Guided Project: Investigating Fandango Movie Ratings

Fandango é uma plataforma de compra de ingressos para cinema, além de também permitir que usuários dêem notas aos filmes. Em 2015 o jornalista Walt Hickey fez uma análise sobre a plataforma e publicou neste artigo.

Em resumo o jornalista evidenciou que a plataforma inflava as notas dadas aos filmes enviesando os dados, comparado aos concorrentes as notas são bem mais altas, mas além disso as notas dadas em estrelas de 0 a 5 exibidas no poster eram maiores que as notas reais, normalmente com meia estrela a mais. A empresa na época se justificou dizendo que havia um bug no site que seria corrigido, mas na verdade removeu os dados reais impossibilitando de confirmar se de fato foi corrigido.

Sabendo disso vamos avaliar dados anteriores e posteriores a esse estudo para entender se houve mudança no sistema de rating do site da Fandango.

Explorando os dados do site Fandango

Vamos utilizar os dados coletados pelo jornalista Walt Hickey, porém referente ao período anterior a análise feita. Os dados se encontram disponíveis neste link.

Um membro do time Dataquest coletou dados de filmes lançados em 2016 e 2017 que são referentes a um período posterior à análise do jornalista, e pode ser encontrado neste link.

```
base_anterior  <- read_csv("fandango_score_comparison.csv")
```

```
## Rows: 146 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): FILM
## dbl (21): RottenTomatoes, RottenTomatoes_User, Metacritic, Metacritic_User, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
glimpse(base_anterior)
```

```
## Rows: 146
## Columns: 22
## $ FILM                         <chr> "Avengers: Age of Ultron (2015)", "Cinderel~
## $ RottenTomatoes               <dbl> 74, 85, 80, 18, 14, 63, 42, 86, 99, 89, 84,~
## $ RottenTomatoes_User          <dbl> 86, 80, 90, 84, 28, 62, 53, 64, 82, 87, 77,~
## $ Metacritic                   <dbl> 66, 67, 64, 22, 29, 50, 53, 81, 81, 80, 71,~
## $ Metacritic_User              <dbl> 7.1, 7.5, 8.1, 4.7, 3.4, 6.8, 7.6, 6.8, 8.8~
## $ IMDB                         <dbl> 7.8, 7.1, 7.8, 5.4, 5.1, 7.2, 6.9, 6.5, 7.4~
## $ Fandango_Stars               <dbl> 5.0, 5.0, 5.0, 5.0, 3.5, 4.5, 4.0, 4.0, 4.5~
## $ Fandango_Ratingvalue         <dbl> 4.5, 4.5, 4.5, 4.5, 3.0, 4.0, 3.5, 3.5, 4.0~
## $ RT_norm                      <dbl> 3.70, 4.25, 4.00, 0.90, 0.70, 3.15, 2.10, 4~
## $ RT_user_norm                 <dbl> 4.30, 4.00, 4.50, 4.20, 1.40, 3.10, 2.65, 3~
## $ Metacritic_norm              <dbl> 3.30, 3.35, 3.20, 1.10, 1.45, 2.50, 2.65, 4~
