Netflix_EDA

Casey Franco

2024-02-16

File Info

I downloaded this file from the following address on Kaggle

It contains information about weekly Netflix user viewership.

Importing and Defining “Value Counts” Function

```r
df <- read.csv("/Users/caseyfranco/Desktop/Data Science Resources/Datasets for Visualization/Netflix Vi

value_counts <- function(x, sort = FALSE) {
  counts <- table(x)
  if (sort) {
    counts <- counts[order(-counts)]
  }
  return(counts)
}

head(df)
```

```
##         week        category weekly_rank                             show_title
## 1 2024-01-07 Films (English)           1                        The Equalizer 3
## 2 2024-01-07 Films (English)           2 Rebel Moon ? Part One: A Child of Fire
## 3 2024-01-07 Films (English)           3                 Leave the World Behind
## 4 2024-01-07 Films (English)           4                 Exodus: Gods and Kings
## 5 2024-01-07 Films (English)           5                                Aquaman
## 6 2024-01-07 Films (English)           6            The Super Mario Bros. Movie
##   season_title weekly_hours_viewed runtime weekly_views
## 1          N/A            26800000  1.8167     14800000
## 2          N/A            25100000  2.2667     11100000
## 3          N/A            18700000  2.3667      7900000
## 4          N/A            18600000  2.5000      7400000
## 5          N/A            16800000  2.3833      7000000
## 6          N/A             8700000  1.5333      5700000
##   cumulative_weeks_in_top_10 is_staggered_launch episode_launch_details
## 1                          1               false
## 2                          3               false
## 3                          5               false
## 4                          1              false
## 5                          1              false
## 6                          6              false
```

Summary Statistics

```r
summary(df)
```

```
##      week             category          weekly_rank    show_title
##  Length:5280        Length:5280        Min.   : 1.0   Length:5280
##  Class :character   Class :character   1st Qu.: 3.0   Class :character
##  Mode  :character   Mode  :character   Median : 5.5   Mode  :character
##                                        Mean   : 5.5
##                                        3rd Qu.: 8.0
##                                        Max.   :10.0
##
##  season_title       weekly_hours_viewed    runtime        weekly_views
##  Length:5280        Min.   :   700000   Min.   : 0.000   Min.   :  600000
##  Class :character   1st Qu.:  6450000   1st Qu.: 1.667   1st Qu.: 1875000
##  Mode  :character   Median : 11555000   Median : 2.117   Median : 3000000
##                     Mean   : 18764227   Mean   : 3.596   Mean   : 4512250
##                     3rd Qu.: 20827500   3rd Qu.: 4.929   3rd Qu.: 5125000
##                     Max.   :571760000   Max.   :20.300   Max.   :44900000
##                                         NA’s   :4080     NA’s   :4080
##  cumulative_weeks_in_top_10 is_staggered_launch episode_launch_details
##  Min.   : 1.000             Length:5280         Length:5280
##  1st Qu.: 1.000             Class :character    Class :character
##  Median : 2.000             Mode  :character    Mode  :character
##  Mean   : 3.118
##  3rd Qu.: 4.000
##  Max.   :30.000
##
```

