of shares and number of loves and a 0.640 correlation between number of comments and number of shares can also be observed.

Moving forward, a further data analysis between the years 2013 to 2017 was also done. Years 2012 and 2018 were disregarded because 2012 only contains Q3 and Q4 data while 2018 contains Q1 and Q2. Thus, can make the data unreliable.

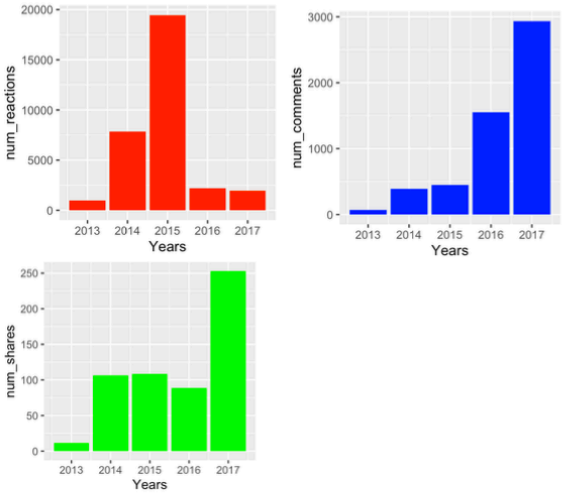

Figure 3 – Bar graph of Years and Number of Reactions, Comments and Shares

The figure 3 above shows the online selling posts from 2013 to 2017. It can be noticed that the year with the highest posts by the 10 Thai fashion and cosmetics online retailer that have the highest reactions was in 2015, the highest comment was from 2017 and the highest share was also 2017.

Further data manipulation and visual analysis was done per Quarter and Months which was the objective of this investigation, in result, a visualization was done in the figure below.

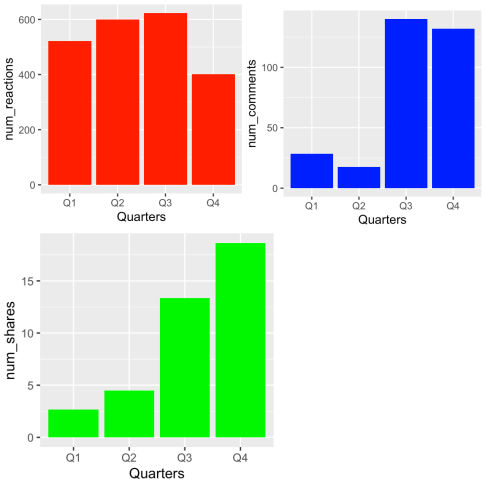

Figure 4 – Bar graph of Numbers of Reactions, Comments and Shares per Quarters

As shown in the figure 4 above, Q2 and Q3 have the highest number of Reactions while Q3 and Q4 have the highest number of comments while Q4 has the highest number of shares. Further analysis was done for the Months as shown on the figure below.

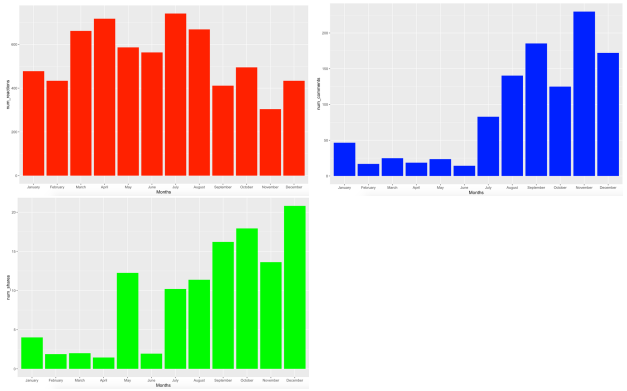

Figure 5 - Bar graph of Numbers of Reactions, Comments and Shares per Months

As shown in the figure 5 above, months March to August have the highest number of reactions and months August to December have the highest number of comments and shares.

Finally, the analysis for the type of posts is shown in the boxplot below.

