ResearchProject

Team members: Vladyslav Humennyy, Volodymyr Kuzma, Svitlana Hovorova.

The Films Research

We chose a dataset from TMDb, containing various information on films from 1880 to 2022 to explore more about the changes that happened to film industry through the years and analyse these **3 main factors:**

1) The average film duration: we suggest that as time goes films become shorter. H0: on average, the duration of films decreases by 0%; H1: the difference is present. We are also interested in testing this metric in different film categories;

2) The quantities of films of different categories produced: we want to analyze the popularity of genres in particular years;

3) The age of leading actors: we assume that in the more recent works, actors are younger. H0: on average, the age of leading actors decreases by 0 years every ten years; H1: the metric is greater;

```
# Needed library
library(BSDA)
```

```
## Loading required package: lattice

##
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':
##
##     Orange
```

```
# Reading the data from data set
df = read.csv("TMDb_Dataset.csv")
```

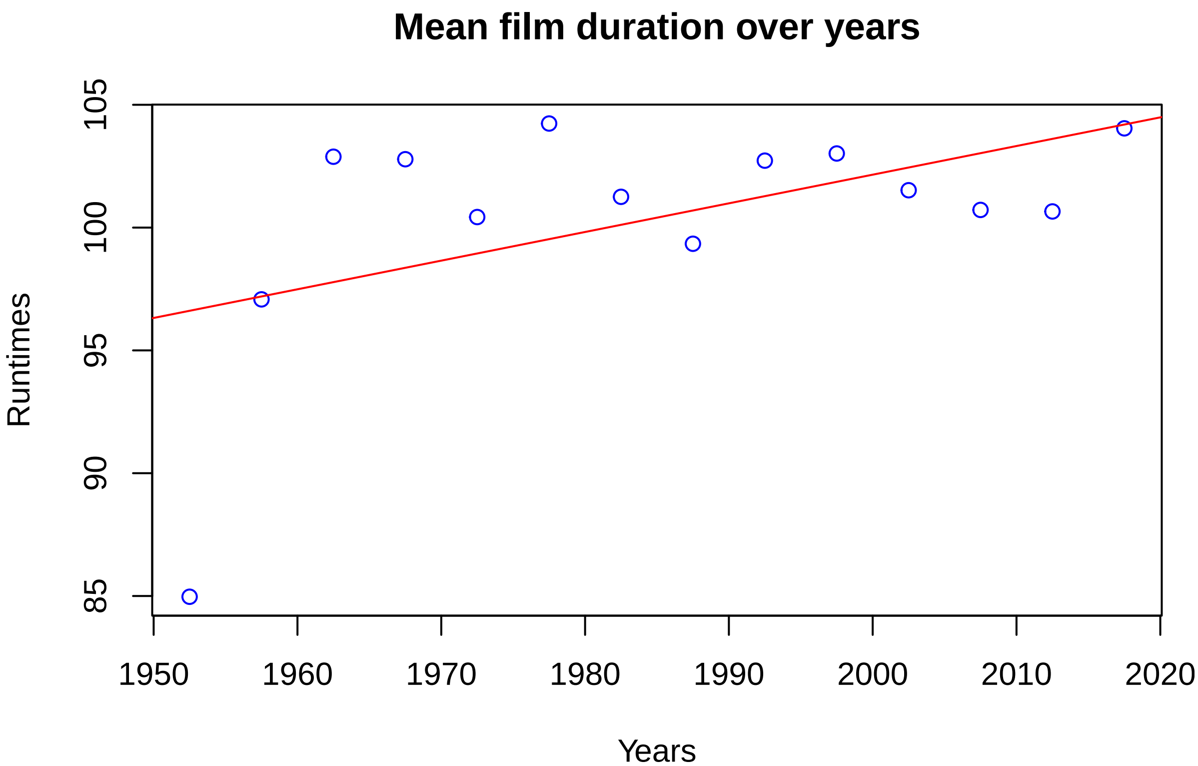
Hypothesis 1: Film Duration

In this part we will take into consideration only films released after 1950, as before that the industry was not really developed and the number of films is not significant.

Here, we test whether the film duration over time actual decreased. The first try was to build a linear regression model of mean film duration over time. **The reason for it is that the linear model can clearly show whether the relation is increasing or decreasing.** We took 5 years periods and built a relation of mean film duration in these periods to the year of film release (for a period of 5 years we took the middle of this period).

```
Years = seq(1950,2015,length = 14)

Runtimes = c()
for (year in Years) {
  runtime = mean(df[strtoi(substring(df$Date, 0, 4)) >= year &
                 strtoi(substring(df$Date, 0, 4)) < year+5, ]$Runtime)
  Runtimes = append(Runtimes, runtime)
}
```

```r
Years = Years + 2.5
regressional_model = lm(Runtimes~Years)
plot(Years, Runtimes, col = "blue", main = "Mean film duration over years")
abline(regressional_model, col = "red")
```

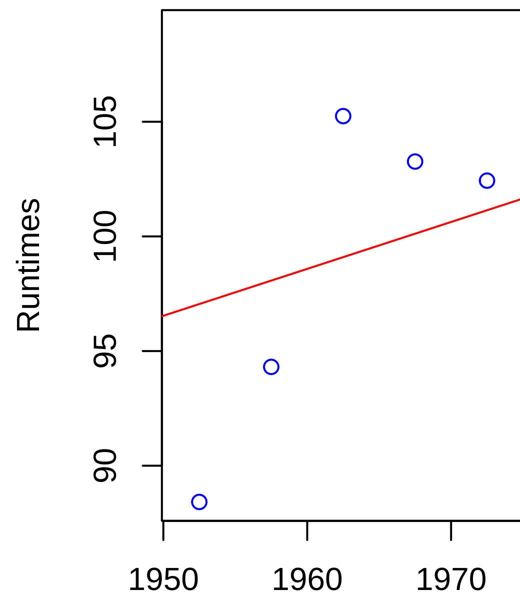



```r
summary(regressional_model)
```

```
##
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6420  -1.2468   0.5114   1.4745   5.1064
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -131.20310  114.49657  -1.146   0.2742
## Years          0.11668    0.05768   2.023   0.0659 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.35 on 12 degrees of freedom
## Multiple R-squared:  0.2543, Adjusted R-squared:  0.1922
## F-statistic: 4.092 on 1 and 12 DF,  p-value: 0.06594
```

The p-value in the test of slope coefficient is hardly bigger than 0.05. The model here states that our starting assumption is wrong: movies are getting longer. **Nevertheless, the main problem is that real relation is far from linear, so we can not make any conclusions here**.

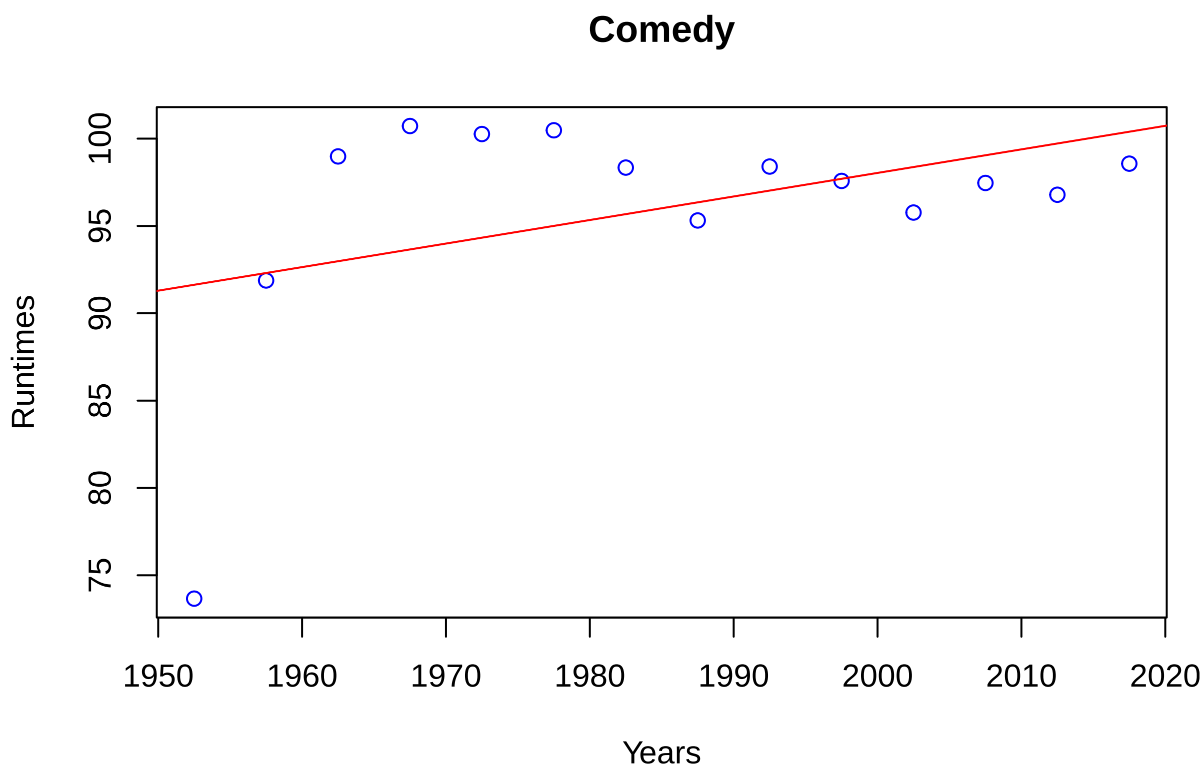
```
Categories = c("Crime", "Comedy", "Action", "Thriller", "Adventure", "Science Fiction", "Drama", "Romanc

for (category in Categories) {
  Years = seq(1950,2015,length = 14)
  temp_df = df[grepl(category, df$Categories), ]

  Runtimes = c()
  for (year in Years) {
    runtime = mean(temp_df[strtoi(substring(temp_df$Date, 0, 4)) >= year &
                   strtoi(substring(temp_df$Date, 0, 4)) < year+5, ]$Runtime)
    Runtimes = append(Runtimes, runtime)
}

  Years = Years + 2.5
  regressional_model = lm(Runtimes~Years)
  plot(Years, Runtimes, col = "blue", main = category)
  abline(regressional_model, col = "red")


  cat("Summary for", category)
  print(summary(regressional_model))
}
```