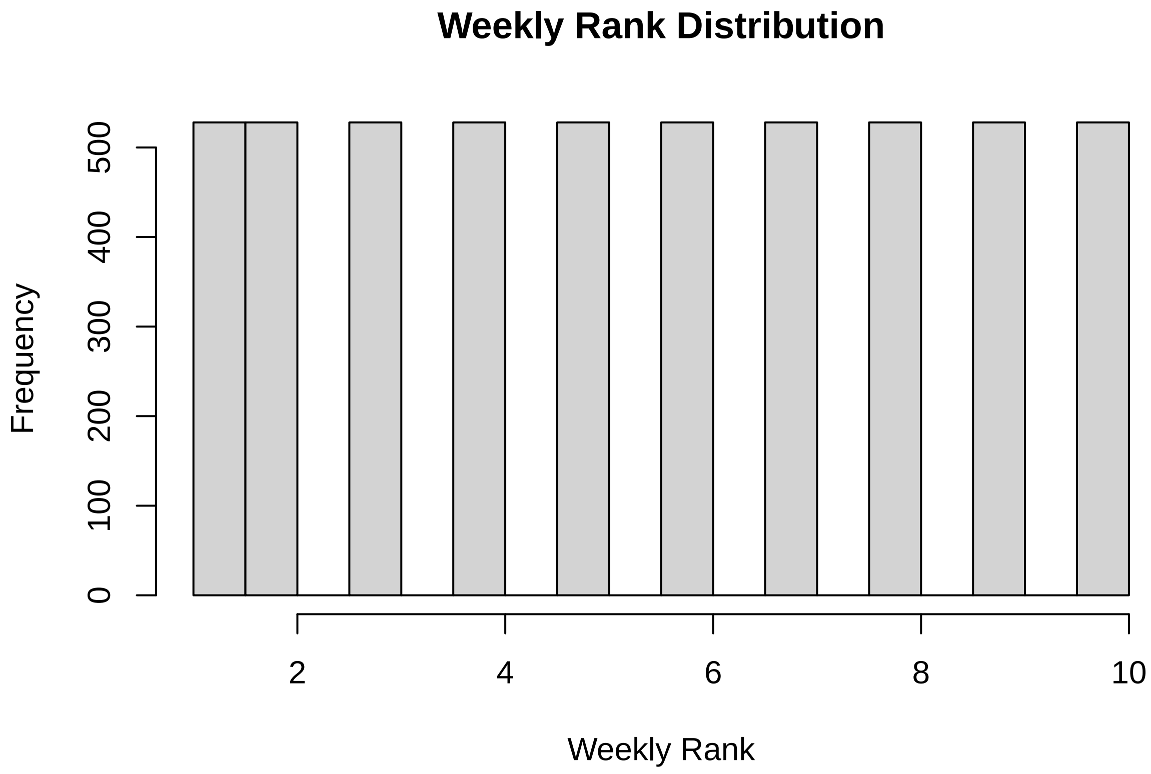
```r
str(df)
```

```
## ’data.frame’:    5280 obs. of  11 variables:
##  $ week                      : chr  "2024-01-07" "2024-01-07" "2024-01-07" "2024-01-07" ...
##  $ category                  : chr  "Films (English)" "Films (English)" "Films (English)" "Films (Eng
##  $ weekly_rank               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ show_title                : chr  "The Equalizer 3" "Rebel Moon ? Part One: A Child of Fire" "Leav
##  $ season_title              : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ weekly_hours_viewed       : int  26800000 25100000 18700000 18600000 16800000 8700000 9800000 860
##  $ runtime                   : num  1.82 2.27 2.37 2.5 2.38 ...
##  $ weekly_views              : int  14800000 11100000 7900000 7400000 7000000 5700000 5500000 520000
##  $ cumulative_weeks_in_top_10: int  1 3 5 1 1 6 7 1 1 4 ...
##  $ is_staggered_launch       : chr  "false" "false" "false" "false" ...
##  $ episode_launch_details    : chr  "" "" "" "" ...
```

Distribution Visualization

I'll plot some histograms to see how the variables are distributed.

```
hist(df$weekly_rank, main = "Weekly Rank Distribution", xlab = "Weekly Rank")
```

