In a [previous post](https://www.mango-solutions.com/blog/gender-diversity-in-the-film-industry), I studied gender diversity in the film industry, I did this by focusing on some key behind-the-camera roles and measuring the evolution of the gender diversity in the last decade. The conclusion was not great: women are under-represented, especially in the most important roles of directors and writers, as these key roles determine the way women are portrayed in front of the camera.

I was curious about the TV series industry too: as the TV series industry is faster paced than the movie industry**, might they might be more open to women?** I decided to have a look.

In this post, as in the [film industry](https://www.mango-solutions.com/blog/gender-diversity-in-the-film-industry) post, the behind-the-camera roles I studied were: **directors**, **writers**, **producers**, **sound teams**, **music teams**, **art teams**, **makeup teams** and **costume teams**.

The whole code to reproduce the following results is available on [GitHub](https://github.com/MangoTheCat/Gender-diversity-series-industry).

**Data Frame Creation – Web Scraping**

All the data I used was gathered from the **IMDb website**: I went through the [100 Most Popular TV Shows](https://www.imdb.com/chart/tvmeter?sort=rk,asc&mode=simple&page=1) (according to the IMDb ratings), and gathered some useful information about these 100 series: I built a data frame which contains the titles of these series, their years of release and their IMDb episode links – the link where we can find all the episodes of a series.

# IMDb 100 most popular TV shows ------------------------------

url <- "https://www.imdb.com/chart/tvmeter?sort=us,desc&mode=simple&page=1"

page <- read\_html(url)

serie\_nodes <- html\_nodes(page, '.titleColumn') %>%

as\_list()

# Series details

serie\_name <- c()

serie\_link <- c()

serie\_year <- c()

for (i in seq\_along(serie\_nodes)){

serie\_name <- c(serie\_name, serie\_nodes[[i]]$a[[1]])

serie\_link <- c(serie\_link, attr(serie\_nodes[[i]]$a, "href"))

serie\_year <- c(serie\_year, serie\_nodes[[i]]$span[[1]])

}

serie\_link <- paste0("http://www.imdb.com",serie\_link)

serie\_year <- gsub("[()]", "", serie\_year)

serie\_episodelist <- sapply(strsplit(serie\_link, split='?', fixed=TRUE),

function(x) (x[1])) %>%

paste0("episodes?ref\_=tt\_eps\_yr\_mr")

# Create dataframe ----------------------------------------------

top\_series <- data.frame(serie\_name, serie\_year, serie\_episodelist, stringsAsFactors = FALSE)

# series\_year was the date of 1st release but we needed the years of release for all the episodes

# I did not manage to gather this information by doing some web scraping.

# I added it manually as it is available on the IMDb episodes links (column serie\_episodelist)

top\_series[20:30, ]

## serie\_name serie\_year

## 20 Legion 2017

## 21 A Series of Unfortunate Events 2017, 2018

## 22 Timeless 2016, 2017, 2018

## 23 Westworld 2016, 2018

## 24 Luke Cage 2016

## 25 MacGyver 2016, 2017, 2018

## 26 Lethal Weapon 2016, 2017, 2018

## 27 Designated Survivor 2016, 2017, 2018

## 28 Bull 2016, 2017, 2018

## 29 This Is Us 2016, 2017, 2018

## 30 Atlanta 2016, 2018

## serie\_episodelist

## 20 http://www.imdb.com/title/tt5114356/episodes?ref\_=tt\_eps\_yr\_mr

## 21 http://www.imdb.com/title/tt4834206/episodes?ref\_=tt\_eps\_yr\_mr

## 22 http://www.imdb.com/title/tt5511582/episodes?ref\_=tt\_eps\_yr\_mr

## 23 http://www.imdb.com/title/tt0475784/episodes?ref\_=tt\_eps\_yr\_mr

## 24 http://www.imdb.com/title/tt3322314/episodes?ref\_=tt\_eps\_yr\_mr

## 25 http://www.imdb.com/title/tt1399045/episodes?ref\_=tt\_eps\_yr\_mr

## 26 http://www.imdb.com/title/tt5164196/episodes?ref\_=tt\_eps\_yr\_mr

## 27 http://www.imdb.com/title/tt5296406/episodes?ref\_=tt\_eps\_yr\_mr

## 28 http://www.imdb.com/title/tt5827228/episodes?ref\_=tt\_eps\_yr\_mr

## 29 http://www.imdb.com/title/tt5555260/episodes?ref\_=tt\_eps\_yr\_mr

## 30 http://www.imdb.com/title/tt4288182/episodes?ref\_=tt\_eps\_yr\_mr

The series\_year column often contains several years. For example, for the series called “This is us”, it means that episodes have been released in 2016, 2017 and 2018. This column will allow me to split the episodes by year of release, and then visualise the gender diversity of the crew for each year.

