Coding Sample

Gabriel Reyes

2025-1-20

Background

This was part of a project that explored the use of collapsible corporation in the form of production companies in the film making business. The collapsible corporation would be used to take advantage of a tax loop that allowed investors to see profit returns in a preferential manner. The data comes from the American Film Institute's AFI Catalog that includes all theatrically American films released from 1893 to 1993. To access the database, we had to use a JupyterLab Python environment within ProQuest. The database stored the data as XML files which required the Python package BeautifulSoup to read XML tag information to convert each film page into a row in a Pandas dataframe.

Additional code samples

Here is a link to my GitHub with additional code samples.

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
```

```r
if(Sys.info()['user']=='gr2757') {

  project <- "/Users/gr2757/Dropbox/Film-Project/Input/AFI_data"
}

film_data_AFI = read.csv(file.path(project,"full_set_AFI.csv"))
producer_data_AFI = read.csv(file.path(project,"full_producers_AFI.csv"))
production_data_AFI = read.csv(file.path(project,"full_production_companies_AFI.csv"))
dist_company_data = read.csv(file.path(project,"AFI_distributors_distributors.csv"))
```

```r
#some of the files have producer, some have producer; this is to standardize it

if ("Producers" %in% colnames(film_data_AFI)) {
  # Rename the column to "producer"
  colnames(film_data_AFI)[colnames(film_data_AFI) == "Producers"] <- "Producer"
}


film_data_AFI <- subset(film_data_AFI, Producer != "")
film_data_AFI <- subset(film_data_AFI, Producer != "[None, None]")
film_data_AFI <- subset(film_data_AFI, Producer != "[None]")
film_data_AFI <- subset(film_data_AFI, Producer != "[]")


film_data_AFI <- subset(film_data_AFI, Production_Companies != "")
film_data_AFI <- subset(film_data_AFI, Production_Companies != "[None, None]")
film_data_AFI <- subset(film_data_AFI, Production_Companies != "[]")
film_data_AFI <- subset(film_data_AFI, Production_Companies != "[None]")

#At this point there is one entry per movie per producer and the
# production companies are in the list format

#dropping non unique entries (removing producers) to only have a data set on films

film_data_AFI = film_data_AFI[!duplicated(film_data_AFI$Title),]

#sticking to US films
film_data_AFI <- film_data_AFI %>%
  filter(country == "United States")
#adding year subset
film_data_AFI$YEAR = as.integer(substring(film_data_AFI$Date,1,4))


#dropping exact date
film_data_AFI = select(film_data_AFI,-Date,-X, Producer, Production_Companies)
```

```r
if ("Producers" %in% colnames(producer_data_AFI)) {
  # Rename the column to "producer"
  colnames(producer_data_AFI)[colnames(producer_data_AFI) == "Producers"] <- "Producer"
}

#renaming blanks
producer_data_AFI["Producer"][producer_data_AFI["Producer"] == ""] <- "Blank Producer"



#adding year
producer_data_AFI$YEAR = substr(producer_data_AFI$Date,1,4)
#dropping production companies and date
producer_data_AFI = subset(producer_data_AFI, select = -c(Production_Companies, Date,X))
#keeping it to US data
producer_data_AFI <- producer_data_AFI %>%
  filter(country == "United States")
#at this point the data is a bunch of films with the attached producers
```

```
if ("Producers" %in% colnames(production_data_AFI)) {
  # Rename the column to "producer"
  colnames(production_data_AFI)[colnames(production_data_AFI) == "Producers"] <- "Producer"
}

#renaming blanks
production_data_AFI["Production_Companies"][production_data_AFI["Production_Companies"] == ""] <- "Blan
production_data_AFI["Production_Companies"][production_data_AFI["Production_Companies"] == "[]"] <- "Bla

#adding year
production_data_AFI$YEAR = substr(production_data_AFI$Date,1,4)
#dropping production companies and date
production_data_AFI = subset(production_data_AFI, select = -c(Producer, Date,X))
#keeping it to US data
production_data_AFI <- production_data_AFI %>%
  filter(country == "United States") %>%
  distinct(Title, Production_Companies, .keep_all =  TRUE)
#at this point the data is a bunch of films with the attached production companies
```

```
#counting how many films each production company made
production_info_AFI = production_data_AFI

