Code ▾

RealGM General Board Analysis

Introduction

In this project, I am analyzing RealGM.com's "Basketball General Board". RealGM is a forum that specializes in NBA discussion. The website has functionality to discuss other sports, but most of its traffic is on the basketball side. On the main general board page, you are able to view topics in order of which topic had the most recent activity (i.e. user posting to topic). The page displays the topic name, the amount of replies, which user had the last post, and when the last post was posted. For this project, the focus will be on topic names and views.

The dataset was web scraped from RealGM's general board's current 5500 pages.

Loading the data

I will be using the package readxl to open the xlsx file from my working directory.

As you will see below, the table has two columns: topic and views.

Hide

```
library(readxl)

rgm <- read_excel("RealGM_Topics_Views_09_24_24.xlsx")

head(rgm)
```

topic <chr>
Official RGM GOAT Debate Thread
Final RGM Ranking of Top 30 NBA Players from countdown voting
2024 NBA Offseason General Discussion Hangout Thread NBA News Thoughts I Don t Know Where To Put This etc
2024 NBA Offseason Free Agency Trade Discussion Thread
28 Days till the season begins best player to wear number 28
Lets Talk Handchecking

6 rows | 1-1 of 2 columns

Hide

```
NA
```

Manipulating the data

To prepare the data for the following analysis, I want to make all of the character strings lowercase and make sure each topic is not null. Some additional string manipulation is done to fix some contractions and left over ISO-8859-1 encoded data.

Hide

```
library(tidyverse)

rgm_filtered <- rgm %>%
  filter(
    is.na(topic) == 0
  )

rgm_clean <- rgm_filtered %>%
  mutate(
    topic = tolower(topic),
    topic = sub("n t ", "nt ", topic),
    topic = sub("o s", "os", topic),
    topic = sub("â ", "", topic),
    topic = sub(" s ", "s ", topic)
  )
```

Does putting LeBron in your topic title increase your view count?

LeBron James is a no-doubt top three player in the NBA all-time no matter who you ask. The other two that could be in this conversation retired over 20 years ago. However, LeBron James started playing 20 years ago and is still playing in the league as a high level.

RealGM was founded in 2000 so it has seen the entirety of LeBron’s career. LeBron dominates headlines, but does he dominate RealGM boards compared to the population?

Hide

```
rgm_lebron <- rgm_clean %>%
  filter(topic %ilike% "lebron")

#mean views for topics with LeBron in topic title
(mean_lebron <- mean(rgm_lebron$views))
```

```
[1] 2022.664
```