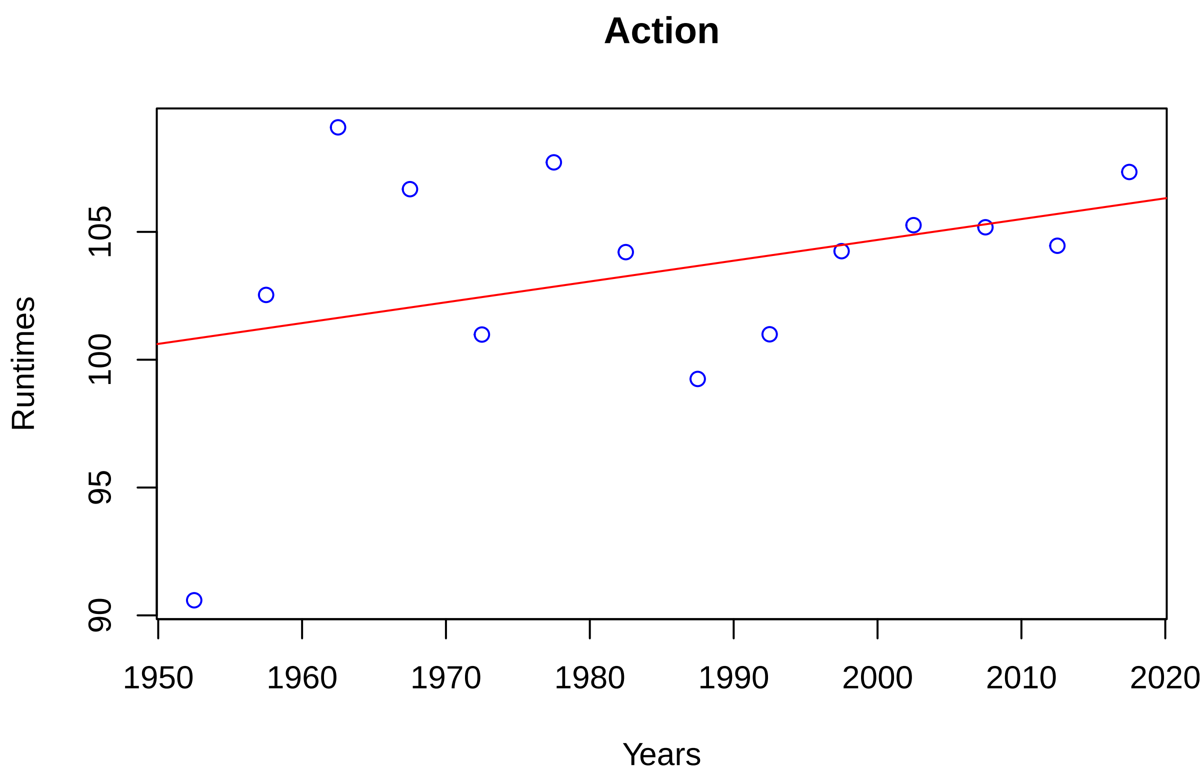
Here you can also see such linear models for different categories of films

```
## Summary for Crime
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

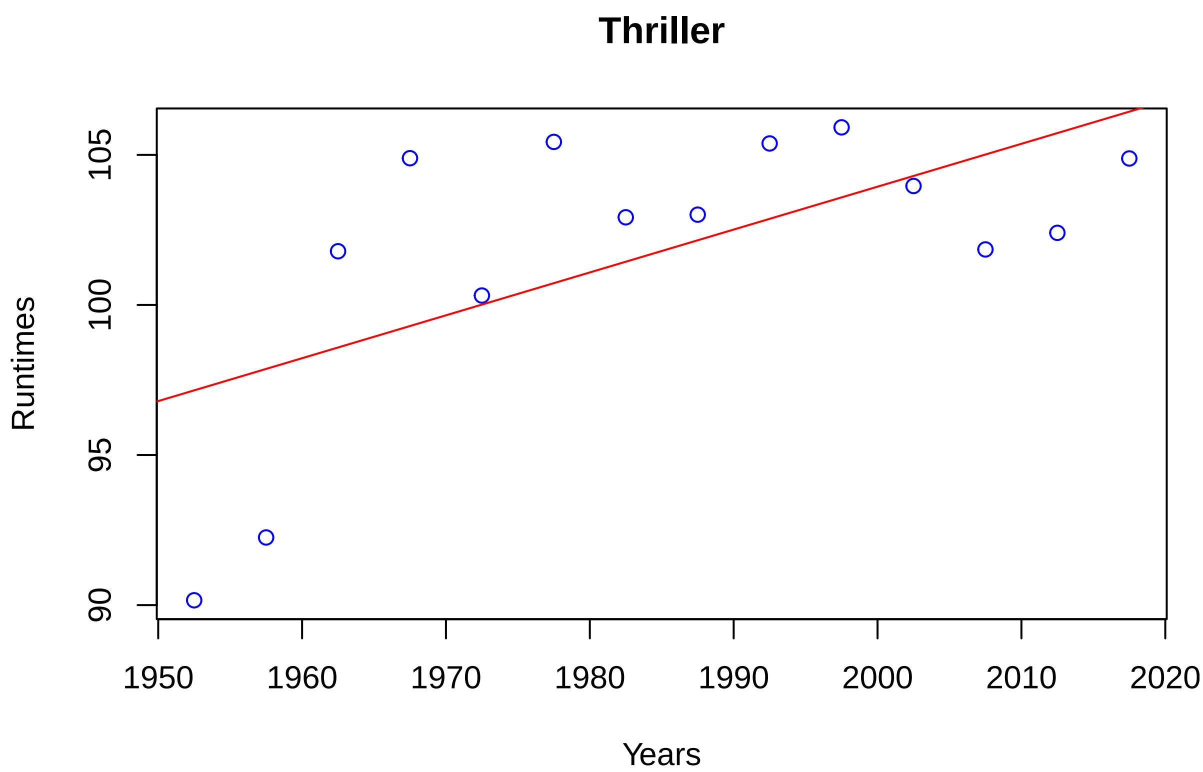
```
## -8.6304 -1.9273 -0.0387  1.8589  6.1515
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -302.26421  104.63720  -2.889  0.01361 *
## Years          0.20451    0.05271   3.880  0.00219 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.975 on 12 degrees of freedom
## Multiple R-squared:  0.5564, Adjusted R-squared:  0.5195
## F-statistic: 15.05 on 1 and 12 DF,  p-value: 0.002188
```



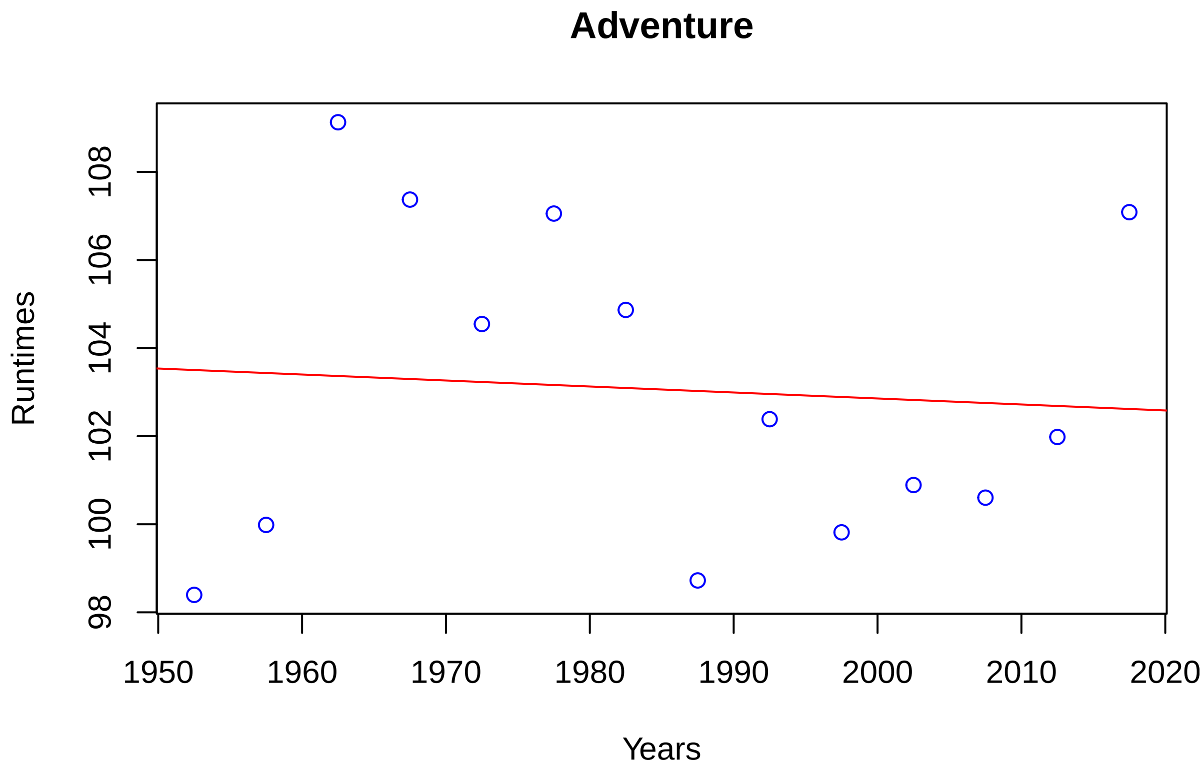