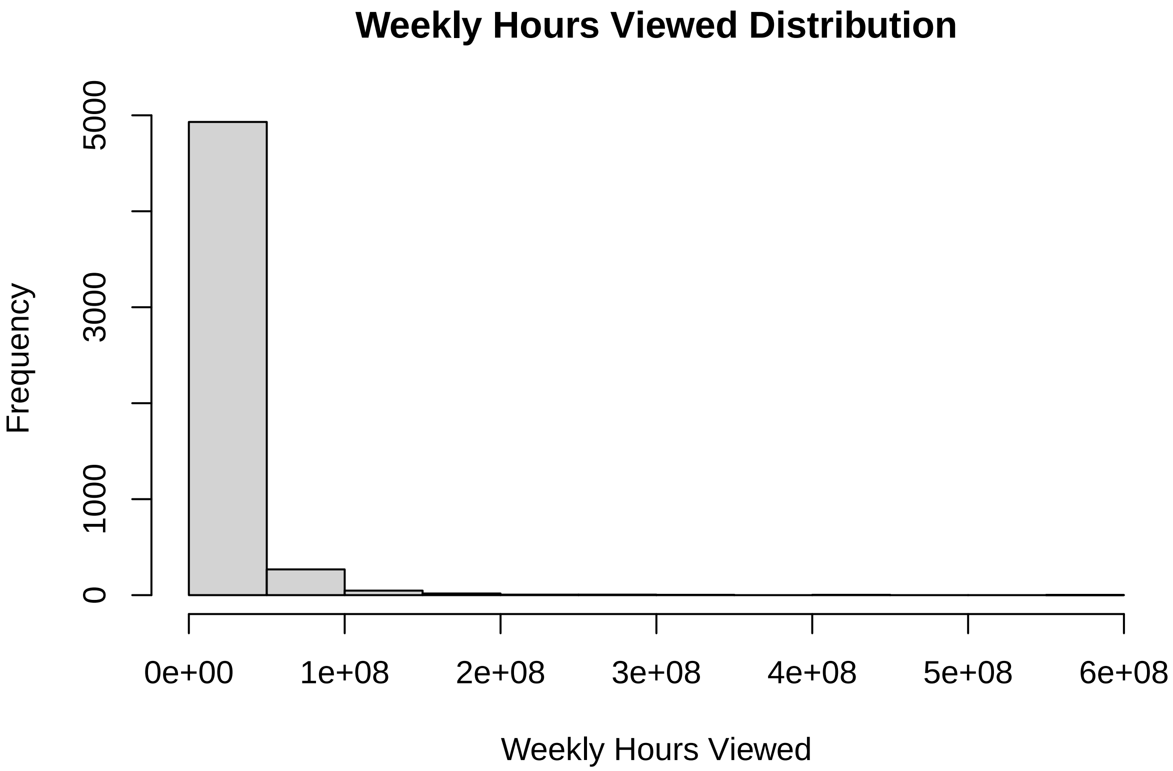


```
value_counts(df$weekly_rank, sort = TRUE)
```

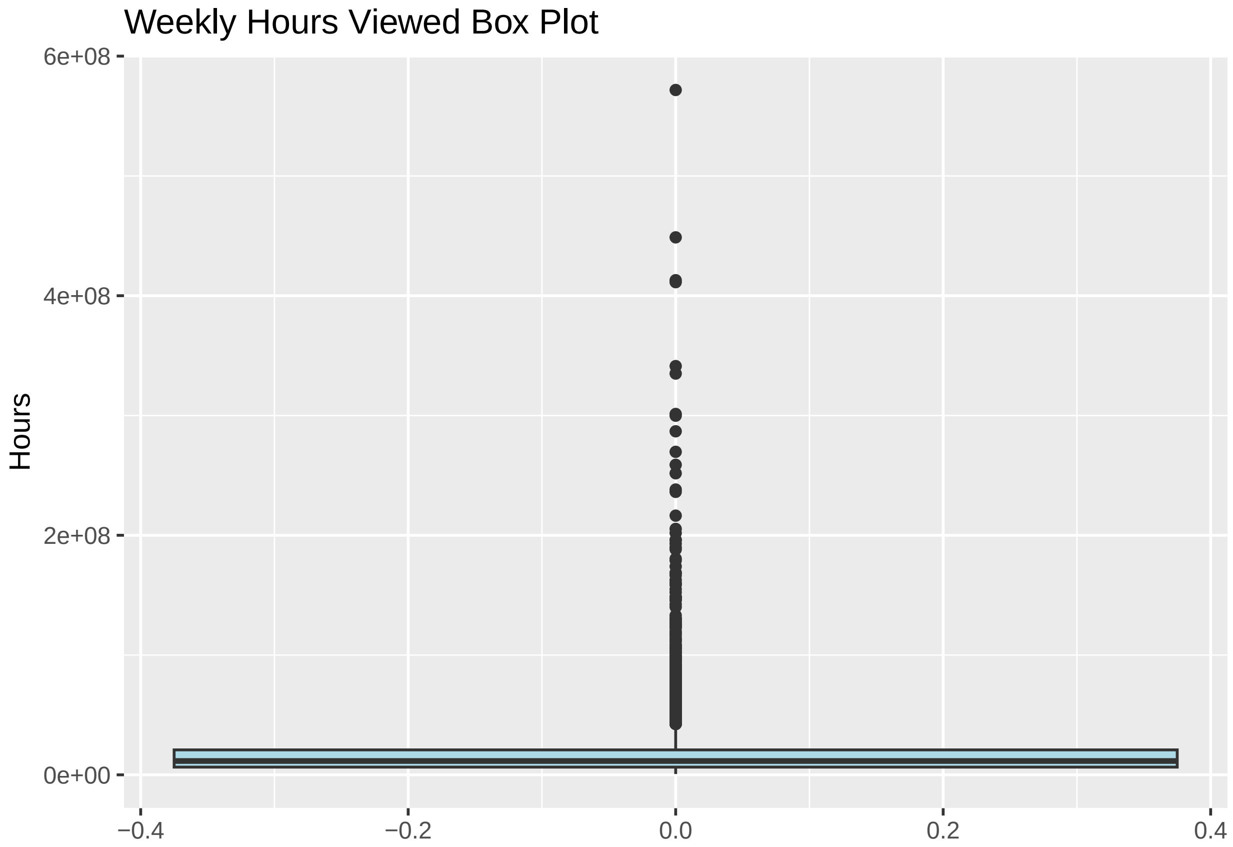
```
## x
##   1   2   3   4   5   6   7   8   9  10 
## 528 528 528 528 528 528 528 528 528 528
```