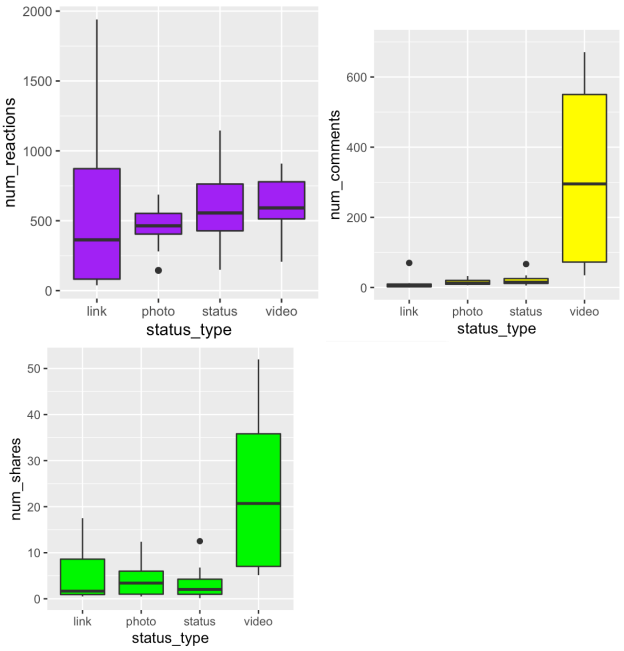

Figure 6 – Boxplot for Type of posts and numbers of Reactions, Comments and Shares.

As shown in the figure 6 above, status and video have higher number of reactions compared to the others while video was the highest in terms of the number of comments and shares.

V. Conclusion

This investigation has successfully achieved its objective which is to identify the quarter and months that have the highest reactions, shares and comments of the Facebook posts from 10 Thai fashion and cosmetics retail sellers. Quarters Q2 and Q3 specifically from months March to August were identified that have the highest number of reactions in Facebook. Quarters Q3 and Q4 specifically from months August to December were identified that have the highest numbers of comments and shares. Moreover, the type of post status and video have higher number of reactions compared to the others while video was the highest in terms of the number of comments and shares.

In addition, it is found that the numbers of reactions and likes as well as the number of shares and loves and number of comments and shares have a linear relationship and can be used to predict each other. Therefore, we can also determine the highest quarters and months for the number of shares.

Although this investigation is only limited to the data provided from 2013 to 2017 of the 10 Thai fashion and Cosmetic retail sellers in Facebook, this can also be used as a base platform for further investigation specially for the latest years where online selling activity is still popular in Thailand

VI. Preferences

[1] “Who are Thailand’s 46 Million Facebook Users?” Retrieved August 2, 2017 from https://www.bangkokpost.com/learning/learning-together/1296218/who-are-thailands-46-million-facebook-users-
[2] Index of /ml/machine-learning-databases/00488. Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/00488/
[3] Facebook Live Sellers in Thailand Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand
[4] RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/. Mode: Desktop Version: 1.2.5019 Release Name: Elderflower

