EDA

2024-06-18

```
dataframe = read.csv("facebook_final_data.csv")
```

Plot the Reactions Ratio by Day of Week

You can also embed plots, for example:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

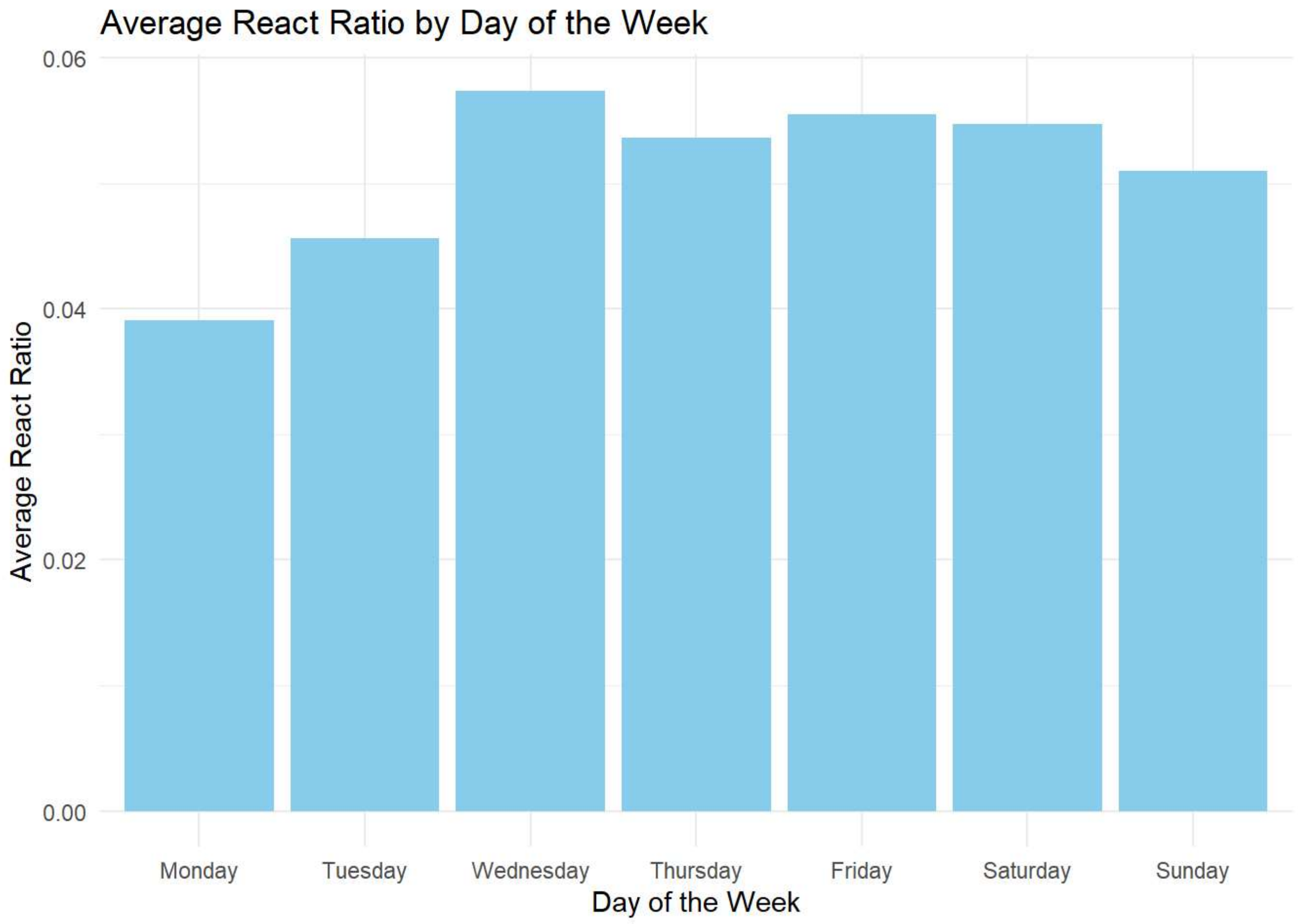
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dataframe <- dataframe %>%
  mutate(react_ratio = reactions/ page_follow)
dataframe <- dataframe %>%
  mutate(publish_day_of_week = factor(publish_day_of_week,
                                      levels = c("Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday", "Sunday")))
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
# Calculate average react_ratio for each publish_day_of_week
average_react_ratio <- dataframe %>%
  group_by(publish_day_of_week) %>%
  summarize(avg_react_ratio = mean(react_ratio, na.rm = TRUE))