It would appear that there is an even distribution of weekly ranks for shows. Exactly 528 entries for each rank. Not much I can garner there except that this dataset is likely curated.

```
library(ggplot2)
hist(df$weekly_hours_viewed, main = "Weekly Hours Viewed Distribution", xlab = "Weekly Hours Viewed")
```



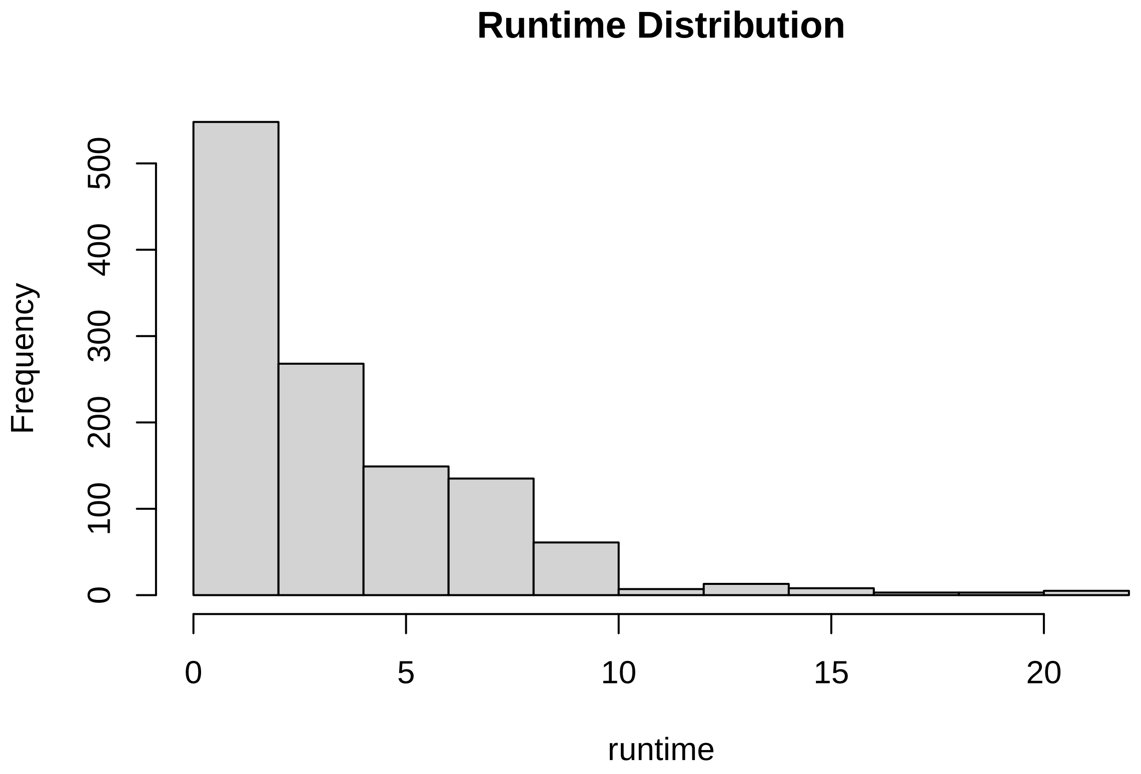
```
ggplot(df, aes(y = weekly_hours_viewed)) + geom_boxplot(fill = "lightblue") + labs(title = "Weekly Hours
```