```
## Summary for Comedy
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.9688  -1.7683  -0.2746   4.7716   7.0660
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -171.59928  170.84118  -1.004    0.335
## Years          0.13482    0.08606   1.567    0.143
##
## Residual standard error: 6.49 on 12 degrees of freedom
## Multiple R-squared:  0.1698, Adjusted R-squared:  0.1006
## F-statistic: 2.454 on 1 and 12 DF,  p-value: 0.1432
```



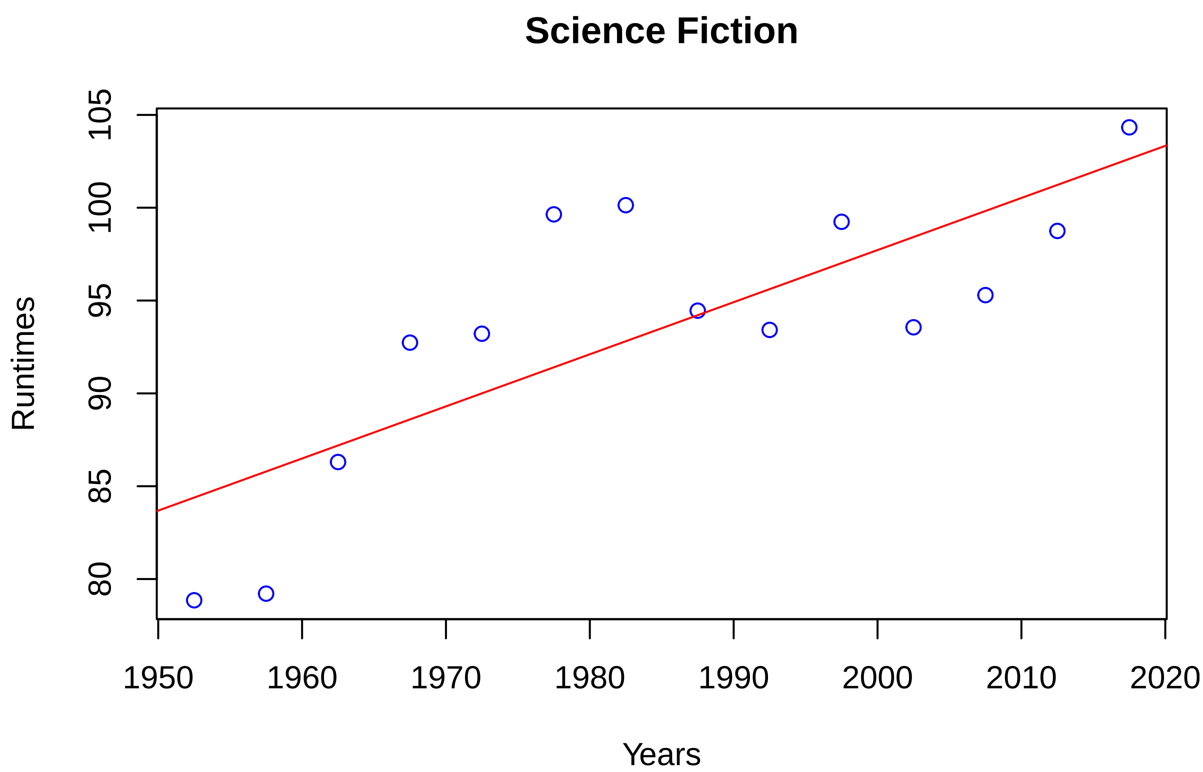

```
## Summary for Action
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2319  -1.4101   0.1286   1.2868   7.4530
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -58.02098  118.30314  -0.490    0.633
## Years         0.08135    0.05960   1.365    0.197
##
## Residual standard error: 4.494 on 12 degrees of freedom
## Multiple R-squared:  0.1344, Adjusted R-squared:  0.06229
## F-statistic: 1.864 on 1 and 12 DF,  p-value: 0.1973
```



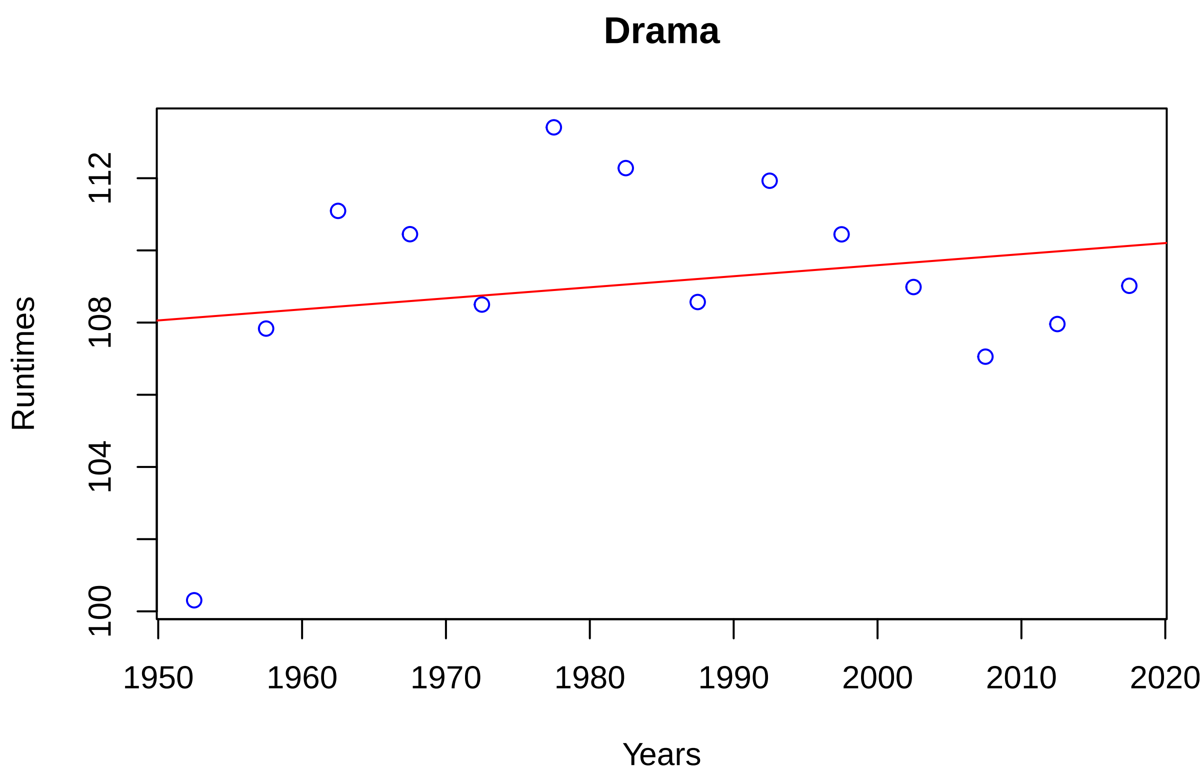

```
## Summary for Thriller
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9933 -2.7611  0.5779  2.4676  5.5926
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -181.79264  102.65358  -1.771   0.1019
## Years          0.14287    0.05171   2.763   0.0172 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.9 on 12 degrees of freedom
## Multiple R-squared:  0.3888, Adjusted R-squared:  0.3378
## F-statistic: 7.633 on 1 and 12 DF,  p-value: 0.01719
```



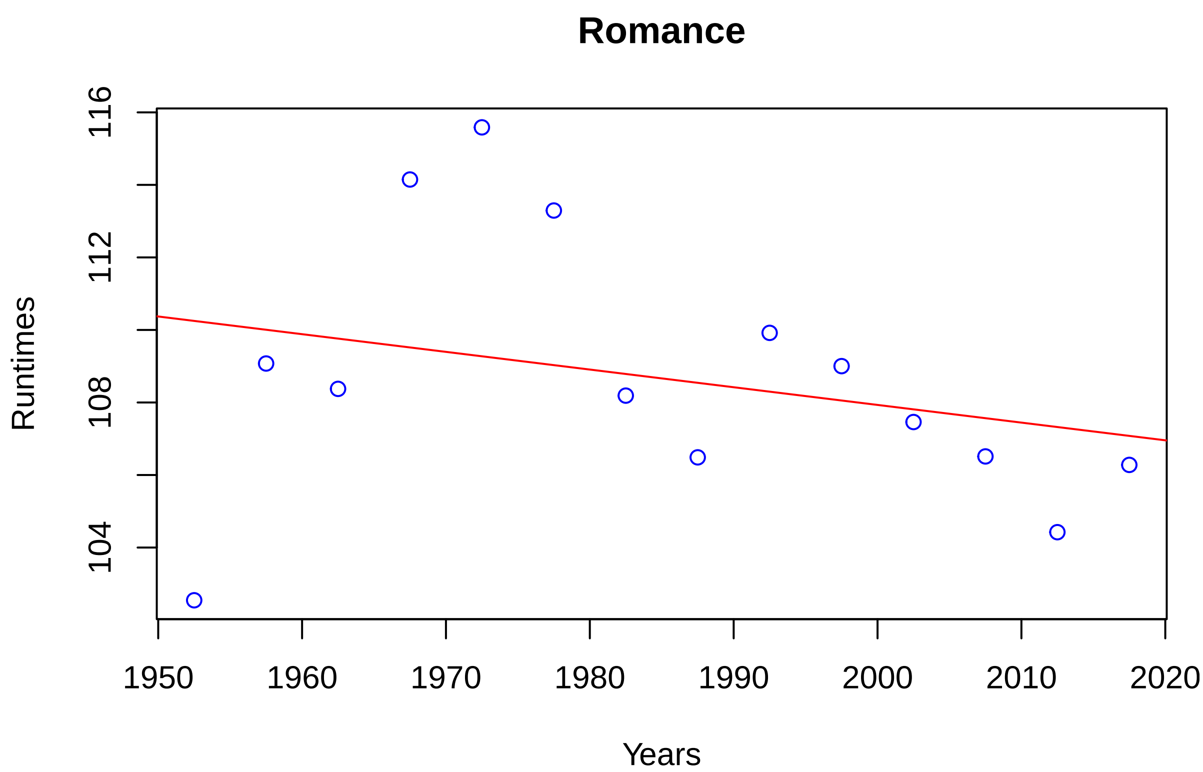

```
## Summary for Adventure
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1057 -2.8423 -0.6381  3.3627  5.7620
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 130.06370   97.57231   1.333    0.207
## Years        -0.01360    0.04915  -0.277    0.787
##
## Residual standard error: 3.707 on 12 degrees of freedom
## Multiple R-squared:  0.006343,   Adjusted R-squared:  -0.07646
## F-statistic: 0.0766 on 1 and 12 DF,  p-value: 0.7867
```