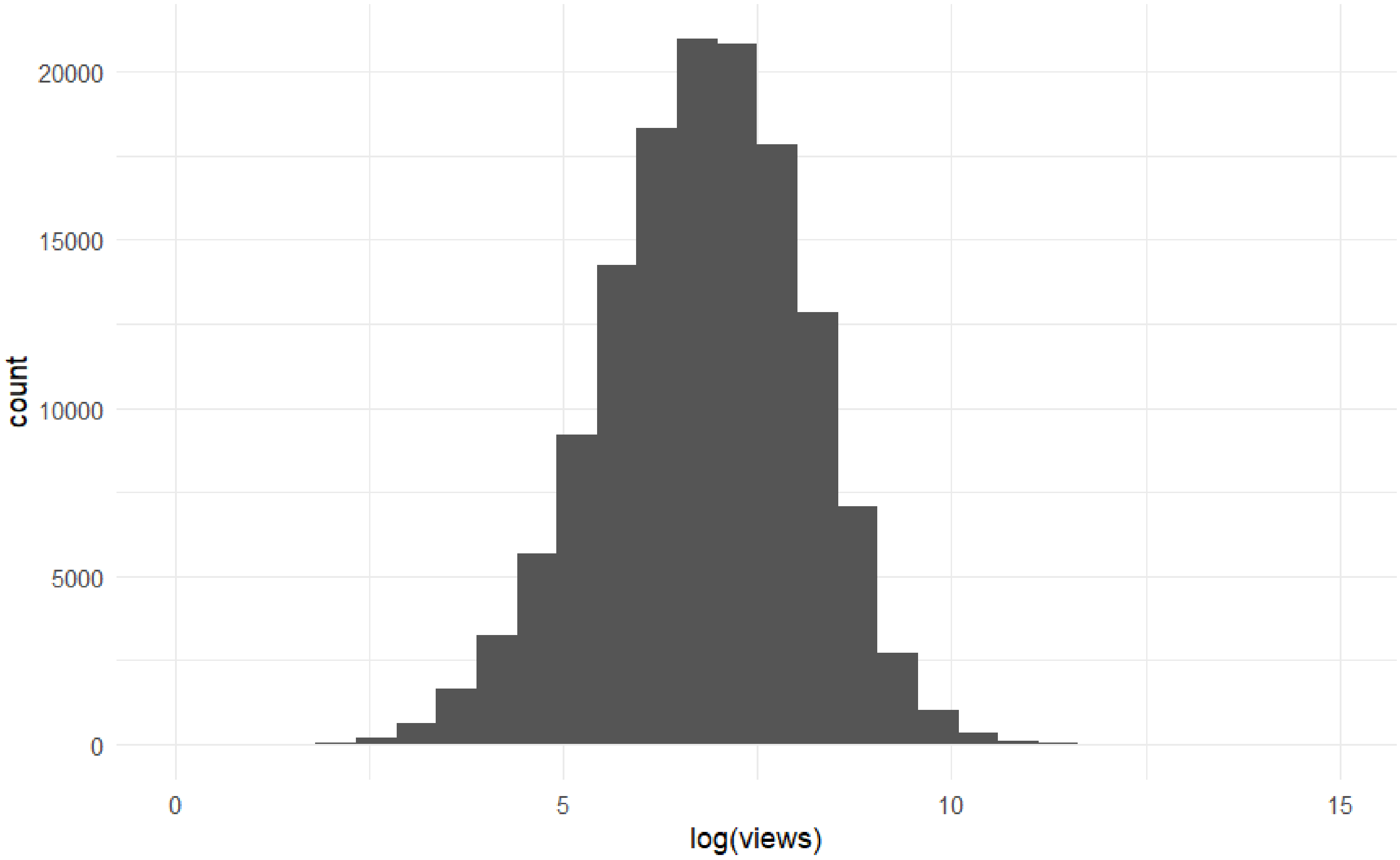
Hide

```
#mean views for all topics
(mean_rgm <- mean(rgm_clean$views))
```

```
[1] 1963.05
```