## $ Metacritic_user_nom          <dbl> 3.55, 3.75, 4.05, 2.35, 1.70, 3.40, 3.80, 3~
## $ IMDB_norm                    <dbl> 3.90, 3.55, 3.90, 2.70, 2.55, 3.60, 3.45, 3~
## $ RT_norm_round                <dbl> 3.5, 4.5, 4.0, 1.0, 0.5, 3.0, 2.0, 4.5, 5.0~
## $ RT_user_norm_round           <dbl> 4.5, 4.0, 4.5, 4.0, 1.5, 3.0, 2.5, 3.0, 4.0~
## $ Metacritic_norm_round        <dbl> 3.5, 3.5, 3.0, 1.0, 1.5, 2.5, 2.5, 4.0, 4.0~
## $ Metacritic_user_norm_round <dbl> 3.5, 4.0, 4.0, 2.5, 1.5, 3.5, 4.0, 3.5, 4.5~
## $ IMDB_norm_round              <dbl> 4.0, 3.5, 4.0, 2.5, 2.5, 3.5, 3.5, 3.5, 3.5~
## $ Metacritic_user_vote_count <dbl> 1330, 249, 627, 31, 88, 34, 17, 124, 62, 54~
## $ IMDB_user_vote_count         <dbl> 271107, 65709, 103660, 3136, 19560, 39373, ~
## $ Fandango_votes               <dbl> 14846, 12640, 12055, 1793, 1021, 397, 252, ~
## $ Fandango_Difference          <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5~
```

```r
base_posterior <- read_csv("movie_ratings_16_17.csv")
```

```
## Rows: 214 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): movie
## dbl (14): year, metascore, imdb, tmeter, audience, fandango, n_metascore, n_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
glimpse(base_posterior)
```

```
## Rows: 214
```

```
## Columns: 15
## $ movie         <chr> "10 Cloverfield Lane", "13 Hours", "A Cure for Wellness",~
## $ year          <dbl> 2016, 2016, 2016, 2017, 2016, 2016, 2016, 2016, 2016, 201~
## $ metascore     <dbl> 76, 48, 47, 43, 58, 76, 54, 34, 60, 38, 59, 53, 81, 25, 3~
## $ imdb          <dbl> 7.2, 7.3, 6.6, 5.2, 6.1, 7.5, 7.4, 6.2, 7.1, 5.0, 7.2, 4.~
## $ tmeter        <dbl> 90, 50, 40, 33, 70, 87, 77, 30, 61, 0, 66, 45, 94, 4, 17,~
## $ audience      <dbl> 79, 83, 47, 76, 57, 84, 79, 50, 66, 27, 71, 16, 82, 22, 5~
## $ fandango      <dbl> 3.5, 4.5, 3.0, 4.5, 3.0, 4.0, 4.5, 4.0, 4.0, 3.5, 4.0, 3.~
## $ n_metascore   <dbl> 3.80, 2.40, 2.35, 2.15, 2.90, 3.80, 2.70, 1.70, 3.00, 1.9~
## $ n_imdb        <dbl> 3.60, 3.65, 3.30, 2.60, 3.05, 3.75, 3.70, 3.10, 3.55, 2.5~
## $ n_tmeter      <dbl> 4.50, 2.50, 2.00, 1.65, 3.50, 4.35, 3.85, 1.50, 3.05, 0.0~
## $ n_audience    <dbl> 3.95, 4.15, 2.35, 3.80, 2.85, 4.20, 3.95, 2.50, 3.30, 1.3~
## $ nr_metascore <dbl> 4.0, 2.5, 2.5, 2.0, 3.0, 4.0, 2.5, 1.5, 3.0, 2.0, 3.0, 2.~
## $ nr_imdb       <dbl> 3.5, 3.5, 3.5, 2.5, 3.0, 4.0, 3.5, 3.0, 3.5, 2.5, 3.5, 2.~
## $ nr_tmeter     <dbl> 4.5, 2.5, 2.0, 1.5, 3.5, 4.5, 4.0, 1.5, 3.0, 0.0, 3.5, 2.~
## $ nr_audience   <dbl> 4.0, 4.0, 2.5, 4.0, 3.0, 4.0, 4.0, 2.5, 3.5, 1.5, 3.5, 1.~
```