```
## Summary for Science Fiction
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5789 -4.0188 -0.3243  2.9638  8.2370
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -463.62097  128.65514  -3.604 0.003622 **
## Years          0.28067    0.06481   4.331 0.000978 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.888 on 12 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.5773
## F-statistic: 18.75 on 1 and 12 DF,  p-value: 0.0009776
```




```
## Summary for Drama
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8299 -1.0006 -0.3526  2.3903  4.5074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.38678   84.62223   0.572    0.578
## Years        0.03060    0.04263   0.718    0.487
##
## Residual standard error: 3.215 on 12 degrees of freedom
## Multiple R-squared:  0.04118,    Adjusted R-squared:  -0.03873
## F-statistic: 0.5153 on 1 and 12 DF,  p-value: 0.4866
```




```
## Summary for Romance
## Call:
## lm(formula = Runtimes ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7029 -1.3032 -0.6987  1.4528  6.3126
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 205.48167   96.10771   2.138   0.0538 .
## Years        -0.04877    0.04841  -1.007   0.3336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.651 on 12 degrees of freedom
## Multiple R-squared:  0.07798,    Adjusted R-squared:  0.001148
## F-statistic: 1.015 on 1 and 12 DF,  p-value: 0.3336
```

The directions of lines for different categories vary significantly, though for some of them the p-value of the testing of slope coefficient is to low to build conclusions based on this estimate.

Though in most cases the p-values for the slope coef. are high, the shown linear graphs **do not realistically represent the relation**, so we decided to **do another test** of the hypothesis and use another simpler yet more accurate method: **divide the films into two categories: those released before 2001 (in the 20th century) and those released after (in the 21st century)**, and then use **t-test** (as we do not have information on the variance) to test $H_0 : \mu_{before} = \mu_{after}$ against $H_1 : \mu_{before} < \mu_{after}$.

```
duration_21 = df[strtoi(substring(df$Date, 0, 4)) >= 2001,]$Runtime
duration_20 = df[strtoi(substring(df$Date, 0, 4)) < 2001 & strtoi(substring(df$Date, 0, 4)) >= 1950,]$Ru
```

```r
t.test(duration_20, duration_21, alternative = "g")
```

```
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -1.9722, df = 21725, p-value = 0.9757
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.249916       Inf
## sample estimates:
## mean of x mean of y
##  100.9782  101.6597
```

```r
t.test(duration_20, duration_21, alternative = "l")
```

```
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -1.9722, df = 21725, p-value = 0.0243
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1130967
## sample estimates:
## mean of x mean of y
##  100.9782  101.6597
```

The results are completely opposite to what we expected: the mean duration of movie is significantly bigger in the new century.

```r
for (category in Categories) {
  temp_df = df[grepl(category, df$Categories), ]

  duration_21 = df[strtoi(substring(temp_df$Date, 0, 4)) >= 2001,]$Runtime
  duration_20 = df[strtoi(substring(temp_df$Date, 0, 4)) < 2001 & strtoi(substring(temp_df$Date, 0, 4))

  cat('Tests for', category)
  print(t.test(duration_20, duration_21, alternative = "g"))
  print(t.test(duration_20, duration_21, alternative = "l"))
}
```

Here you can also see test results for different categories of films

```
## Tests for Crime
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -2.0551, df = 22895, p-value = 0.9801
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.381386       Inf
## sample estimates:
## mean of x mean of y
##   99.2909  100.0582
```

```
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -2.0551, df = 22895, p-value = 0.01994
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1531256
## sample estimates:
## mean of x mean of y
##   99.2909  100.0582
##
## Tests for Comedy
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -0.88621, df = 23267, p-value = 0.8122
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.9392339        Inf
## sample estimates:
## mean of x mean of y
##  99.48734  99.81619
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -0.88621, df = 23267, p-value = 0.1878
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.2815377
## sample estimates:
## mean of x mean of y
##  99.48734  99.81619
##
## Tests for Action
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -1.6918, df = 23442, p-value = 0.9547
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.214705       Inf
## sample estimates:
## mean of x mean of y
##  99.30949  99.92538
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -1.6918, df = 23442, p-value = 0.04534
```

```
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01708202
## sample estimates:
## mean of x mean of y
##  99.30949  99.92538
##
## Tests for Thriller
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -0.45012, df = 18826, p-value = 0.6737
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.800126       Inf
## sample estimates:
## mean of x mean of y
##  99.45429  99.62620
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -0.45012, df = 18826, p-value = 0.3263
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4563147
## sample estimates:
## mean of x mean of y
##  99.45429  99.62620
##
## Tests for Adventure
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = 3.6789, df = 23856, p-value = 0.0001174
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7426223       Inf
## sample estimates:
## mean of x mean of y
## 100.23325  98.89006
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = 3.6789, df = 23856, p-value = 0.9999
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.943763
## sample estimates:
## mean of x mean of y
## 100.23325  98.89006
```

```
##
## Tests for Science Fiction
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = 0.78019, df = 24516, p-value = 0.2176
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3134872        Inf
## sample estimates:
## mean of x mean of y
##  99.75044  99.46760
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = 0.78019, df = 24516, p-value = 0.7824
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.8791715
## sample estimates:
## mean of x mean of y
##  99.75044  99.46760
##
## Tests for Drama
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -2.1778, df = 22979, p-value = 0.9853
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.424232       Inf
## sample estimates:
## mean of x mean of y
##  99.19103 100.00241
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -2.1778, df = 22979, p-value = 0.01472
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1985217
## sample estimates:
## mean of x mean of y
##  99.19103 100.00241
##
## Tests for Romance
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -0.77365, df = 22239, p-value = 0.7804
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.9118575        Inf
## sample estimates:
## mean of x mean of y
##  99.34345  99.63513
##
##
##  Welch Two Sample t-test
##
## data:  duration_20 and duration_21
## t = -0.77365, df = 22239, p-value = 0.2196
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.3284913
## sample estimates:
## mean of x mean of y
##  99.34345  99.63513
```

As we see, the only category in which the duration decreased as we expected is **Adventure films** as the p-value of the test, where the H_1 is that the duration decreased, is very low, meaning that we should reject H_0.

Overall, in this part we took a chance to use linear regression and t-test to find out that the duration of films actually did not decrease from the previous century's last decades.
Now, let's move to the second hypothesis

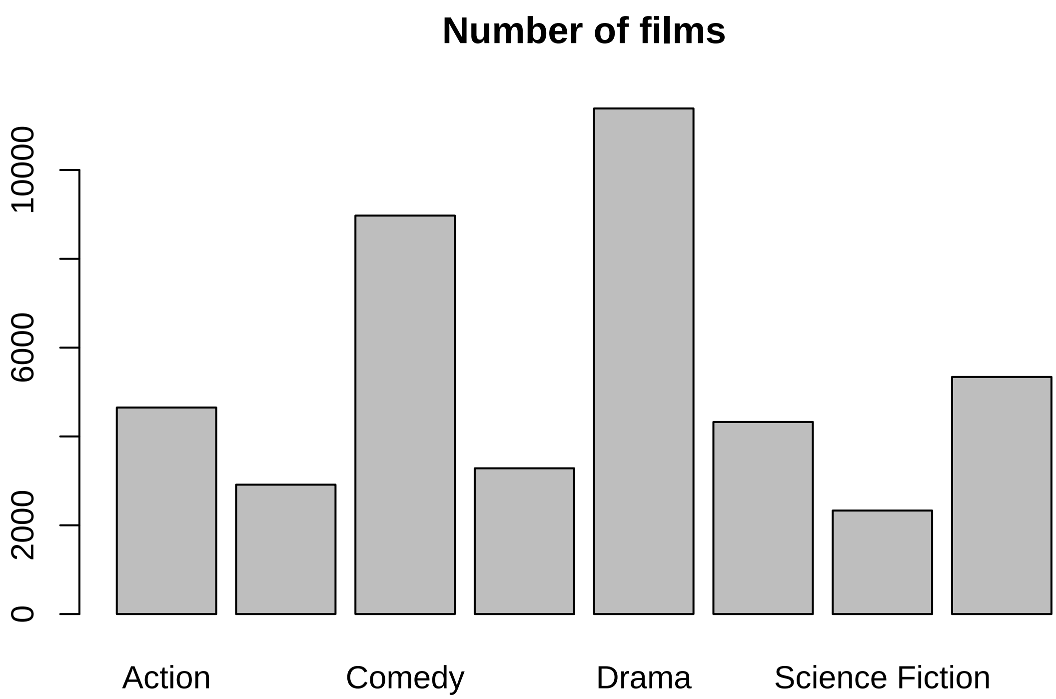
Part 2: here we want to analyze different film categories and see how their popularities are distributed

Firstly, that's the quantities of all films from dataframe corresponding to the genre:

```
Categories = c("Crime", "Comedy", "Action", "Thriller", "Adventure", "Science Fiction", "Drama", "Romanc
categories_vec <- c()
for (category in Categories){
  for (row in df$Categories){
    if (grepl(category, row)){
      categories_vec = append(categories_vec, category)
    }
  }
}
categories_counts <- as.data.frame(table(categories_vec))
categories_counts
```