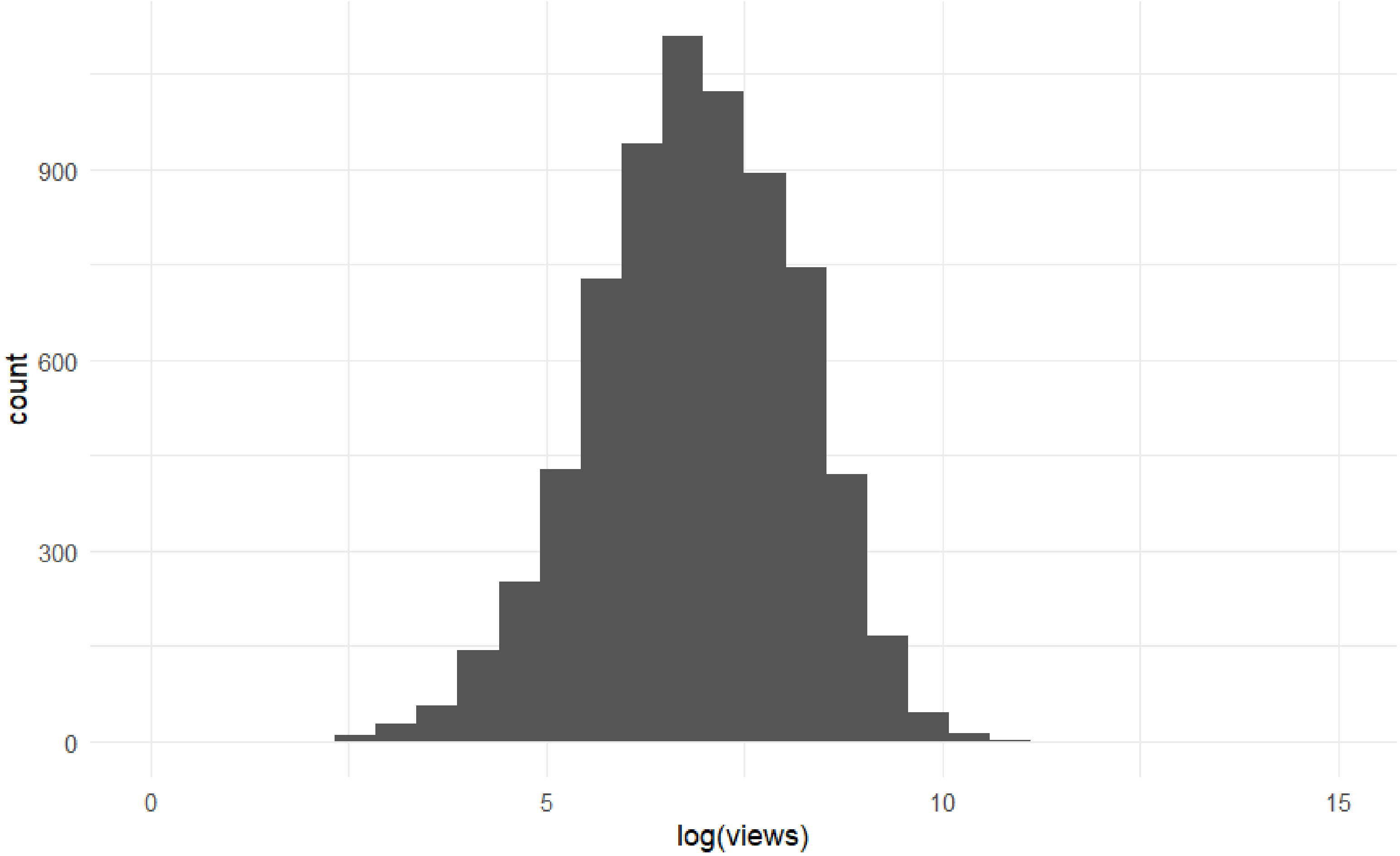
Hide

```
ggplot(rgm_clean, aes(log(views))) +
  geom_histogram() +
  theme_minimal() +
  xlim(x = c(0, 15))
```



Hide

```
ggplot(rgm_lebron, aes(log(views))) +
  geom_histogram() +
  theme_minimal() +
  xlim(x = c(0, 15))
```



LeBron’s average views is higher than the overall population’s views, but is it *significantly* higher.

To find out, we will conduct a t-test and find the p-value and compare it to our alpha value of 0.05.

Hide

```
alpha <- .05

sd_rgm <- sd(rgm_clean$views)
n_rgm <- nrow(rgm_clean)
sd_lebron <- sd(rgm_lebron$views)
n_lebron <- nrow(rgm_lebron)

standard_error <- sqrt((sd_lebron^2/n_lebron) + (sd_rgm^2/n_rgm))
t_rgm_lebron <- (mean_lebron - mean_rgm)/standard_error
df <- n_rgm + n_lebron - 2
p_rgm <- pt(t_rgm_lebron, df = df, lower.tail = FALSE)

ifelse(p_rgm > alpha, "Fail to reject null hypothesis", "Reject null hypothesis")
```

```
[1] "Fail to reject null hypothesis"
```

Despite LeBron’s average views being higher, it is significant enough to reject the null hypothesis that is higher.

I have a way to cheer him up.

Who is the GOAT (Greatest of All Time) of views on RealGM? LeBron or Jordan

Remember when I said LeBron is consensus top three all time? Well, Jordan is his main competitor for the number one spot.

Off the court, which player gets more views?

Hide

```
rgm_mj <- rgm_clean %>%
  filter(topic %ilike% "michael jordan" | topic %ilike% " mj ")

(mean_mj <- mean(rgm_mj$views))
```

```
[1] 1879.21
```

Hide

```
mean_lebron
```

```
[1] 2022.664
```

Looks like LeBron has him beat out in this race! However, is it significant difference to say they are different.

Our null hypothesis for this experiment will be that LeBron and MJ get the same amount of views, but our alternative hypothesis will be that they are significantly different.

Same as before, we will do a t-test and compare its p-value to an alpha of 0.05.

