EG Data Analyst Assessment

Raymond Lee

2023-05-24

Preprocessing

```r
# load packages
library(readxl)
library(tidyverse)
library(janitor)
library(lubridate)
```

```r
# read in data
social_data = read_excel('social_data.xlsx')
social_data = clean_names(social_data)
```

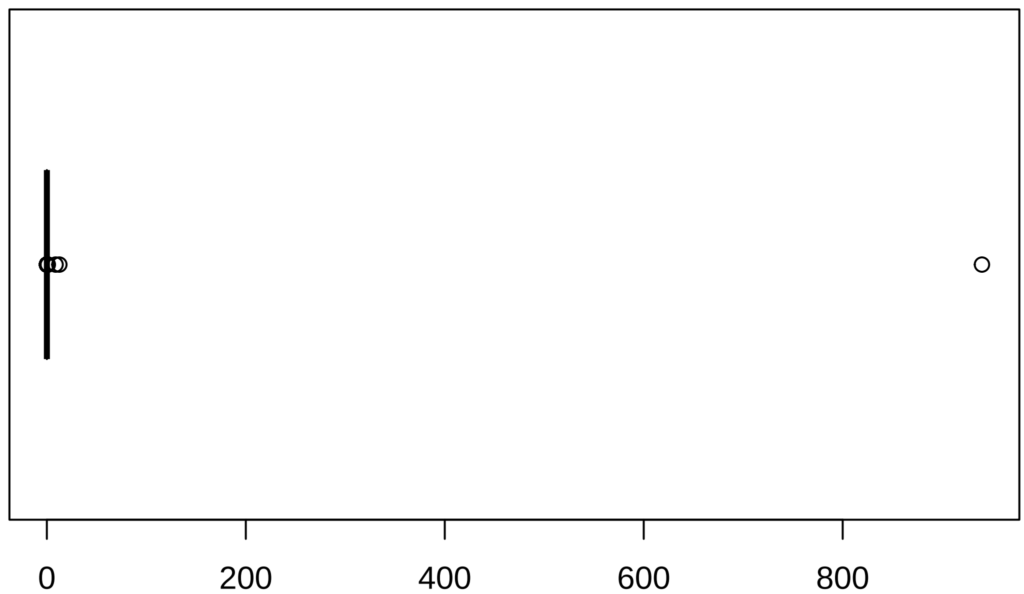
Impressions are the number of times the post is seen, and engagements are the number of times people interacted with the post. Therefore, impressions must exist for engagements to occur. Also, not every person interacts with a post that they see, so we can view engagements as a ratio to impressions. I decided to drop all observations with 0 impressions since that either means the post was never made or nobody saw it. In either case, the engagement rate would be undefined.

```r
social_data = social_data[social_data$total_impressions != 0, ]
```

I then calculated the engagement rate for each post and looked at its distribution. I rounded engagement rates to the nearest percent.

```r
social_data$engagement_rate = social_data$total_engagements/social_data$total_impressions
social_data$engagement_rate = round(social_data$engagement_rate, digit=2)

boxplot(social_data$engagement_rate, horizontal=TRUE)
```




```
outliers = sort(boxplot(social_data$engagement_rate, plot=FALSE)$out)
unique(outliers)
```

```
##  [1]   0.14   0.15   0.16   0.17   0.18   0.19   0.20   0.21   0.22   0.23
## [11]   0.24   0.25   0.26   0.27   0.28   0.29   0.30   0.31   0.32   0.34
## [21]   0.35   0.36   0.37   0.38   0.40   0.41   0.43   0.45   0.52   0.56
## [31]   0.58   0.59   0.60   1.00   8.80  12.59 940.00
```

The distribution of engagement rates is heavily skewed right. It seems that an engagement rate for a post is considered an outlier if it is at least 14%. I decided to leave these values in the dataset since it would give us information on the kind of posts that generate particularly high engagement rates (although with a relatively small sample size) and because the first question of the assessment asks about the probability of a 15% engagement rate as well.