# Plot the bar chart
ggplot(average_react_ratio, aes(x = publish_day_of_week, y = avg_react_ratio)) +
  geom_bar(stat = "identity", fill = "skyblue" ) +
  labs(title = "Average React Ratio by Day of the Week",
       x = "Day of the Week",
       y = "Average React Ratio") +
  theme_minimal()
```



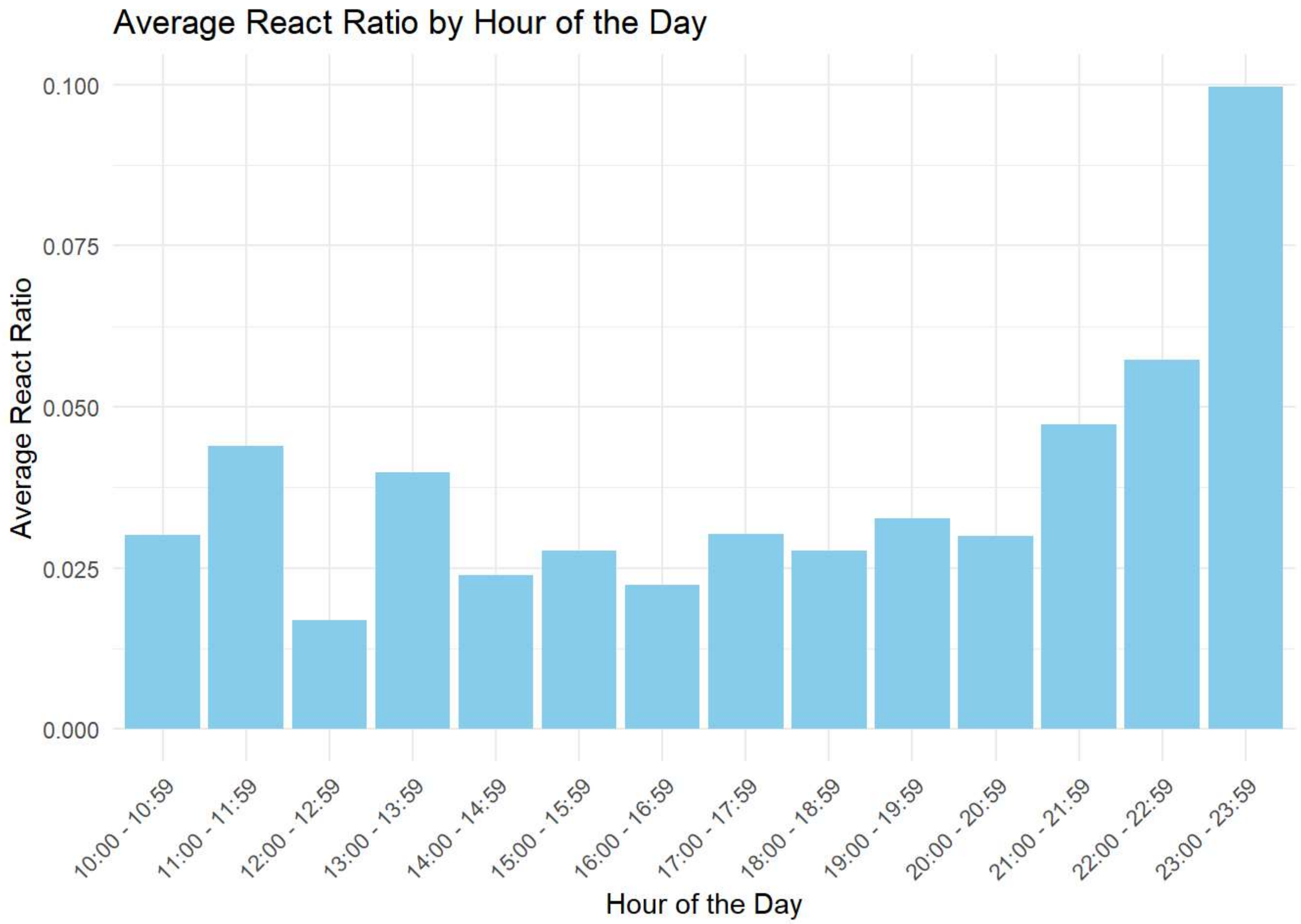
Plot the Reactions Ratio by hours in day