**List Creation – Web Scraping**

At this stage, I just had some global information on the 100 series. The next step was to go through the IMDb links gathered in the column series\_episodelist of my top\_series data frame, which gives me access to all the series episodes split by year of release. I did some **web scraping** on these links and built a list which gathered:

* the names of the 100 most popular TV shows
* for each series, the different years of release
* for each year, the names of the episodes which have been released
* for each episode, the names of the people whose job was included in one of the categories I listed above (directors, writers, …, costume teams)

### Create series list

series\_list <- list()

# FOCUS ON EACH SERIES -----------------------------------------------------------------

for (r in seq\_len(nrow(top\_series))) {

serie\_name <- top\_series[r, "serie\_name"]

print(serie\_name)

# Years of release for each serie

list\_serieyear <- as.list(strsplit(top\_series[r, "serie\_year"], split = ", ")[[1]])

# List of IMDb links where we find all the episodes per year of release

link\_episodelist\_peryear <- list()

episodes\_list\_peryear <- list()

# FOCUS ON EACH YEAR OF REALEASE FOR THIS SERIE -------------------------------------

for (u in seq\_along(list\_serieyear)){

year <- list\_serieyear[[u]]

print(year)

link\_episodelist\_yeari <- strsplit(top\_series[r, "serie\_episodelist"], split='?', fixed=TRUE)[[1]][1] %>%

paste0("?year=", year, collapse = "")

link\_episodelist\_peryear[[u]] <- link\_episodelist\_yeari

# FOCUS ON EACH EPISODE FOR THIS YEAR OF RELEASE ----------------------------------

for (l in seq\_along(link\_episodelist\_peryear)){

page <- read\_html(link\_episodelist\_peryear[[l]])

episodes\_nodes <- html\_nodes(page, '.info') %>%

as\_list()

episode\_name <- c()

episode\_link <- c()

for (t in seq\_along(episodes\_nodes)){

episode\_name <- c(episode\_name, episodes\_nodes[[t]]$strong$a[[1]])

episode\_link <- c(episode\_link, attr(episodes\_nodes[[t]]$strong$a, "href"))

}

episode\_link <- paste0("http://www.imdb.com",episode\_link)

episode\_link <- sapply(strsplit(episode\_link, split='?', fixed=TRUE),

function(x) (x[1])) %>%

paste0("fullcredits?ref\_=tt\_ql\_1")

episode\_name <- sapply(episode\_name,

function(x) (gsub(pattern = "\\#", replacement = "", x))) %>% # some names = "Episode #1.1"

as.character()

# GATHER THE NAME OF THE EPISODE, ITS YEAR OF RELEASE AND ITS FULL CREW LINK ----

episodes\_details\_peryear <- data.frame(year = year,

episode\_name = episode\_name,

episode\_link = episode\_link,

stringsAsFactors = FALSE)

}

# FOCUS ON EACH FULL CREW LINK ----------------------------------------------------

for (e in seq\_len(nrow(episodes\_details\_peryear))){

print(episodes\_details\_peryear[e, "episode\_link"])

episode\_page <- read\_html(episodes\_details\_peryear[e, "episode\_link"])

episode\_name <- episodes\_details\_peryear[e, "episode\_name"]

# GATHER ALL THE CREW NAMES FOR THIS EPISODE -------------------------------------

episode\_allcrew <- html\_nodes(episode\_page, '.name , .dataHeaderWithBorder') %>%

html\_text()

episode\_allcrew <- gsub("[\n]", "", episode\_allcrew) %>%

trimws() #Remove white spaces

# SPLIT ALL THE CREW NAMES BY CATEGORY -------------------------------------------

episode\_categories <- html\_nodes(episode\_page, '.dataHeaderWithBorder') %>%

html\_text()

episode\_categories <- gsub("[\n]", "", episode\_categories) %>%

trimws() #Remove white spaces

## MUSIC DEPT -----------------------------------------------------------------------

episode\_music <- c()

for (i in 1:(length(episode\_allcrew)-1)){

if (grepl("Music by", episode\_allcrew[i])){

j <- 1

while (! grepl(episode\_allcrew[i], episode\_categories[j])){

j <- j+1

}

k <- i+1

while (! grepl(episode\_categories[j+1], episode\_allcrew[k])){

episode\_music <- c(episode\_music, episode\_allcrew[k])

k <- k+1

}

}

}

for (i in 1:(length(episode\_allcrew)-1)){

if (grepl("Music Department", episode\_allcrew[i])){

# Sometimes music dept is last category

if (grepl ("Music Department", episode\_categories[length(episode\_categories)])){

first <- i+1

for (p in first:length(episode\_allcrew)) {

episode\_music <- c(episode\_music, episode\_allcrew[p])