#seeing how many films were produced by a company
production_counts_AFI = production_info_AFI %>%
  group_by(Production_Companies) %>%
  summarise(number_of_films = n())

#first film, last film and how long the company was active
company_life_AFI = production_info_AFI %>%
  group_by(Production_Companies) %>%
  summarise(company_life = max(as.numeric(YEAR)) - min(as.numeric(YEAR)),
            First_Apperance = min(as.numeric(YEAR)),
            Last_Apperance = max(as.numeric(YEAR)))

#count how many production companies there are per film
production_data_AFI <- production_data_AFI %>%
  group_by(Title) %>%
  mutate(ProductionCompanyCount = n_distinct(Production_Companies)) %>%
  ungroup() %>%
  mutate(BlankProduction = ifelse(Production_Companies == "Blank Production", 1, 0)) #specifying whethe

production_info_AFI <- production_counts_AFI %>%
  left_join(company_life_AFI, by = "Production_Companies")


#creating column for suspicion
production_info_AFI <- production_info_AFI %>%
  group_by(Production_Companies) %>%
  mutate(
    production_suspicion = case_when(
      company_life > 1  & number_of_films > 0 ~ 0, # 0 if they have made multiple movies and have a com
      number_of_films == 1 ~ 1,                    # 1 if they have only made 1 movie
      company_life == 1 & number_of_films > 1 ~ 2, # 2 if they have made multiple movies but all in two
      company_life == 0 & number_of_films > 1 ~ 3, # 3 if they have made multiple movies but all in one
    )
  ) %>%
  ungroup()
```

```
#counting how many films each production company made
producer_info_AFI = producer_data_AFI

#seeing how many films were produced by a company
producer_counts_AFI = producer_info_AFI %>%
  group_by(Producer) %>%
  summarise(number_of_films = n())

#first film, last film and how long the company was active
career_span_AFI = producer_info_AFI %>%
  group_by(Producer) %>%
  summarise(career_span = max(as.numeric(YEAR)) - min(as.numeric(YEAR)),
            First_Apperance = min(as.numeric(YEAR)),
            Last_Apperance = max(as.numeric(YEAR)))

#counting how many films a producer made
producer_data_AFI <- producer_data_AFI %>%
  group_by(Title) %>%
  mutate(Producer_Count = n_distinct(Producer)) %>%
  ungroup() %>%
  mutate(BlankProducer = ifelse(Producer == "Blank Producer", 1, 0)) #specifying whether there was a bl


producer_info_AFI <- producer_counts_AFI %>%
  left_join(career_span_AFI, by = "Producer")


#creating column for suspicion
producer_info_AFI <- producer_info_AFI %>%
  group_by(Producer) %>%
  mutate(
    producer_suspicion = case_when(
      career_span > 1  & number_of_films > 0 ~ 0,  # 0 if they have made multiple movies and have a car
      number_of_films == 1 ~ 1,                    # 1 if they have only made 1 movie
      career_span == 1 & number_of_films > 1 ~ 2,  # 2 if they have made multiple movies but have a car
      career_span == 0 & number_of_films > 1 ~ 3  # 3 if they have made multiple movies but all in one
    )
  ) %>%
  ungroup()
```

```
#subset to only have the two things need to merge
producers_info_subset = producer_info_AFI[, c("Producer", "producer_suspicion")]

#merging suspicion level and film data
film_suspicion_producers <- merge(producer_data_AFI, producers_info_subset, by = "Producer", all.x = TRU

#picking the highest values and attaching the that to the film
film_suspicion_producers = film_suspicion_producers %>%
  group_by(Title) %>%
  slice(which.max(producer_suspicion)) %>%
  ungroup() %>%
  mutate(producer_suspicion = ifelse(BlankProducer == 1, 4, producer_suspicion))


#subset to only have the two things need to merge
production_info_subset = production_info_AFI[, c("Production_Companies", "production_suspicion")]

#merging suspicion level and film data
film_suspicion_production <- merge(production_data_AFI, production_info_subset, by = "Production_Compan
#picking the highest values and attaching the that to the film
film_suspicion_production = film_suspicion_production %>%
  group_by(Title) %>%
  slice(which.max(production_suspicion)) %>%
  ungroup() %>%
  mutate(production_suspicion = ifelse(BlankProduction == 1, 4, production_suspicion))


