Санкт-Петербургский государственный университет

Прикладная математика, программирование и искусственный интеллект

Отчет по учебной практике 1 (научно-исследовательской работе) (семестр 1)

Прохождение курса по R на Stepik.

Выполнила:

Дмитроченко Елизавета Дмитриевна,

группа 22.Б04-мм

Научный руководитель:

Кандидат физико-математических наук,

доцент

Голяндина Нина Эдуардовна.

Кафедра

Статистического моделирования

Работа выполнена на отличном уровне и может быть зачтена с оценкой А.

Санкт-Петербург

2022

Содержание

Введение

В качестве учебной практики я проходила курс по программирования на R на платформе Stepik.

Задачи моей работы были следующие:

1.) Освоить основные методы программирования на R.
2.) Научиться работать в среде разработки R Studio.
3.) Применить полученные знания к анализу массива films.

Данный курс состоял из трёх модулей

1.) Базовые структуры и понятия
2.) Продвинутые структуры
3.) Продвинутое программирование

В процессе изучения первого модуля я познакомилась с основными особенностями языка R, научилась работать с векторами и векторизованными функциями.

В ходе работы со вторым модулем я узнала о более сложных структурах, таких как матрицы и списки, дата фреймы. Также в этом блоке мне было необходимо провести работу с таблицей Avian Habitat. Сперва с помощью встроенных функций языка я преобразовала данные таблицы в дата фрейм, а затем проанализировала их.

В третьем модуле мы вновь вернулись к изучению функций, но уже на более продвинутом уровне. Я научилась писать собственные функции для решения конкретных задач. Также в этом модуле был дан краткий обзор таких библиотек как tidyr и dplyr.

Весь курс сопровождался большим количеством упражнений и практических заданий, с которыми мне удалось справиться. Для выполнения некоторых из них я обращалась к дополнительным источникам (в частности книга Hadley Wickham “Advanced R” и книга Richard Cotton “ Learning R”).

Данный курс я закончила на отлично. Сертификат, подтверждающий это находится в приложении.

В основной части работы я продемонстрирую своё умение применять знания, полученные в процессе прохождения курса.

Анализ данных по фильмам из IMDB

Дмитроченко Елизавета

2022-12-16

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)

library(stringi)
library(stringr)
library(tidyr)
library(dplyr)
```

Читаем данные из файла в дата фрейм

```
films<-read.csv(sep=";",
                header=T,quote=""
                ,file="C:/Users/dmitr/Desktop/imdb dataset 500.csv")
```

С помощью этих команд мы получаем общие сведения о нашей таблице. Мы имеем 7 переменных, 501 значение для каждой. С помощью последней команды получаем максимальное, минимальное и среднее значение каждой переменной. Теперь мы знаем, что изучаемые нами фильмы были выпущены в период с 2017 по 2019 год. Длительность самого короткого фильма составила 81 минуту. А самый длинный фильм из нашего перечня идёт 3 часа и 1 минуту. Интересно узнать название этих фильмов. Попробуем это сделать.

```
str(films)

## 'data.frame':    501 obs. of  7 variables:
##  $ movie_name    : chr  "47 Meters Down: Uncaged" "A Dog's Journey" "A Dog's
Way Home" "A Hidden Life" ...
##  $ released_year : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ genre         : chr  "Adventure, Drama, Horror" "Adventure, Comedy, Drama"
"Adventure, Drama, Family" "Biography, Drama, War" ...
##  $ duration_min  : int  90 109 96 173 109 92 123 105 112 128 ...
##  $ user_rating   : num  5.1 7.4 6.7 7.3 4.3 6.9 7.1 5.4 5.9 7.1 ...
##  $ critics_rating: int  43 43 50 76 39 44 80 30 51 53 ...
##  $ votes         : int  2903 8362 9493 453 2731 2225 26831 18023 611 133838 .
..

head(films)

##                movie_name released_year                    genre duration_min
## 1 47 Meters Down: Uncaged          2019 Adventure, Drama, Horror           90
## 2         A Dog's Journey          2019 Adventure, Comedy, Drama          109
## 3        A Dog's Way Home          2019 Adventure, Drama, Family           96
## 4           A Hidden Life          2019    Biography, Drama, War          173
## 5  A Madea Family Funeral          2019                   Comedy          109
## 6 A Rainy Day in New York          2019          Comedy, Romance           92
##   user_rating critics_rating votes
```

```
## 1         5.1             43  2903
## 2         7.4             43  8362
## 3         6.7             50  9493
## 4         7.3             76   453
## 5         4.3             39  2731
## 6         6.9             44  2225

summary(films)