}

} else {

j <- 1

while (! grepl(episode\_allcrew[i], episode\_categories[j])){

j <- j+1

}

k <- i+1

while (! grepl(episode\_categories[j+1], episode\_allcrew[k])){

episode\_music <- c(episode\_music, episode\_allcrew[k])

k <- k+1

}

}

}

}

if (length(episode\_music) == 0){

episode\_music <- c("")

}

## IDEM FOR OTHER CATEGORIES ----------------------------------------------------------

## EPISODE\_INFO CONTAINS THE EPISODE CREW NAMES ORDERED BY CATEGORY -------------------

episode\_info <- list()

episode\_info$directors <- episode\_directors

episode\_info$writers <- episode\_writers

episode\_info$producers <- episode\_producers

episode\_info$sound <- episode\_sound

episode\_info$music <- episode\_music

episode\_info$art <- episode\_art

episode\_info$makeup <- episode\_makeup

episode\_info$costume <- episode\_costume

## EPISODES\_LIST\_PER\_YEAR GATHERS THE INFORMATION FOR EVERY EPISODE OF THE SERIE-------

## SPLIT BY YEAR OF RELEASE --------------------------------------------------------

episodes\_list\_peryear[[year]][[episode\_name]] <- episode\_info

}

## SERIES\_LIST GATHERS THE INFORMATION FOR EVERY YEAR AND EVERY SERIE -------------------

series\_list[[serie\_name]] <- episodes\_list\_peryear

}

}

Let’s have a look at the information gathered in series\_list. Here are some of the names I collected:

## - Black Mirror, 2011

## Episode: The National Anthem

## Director: Otto Bathurst

## - Black Mirror, 2017

## Episode: Black Museum

## Director: Colm McCarthy

## - Game of Thrones, 2011

## Episode: Winter Is Coming

## Music team: Ramin Djawadi, Evyen Klean, David Klotz, Robin Whittaker, Michael K. Bauer, Brandon Campbell, Stephen Coleman, Janet Lopez, Julie Pearce, Joe Rubel, Bobby Tahouri

## - Game of Thrones, 2017

## Episode: Dragonstone

## Music team: Ramin Djawadi, Omer Benyamin, Evyen Klean, David Klotz, William Marriott, Douglas Parker, Stephen Coleman

What we can see is that for the same series the crew changes depending on the episode we consider.

**Gender Determination**

Now that I had all the names gathered in the series\_list, I needed to determine the genders. I used the same package as in my previous post on the film industry: **GenderizeR**, which “uses genderize.io API to predict gender from first names”. More details on this package and the reasons why I decided to use it are available in my [previous post.](https://www.mango-solutions.com/blog/gender-diversity-in-the-film-industry)

With this R package, I was able to determine for each episode the number of males and females in each category of jobs:

* the number of male directors,
* the number of female directors,
* the number of male producers,
* the number of female producers,
* the number of males in costume team,
* the number of females in costume team.

Here is the code I wrote:

### Genderize our lists of names

# for each serie

for (s in seq\_along(series\_list) ){

print(names(series\_list[s])) # print serie name

# for each year

for (y in seq\_along(series\_list[[s]])){

print(names(series\_list[[s]][y])) # print serie year

# for each episode

for (i in seq\_along(series\_list[[s]][[y]])){

print(names(series\_list[[s]][[y]][i])) # print serie episode

# Genderize directors -----------------------------------------------------

directors <- series\_list[[s]][[y]][[i]]$directors

if (directors == ""){

directors\_gender <- list()

directors\_gender$male <- 0

directors\_gender$female <- 0

series\_list[[s]][[y]][[i]]$directors\_gender <- directors\_gender

}

else{

# Split the firstnames and the lastnames

# Keep the firstnames

directors <- strsplit(directors, " ")

l <- c()

for (j in seq\_along(directors)){

l <- c(l, directors[[j]][1])