```
  dataframe <- dataframe %>%
  mutate(publish_hour = as.numeric(substr(publish_time, 1, 2)))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `publish_hour = as.numeric(substr(publish_time, 1, 2))`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
dataframe_clean <- dataframe %>%
  filter(!is.na(publish_hour))

# Calculate average react_ratio for each publish_hour
average_react_ratio_hourly <- dataframe_clean %>%
  group_by(publish_hour) %>%
  summarize(avg_react_ratio = mean(react_ratio, na.rm = TRUE)) %>%
  ungroup()

# Create a complete sequence of hours from 0 to 23
all_hours <- data.frame(publish_hour = 0:23)

# Join with the calculated averages to ensure all hours are represented
average_react_ratio_hourly <- all_hours %>%
  left_join(average_react_ratio_hourly, by = "publish_hour") %>%
  mutate(avg_react_ratio = ifelse(is.na(avg_react_ratio), 0, avg_react_ratio))
```

```
average_react_ratio_10_to_23 <- average_react_ratio_hourly %>%
  filter(publish_hour >= 10 & publish_hour <= 23)

ggplot(average_react_ratio_10_to_23, aes(x = factor(publish_hour), y = avg_react_ratio)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Average React Ratio by Hour of the Day",
       x = "Hour of the Day",
       y = "Average React Ratio") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_discrete(breaks = 0:23, labels = sprintf("%02d:00 - %02d:59", 0:23, 0:23))
```



Confirm the non-linear relationship between followers and reactions

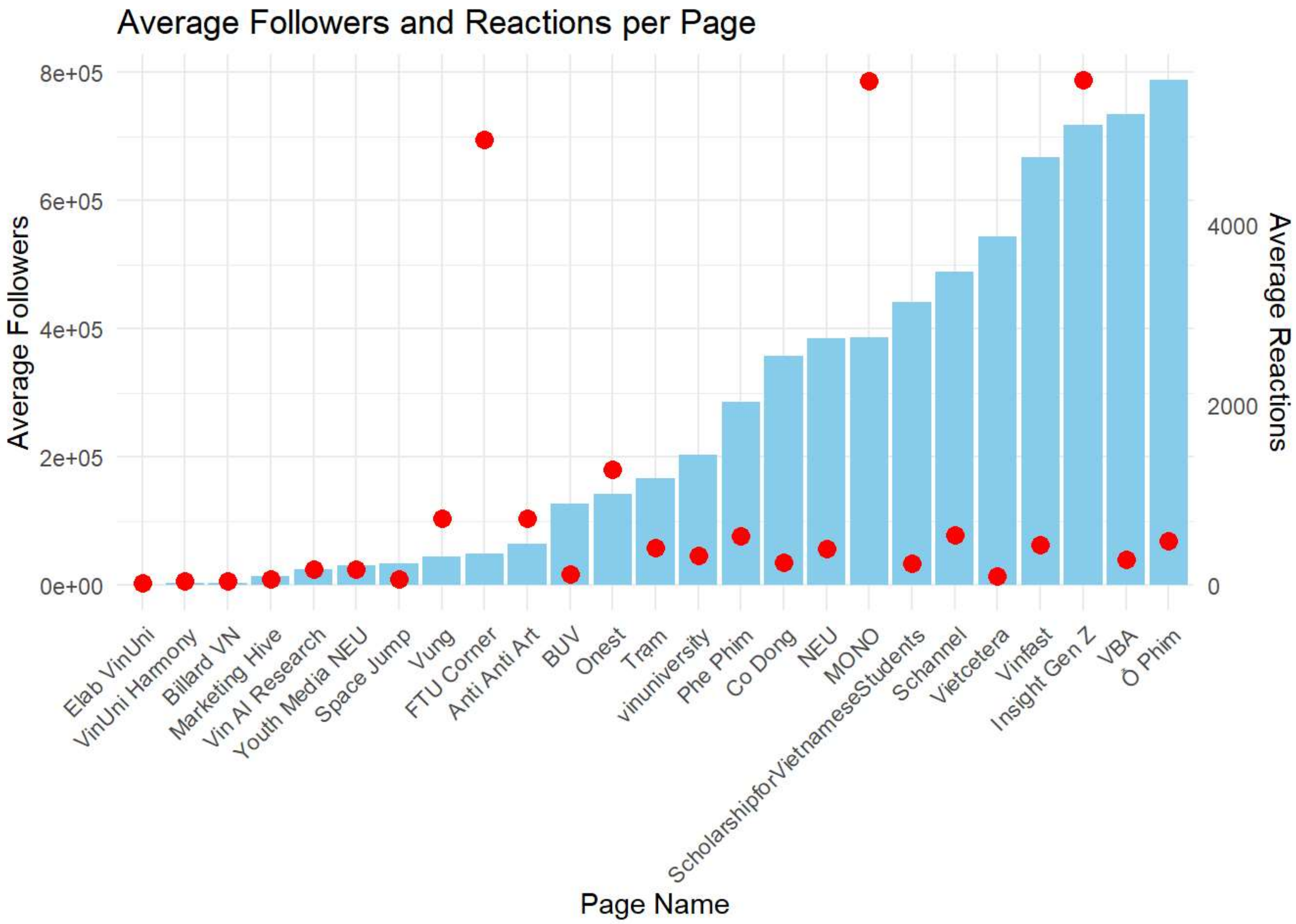
```
# Calculate average reactions per page
avg_dataframe <- dataframe %>%
  group_by(page_name) %>%
  summarize(avg_reactions = mean(reactions))
dataframe <- dataframe %>%
  arrange(page_follow)
```