It appears "Weekly Hours Viewed" skews heavily to the lower end of the distribution. This tells me again the data has likely been curated to only include many "average" weeks and a few outliers on the high end of the distribution. It would be strange that the lowest values would be so overly represented. One would think this, as a continuous data category would form a normal distribution.

I suspect the under-performing weeks were trimmed from the dataset.

```r
hist(df$runtime, main = "Runtime Distribution", xlab = "runtime")
```

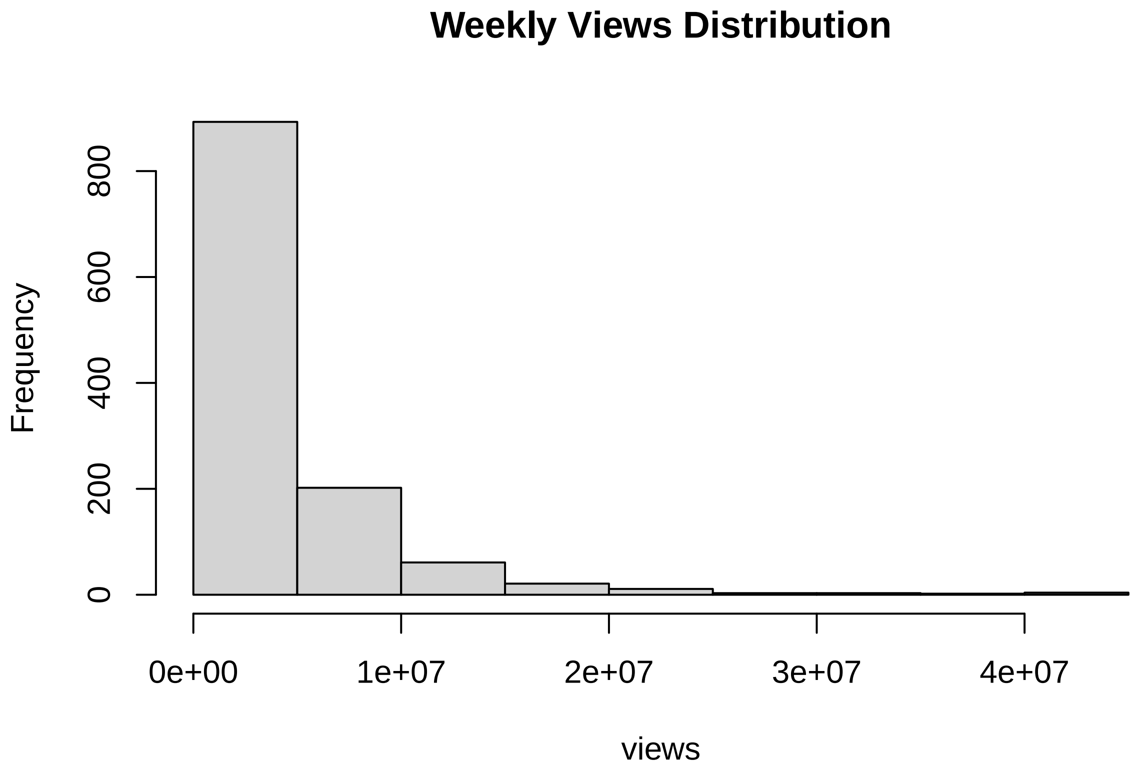


```r
summary(df$runtime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.667   2.117   3.596   4.929  20.300    4080
```