Hide

```
sd_mj <- sd(rgm_mj$views)
n_mj <- nrow(rgm_mj)

standard_error_lmj <- sqrt((sd_lebron^2/n_lebron) + (sd_mj^2/n_mj))
t_lmj <- (mean_lebron - mean_mj)/standard_error
df_lmj <- n_mj + n_lebron - 2
p_lmj <- 2*pt(t_lmj, df = df_lmj, lower.tail = FALSE)

p_lmj
```

```
[1] 0.0008880342
```

Hide

```
result <- ifelse(p_lmj > alpha, "Fail to reject null hypothesis", "Reject null hypothesis")

if(result == "Fail to reject null hypothesis") {
  break
} else if(mean_mj > mean_lebron) {
  print("Michael Jordan is the RealGM GOAT")
} else {
  print("LeBron James is the RealGM GOAT")
}
```

```
[1] "LeBron James is the RealGM GOAT"
```

Looks like R said it best. LeBron James is the GOAT of RealGM.

Which words in topic titles are most frequent?

If you’ve ever browsed social media, you may see the same topics brought up over and over. RealGM is no different. Lets see which topics get discussed to death.

To do this, I will be splitting the topics into separate words and counting how many times they show up.

Hide

```
count_words <- rgm_clean %>%
  separate_rows(topic, sep = ' ') %>%
  group_by(topic) %>%
  summarize(
    word_count = n()
  ) %>%
  mutate(word = topic) %>%
  select(word, word_count) %>%
  arrange(desc(word_count))

head(count_words)
```

word <chr>	word_count <int>
the	49664
to	19508
in	17622
a	16066
is	15353
nba	15041

6 rows

Well a lot of those words are un-fun for these discussions, but make sense. For more fun, I am going to take out what I deem as “generic words” until the top ten list of words are basketball related.

Hide

```
generic_words <- c("the", "to", "in", "a", "is", "of", "for", "and", "on", "with", "you", "be", "will", "al
l", "what", "this", "are", "how", "who", "or", "if", "would", "do", "have")

'%notin%' <- Negate('%in%')

count_words_ungeneric <- count_words %>%
  filter(word %notin% generic_words) %>%
  head(10)

count_words_ungeneric
```