##   movie_name        released_year     genre            duration_min
##  Length:501         Min.   :2017   Length:501         Min.   : 81.0
##  Class :character   1st Qu.:2018   Class :character   1st Qu.: 95.0
##  Mode  :character   Median :2018   Mode  :character   Median :104.0
##                     Mean   :2018                      Mean   :107.1
##                     3rd Qu.:2019                      3rd Qu.:116.0
##                     Max.   :2019                      Max.   :181.0
##                                                       NA's   :1
##   user_rating    critics_rating      votes
##  Min.   :2.900   Min.   : 8.00   Min.   :    20
##  1st Qu.:5.600   1st Qu.:43.00   1st Qu.:  2382
##  Median :6.200   Median :55.00   Median : 11644
##  Mean   :6.175   Mean   :54.74   Mean   : 43434
##  3rd Qu.:6.900   3rd Qu.:67.00   3rd Qu.: 44489
##  Max.   :9.500   Max.   :96.00   Max.   :709113
##

films[is.na(films)]<-0
```

Таким образом, мы узнали, что самый длинный фильм - это “Avengers: Endgame” (“Мстители финал”), а самых коротких фильмов оказалось два: “Piercing”,“A Happening of Monumental Proportions”. Заметим, что Мстители довольно известный фильм и у меня не возникло трудностей с перевод его названия на русский язык, в то время как два самых коротких фильма мне неизвестны. Данный факт побуждает нас узнать, как были оценены эти фильмы зрителями (возможно, существует некоторая связь между длительностью фильма и зрительской оценкой)

```
films[films$duration_min==max(films$duration_min),]

##           movie_name released_year                      genre duration_min
## 17 Avengers: Endgame          2019 Action, Adventure, Sci-Fi          181
##    user_rating critics_rating  votes
## 17         8.6             78 562071

films[films$duration_min==min(films$duration_min[films$duration_min!=min(films$d
uration_min)]),]

##                                 movie_name released_year
## 317                               Piercing          2018
## 440 A Happening of Monumental Proportions          2017
##                          genre duration_min user_rating critics_rating votes
## 317 Horror, Mystery, Thriller           81         5.6             63  4096
## 440             Comedy, Drama           81         4.6             35   507
```

Ранее мы уже находили среднюю пользовательскую оценку фильмов. Она составляет 6.2. То есть оценка самого длинного фильма на 2.4 выше средней, а оценка самых коротких фильмов ниже средней. Теперь с помощью сортировки попытаемся узнать есть ли какая-нибудь связь между длительностью фильма и оценкой пользователей Действительно, фильмы, которые занимают большее количество времени, имеют более высокую оценку, в то время как фильмы с самой низкой оценкой длятся в среднем 1.5 часа.

```
films<-films[order(films$user_rating,decreasing=TRUE),]
head(films[,c(1,4,5)])
```

```
##                            movie_name duration_min user_rating
## 59                           Joker I           122         9.5
## 17                 Avengers: Endgame           181         8.6
## 176           Avengers: Infinity War           149         8.5
## 207                        Capharnaüm           126         8.4
## 352 Spider-Man: Into the Spider-Verse           117         8.4
## 487                           Coco I           105         8.4
```

```
tail(films[,c(1,4,5)])
```

```
##                       movie_name duration_min user_rating
## 357                     Supercon          100         3.6
## 34                   Dead Water            0         3.3
## 346                 Slender Man           93         3.2
## 57                Jacob's Ladder           89         3.1
## 239                 Future World           90         3.1
## 112 The Haunting of Sharon Tate           94         2.9
```

Итак, мы узнали мнение зрителей о наших фильмах. Посмотрим, что думают критики. Оказывается мнения профессионалов и обычных посетителей кинотеатров разнятся. В последних двух полученных сводках практически нет совпадений.

```
films<-films[order(films$critics_rating,decreasing=TRUE),]
head(films[,c(1,4,5,6)])
```

```
##          movie_name duration_min user_rating critics_rating
## 329            Roma          135         7.8             96
## 69   Marriage Story          136         7.8             95
## 500         Dunkirk          106         7.9             94
## 129    The Souvenir          120         6.6             92
## 374   The Favourite          119         7.6             90
## 111 The Farewell I           100         8.1             89
```