#more subset
film_suspicion_production = film_suspicion_production[, c("Title", "production_suspicion", "BlankProduc

#left joining and adding a total
master_film_suspicion <- film_suspicion_producers %>%
  left_join(film_suspicion_production, by = "Title") %>%
  select(-Producer) %>%
  mutate(total_suspicion = producer_suspicion + production_suspicion)
```

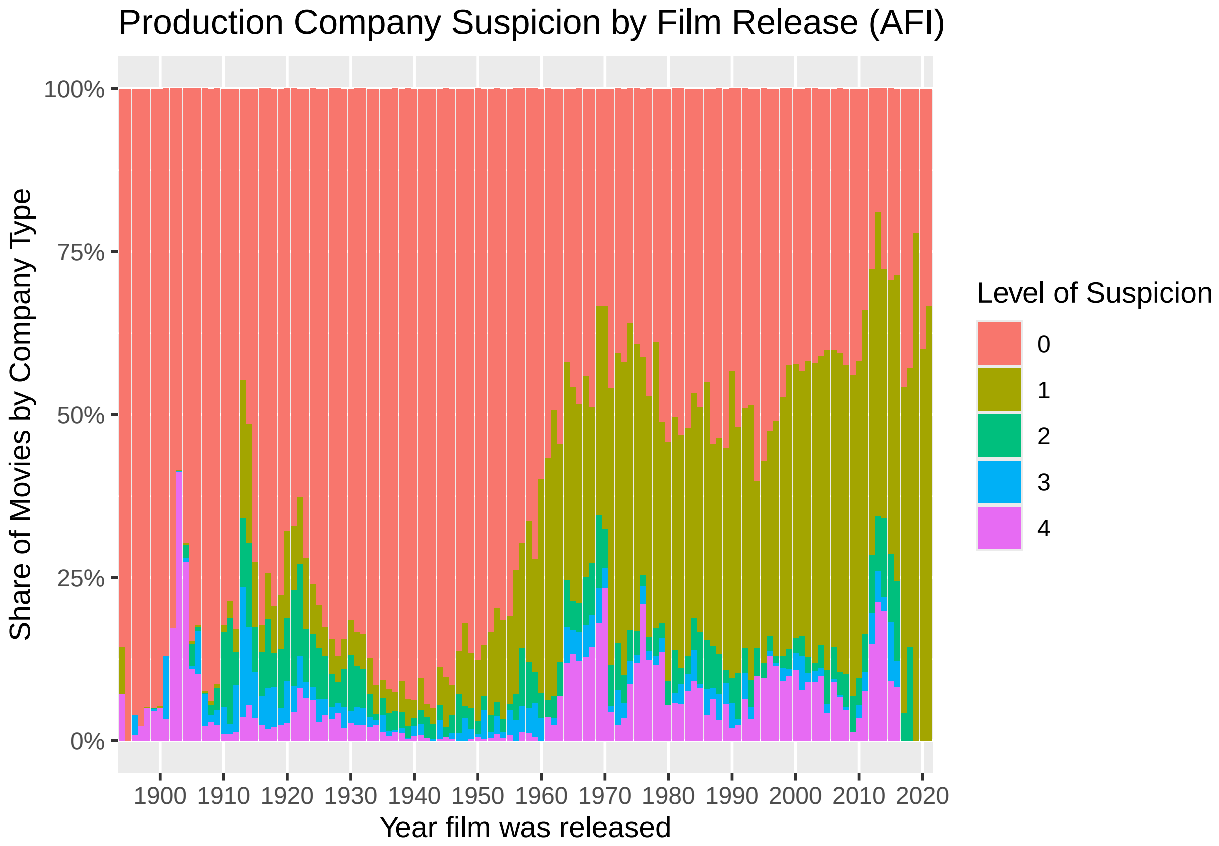