However, I dropped the observation with 1 impression and 1 engagement since that does not tell us much. I also dropped the observation with 1 impression and 940 engagements since it heavily skews our data, is not a typical phenomenon, and is most likely an entry error. I find it hard to believe a single impression led to almost a thousand engagements.

```
social_data = social_data[!social_data$engagement_rate==940, ]
social_data = social_data[!social_data$engagement_rate==1, ]
```

Problem 1

```r
median(social_data$engagement_rate)
```

```
## [1] 0.03
```

```r
length(which(social_data$engagement_rate == .15)) / length(social_data$engagement_rate)
```

```
## [1] 0.01061105
```

The typical engagement rate is 3%. I calculated the median since it is more resistant to outliers and skewed data than the mean. The likelihood of achieving a 15% engagement rate is about 1%.

Problem 2

I created columns that identify what day of the week the published date was and whether the published time was in the morning or afternoon (morning being 12pm). Then, I calculated the average engagement rates by day of the week and by time. I decided to use the mean instead of the median here because I want to see whether any particular day or time has some particularly high engagement rates. This would give me more information on how the day or time affects engagement rate.

```r
social_data$day = tolower(wday(social_data$published_date, label=TRUE))
social_data$time = ifelse(hour(social_data$published_date)<12, 'morning', 'afternoon')
```

```r
social_data %>%
  group_by(day) %>%
  summarize_at(vars(engagement_rate), list(avg_engagement_rate=mean)) %>%
  arrange(desc(avg_engagement_rate))
```

```
## # A tibble: 7 x 2
##   day   avg_engagement_rate
##   <chr>               <dbl>
## 1 thu                0.0967
## 2 tue                0.0655
## 3 sun                0.0588
## 4 fri                0.0509
## 5 mon                0.0453
## 6 wed                0.0446
## 7 sat                0.0378
```

```r
social_data %>%
  group_by(time) %>%
  summarize_at(vars(engagement_rate), list(avg_engagement_rate=mean)) %>%
  arrange(desc(avg_engagement_rate))
```

```
## # A tibble: 2 x 2
##   time      avg_engagement_rate
##   <chr>                   <dbl>
## 1 morning                0.0776
## 2 afternoon              0.0495
```

Posts that are made on Thursdays have relatively higher engagement rates whereas posts that are made on Saturdays have relatively lower engagement rates. Also, posts that are made in the morning have higher engagement rates. The existence of some particularly high engagement rates for Thursdays and the morning would cause this.

Problem 3

I filtered the dataset for just the accounts for games. I then calculated number of posts, average engagement rate, and average impressions by account. Again, I used the mean to get more information on how the games are related to engagement rates.

```
social_data_games = social_data[social_data$account %in% c('CSGO', 'DOTA2', 'Valorant'), ]
social_data_games %>%
  group_by(account) %>%
  summarize_at(vars(engagement_rate, total_impressions), list(avg=mean, count=length)) %>%
  arrange(desc(engagement_rate_avg))
```

```
## # A tibble: 3 x 5
##   account  engagement_rate_avg total_impressions_avg engagement_rate_c~1 total~2
##   <chr>                  <dbl>                 <dbl>               <int>   <int>
## 1 Valorant              0.0532                  676.                  34      34
## 2 DOTA2                 0.0494                 2664.                 698     698
## 3 CSGO                  0.0413                11686.                 198     198
## # ... with abbreviated variable names 1: engagement_rate_count,
## #   2: total_impressions_count
```