Nothing particularly revealing about runtimes other than it would appear the vast majority of representations are films or series with less than 10 cumulative runtime hours.

```r
hist(df$weekly_views, main = "Weekly Views Distribution", xlab = "views")
```



View numbers seem to follow a similar pattern. The existence of high outliers indicates weeks of extremely high viewership. Would be interesting to identify these weeks.

```
df[which.max(df$weekly_views), ]
```

```
##           week        category weekly_rank             show_title season_title
## 121 2023-12-17 Films (English)           1 Leave the World Behind          N/A
##     weekly_hours_viewed runtime weekly_views cumulative_weeks_in_top_10
## 121           106200000  2.3667     44900000                          2
##     is_staggered_launch episode_launch_details
## 121               false
```

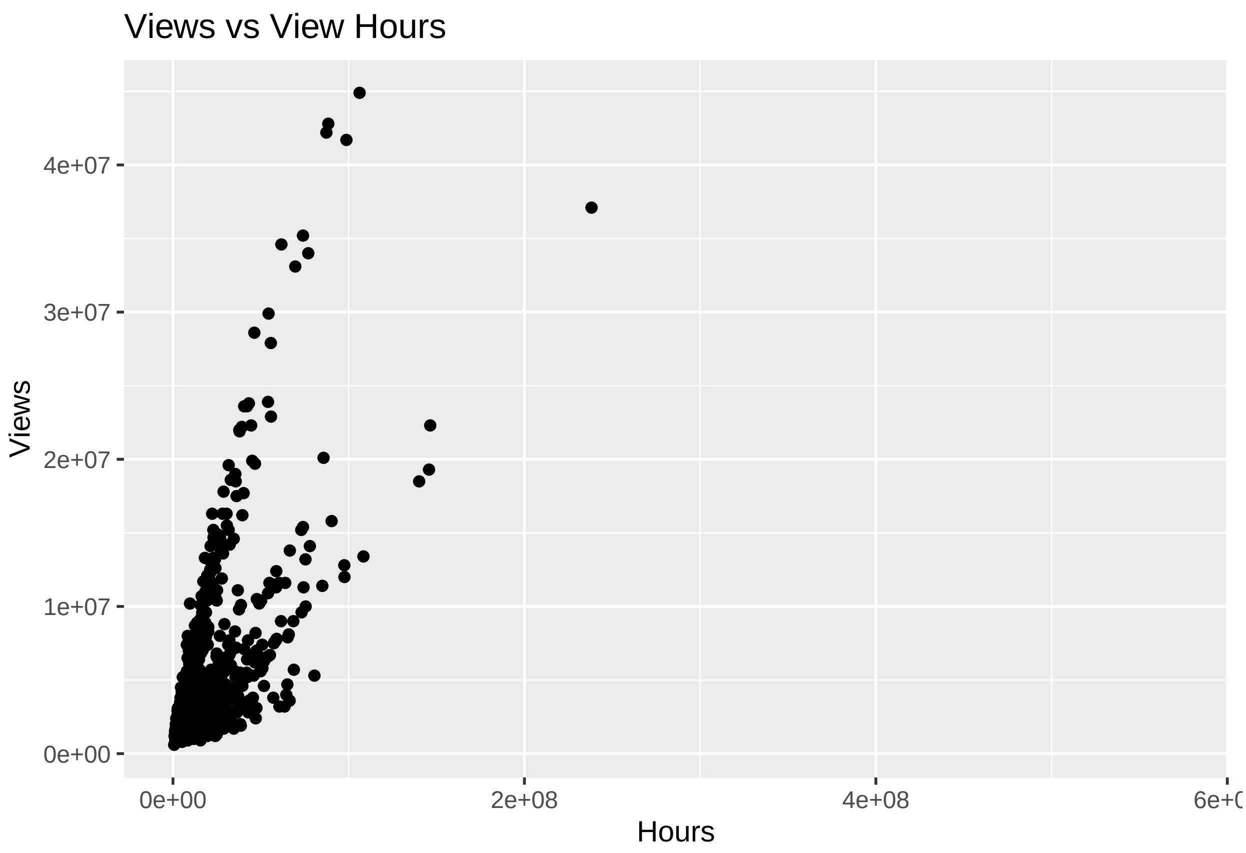
The week of 2023-12-17 appears to be the maximum for viewership. This should be taken with a grain of salt considering viewership information does not go back further than June of 2023.

