

Facebook_EDA

2023-09-29

Set Working Directory

First, set the working directory to the location where your data is stored on your computer. Uncomment and modify the `setwd` line as needed.

```
setwd("~/Desktop/MA575/GroupProject/facebook-metrics")
```

Data Import and Exploration

Read Data

Let's start by reading the data from a CSV file using a semicolon as the delimiter.

```
# Read data from a CSV file using semicolon as delimiter
fc_data <- read.csv("./data/dataset_Facebook.csv", header=TRUE, as.is=TRUE, sep=';')
```

Check Missing Values

```
# Summary of each column
summary(fc_data)
```

```
## Page.total.likes      Type      Category      Post.Month
## Min.   : 81370      Length:500      Min.   :1.00      Min.   : 1.000
## 1st Qu.:112676      Class :character 1st Qu.:1.00      1st Qu.: 4.000
## Median :129600      Mode  :character Median :2.00      Median : 7.000
## Mean   :123194                      Mean   :1.88      Mean   : 7.038
## 3rd Qu.:136393                      3rd Qu.:3.00      3rd Qu.:10.000
## Max.   :139441                      Max.   :3.00      Max.   :12.000
##
## Post.Weekday      Post.Hour      Paid      Lifetime.Post.Total.Reach
## Min.   :1.00      Min.   : 1.00      Min.   :0.0000      Min.   : 238
## 1st Qu.:2.00      1st Qu.: 3.00      1st Qu.:0.0000      1st Qu.: 3315
## Median :4.00      Median : 9.00      Median :0.0000      Median : 5281
## Mean   :4.15      Mean   : 7.84      Mean   :0.2786      Mean   : 13903
## 3rd Qu.:6.00      3rd Qu.:11.00     3rd Qu.:1.0000      3rd Qu.: 13168
## Max.   :7.00      Max.   :23.00     Max.   :1.0000      Max.   :180480
##
##                      NA's      :1
## Lifetime.Post.Total.Impressions Lifetime.Engaged.Users Lifetime.Post.Consumers
## Min.   : 570                      Min.   : 9.0          Min.   : 9.0
## 1st Qu.: 5695                      1st Qu.: 393.8        1st Qu.: 332.5
```

```

## Median : 9051 Median : 625.5 Median : 551.5
## Mean : 29586 Mean : 920.3 Mean : 798.8
## 3rd Qu.: 22086 3rd Qu.: 1062.0 3rd Qu.: 955.5
## Max. :1110282 Max. :11452.0 Max. :11328.0
##
## Lifetime.Post.Consumptions
## Min. : 9.0
## 1st Qu.: 509.2
## Median : 851.0
## Mean : 1415.1
## 3rd Qu.: 1463.0
## Max. :19779.0
##
## Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
## Min. : 567
## 1st Qu.: 3970
## Median : 6256
## Mean : 16766
## 3rd Qu.: 14860
## Max. :1107833
##
## Lifetime.Post.reach.by.people.who.like.your.Page
## Min. : 236
## 1st Qu.: 2182
## Median : 3417
## Mean : 6585
## 3rd Qu.: 7989
## Max. :51456
##
## Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
## Min. : 9.0
## 1st Qu.: 291.0
## Median : 412.0
## Mean : 610.0
## 3rd Qu.: 656.2
## Max. :4376.0
##
## comment like share Total.Interactions
## Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.0
## 1st Qu.: 1.000 1st Qu.: 56.5 1st Qu.: 10.00 1st Qu.: 71.0
## Median : 3.000 Median : 101.0 Median : 19.00 Median : 123.5
## Mean : 7.482 Mean : 177.9 Mean : 27.27 Mean : 212.1
## 3rd Qu.: 7.000 3rd Qu.: 187.5 3rd Qu.: 32.25 3rd Qu.: 228.5
## Max. :372.000 Max. :5172.0 Max. :790.00 Max. :6334.0
## NA's :1 NA's :4