VII. Appendix

R. Code used for this investigative analysis.

```
library(dplyr)
library(ggplot2)
library(tidyr)

#Read Data
Data = read.table("Live.csv", header = T, sep =
",",na.strings = "")
summary(Data)

#Select Necessary variables and data / Delete
Columns1,2,3,4
Data = Data %>%
select(status_id,status_type,status_published,

num_reactions,num_comments,num_shares,
num_likes,num_loves,num_wows,
                num_hahas,num_sads,num_angrys)
str(Data)

#Assign Variable Types manually
Data$status_published =
as.POSIXct(Data$status_published , format =
"%m/%d/%Y %H:%M")
Data$num_reactions =
as.numeric(Data$num_reactions)
Data$num_comments =
as.numeric(Data$num_comments)
Data$num_shares = as.numeric(Data$num_shares)
Data$num_likes = as.numeric(Data$num_likes)
Data$num_loves = as.numeric(Data$num_loves)
Data$num_wows = as.numeric(Data$num_wows)
Data$num_hahas = as.numeric(Data$num_hahas)
Data$num_sads = as.numeric(Data$num_sads)
Data$num_angrys = as.numeric(Data$num_angrys)
str(Data)
```

```
#Added new Variables Quarters,Months and Years
Data = Data %>% mutate(Quarters =
quarters.POSIXt(status_published)) %>%
  mutate(Months =
months.POSIXt(status_published)) %>%
  mutate(Years =
as.numeric(format(Data$status_published,'%Y')))

#Assign type for new Variables
Data$Years = as.numeric(Data$Years)
Data$Quarters = as.factor(Data$Quarters)
Data$Months = factor(Data$Months, levels =
c("January","February","March","April",

"May","June","July","August","September",
                                    "October","November",
"December"))
#Eliminate rows with Year 2012 and 2018
Data = Data %>% filter(Years > 2012 ) %>%
filter(Years < 2018)
summary(Data)

#Data Exploration - Pearson Correlation
DataRequired = c(
"num_reactions","num_comments","num_shares","
num_likes","num_loves","num_wows","num_hahas
","num_sads","num_angrys")
Data2 = Data[ ,DataRequired]
cor(Data2, method ="pearson")

#Data Visualization - scatterplot and Linear
Regression
ggplot(Data2, aes(num_reactions,num_likes)) +
geom_point(color="Red") + geom_smooth(method
= "lm") +
  labs(title = "Scatterplot between Number of
Reaction and Likes")
ggplot(Data2, aes(num_comments,num_shares)) +
geom_point(color="Blue") + geom_smooth(method
= "lm") +
labs(title = "Scatterplot between Number of
Comments and Shares")
ggplot(Data2, aes(num_loves,num_shares)) +
geom_point(color="Green") +
geom_smooth(method = "lm") +
labs(title = "Scatterplot between Number of Loves
and Shares")

#Data Manipulation - Group data set by
Years,Quarters, Months and Means of the Number
of Reactions, comments and shares
Data_Years = Data %>% group_by(Years,
Quarters, Months, status_type) %>%
  summarise(num_reactions=mean(num_reactions),
            num_comments = mean(num_comments),
            num_shares = mean(num_shares))

#Bargraph of Number of Reactions per Years and
Quarters
ggplot(Data_Years,aes(Years, num_reactions)) +
geom_col(fill = "Red")
ggplot(Data_Years,aes(Years, num_comments)) +
geom_col(fill = "Blue")
ggplot(Data_Years,aes(Years, num_shares)) +
geom_col(fill = "Green")

# Quarters Data
Data_Quarters = Data_Years %>% group_by(
Quarters,Months, status_type) %>%
  summarise(num_reactions=mean(num_reactions),
            num_comments = mean(num_comments),
            num_shares = mean(num_shares))

#Bar graph for Quarters Data
ggplot(Data_Quarters,aes(Quarters,
num_reactions)) + geom_col(fill = "Red")
ggplot(Data_Quarters,aes(Quarters,
num_comments)) + geom_col(fill = "Blue")
ggplot(Data_Quarters,aes(Quarters, num_shares)) +
geom_col(fill = "Green")

#Boxplot for status_type
ggplot(Data_Type, aes(status_type,num_reactions))
+ geom_boxplot(fill="Purple")
ggplot(Data_Type,
aes(status_type,num_comments)) +
geom_boxplot(fill="Yellow") +
  coord_cartesian(ylim =
quantile(y=num_comments,c(0,700)))
ggplot(Data_Type, aes(status_type,num_shares)) +
geom_boxplot(fill="Green")

#Months Data
Data_Months = Data_Years %>%
group_by(Months, status_type) %>%
  summarise(num_reactions=mean(num_reactions),
            num_comments = mean(num_comments),
```

```
        num_shares = mean(num_shares))

#Bar graph for Months Data
ggplot(Data_Months,aes(Months, num_reactions))
+ geom_col(fill = "Red") +
facet_wrap("status_type")
ggplot(Data_Months,aes(Months, num_comments))
+ geom_col(fill = "Blue") +
facet_wrap("status_type")
ggplot(Data_Months,aes(Months, num_shares)) +
geom_col(fill = "Green") +
facet_wrap("status_type")
```