```
page_df <- dataframe[, c("page_name", "page_follow")]
page_df <- page_df %>%
  group_by(page_name) %>%
  summarise(avg_followers = mean(page_follow))
page_df <- page_df %>%
  arrange(avg_followers)
page_df_subset <- page_df[1:25, ]
```

```r
library(ggplot2)
joined_df <- inner_join(page_df_subset, avg_dataframe, by = "page_name")
a <- ggplot(joined_df, aes(x = reorder(page_name, avg_followers), y = avg_followers)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Average Followers per Page (Subset)",
       x = "Page Name",
       y = "Average Followers") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
# Calculate necessary limits for the secondary y-axis
reactions_max <- max(joined_df$avg_reactions, na.rm = TRUE)
reactions_min <- min(joined_df$avg_reactions, na.rm = TRUE)

# Create the plot
dual_plot <- ggplot(joined_df, aes(x = reorder(page_name, avg_followers))) +
  geom_bar(aes(y = avg_followers), stat = "identity", fill = "skyblue") +
  geom_point(aes(y = avg_reactions * (max(joined_df$avg_followers) / reactions_max)),
             color = "red", size = 3) +  # Adjust size for scatter points
  scale_y_continuous(name = "Average Followers",
                     sec.axis = sec_axis(~ . * (reactions_max / max(joined_df$avg_followers)),
                                         name = "Average Reactions")) +
  labs(title = "Average Followers and Reactions per Page",
       x = "Page Name",
       y = "Average Followers") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Print the dual-axis plot
print(dual_plot)
```

```
## Warning: Use of `joined_df$avg_followers` is discouraged.
## ℹ Use `avg_followers` instead.
```


Average Followers and Reactions per Page
Average Followers
8e+05
6e+05
4e+05
2e+05
0e+00
Average Reactions
4000
2000
0
Elab VinUni
VinUni Harmony
Billard VN
Marketing Hive
Vin AI Research
Youth Media NEU
Space Jump
Vung
FTU Corner
Anti Anti Art
BUV
Onest
Trạm
vinuniversity
Phe Phim
Co Dong
NEU
MONO
ScholarshipforVietnameseStudents
Schannel
Vietcetera
Vinfast
Insight Gen Z
VBA
Ổ Phim
Page Name