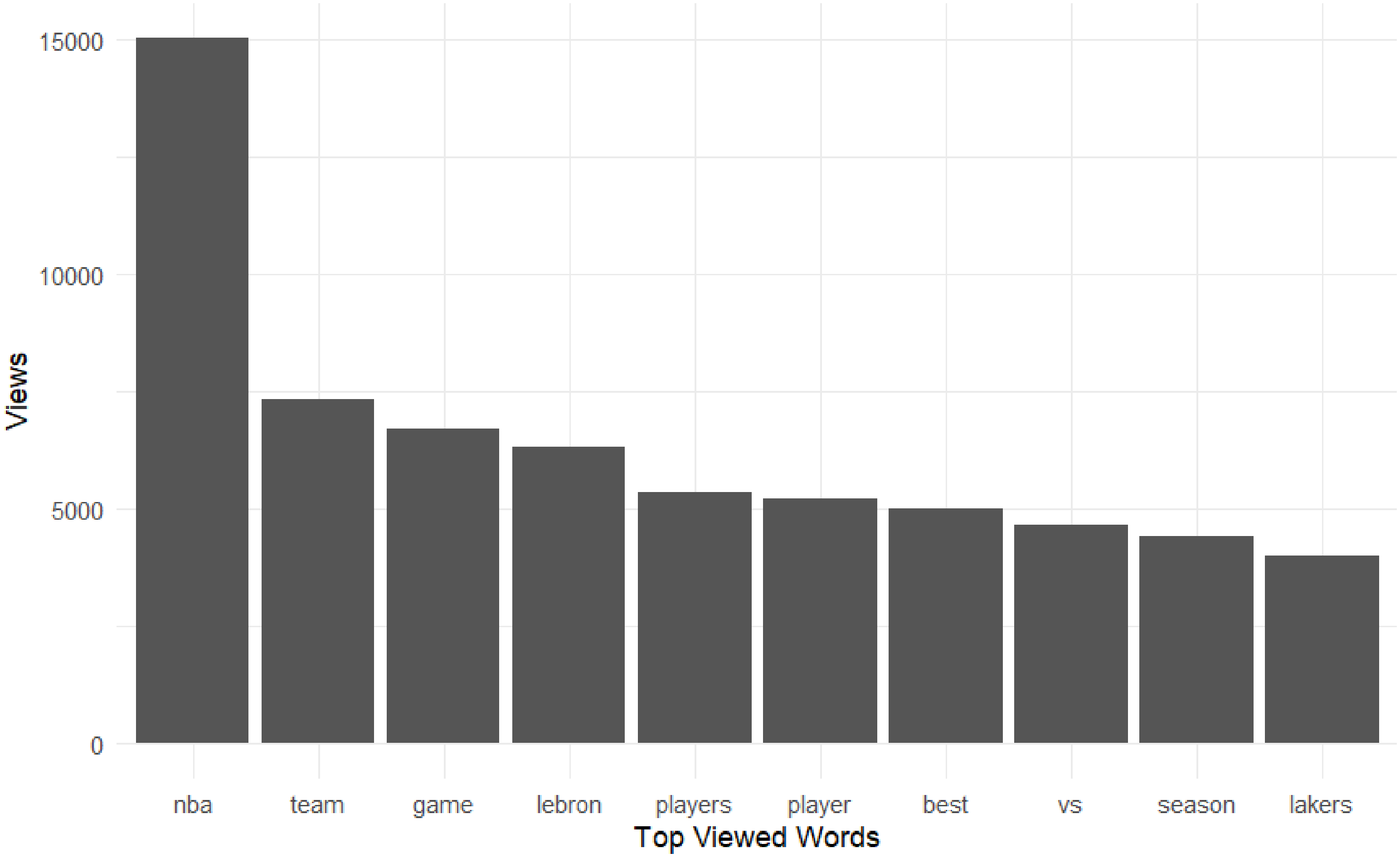
word <chr>	word_count <int>
nba	15041

word <chr>	word_count <int>
team	7333
game	6703
lebron	6333
players	5362
player	5197
best	4999
vs	4647
season	4405
lakers	3990

1-10 of 10 rows

Hide

```
ggplot(count_words_ungeneric, aes(reorder(word, -word_count), word_count)) +
  geom_col() +
  theme_minimal() +
  xlab("Top Viewed Words") +
  ylab("Views")
```



That is more like it. Looks like our friend LeBron made it on to the list as well.

Most of these words are still “generic” terms when it comes to basketball, but they show some of the topics that may be discussed on the General Board. RealGM loves to talk about the Lakers (with or without LeBron), they love to declare players or teams as the best, and they love to compare players or teams with vs.

Which teams have the most views per word?

In 2024, the Boston Celtics won the NBA Championship. However, which team all-time has the most engagement every time they are brought up.

One issue that we will solve with some manipulation is that the Trail Blazers are two words, so before we split up each string we are going to make them the Trailblazers.