The amount of absent information is starting to impact EDA.

I'm curious about the relationship between viewing hours and the number of views.

```
ggplot(data = df, aes(x = weekly_hours_viewed, y = weekly_views)) + geom_point() + labs(title = "Views
```

```
## Warning: Removed 4080 rows containing missing values (‘geom_point()‘).
```



Obvious correlation there. What others are there?

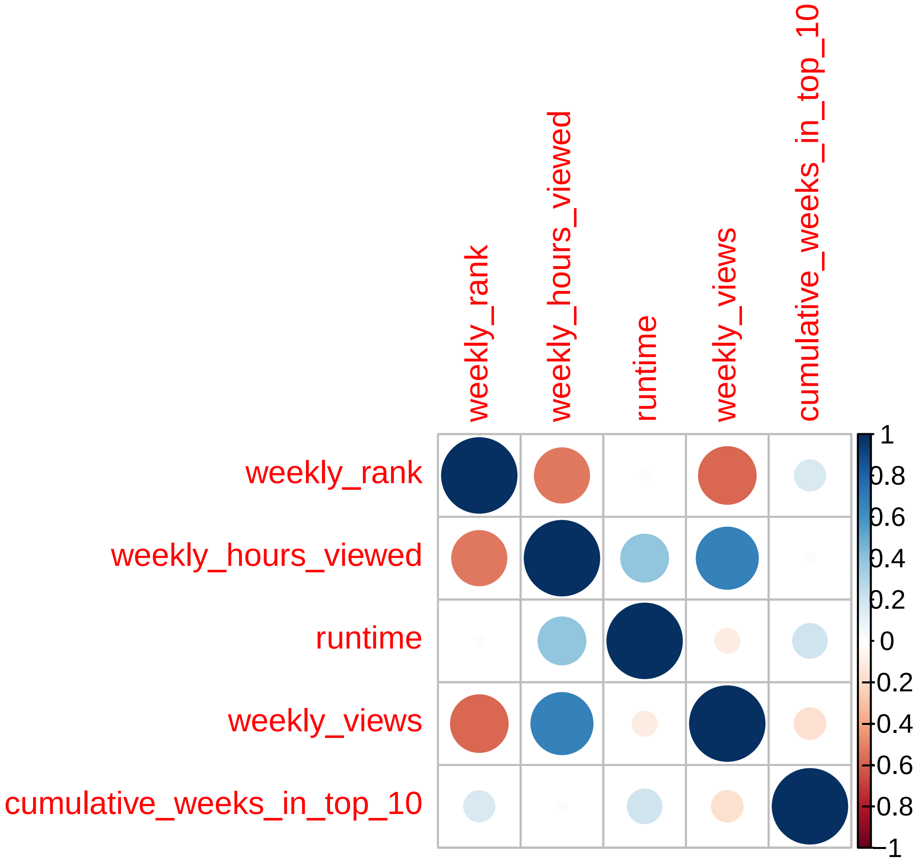
Correlation Matrix

```r
cor_matrix <- cor(df[, sapply(df, is.numeric)], use = "complete.obs")

library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
corrplot(cor_matrix, method = "circle")
```



Here we see the correlation between views and view hours visualized.

More interestingly, there appears to be a negative correlation between Weekly Views/Weekly View Hours and Weekly Rank. Meaning it would appear that as rank increases, view hours and viewership tends to decrease.

This, while strange on face-value, is explained by lower ranking indicating a higher position. A rank of 1 is superior to rank 10. Thus, as rank “decreases,” viewership increases.