Section of graphs:

Constructed variable of movie suspicion:

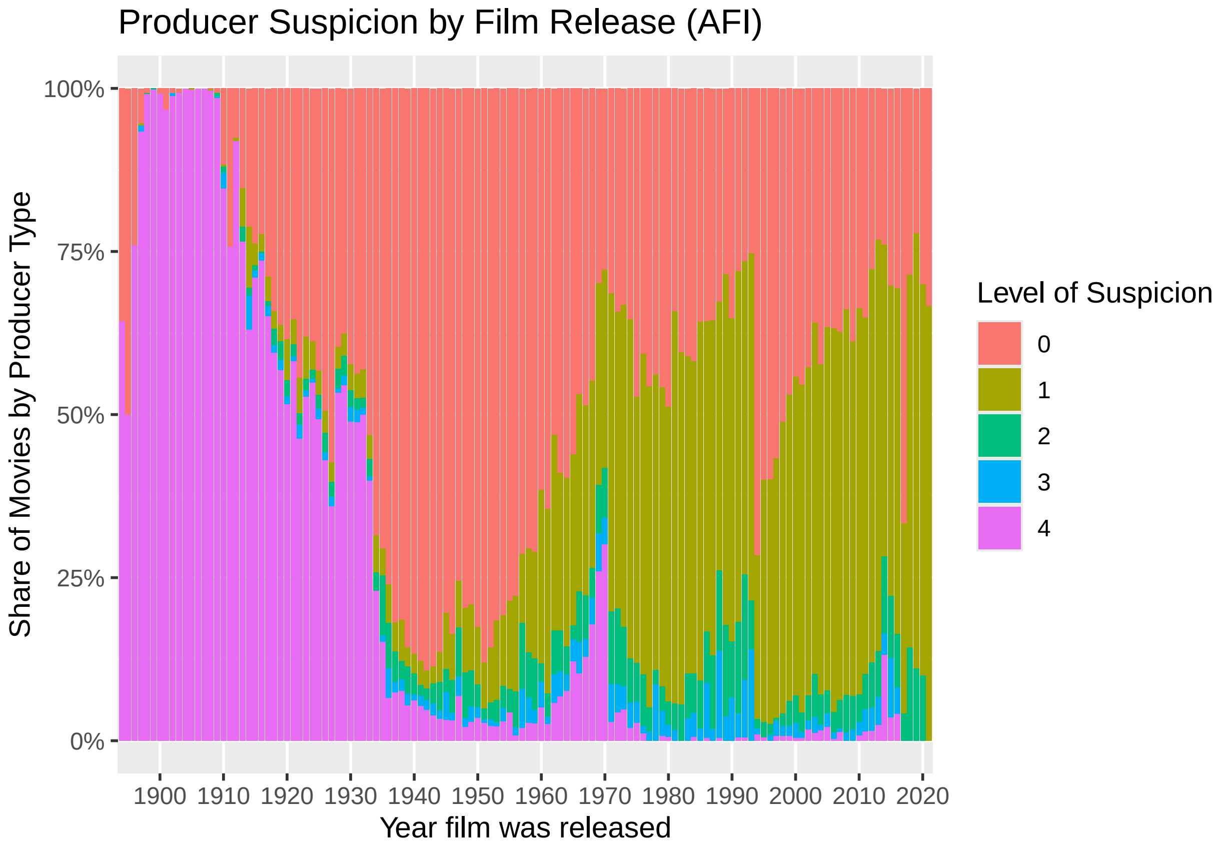
We created a constructed variable to be able to graph how likely a producer/production company acted in a manner that would lead us to believe it could be for tax purposes. We used the parameters of career length and quantity of credited films.

Here is the key for level of suspicion:

```
0: A production company/producer worked on more than one film and year.

1: A production company/producer only made one movie ever.

2: A production company/producer made more than one movie been only had a career length of 2 years.

3: A production company/producer made multiple movies in only just one year.

4: The data point has a blank for production company/producer.
```

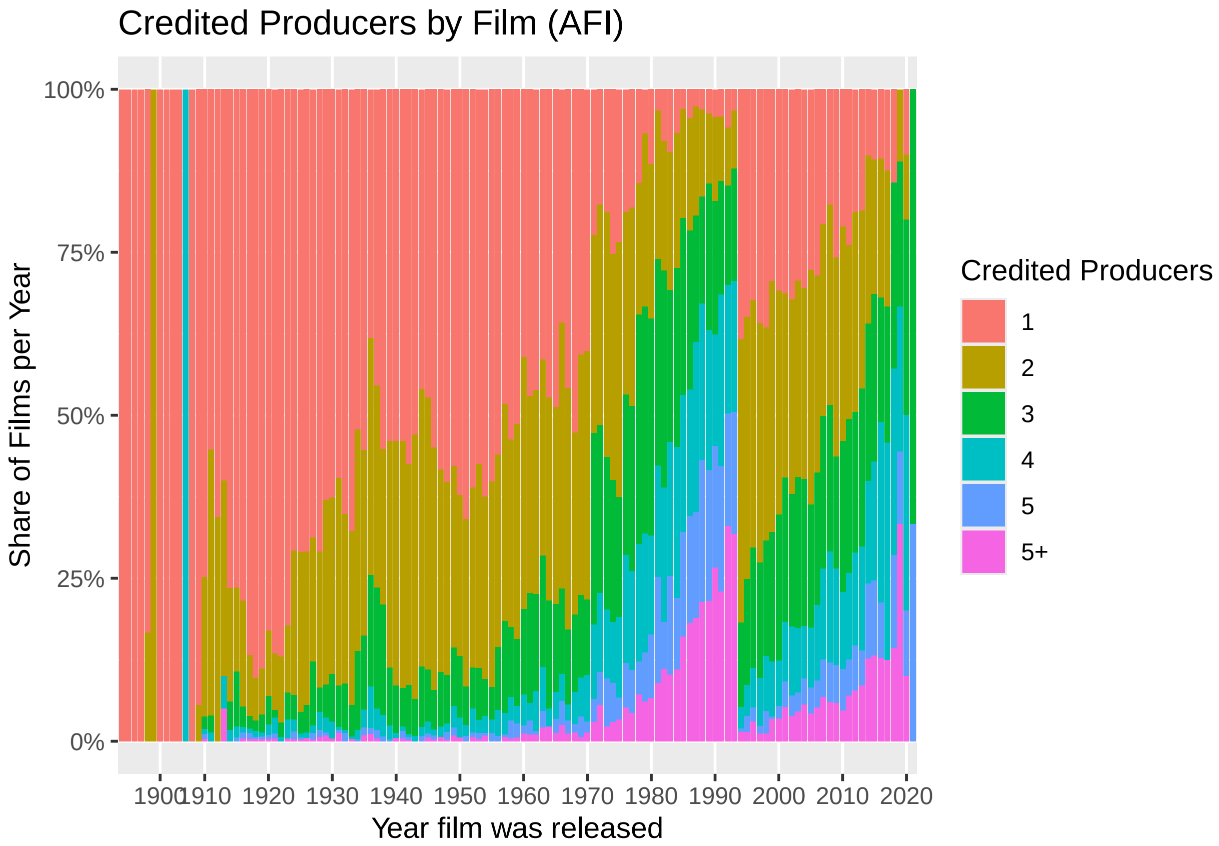


The target time of interest is 1980 to 1990 and there doesn't seem to be any pattern of growth in high suspicion coded production companies. Starting in around 1950s there is a increase in production companies only producing one film in one year. There is also a spike of 1910s, however that is due to the graph being measured in proportion, very few films were released in the 1910s; likely due to WW1.

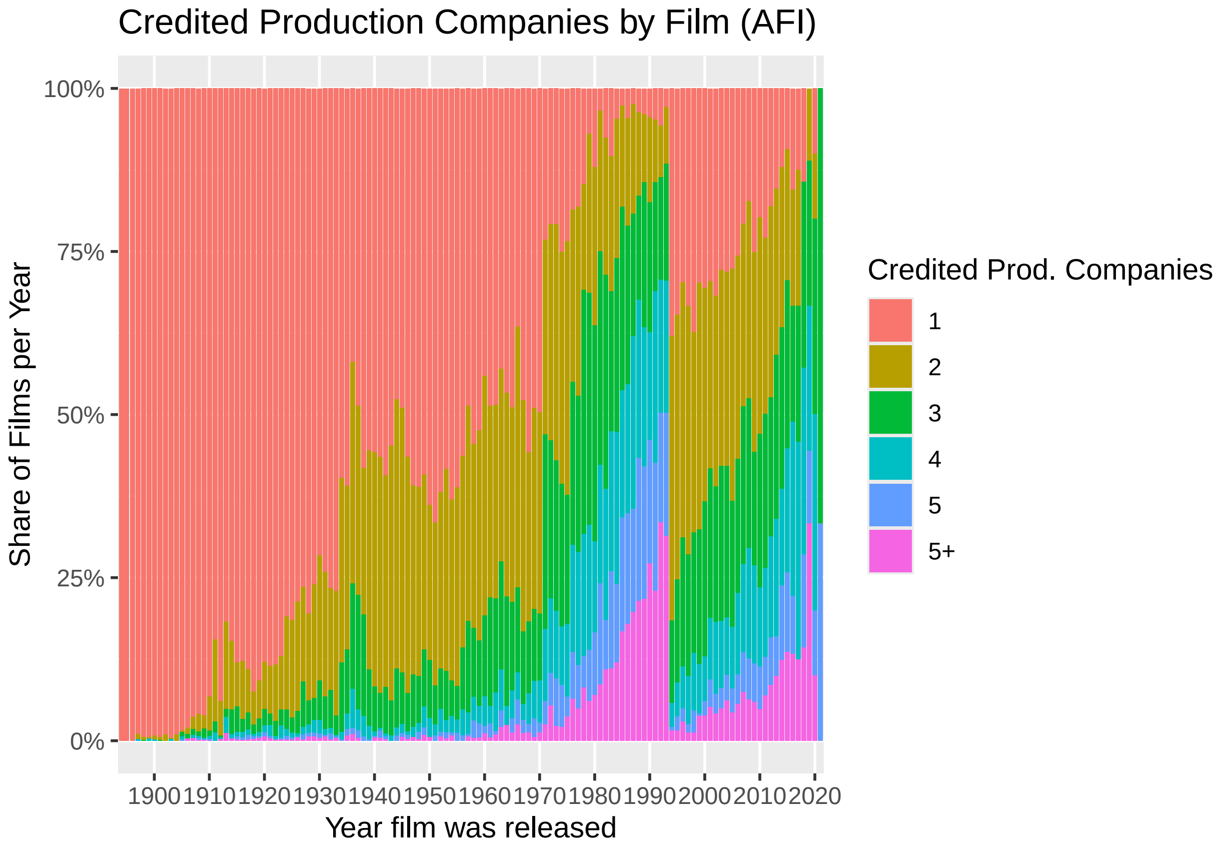