```

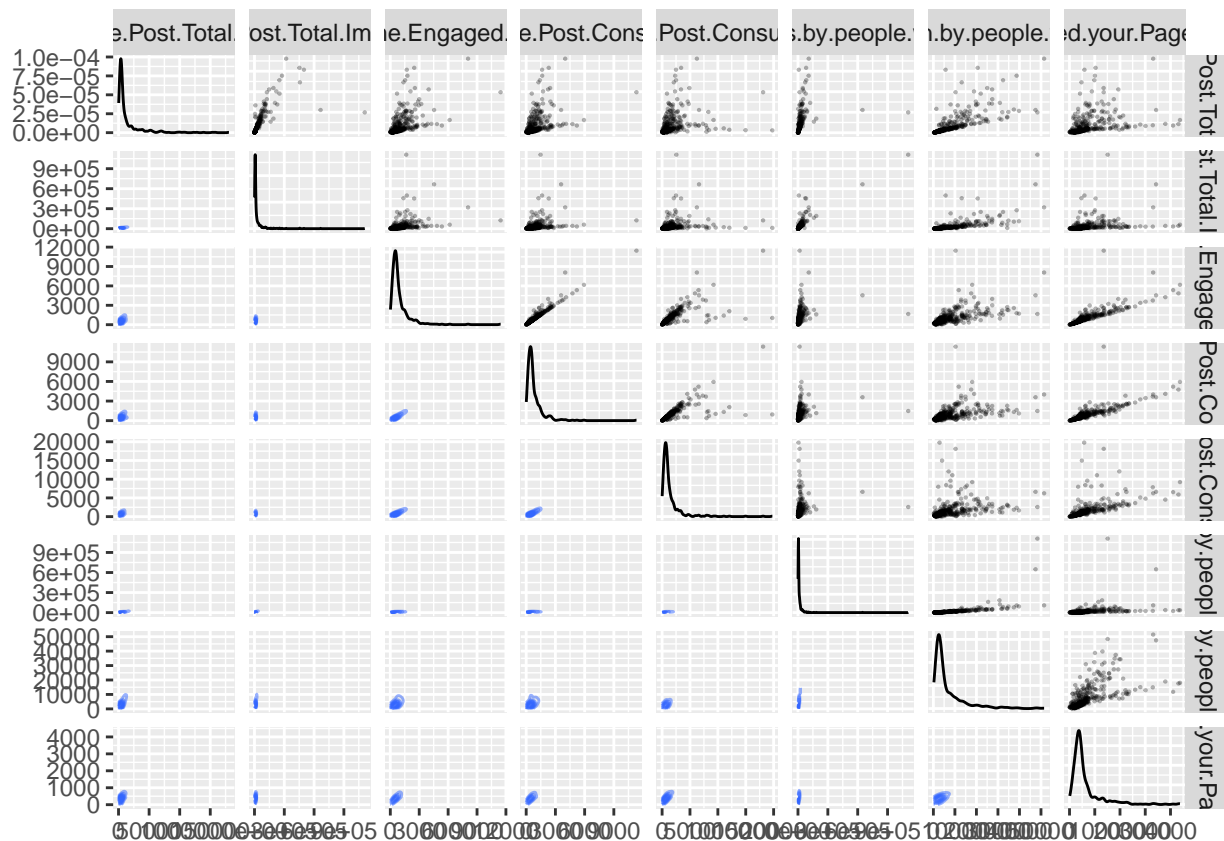
- NA values in variable “Paid”, “Like”, “share”.
- In total 500 observations and 19 variables including categorical variables.
- Original dataset is clean enough for later analysis.