I also looked for any differences in media type and account type across the games.

```
table(social_data_games$account, social_data_games$media_type, social_data_games$account_type)
```

```
## , ,  = TWITTER
##
##
##            Link Mixed Photo Text Video
##   CSGO        2     0   131   27    38
##   DOTA2       9     2   305  283    94
##   Valorant    0     0     0    0     0
##
## , ,  = YOUTUBE
##
##
##            Link Mixed Photo Text Video
##   CSGO        0     0     0    0     0
##   DOTA2       0     0     0    0     5
##   Valorant    0     0     0    0    34
```

```
social_data_games %>%
  group_by(account_type) %>%
  summarize_at(vars(engagement_rate, total_impressions), list(avg=mean)) %>%
  arrange(desc(engagement_rate_avg))
```

```
## # A tibble: 2 x 3
##   account_type engagement_rate_avg total_impressions_avg
##   <chr>                      <dbl>                 <dbl>
## 1 YOUTUBE                   0.0528                 2346
## 2 TWITTER                   0.0476                 4607.
```

DOTA2 has the most posts made, then CSGO, then Valorant. Average engagement rate is highest for Valorant, then DOTA2, and then CSGO. However, average impression is highest for CSGO, then DOTA2, then Valorant (the reverse order).

CSGO posts garner much more views than DOTA2 posts despite having less posts made. This may be due to CSGO having a relatively larger fanbase/following. The average engagement rate is lowest for CSGO, but this is most likely due to its much higher average impressions rather than its content being uninteresting or not engaging. DOTA2 does have a higher average engagement rate than CSGO, but this may be due to its smaller average impressions (although it could be that DOTA2's posts prompt more engagement too). I would suggest trying to recruit a larger following for the DOTA2 team.

DOTA2 also has the most Twitter content. I would not necessarily suggest to focus less on DOTA2, but I would suggest making more content on Twitter for CSGO since their posts get viewed a lot. I also would suggest making Twitter posts for Valorant. I was surprised to see that no posts were made on Twitter for Valorant since it is currently one of the most popular competitive games. CSGO's social performance (and its popularity as a similar FPS game) and the fact that Valorant's engagement rate is the highest (although with the smallest average impressions and least amount of posts made) gives some support to the idea that Twitter posts for Valorant would help its social performance.

More content can be made on YouTube for DOTA2 and CSGO as well (CSGO has no YouTube content). Although it could be due to average YouTube impressions being relatively smaller, the average engagement rate for YouTube is greater than that of Twitter. However, YouTube as a platform simply may not be something that EG focuses on since there is less content on YouTube overall compared to Twitter. If there is a reason for making YouTube content mostly for Valorant, I would then suggest to make much more YouTube content for it.

Problem 4

I looked at the average engagement rates by media type. Again, I used the mean to get more information on how engagement rates may spike with certain media types.

```r
social_data %>%
  group_by(media_type) %>%
  summarize_at(vars(engagement_rate), list(avg_engagement_rate=mean, count=length)) %>%
  arrange(desc(avg_engagement_rate))
```

```
## # A tibble: 7 x 3
##   media_type avg_engagement_rate count
##   <chr>                    <dbl> <int>
## 1 Album                   0.4        1
## 2 Mixed                   0.108      5
## 3 Photo                   0.0756  1243
## 4 Video                   0.0535   886
## 5 Text                    0.0409   548
## 6 Carousel                0.0378     9
## 7 Link                    0.0371    41
```

Album type has the highest average engagement rate, but this is based on only 1 post. When we consider the fact that photo type had the third highest engagement rate despite having the most posts, we can say that photo type performs the best, especially since mixed type has the second highest engagement rate but only has 5 posts.

Problem 5

I looked at average engagement rate and impressions by campaign to see how the campaign influences social performance.

```
social_data %>%
  group_by(campaign_name) %>%
  summarize_at(vars(engagement_rate, total_impressions), list(avg=mean)) %>%
  arrange(desc(engagement_rate_avg))
```

```
## # A tibble: 4 x 3
##   campaign_name        engagement_rate_avg total_impressions_avg
##   <chr>                              <dbl>                 <dbl>
## 1 N/A                               0.0876                 9123.
## 2 Community Engagement              0.0577                17262.
## 3 Evergreen                         0.0353                 8253.
## 4 Evil Exhibited                    0.0286                14612.
```

The Community Engagement campaign has the highest engagement rate despite having the highest average impressions, which is impressive. It garners the most views on average with its posts, and the posts generate the most engagement on average. However, the posts with missing campaign name had the highest average engagement rate overall.