The producer data results is much more noisy, from 1900s to around 1935 there are a lot films in the data that are missing producer credits. Starting at 1940s there is a trend of producers only ever produced one film. This goes all the way to roughly 1990, then there is a recent and goes back up to 75% of producers only producing one film. There is some other levels of suspicion in producers from 1970 to 1990, but it is a small share of producers that either only produced two films in two years or produced multiple films in on year. Overall this a less productive graph as there is more variance in producers than production films. In the data there is roughly 80,000 producers and 60,000 production companies in the data.

Count of producers / production companies per film



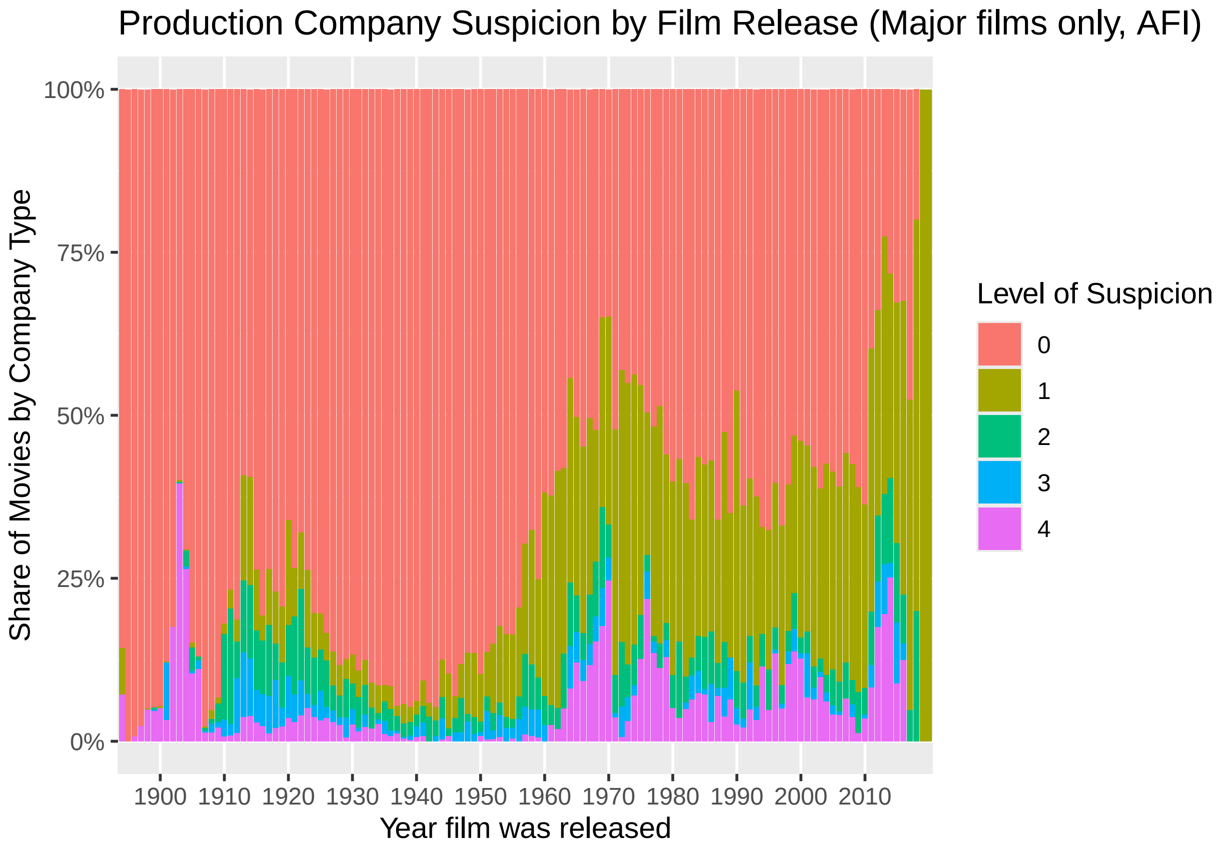
Generally there is a trend of increasing producers per film, after 1970 there is only a small proportion of films that only had a single producer. There is a harsh change in trend in around 2004 where over 25% of