```
##    categories_vec  Freq
## 1          Action  4651
## 2       Adventure  2914
## 3          Comedy  8974
## 4           Crime  3283
## 5           Drama 11387
## 6         Romance  4327
## 7 Science Fiction  2332
## 8        Thriller  5341
```

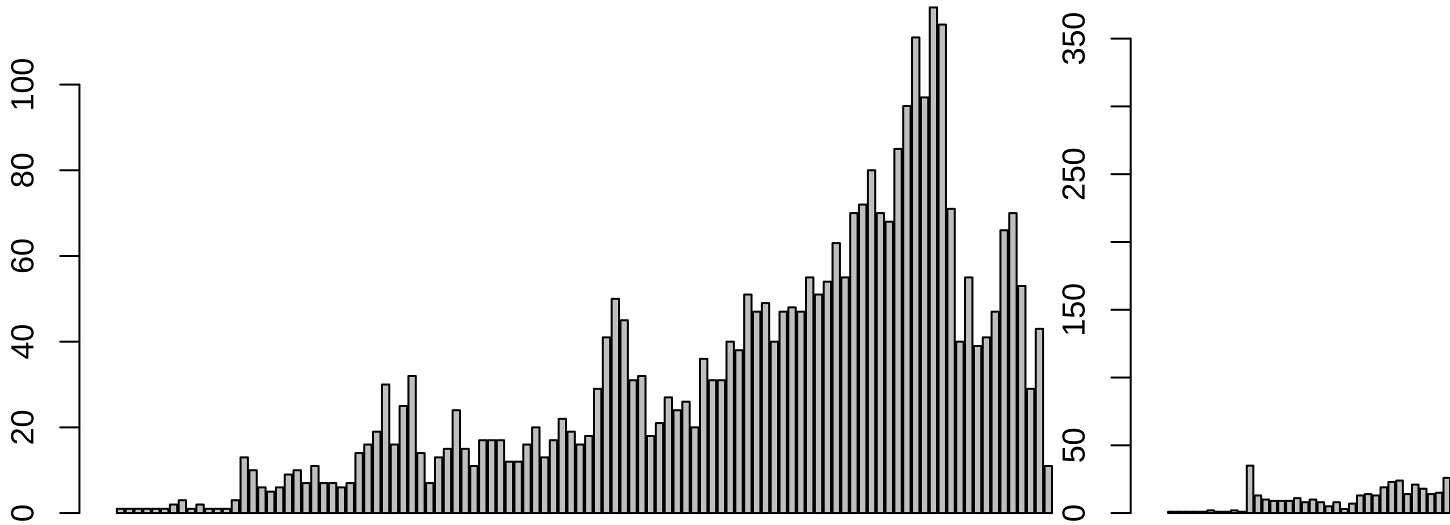
```
barplot(categories_counts$Freq, main = "Number of films", names = categories_counts$categories_vec)
```



Let's now see how popularit of different genres changed over years (xlab depicts the year when there was the biggest number of films of such category produced)

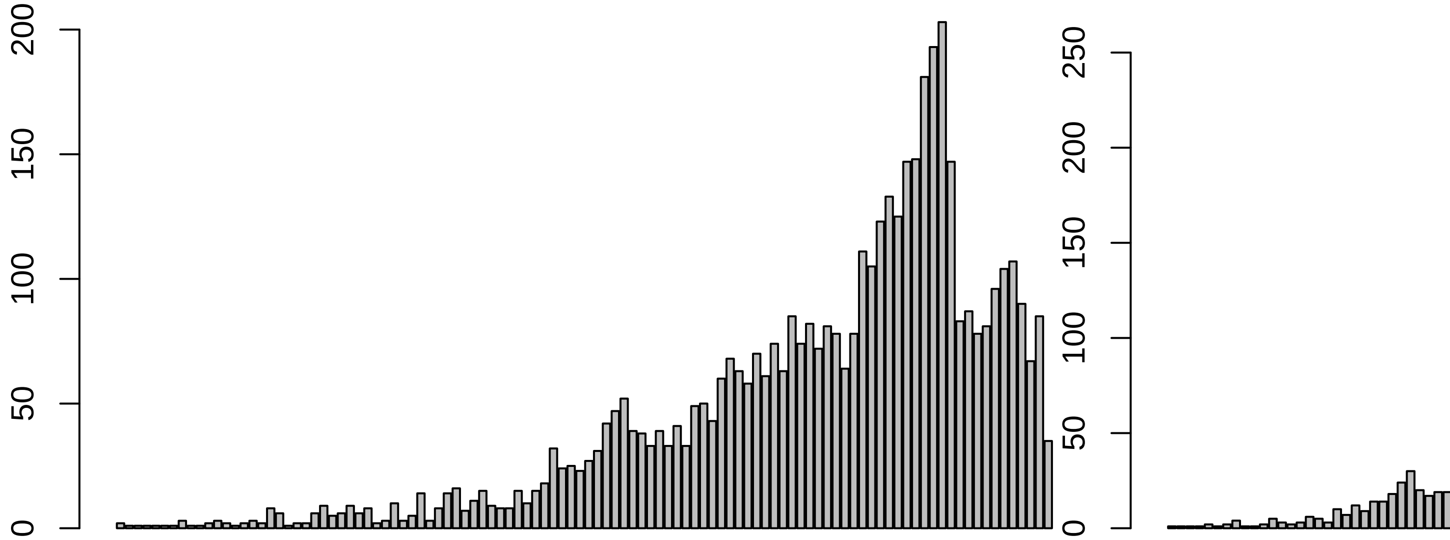
```
years <- strtrim(df$Date, 4)
years <- sort(union(years, years))
popularity_df <- data.frame(years)
for (category in Categories){
  popularity_df[category] <- 0
}
for (category in Categories){
  curr_df <- df[grepl(category, df$Categories),]
  years <- strtrim(curr_df$Date, 4)
  curr_counts <- as.data.frame(table(years))
  most_prod_y <- curr_counts[curr_counts$Freq == max(curr_counts$Freq),]$years
  barplot(curr_counts$Freq,main=category, xlab = most_prod_y)
}
```

Crime

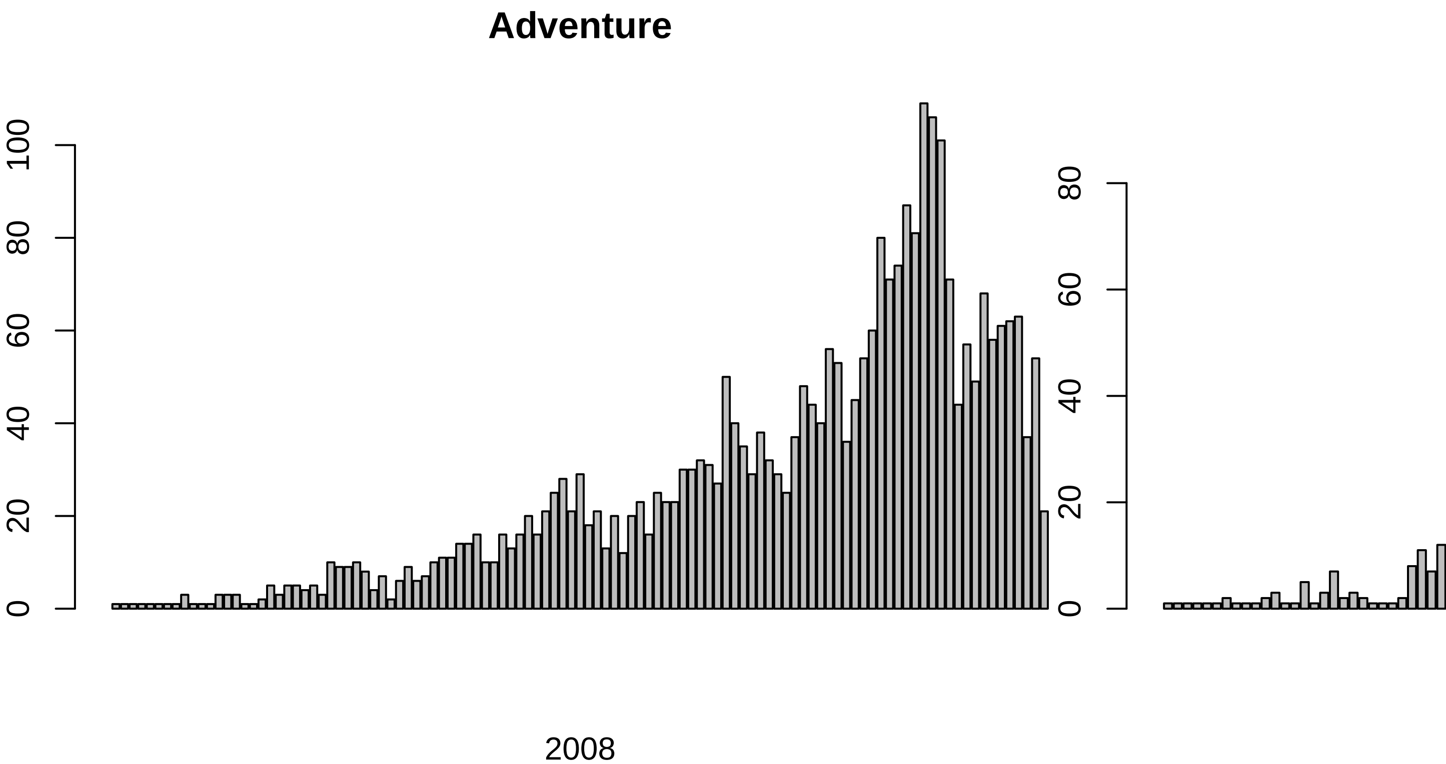


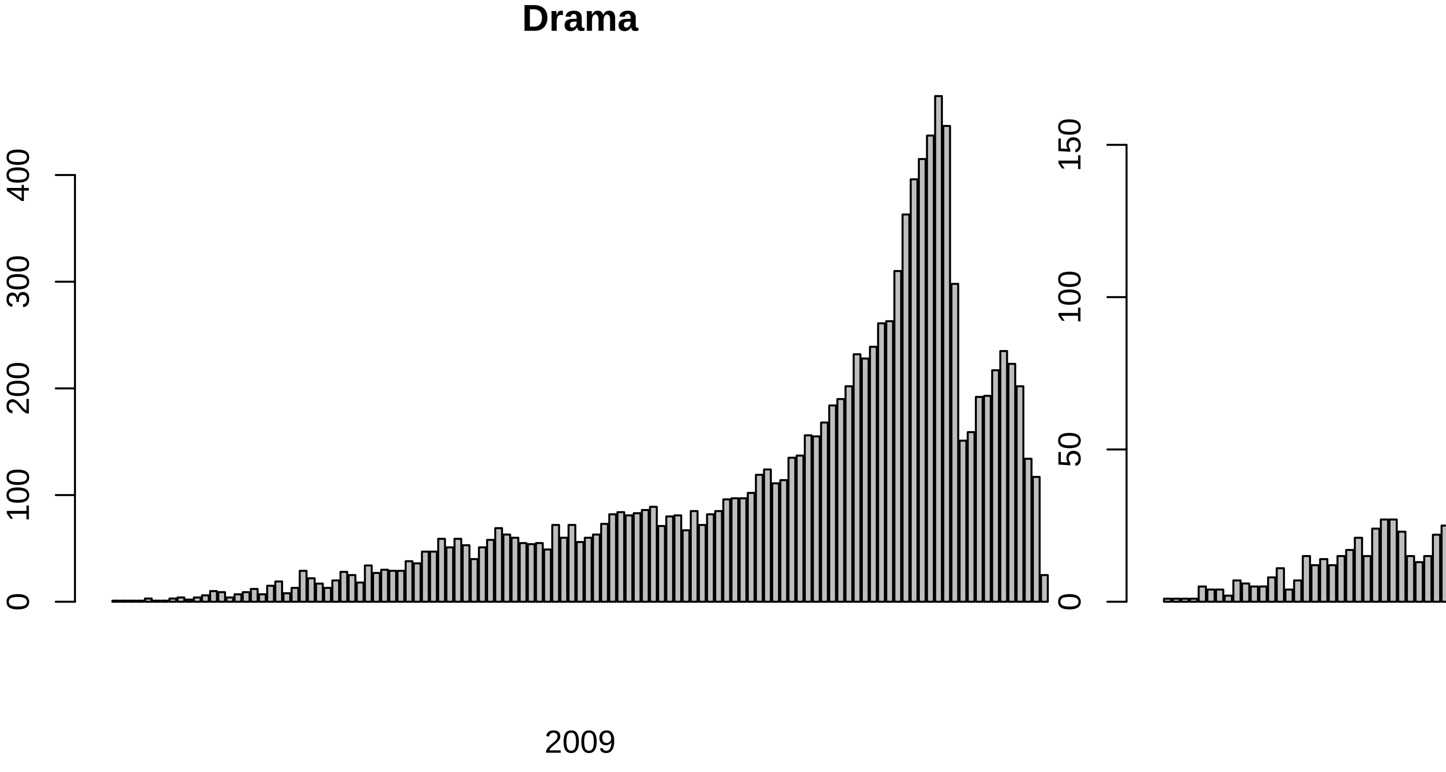
2009

Action



2010





As plots show, there just was growth in number of films present, and it also is not linear. So, roughly speaking, there is no particular trend in in genres preferences of producers changes: the years with the biggest number of films is 2009 +- 1 year for all the categories.

Hypothesis 3: Age of Leading Actors

In that hypothesis we test the relation between the year of production and the age of leading actor. In our hypothesis we expect, that in more recent films actors were more mature than in more recent ones. For that we would combine the data from .csv file, that was used in previous tests and from file Birth_Actors.csv, which has the data about dates of birth of actors. At first we import that file and store its data in dataframe.