Problem 6

I looked at average engagement rates by game, account type, media type, day, and time.

```
social_data_games %>%
  group_by(account, account_type, day, time) %>%
  summarize_at(vars(engagement_rate), list(avg_engagement_rate=mean, count=length)) %>%
  arrange(desc(avg_engagement_rate))
```

```
## # A tibble: 42 x 6
## # Groups:   account, account_type, day [24]
##    account  account_type day   time      avg_engagement_rate count
##    <chr>    <chr>        <chr> <chr>                   <dbl> <int>
##  1 Valorant YOUTUBE      fri   afternoon              0.0833     3
##  2 DOTA2    TWITTER      mon   afternoon              0.0788    16
##  3 Valorant YOUTUBE      fri   morning                0.0775     4
##  4 DOTA2    YOUTUBE      fri   morning                0.07       1
##  5 DOTA2    TWITTER      tue   afternoon              0.0698    63
##  6 CSGO     TWITTER      tue   morning                0.0667    15
##  7 CSGO     TWITTER      thu   morning                0.0642    19
##  8 Valorant YOUTUBE      mon   afternoon              0.0633     3
```

```
## 9 Valorant YOUTUBE     wed   afternoon              0.0625     4
## 10 CSGO     TWITTER     sat   morning                0.0611    18
## # ... with 32 more rows
```

First, I would focus on making posts on Twitter that involve photos since photos performed the best despite having the most posts. Perhaps it would be a good idea to make posts that utilize GIFs as well since videos had the most impressions and posts made after photos. GIFs would be a good middle ground between a photo and video. If achieving high engagement rate is the goal, I would try to avoid making posts that only consist of text.

For CSGO Twitter, I would make posts on Tuesday morning, Thursday morning, or Saturday morning since they have the highest average engagement rates. These days and times have similarly high engagement rates and similar numbers of posts made as well. For DOTA2 Twitter, I would make posts on Monday afternoon and Tuesday afternoon. DOTA2 Twitter posts that are made on Monday afternoon have especially higher engagement rates, although based on a smaller post count compared to Tuesday afternoon. Valorant YouTube does not have that many videos posted, but videos posted on Friday afternoon had the highest engagement rate.

Also, like I mentioned above, if it true that the following/fanbase for DOTA2 is smaller, I would first focus on making posts that garner views to increase it. That would mean making posts that catch people's attention and end up on people's feed more. Then, we can make posts that prompt engagement in the larger fanbase.

Problem 7

Although it is based on a limited sample size, the engagement rates for the Valorant YouTube channel suggest that it is doing well. I would definitely expand to short video platforms like TikTok, YouTube Shorts, and/or Instagram Reels. For example, highlights from Valorant YouTube videos and other attention-grabbing content can be posted on these platforms. Personally, much of my exposure to Valorant content comes from short videos that would be found on these platforms. Also, I believe there is less effort involved in watching/sharing short videos because there is less commitment needed to watch the whole video and stay focused. Watching short videos may also prompt people to watch the full length videos on YouTube.

Creating a Twitter account for Valorant would also help since Twitter has served well for the other games. Content on the YouTube channel can be cross-posted on it as well, most likely in the form of photos and GIFs rather than links. However, if we want to focus on only having a YouTube channel, we should definitely tap into the potential of short videos.

DOTA2 does not have much YouTube content, but based on this limited sample size, I would consider making more videos since they have relatively high engagement rate compared to other games and their accounts. CSGO does not have any YouTube content, but I would first try expanding into short videos before creating a YouTube channel for full length videos.

Also, I believe doing giveaways would definitely boost engagement rates if we make it so that people can enter by sharing or liking posts. Making posts that incorporate current trends (gameplay, memes, quotes/references, etc.) for each game would also help expand presence in general.