Preparando os dados para análise

Selecionando colunas de interesse

```
base_anterior <- base_anterior %>%
  select(FILM, Fandango_Stars, Fandango_Ratingvalue, Fandango_votes, Fandango_Differe

base_posterior <- base_posterior %>%
  select(movie, year, fandango)
```

Lendo o README de cada repositório é possível notar que as amostras não foram aleatórias, e que na verdade tiveram critérios específicos para seleção e em alguns caso não muito claros.

A base de Walt Hickey tem apenas filmes com votos em algumas plataformas concorrentes à Fandango, e pelo menos 30 votos na Fandango. Os dados foram extraídos em agosto de 2015.

A outra base tem filmes populares com número significante de votos (mas não revela quantos votos) e filmes lançados a partir de 2016.

Apesar de não ser o cenário ideal para coletas dos dados, podemos mudar ligeiramente nosso objetivo, analisar e tentar extrair resultados relevantes. Pois uma nova coleta de dados da forma necessária para atingir o objetivo inicial seria quase impossível. Sendo assim, vamos então observar se houve mudança de comportamento entre as notas dos anos de 2015 e 2016.

O primeiro passo é tentar isolar na primeira base apenas filmes do ano de 2015.

```
#isolando 2015
base_2015 <- base_anterior %>%
  mutate(year = str_sub(FILM, -5, -2)) %>%
  filter(year == 2015)

#conferindo se funcionou
base_2015$year %>% table()
```

```
## .
## 2015
##  129
```

```
#limpeza do que não é mais necessário
rm(base_anterior)

#resultado
base_2015 %>% head()
```

```
## # A tibble: 6 x 6
##   FILM    Fandango_Stars Fandango_Ratingvalue Fandango_votes Fandango_Difference
##   <chr>            <dbl>                <dbl>          <dbl>                <dbl>
## 1 Avenge~            5                  4.5          14846                  0.5
## 2 Cinder~            5                  4.5          12640                  0.5
## 3 Ant-Ma~            5                  4.5          12055                  0.5
## 4 Do You~            5                  4.5           1793                  0.5
## 5 Hot Tu~            3.5                3             1021                  0.5
## 6 The Wa~            4.5                4              397                  0.5
## # i 1 more variable: year <chr>
```

Seguindo a mesma ideia, vamos filtrar apenas 2016 na segunda base.

```
#isolando 2016
base_2016 <- base_posterior %>%
  filter(year == 2016)

#conferindo se funcionou
base_2016$year %>% table()
```

```
## .
## 2016
##  191
```

```r
#limpeza do que não é mais necessário
rm(base_posterior)

#resultado
base_2016 %>% head()
```

```
## # A tibble: 6 x 3
##   movie                       year fandango
##   <chr>                      <dbl>    <dbl>
## 1 10 Cloverfield Lane         2016      3.5
## 2 13 Hours                    2016      4.5
## 3 A Cure for Wellness         2016      3
## 4 A Hologram for the King     2016      3
## 5 A Monster Calls             2016      4
## 6 A Street Cat Named Bob      2016      4.5
```

O exercício sugere que investiguemos a popularidade dos filmes presentes na segunda base já que o critério de popularidade não foi tão claro igual a primeira baseada em 30 votos. No entanto uma vez que o site só funciona nos EUA, não vamos conseguir avaliar a representatividade e assumir que está ok para seguir com o exercício.

Outro ponto importante de ser analisado é se a nota continua usando a mesma métrica de 2015. Usando o summary é possível notar que nenhuma nota ultrapassou 5, onde chegamos a conclusão que o sistema ainda consiste em estrelas de 1 a 5. No entanto já é possível notar uma diferença interessante na menor nota, que em 2016 chegou a 2,5. Esse já pode ser um indício que o site mudou de comportamento.

```r
base_2015$Fandango_Stars %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   3.500   4.000   4.085   4.500   5.000
```

```r
base_2016$fandango %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.500   3.500   4.000   3.887   4.250   5.000
```

Analisando os dados

Existem muitas maneiras de analisar os dados, mas de forma geral, podemos num gráfico desenhar a distribuição de ambos os anos para ter uma ideia de como se comporta.