of films listed only had a single producer. Then the trend continues for the next 20 years. This might be a data anomaly as it is not consistent on who is credited as a producer. Comparing some of the films in the database, some of the producers are credited as assistant producers or executive producers on IMDb. It seems that it was a classification inconsistency that gives this harsh change in trend, which makes this graph less conclusive.



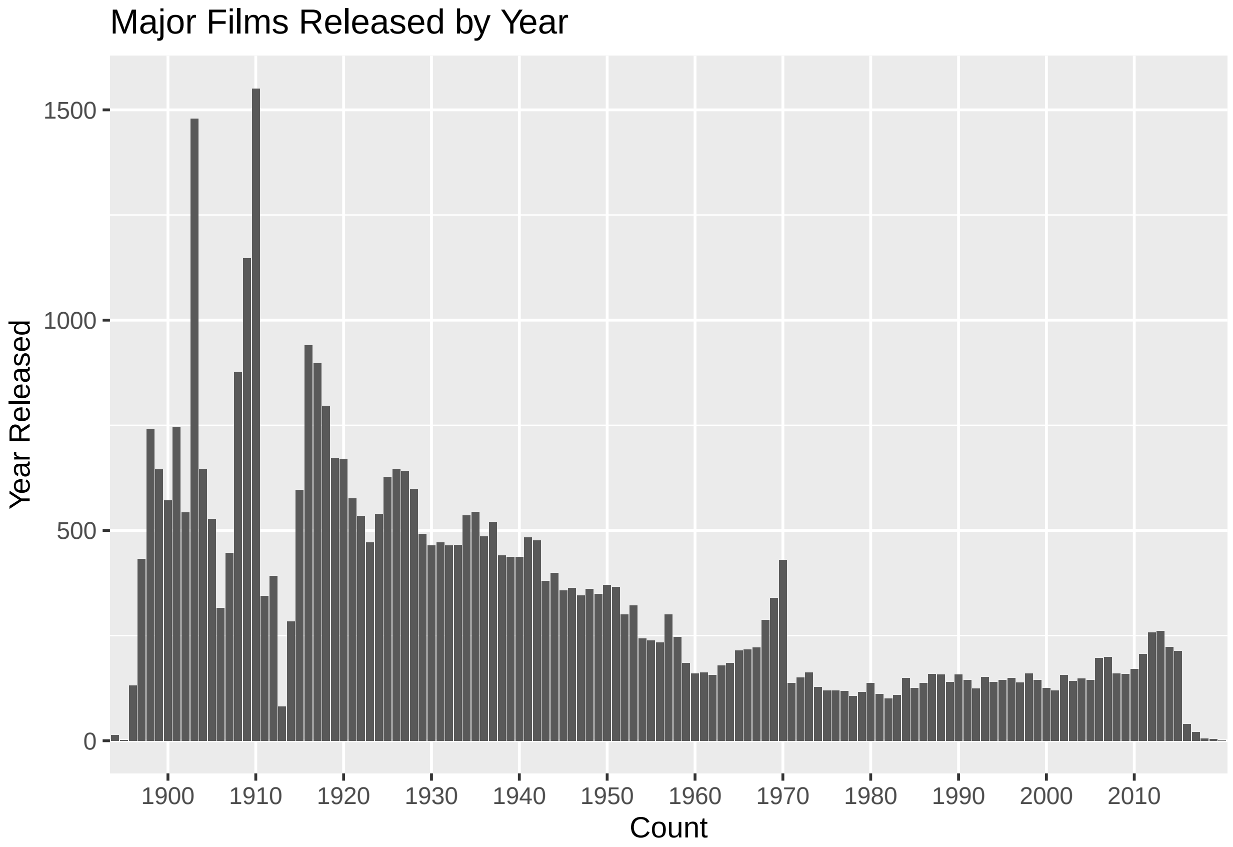
The graph for production companies follows a very similar pattern to the previous graph of credited producers. The database classifies all production companies as production companies, despite there being more than one type of production company. There are production companies as production companies, and there are production companies credited as in association with. The data does not carry this distinction which makes counting production companies less exact.