```
# Read the data from data set about year of birth of actors
dob_df = read.csv("Birth_Actors.csv")
```

Here we create the test that would test our ages of actors over years. But besides general age of all leading actors, we would divide male and female actors to plot them at the same graph and see how their mean ages differ. After plotting the information we show the summary of linear model for our results.

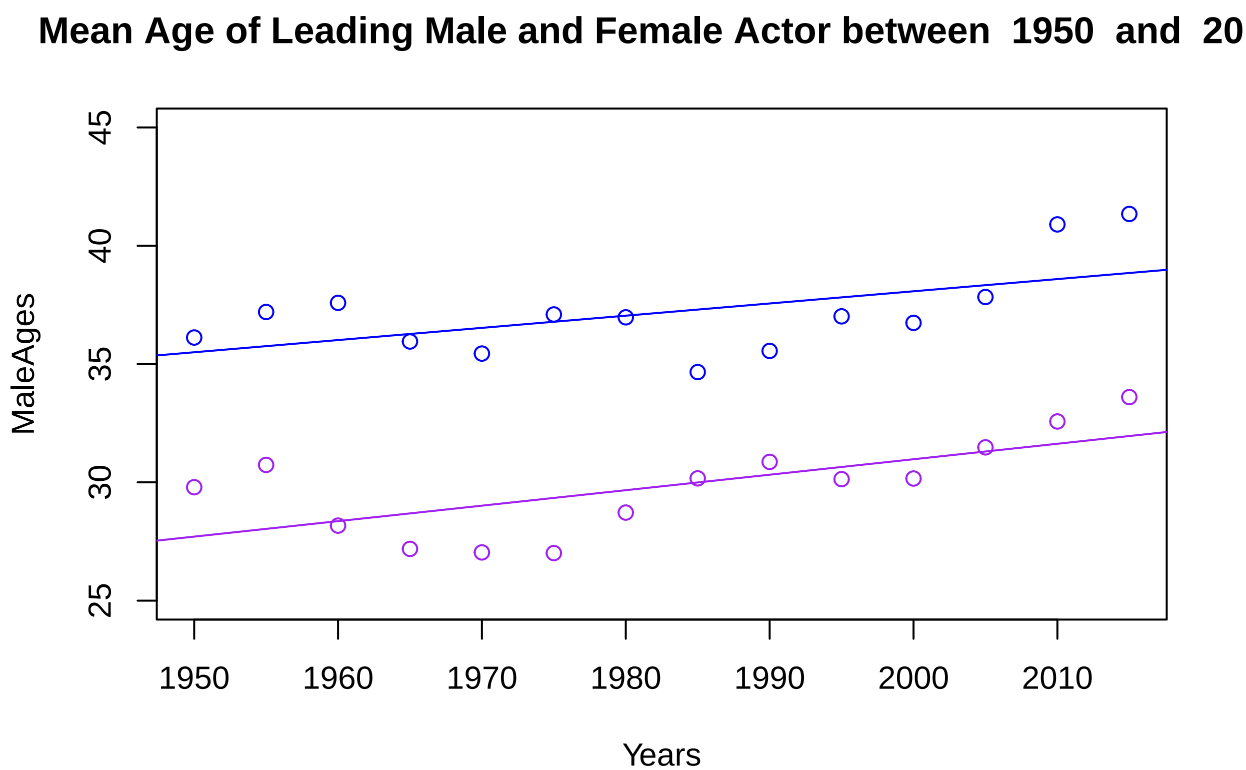
```
age_of_actors_function <- function(start_of_period, end_of_period, step){
  number_of_period = ((end_of_period - start_of_period) / step) + 1

  Years = seq(start_of_period,end_of_period,length = number_of_period)

  print(Years)

  Ages = c()
  MaleAges = c()
  FemaleAges = c()
  for (year in Years) {
    year_ages = c()
    cast_of_films = df[strtoi(substring(df$Date, 0, 4)) >= year &
                     strtoi(substring(df$Date, 0, 4)) < year+step, ]$Cast

    number_of_actors = 0
    age_of_actors = 0
    number_of_male_actors = 0
    age_of_male_actors = 0
    number_of_female_actors = 0
    age_of_female_actors = 0

    for (cast in cast_of_films){
      start_index = unlist(gregexpr("id': ", cast))[1]
      id_with_junk = substr(cast, start_index + 5, start_index + 12)
      id = strtoi(substr(id_with_junk, 1, unlist(gregexpr(",", id_with_junk))[1] - 1))
      gender_start_index = unlist(gregexpr("gender': ", cast))[1]
      gender = strtoi(substr(cast, gender_start_index + 9, gender_start_index + 9))
      row = which(grepl(id, dob_df$Id))[1]
      if (!is.na(row)){
        if (gender == 2){
            number_of_male_actors = number_of_male_actors + 1
            age_of_male_actors = age_of_male_actors + (year - strtoi(dob_df$Birth[row]))
        }
        else{
            number_of_female_actors = number_of_female_actors + 1
            age_of_female_actors = age_of_female_actors + (year - strtoi(dob_df$Birth[row]))
        }
        number_of_actors = number_of_actors + 1
        age_of_actors = age_of_actors + (year - strtoi(dob_df$Birth[row]))
      }
    }
```

```
    Ages = append(Ages, age_of_actors / number_of_actors)
    MaleAges = append(MaleAges, age_of_male_actors / number_of_male_actors)
    FemaleAges = append(FemaleAges, age_of_female_actors / number_of_female_actors)
    #cat("=============================================================================
    #cat("For year ", year, " the average actor age is ", age_of_actors / number_of_actors, "\n")
    #cat("For year ", year, " the average male actor age is ", age_of_male_actors / number_of_male_acto
    #cat("For year ", year, " the average female actor age is ", age_of_female_actors / number_of_femal
  }


  plot(Years, MaleAges, col = "blue", main = paste("Mean Age of Leading Male and Female Actor between "
  points(Years, FemaleAges, col = "purple")

  regressional_model_male = lm(MaleAges~Years)
  regressional_model_female = lm(FemaleAges~Years)
  abline(regressional_model_male, col = "blue")
  abline(regressional_model_female, col = "purple")
  print(summary(regressional_model_male))
  print(summary(regressional_model_female))

  plot(Years, Ages, col="blue", main = paste("Mean Age of Leading Actor between ", toString(start_of_pe