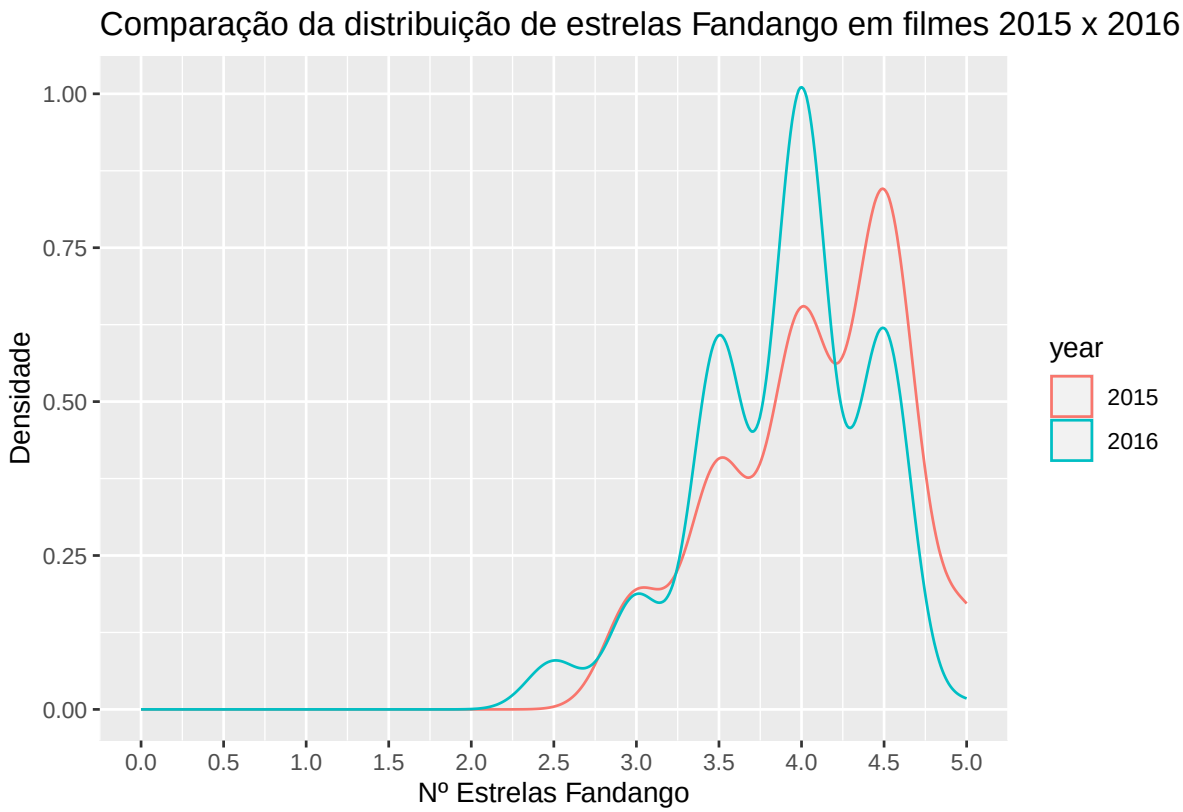
Primeiro unificando as bases para facilitar a plotagem.

```
base_completa <- rbind(
  base_2015 %>% select(movie = FILM,
                       year,
                       fandango = Fandango_Stars),
  base_2016
)
base_completa %>% sample_n(10)
```

```
## # A tibble: 10 x 3
##    movie                                          year  fandango
##    <chr>                                          <chr>    <dbl>
##  1 It Follows (2015)                              2015       3
##  2 My Big Fat Greek Wedding 2                     2016       4
##  3 Black Sea (2015)                               2015       4
##  4 Keeping Up with the Joneses                    2016       3.5
##  5 Pixels (2015)                                  2015       4.5
##  6 Fantastic Beasts and Where to Find Them        2016       4.5
##  7 Mr. Church                                     2016       4.5
##  8 Sinister 2 (2015)                              2015       3.5
##  9 Split                                          2016       4
## 10 Teenage Mutant Ninja Turtles: Out of the Shadows 2016       4
```

Enfim gerando o gráfico de forma a respeitar a escala de 0 a 5 estrelas de meia em meia estrela. Conseguimos notar que de fato o comportamento de 2016 mudou, e os filmes ficaram ligeiramente com menor nota que se comparado ao ano de 2015. Isso pode indicar que de fato foi corrigido o viés levantado no estudo do jornalista.

```
base_completa %>%
  ggplot(aes(x=fandango,color=year))+
  geom_density()+
  labs(x = "Nº Estrelas Fandango",
       y = "Densidade",
       title = "Comparação da distribuição de estrelas Fandango em filmes 2015 x 2016
  scale_x_continuous(breaks = seq(0,5, by = 0.5),
                     limits = c(0, 5))
```



Mas agora vamos investigar mais a fundo algumas informações, será que os dados são comparáveis? Vamos investigar algumas métricas.

```
#calculo para moda
mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

stats <-
base_completa %>%
  group_by(year) %>%
  summarise(n    = n(),
            mean = mean(fandango),
            min  = min(fandango),
            max  = max(fandango),
            mode = mode(fandango),
            median = median(fandango))

stats
```

```
## # A tibble: 2 x 7
```

```
##   year      n  mean   min   max  mode median
##   <chr> <int> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 2015    129  4.09   3       5   4.5      4
## 2 2016    191  3.89   2.5     5   4        4
```

com o resultado acima é possível notar tanto a média, moda e valor mínimo que houve uma queda na nota, reforçando o comportamento que vimos no gráfico. O número de filmes sendo comparados também é justo, a base de 2016 tem ainda mais filmes que a base anterior, então podemos descartar uma distorção nos números por falta de dados.