```
### narrowing down the data to be distinct companies ###
#adding NA's to empty rows
dist_company_data["distribution"][dist_company_data["distribution"] == ""] <- NA

dist_company_data = subset(dist_company_data, select = -c(Production_Companies, Date,X, Producers, GOID
#keeping it to US data
dist_company_data <- dist_company_data %>%
  filter(country == "United States")



dist_company_info = dist_company_data

dist_counts = dist_company_info %>%
  group_by(distribution) %>%
  summarise(films_distributed = n())



film_suspicion_producers <- merge(producer_data_AFI, producers_info_subset, by = "Producer", all.x = TRU
dist_company_info = merge(dist_company_info, dist_counts, by = "distribution", all.x = TRUE)

#making whether a film was distributed by a large distributor by whether it was larger than 50 or small
dist_company_info$major <- ifelse(dist_company_info$films_distributed
                                  %in% dist_company_info$films_distributed[dist_company_info$films_dist

#sub-setting the two columns we need
major_subset = dist_company_info[, c("Title", "major")]
#dropping NAs for now
major_subset <- na.omit(major_subset)

#merging with the whole data set of suspicion level
major_master_film_sus <- merge(master_film_suspicion, major_subset, by = "Title", all.x = TRUE)
#only keeping where major == true
major_master_film_sus <- major_master_film_sus[major_master_film_sus$major == 1, ]
#now this data set is only films that were distributed by distribution companies that have distributed
```

Major releases only



To consider a film to be major it must have been associated with at least one production company that has produced more than 50 films. The trend of increasing use of single use production still exists. This has to go with using the highest level of suspicion for production companies for films with multiple production companies. Most of the films are still produced by production companies that have produced more than one film, so they will be included in the data, then the highest level of suspicion will be used for the graphing.



Mechanically there will be less major films for films released recently, as it will take time a for a new production company to produce 50 films.

Conclusion

The collapsible corporation to take advantage of the tax break doesn't seem to show in the data, we should see a large share of high level of type three production companies, as there are costs to creating a new corporation. To use a collapsible corporation only once is an expensive way to take an advantage of the tax break. Either this tax break was not used as much as we thought or the data is not capturing the tax break. Unfortunately this data is the best we had access to and we can't make any further conclusions.A better constructed variable would be to incorporate film budgets. However this should not skew too much of the data as the target time frame is going to be the 80s, so there is a good amount of data before and after the target time frame. Given the limitations of data, we decided to move on from this project.