  regressional_model = lm(Ages~Years)
  abline(regressional_model, col = "red")
  print(summary(regressional_model))
}
```

Here, we test whether the age of leading actor over time actualy decreased. We build the linear model for films from 1950 to 2015 with a period of 5 years. The reason to limit our data set to such year is because of development of industry, which was high enough only in 50's. We took 5 years periods and built a relation of mean age of actor in these periods to the year of film release (for a period of 5 years we took the middle of this period).

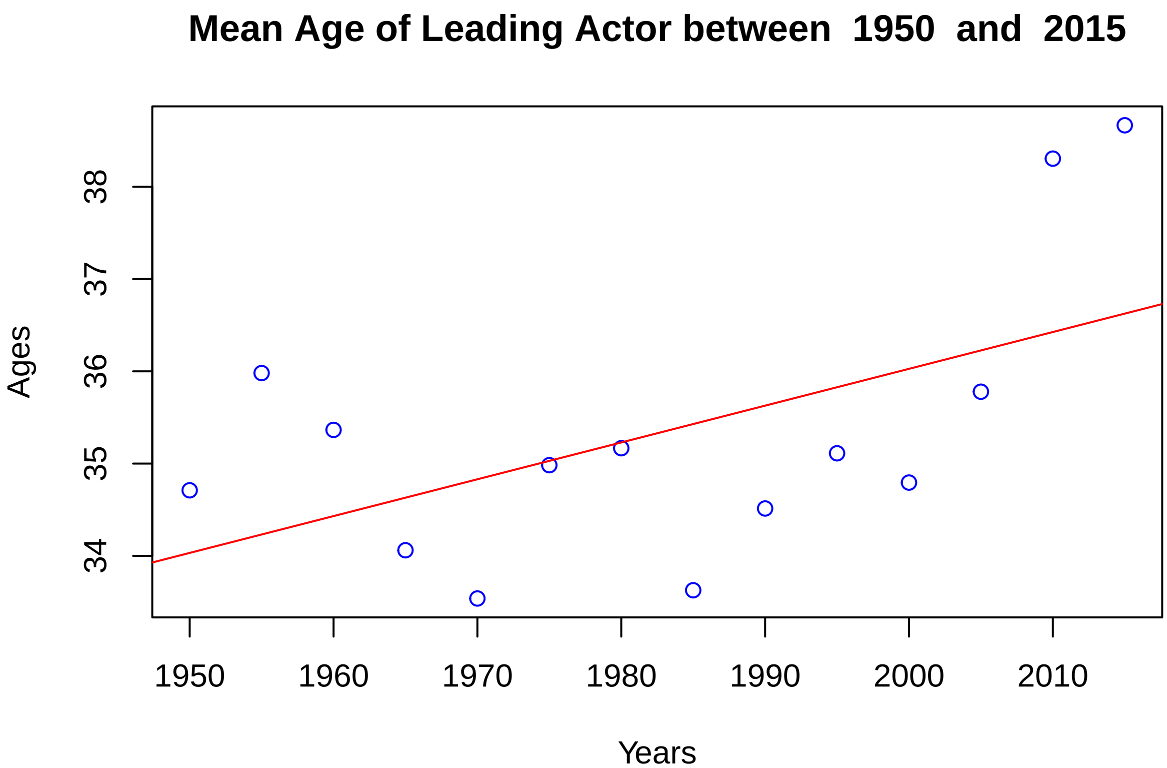
```
age_of_actors_function(1950, 2015, 5)
```

```
##  [1] 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015
```




```
##
## Call:
## lm(formula = MaleAges ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6414 -1.0152 -0.1933  1.2397  2.4947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -65.00772   42.63906  -1.525   0.1533
## Years         0.05154    0.02151   2.397   0.0337 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.622 on 12 degrees of freedom
## Multiple R-squared:  0.3237, Adjusted R-squared:  0.2673
## F-statistic: 5.743 on 1 and 12 DF,  p-value: 0.03373
##
##
## Call:
## lm(formula = FemaleAges ~ Years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33028 -0.91327 -0.00783  0.84221  2.70294
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -99.81442   40.90683  -2.440  0.03116 *
```

```
## Years          0.06540    0.02063   3.169  0.00808 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.556 on 12 degrees of freedom
## Multiple R-squared:  0.4557, Adjusted R-squared:  0.4103
## F-statistic: 10.05 on 1 and 12 DF,  p-value: 0.008078
```




```
##
## Call:
## lm(formula = Ages ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8014 -1.0150 -0.2549  0.8703  2.0411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43.77747   34.62155  -1.264   0.2301
## Years         0.03990    0.01746   2.285   0.0413 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.317 on 12 degrees of freedom
## Multiple R-squared:  0.3032, Adjusted R-squared:  0.2451
## F-statistic: 5.221 on 1 and 12 DF,  p-value: 0.0413
```

From the data that we can see at our plots and summaries of linear models we can assume, that there is linear dependency between the year of production and the age of leading actor. But, surprisingly, it is the opposite to the one, that was predicted before the experiment. We can see, that the relation between the year of production and the age is straight, in contrast to our hypothesis. Also, we have p-value less than 0.05

for linear model, which is very good statistics and we should stick to it.

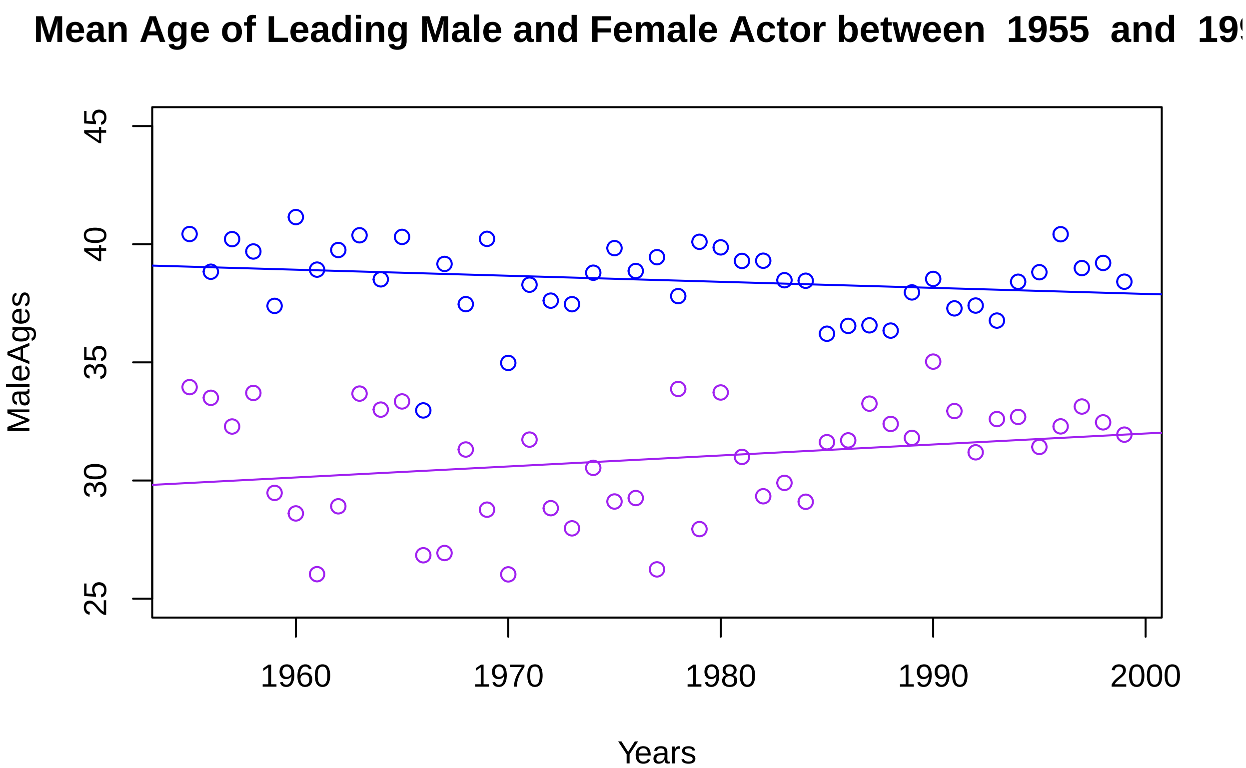
Besides, the plot of male and female ages shows, that they behave very similar to general trend. But there is one interesting, though predictable, detail - difference between mean ages for same periods is **more than 5 years** in average. The reason for it – **SOCIETY!**

Now we would test the 20th (1950 to 1999) and 21st (2000 to 2015) centuries of film industries and their year to age relation.

The first test would be for 20th century.

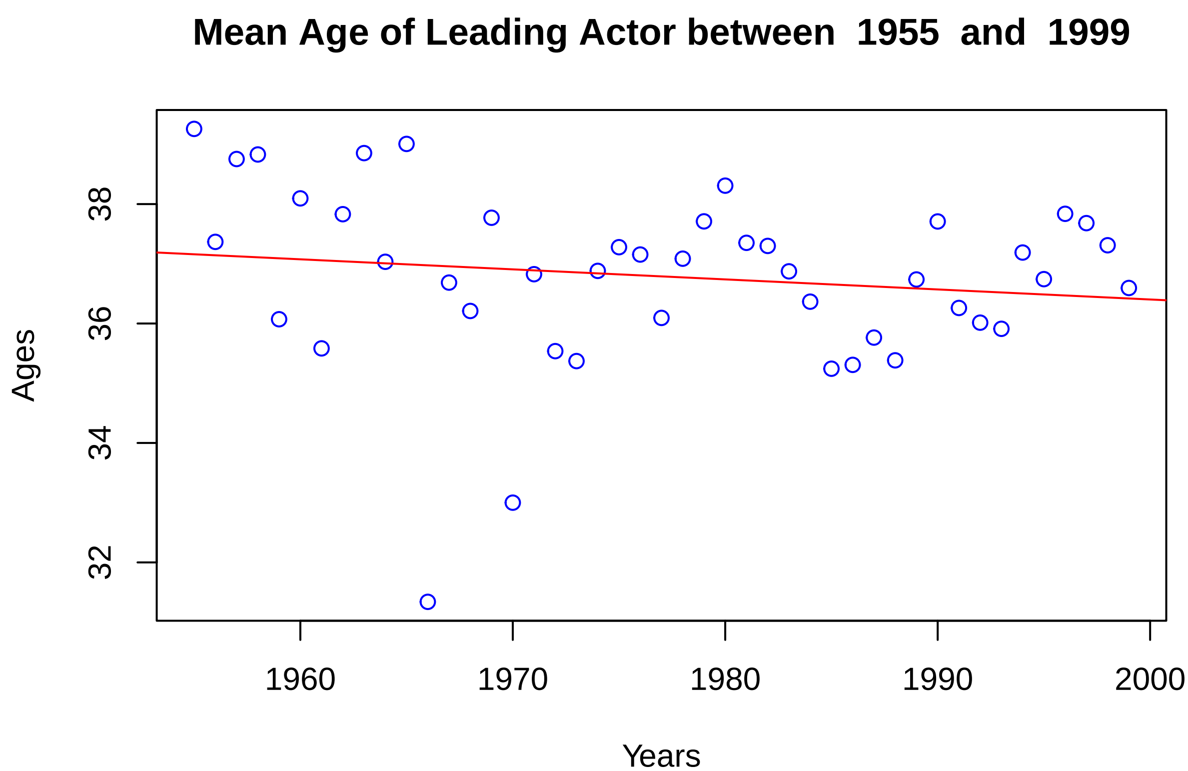
```
age_of_actors_function(1955, 1999, 1)
```