Data Visualization

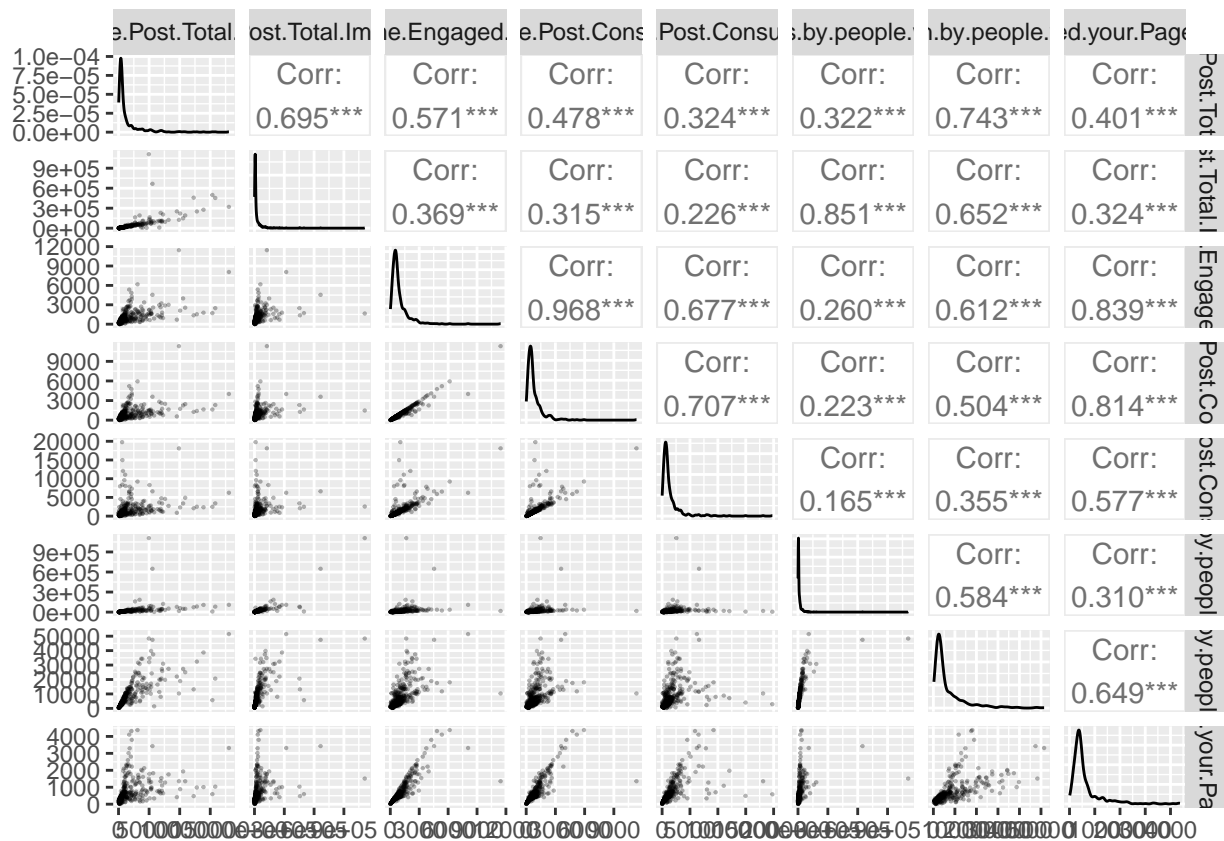
Scatterplot Matrix using ggplot2

```
temp_data <- fc_data %>% select(starts_with("Lifetime"))
```

```
ggpairs(temp_data,
  lower=list(continuous=wrap("density", alpha=0.5), combo="box"),
  upper=list(continuous=wrap("points", alpha=0.3, size=0.1)))
```



```
ggpairs(temp_data,
  lower=list(continuous=wrap("points", alpha=0.3, size=0.1)),
  upper=list(continuous=wrap("cor", size=4)))
```



- High correlation between Lifetime.Post.Consumptions and Lifetime.Engaged.Users
- Lifetime.Post.Consumptions: The number of clicks anywhere in a post.
- Lifetime.Engaged.Users: The number of people who clicked anywhere in a post (unique users).

Find Outliers

```
Q1_consumption<-509.2
Q3_consumption<-1463.0

Q1_Engaged<-393.8
Q3_Engaged<-1062.0

data_2col <- temp_data %>%
  filter(Lifetime.Post.Consumptions>= Q1_consumption & Lifetime.Post.Consumptions <= Q3_consumption,
         Lifetime.Engaged.Users>=Q1_Engaged & Lifetime.Engaged.Users<=Q3_Engaged ) %>%
  select(Lifetime.Engaged.Users,Lifetime.Post.Consumptions)
```

Scatterplot

```
plot(data_2col$Lifetime.Post.Consumptions,data_2col$Lifetime.Engaged.Users,
     pch=19, cex=0.2)
abline(lsfilt(x=data_2col$Lifetime.Post.Consumptions,y=data_2col$Lifetime.Engaged.Users), col="blue")
```