}

directors <- l

serie\_directors\_male <- 0

serie\_directors\_female <- 0

# Genderize every firstname and count the number of males and females

for (p in seq\_along(directors)){

directors\_gender <- genderizeAPI(x = directors[p], apikey = "233b284134ae754d9fc56717fec4164e")

gender <- directors\_gender$response$gender

if (length(gender)>0 && gender == "male"){

serie\_directors\_male <- serie\_directors\_male + 1

}

if (length(gender)>0 && gender == "female"){

serie\_directors\_female <- serie\_directors\_female + 1

}

}

# Put the number of males and females in series\_list

directors\_gender <- list()

directors\_gender$male <- serie\_directors\_male

directors\_gender$female <- serie\_directors\_female

series\_list[[s]][[y]][[i]]$directors\_gender <- directors\_gender

}

# Same code for the 7 other categories -----------------------------------

}

}

}

}

Here are some examples of numbers of male and female I collected:

## Black Mirror, 2011

## Episode: The National Anthem

## Number of male directors: 1

## Number of female directors: 0

##

## Black Mirror, 2017

## Episode: Black Museum

## Number of male directors: 1

## Number of female directors: 0

##

## Game of Thrones, 2011

## Episode: Winter Is Coming

## Number of male in music team: 8

## Number of female in music team: 3

##

## Game of Thrones, 2017

## Episode: Dragonstone

## Number of male in music team: 7

## Number of female in music team: 0

##

**Percentages Calculation**

With these numbers gathered in my list, I then calculated the percentages of women in each job category, for each year between 2007 and 2018. I gathered these figures in a data frame called percentages:

## year directors writers producers sound music art makeup

## 1 2018 22.69693 25.06514 27.87217 12.247212 23.25581 36.93275 73.10795

## 2 2017 20.51948 28.20016 27.28932 10.864631 25.46912 29.90641 71.41831

## 3 2016 17.13456 24.51189 27.93240 11.553444 25.03117 30.98003 71.74965

## 4 2015 16.14764 19.42845 26.43828 11.214310 22.16505 29.83354 69.50787

## 5 2014 18.38624 20.88644 27.59163 10.406150 22.21016 30.11341 69.97544

## 6 2013 14.94413 19.60432 28.15726 10.504896 23.29693 29.01968 69.01683

## 7 2012 15.60694 19.82235 29.66566 10.685681 21.45378 26.74160 67.47677

## 8 2011 13.95349 17.60722 26.73747 11.296882 17.11185 25.61805 64.81795

## 9 2010 15.95745 17.05882 27.38841 11.264644 16.51376 24.14815 65.33004

## 10 2009 16.49123 18.90496 28.79557 8.498350 21.72285 26.11128 68.15961

## 11 2008 17.87440 16.62088 29.05844 7.594264 18.74405 23.46251 68.39827

## 12 2007 21.15385 21.78771 30.12798 9.090909 19.23077 21.66124 63.03502

## costume

## 1 77.24853

## 2 81.34648

## 3 79.35358

## 4 76.48649

## 5 76.62972

## 6 74.74791

## 7 77.35247

## 8 77.46315

## 9 77.67380

## 10 79.56332

## 11 80.53191

## 12 79.24720

**Gender Diversity in 2017: TV Series Industry VS Film Industry**

Based on this data frame, I created some bar plots to visualise the gender diversity of each job category for each year. Here is the code I wrote to create the bar plot for 2017, which compares the TV series industry to the film industry.

### Barplot 2017

# Data manipulation -------------------------------------------------------------

# Import our movies dataset

percentages\_movies <- read.csv("percentages\_movies.csv")

percentages\_movies <- percentages\_movies[ , -1]

# Change column names for movie and serie dataframes

colnames(percentages\_movies) <- c("year", "directors", "writers", "producers", "sound", "music", "art", "makeup", "costume")

colnames(percentages) <- c("year", "directors", "writers", "producers", "sound", "music", "art", "makeup", "costume")

# From wide to long dataframes

percentages\_movies\_long <- percentages\_movies %>%

gather(key = category, value = percentage, -year)

percentages\_long <- percentages %>%

gather(key = category, value = percentage, -year)

# Add a column to these dataframes: movie or film ?

percentages\_movies\_long$industry <- rep("Film industry", 88)

percentages\_long$industry <- rep("Series industry", 96)

# Combine these 2 long dataframes

percentages\_movies\_series <- bind\_rows(percentages\_long, percentages\_movies\_long)

# Filter with year=2017

percentages\_movies\_series\_2017 <- percentages\_movies\_series %>%

filter(year == 2017)