```
##  [1] 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969
## [16] 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984
## [31] 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
```




```
##
## Call:
## lm(formula = MaleAges ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8017 -0.8426  0.3552  1.0158  2.4228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89.13300   35.72089   2.495   0.0165 *
## Years       -0.02562    0.01807  -1.418   0.1634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.574 on 43 degrees of freedom
## Multiple R-squared:  0.04466,	Adjusted R-squared:  0.02245
## F-statistic:  2.01 on 1 and 43 DF,  p-value: 0.1634
##
##
## Call:
## lm(formula = FemaleAges ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6810 -1.7160  0.3255  1.3683  4.0551
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.00052   54.12486  -1.127   0.2660
## Years         0.04650    0.02738   1.698   0.0967 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.385 on 43 degrees of freedom
## Multiple R-squared:  0.06286,	Adjusted R-squared:  0.04107
## F-statistic: 2.884 on 1 and 43 DF,  p-value: 0.09667
```



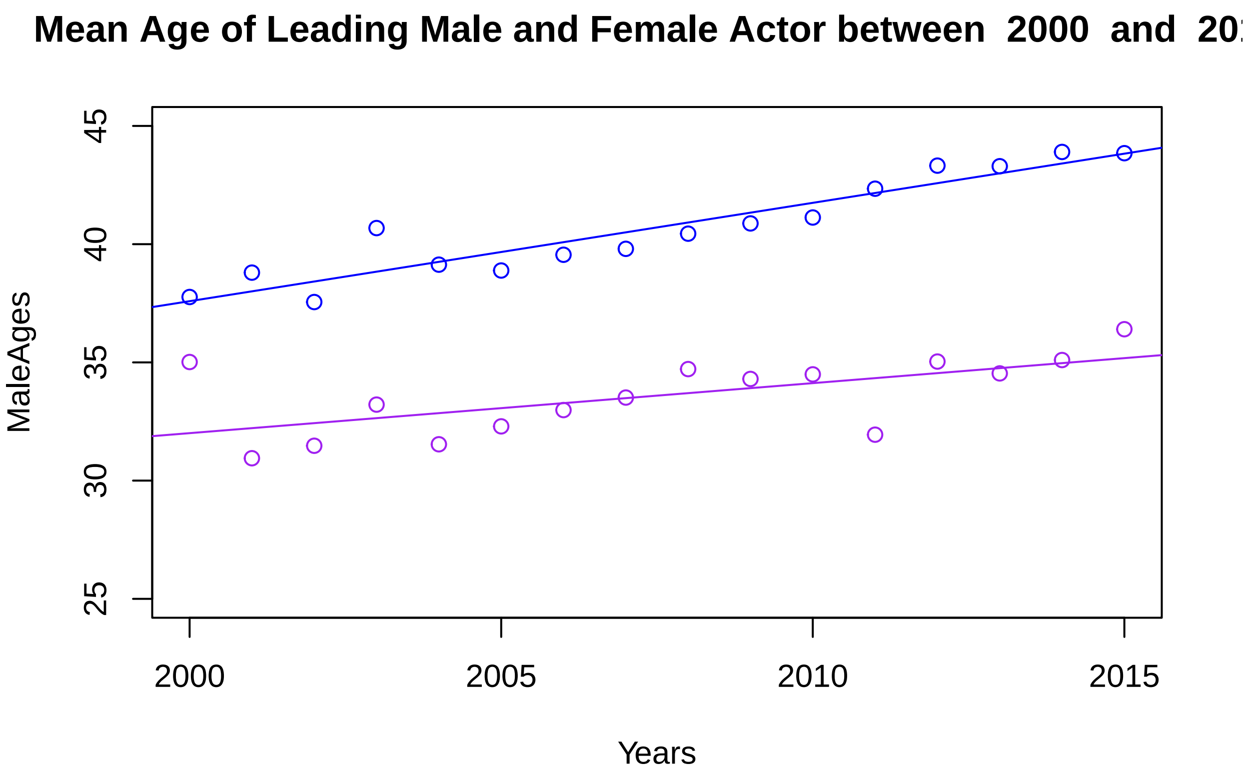

```
##
## Call:
## lm(formula = Ages ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6347 -0.6954  0.1854  0.8744  2.0994
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.03382   33.30337   2.103   0.0414 *
## Years       -0.01682    0.01685  -0.998   0.3237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.468 on 43 degrees of freedom
## Multiple R-squared:  0.02265,    Adjusted R-squared:  -7.925e-05
## F-statistic: 0.9965 on 1 and 43 DF,  p-value: 0.3237
```

At that period we can see, that we have slight decrease in the mean age of male actors towards the end of century, though the average age of female actresses a little increased. But both changes are insignificant. We bound that with the high flow of actors in the industry due to the rapid growth of it in that period. And because of it we have the mean value very similar at the whole period.

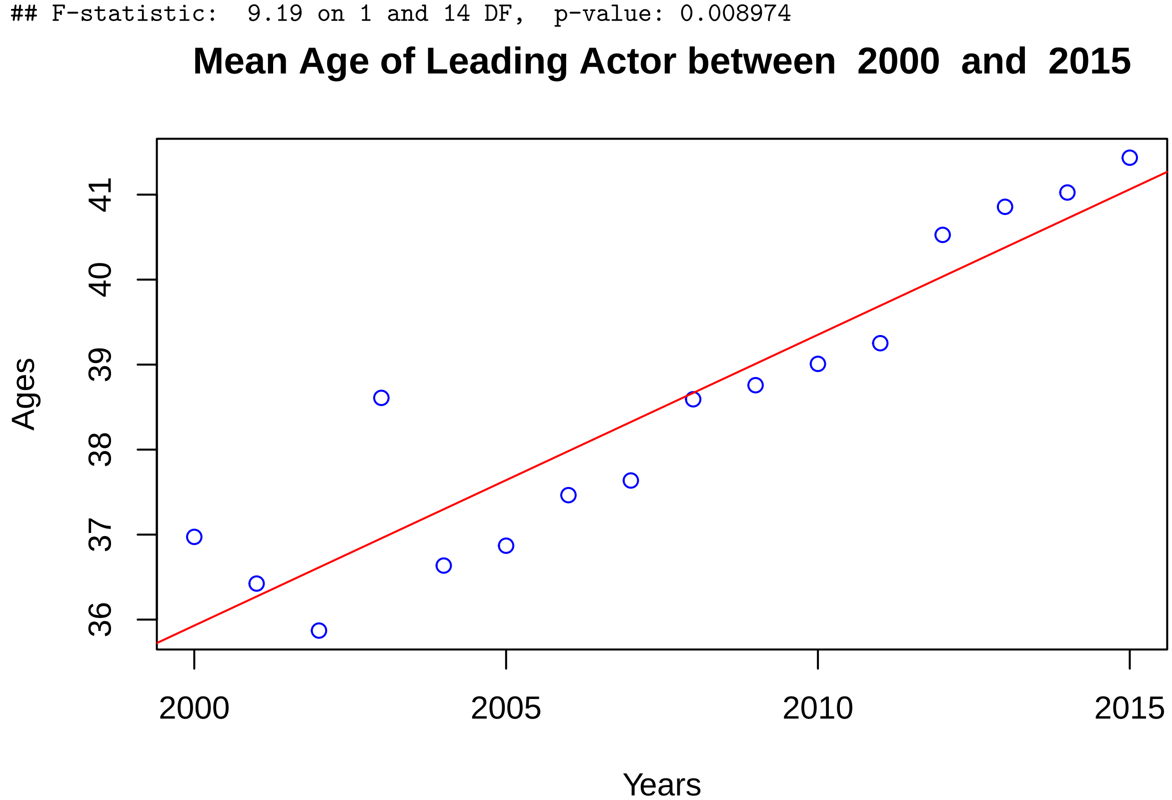
```
age_of_actors_function(2000, 2015, 1)
```

```
##  [1] 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
## [16] 2015
```




```
##
## Call:
## lm(formula = MaleAges ~ Years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86776 -0.55444 -0.04905  0.34735  1.84430
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -794.30376   81.97127   -9.69 1.38e-07 ***
## Years          0.41595    0.04083   10.19 7.43e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7529 on 14 degrees of freedom
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.8726
## F-statistic: 103.8 on 1 and 14 DF,  p-value: 7.425e-08
##
##
## Call:
## lm(formula = FemaleAges ~ Years)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39202 -0.81726  0.07771  0.51028  3.00810
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -391.03345  140.07332  -2.792  0.01442 *
## Years          0.21152    0.06977   3.031  0.00897 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.287 on 14 degrees of freedom
## Multiple R-squared:  0.3963, Adjusted R-squared:  0.3532
## F-statistic:  9.19 on 1 and 14 DF,  p-value: 0.008974
```




```
##
```

```
## Call:
## lm(formula = Ages ~ Years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7710 -0.5545 -0.1633  0.3989  1.6527
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -648.43860   78.37463  -8.274 9.24e-07 ***
## Years          0.34218    0.03904   8.765 4.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7199 on 14 degrees of freedom
## Multiple R-squared:  0.8459, Adjusted R-squared:  0.8348
## F-statistic: 76.82 on 1 and 14 DF,  p-value: 4.658e-07
```

For the 21st century we see the most 'linear' graphs. P-values of all our graphs are very-very small. In fact, they are almost perfect and from the result of our linear model we can say, that there is indeed the relation between age and year. In recent years mean age spiked in comparison to previous decades. There may be several reasons for that: demand on older actors, lack of actor flow (stars occupy their places for years) general aging of people in that industry.

Conclusions

In that research we studied the data about film industry at very long period (almost from the start of its development). We started our research with very different view at the movies and statistics about them. 2 of 3 our initial hypothesis were opposite to the reality, which is a very cool result in fact. With help of statistics we found the truth about average film length, which did not decrease with the