```
tail(films[,c(1,4,5,6)])
```

```
##                       movie_name duration_min user_rating critics_rating
## 460              Armed Response           93         3.7             13
## 54                  I Hate Kids           89         6.1             12
## 140                   Unplanned          109         5.8             10
## 239                 Future World           90         3.1             10
## 357                     Supercon          100         3.6              9
## 112 The Haunting of Sharon Tate           94         2.9              8
```

Далее формируем переменную, в которой будем хранить полный перечень жанров.

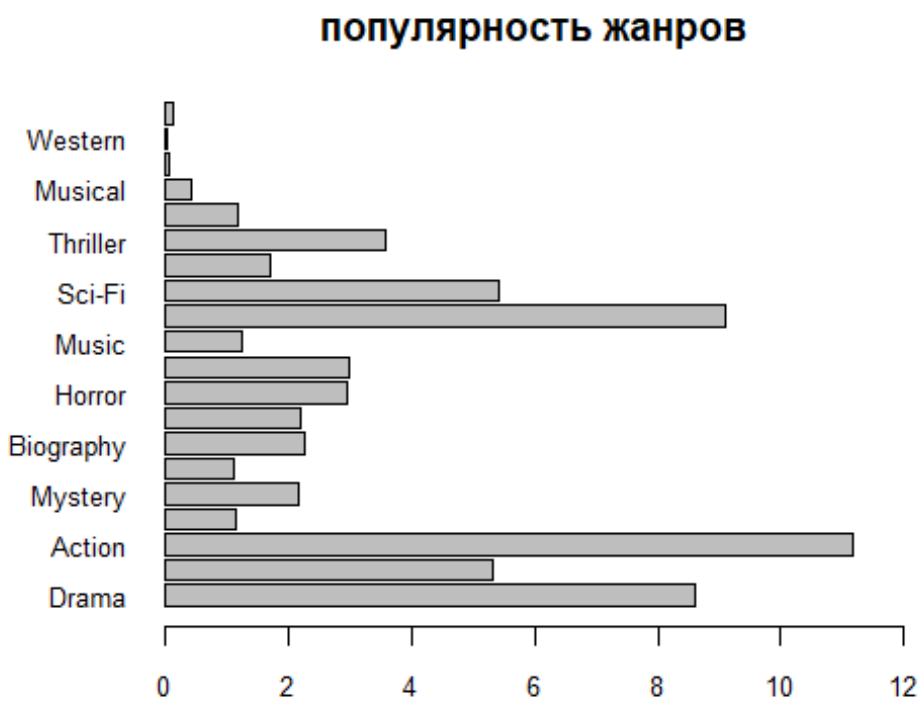
```
film_types<-unique(unlist(strsplit(films$genre,", ")))
```

Создадим дата фрейм, в котором сгруппируем все фильмы по жанрам. Для каждого жанра найдём количество фильмов, относящихся к нему, общее количество голосов за фильмы в этом жанре, среднюю оценку критиков и пользователей.

```
sum_of_votes<-NULL
mean_of_using_rating<-NULL
mean_of_critics_rating<-NULL
kol_of_films<-NULL
for (i in 1:length(film_types))
{
  kol_of_films<-c(kol_of_films,sum(str_count(films$genre,film_types[i])))
  sum_of_votes<-c(sum_of_votes,sum(films[grepl(film_types[i],films$genre),7]))
  mean_of_using_rating<-c( mean_of_using_rating,mean(films[grepl(film_types[i],f
ilms$genre),5]))
  mean_of_critics_rating<-c( mean_of_critics_rating,mean(films[grepl(film_types[
i],films$genre),6]))
}
genre_table<-data.frame(film_types=film_types,kol_of_films=kol_of_films,sum_of_v
otes=sum_of_votes,
                        mean_of_using_rating=mean_of_using_rating,
                        mean_of_critics_rating=mean_of_critics_rating
)
genre_table$sum_of_votes_mln<-sum_of_votes/1000000
```

Пользуясь нашим новым дата фреймом, создадим график, отражающий популярность каждого жанра фильм Изучив его, мы можем сделать вывод, что наибольшей популярностью обладают фильмы жанров Action, Adventure, Drama, наименьшей - Western, War, Sport.