# Data visualisation -------------------------------------------------------------

percentages\_movies\_series\_2017$percentage <- as.numeric(format(percentages\_movies\_series\_2017$percentage,

digits = 2))

bar\_2017 <- ggplot(percentages\_movies\_series\_2017, aes(x = category,

y = percentage,

group = category,

fill = category)) +

geom\_bar(stat = "identity") +

facet\_wrap(~industry) +

coord\_flip() + # Horizontal bar plot

geom\_text(aes(label = percentage), hjust=-0.1, size=3) +

theme(axis.text.x=element\_blank(),

axis.ticks.x=element\_blank(),

axis.text.y=element\_blank(),

axis.ticks.y=element\_blank(),

axis.title.y=element\_blank(),

plot.title = element\_text(hjust = 0.5), # center the title

legend.title=element\_blank()) +

labs(title = paste("Percentages of women in 2017"),

x = "",

y = "Percentages") +

guides(fill = guide\_legend(reverse=TRUE)) + # reverse the order of the legend

scale\_fill\_manual(values = brewer.pal(8, "Spectral")) # palette used to fill the bars and legend boxs

I have built a simple shiny app which gives access to the bar plots for each year between 2007 and 2017.

Let’s analyse the graph of the year 2017. If we only focus on the TV series figures, we see that sound teams show the lowest female occupation, with less than 11%. It is followed by the **role of director with 20.5%.** Then, we can see that **between 25% and 30% of the roles of writers, producers, music teams and art teams are taken by women**. Thus, women are still under-represented in the TV series industry. However, even if series figures show little gender diversity in the above job categories, they are better than the film industry ones, especially for the key roles of **directors, writors and producers**, which are respectively **5.7, 3 and 1.2 times higher for the series industry than for the film industry.** The last thing to notice is that as in the film industry, the series industry graph shows a representativeness gap between the above roles and the jobs of **make-up artists and costume designers, among which more than 70% of the roles are taken by women.**

**Evolution of the Gender Diversity: TV Series Industry VS Film Industry**

Let’s have a look at the evolution of the gender diversity in these two industries in the last decade.

### Evolution plot

# year as date

percentages\_movies\_series\_ymd <- percentages\_movies\_series %>%

subset(year != 2018)

percentages\_movies\_series\_ymd$year <- ymd(percentages\_movies\_series\_ymd$year, truncated = 2L)

# Data visualisation

evolution <- ggplot(percentages\_movies\_series\_ymd, aes(x = year,

y = percentage,

group = category,

colour = category)) +

geom\_line(size = 2) +

facet\_wrap(~industry) +

theme(panel.grid.minor.x = element\_blank(),

plot.title = element\_text(hjust = 0.5)) + # center the title

scale\_x\_date(date\_breaks = "2 year", date\_labels = "%Y") +

scale\_color\_manual(values = brewer.pal(8, "Set1")) +

labs(title = "Percentages of women from 2007 to 2017\n Film industry VS serie industry",

x = "",

y = "Percentages")

The first thing I noticed is that for both the film and series industries, the representation gap between the roles of make-up artists and costume designers and the other ones had not decreased since 2007.

The fact that the roles of directors, writers and producers are more open to women in the TV series industry than in the film one is easy to visualise with this graph, and we can see that it has been the case at least since 2007 (and probably before). Besides, since 2007 the series industry has been more diversified in terms of gender for all the categories I studied, except for the sound roles.

I also noticed that **since 2010/2011, in the TV series industry, almost all the categories tend to be more diversified in terms of gender**. The only exceptions are the roles of producers (percentages are generally decreasing slightly since 2007), sound teams (no improvement has been achieved since 2010) and costume teams (the trend has been positive only since 2013). Apart from that, **there is a positive trend for the TV series industry, which is not the case for the film industry**.

This trend is significant for some roles: writers, music teams, art teams and make-up teams percentages in the series industry have increased by 5 to 10% in the last decade**.** But if we look at the role of **directors**, the percentage of women has also increased by 5% since 2011, but the percentage reached in 2017 is essentially the same as the one reached in 2007, just as for the film industry. Let’s hope that the trend seen since 2011 for directors will continue.

**Conclusion**

This study has definitely shown that **the TV series industry is more diversified in terms of gender than the film industry, especially for the key roles of directors and writers**.

However even if the series percentages are better than the film ones, **women are still under-represented in the TV series industry** as the same regrettable analysis has been echoed: the only jobs which seem open to women are the stereotyped female jobs of make-up artists and costume designers. In all the other categories, the percentages of women in the series industry never reach more than 30%.

But contrary to the film industry, the TV series one is actually evolving in the right direction: since 2011, a positive trend has been happening for directors and writers. This evolution is encouraging for the future and suggests that powerful female characters, such as Daenerys Targaryen from Game of Thrones, are coming on TV screens.