Hide

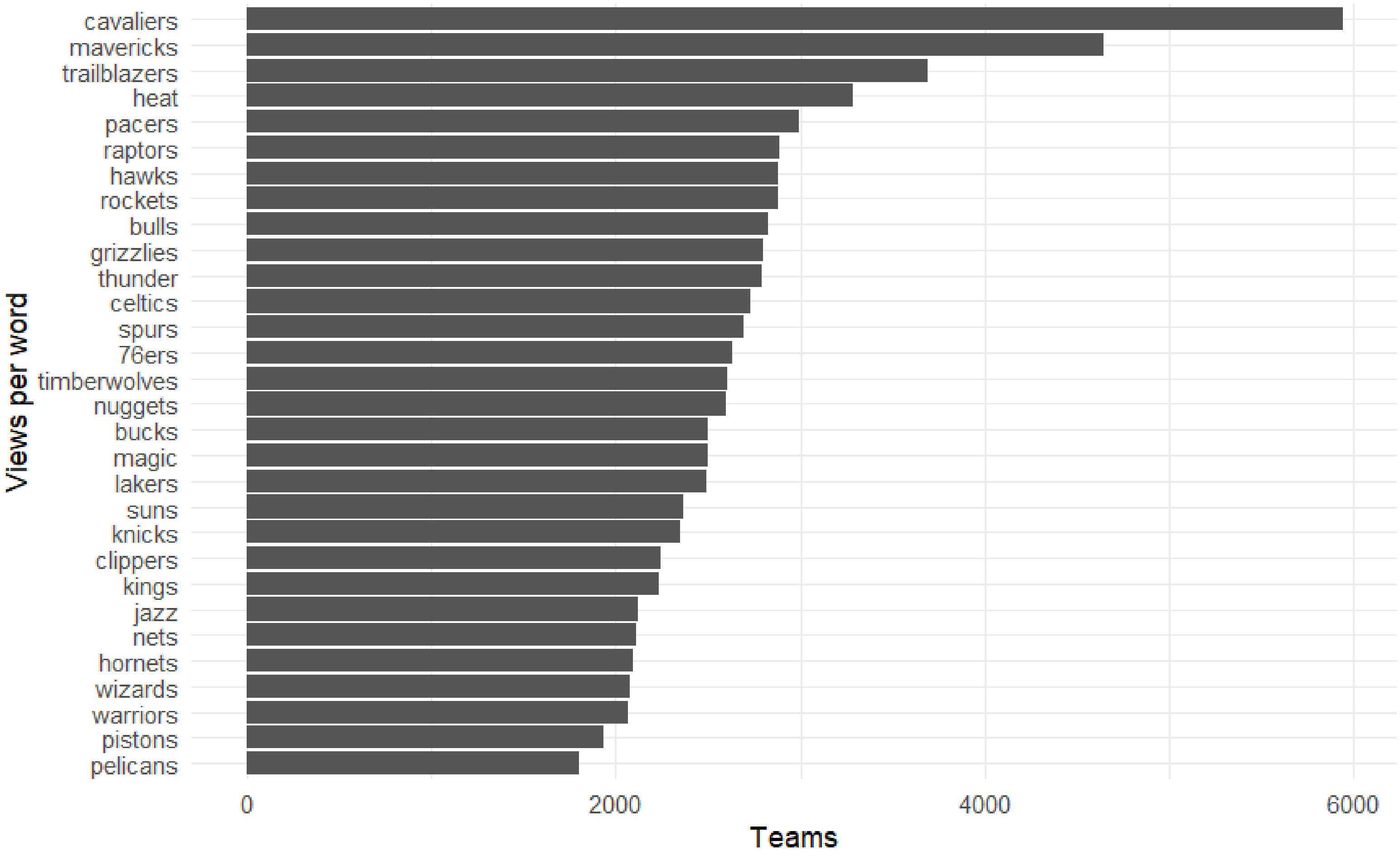
```
teams <- c("heat", "wizards", "raptors", "76ers", "magic", "knicks", "bucks", "hawks", "pacers", "cavalier
s", "celtics", "nets", "bulls", "hornets", "pistons", "grizzlies", "nuggets", "mavericks", "timberwolves",
"pelicans", "warriors", "thunder", "lakers", "clippers", "suns", "trailblazers", "kings", "spurs", "jazz",
"rockets")

rgm_tb <- rgm_clean %>%
  mutate(topic = sub("trail blazers", "trailblazers", topic))

view_words <- rgm_tb %>%
  separate_rows(topic, sep = ' ') %>%
  group_by(topic) %>%
  summarize(
    word_count = n(),
    views = sum(views),
    views_per_word = views/word_count
  ) %>%
  filter(topic %in% teams) %>%
  mutate(word = topic) %>%
  select(word, word_count, views_per_word) %>%
  arrange(desc(views_per_word))

view_words %>% print(30)
```

word	word_count	views_per_word
<chr>	<int>	<dbl>
cavaliers	293	5945.007
mavericks	327	4645.498
trailblazers	104	3695.442
heat	1843	3284.600
pacers	574	2996.434
raptors	1038	2891.071
hawks	637	2882.749
rockets	1225	2878.277
bulls	1113	2828.216
grizzlies	418	2803.203

1-10 of 30 rows Previous **1** 2 3 Next

Hide

```
ggplot(view_words, aes(reorder(word, views_per_word), views_per_word)) +
  geom_col() +
  theme_minimal() +
  coord_flip() +
  ylab("Teams") +
  xlab("Views per word")
```



As a Cavaliers fan, can I consider this an NBA championship? We may not get discussed very much, but when we do, we make good discussion.

Sadly, the Pelicans are cemented as the least engaging team by RealGM fans.