```
par(mar=c(3,5,3,1))
barplot(genre_table$sum_of_votes_mln,names.arg=genre_table$film_types,
        main="популярность жанров",
        xlab="общее количество голосов по каждому жанру,млн",horiz = TRUE,las=1,
        xlim=c(0,12),
        width=5,
        cex.axis=0.8, cex.names=0.8
        )
```

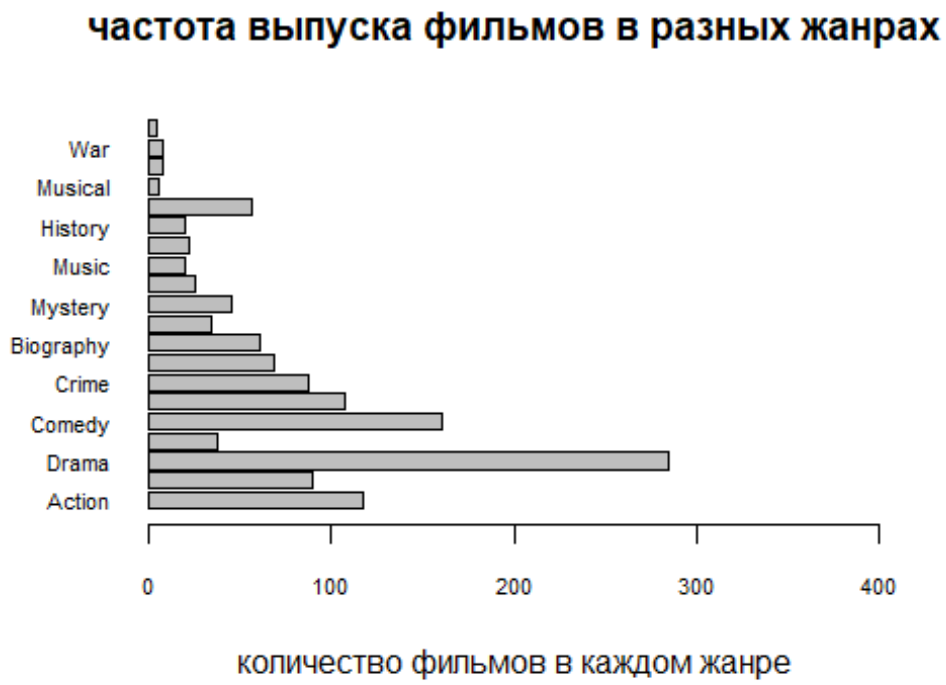


Мы знаем, что больше всего голосов набрали фильмы, снятые в жанрах Action, Adventure, Drama. Тогда давайте посмотрим, какое количество фильмов из нашего перечня принадлежит этим жанрам и сравним наш результат со средним значением. Скомпилировав написанный ниже код, мы видим, что количество фильмов, снятых в наиболее популярных жанрах значительно выше среднего показателя по таблице. Это говорит о том, что современные режиссёры активно подстраиваются под запросы потребителей.

```
genre_table<-genre_table[order(genre_table$sum_of_votes_mln,decreasing=TRUE),]
head(genre_table[,c(1,2,3)],3)

##    film_types kol_of_films sum_of_votes
## 3      Action          117     11154703
## 12  Adventure           89      9083913
## 1       Drama          284      8607496

mean(genre_table$kol_of_films)

## [1] 62.8
```

Теперь зайдём с другой стороны: создадим график, который показывает количество фильмов, выпущенных в разных жанрах. Изучив его, можем заметить, что кроме Drama и Action большое число фильмов принадлежит жанру Comedy.

```
barplot(genre_table$kol_of_films,names.arg=genre_table$film_types,
        main="частота выпуска фильмов в разных жанрах",
        xlab="количество фильмов в каждом жанре",horiz = TRUE,las=1,
        xlim=c(0,400), cex.axis=0.7, cex.names=0.7)
```



Добавим в дата фрейм films столбец, который будет отвечать за разницу в оценке критиков и пользователей. Таким образом, мы сможем узнать названия фильмов, для которых разница в оценке критиков и пользователей была максимальной/минимальной. Заметим, что наибольшее разногласие между пользователями и критиками произошло в процессе оценки следующих фильмов: “Roma”,“Marriage Story”,“Dunkirk’.

```
films$difference_between_critics_and_users_rating<-films$critics_rating-films$us
er_rating
head(films[order(films$difference_between_critics_and_users_rating,decreasing=TR
UE),c(1,8)])
```

```
##         movie_name difference_between_critics_and_users_rating
## 329           Roma                                        88.2
## 69  Marriage Story                                        87.2
## 500        Dunkirk                                        86.1
## 129   The Souvenir                                        85.4
## 374  The Favourite                                        82.4
## 227   Eighth Grade                                        81.6
```

```
tail(films[order(films$difference_between_critics_and_users_rating,decreasing=TR
UE),c(1,8)])
```

```
##                     movie_name difference_between_critics_and_users_rating
## 460             Armed Response                                         9.3
## 239               Future World                                         6.9
## 54                 I Hate Kids                                         5.9
## 357                   Supercon                                         5.4
```

```
## 112 The Haunting of Sharon Tate                                    5.1
## 140                    Unplanned                                   4.2
```

В завершении исследования современных фильмов создадим ещё один дата фрейм, в нём посчитаем количество фильмов, выпущенных в каждом году и их суммарную длительность. Теперь мы видим, что больше всего фильмов было выпущено в 2018 году. Их суммарная длительность составила 30540 минут. Примерно вполовину меньше было выпущено фильмов в 2019 году, и ещё вполовину меньше в 2017.

```
kol_2019<-films$released_year==2019
kol_2018<-films$released_year==2018
kol_2017<-films$released_year==2017
year_statistic<-data.frame(year=c(2019,2018,2017),
                           number_of_films=c(table(kol_2019)[2],table(kol_2018)[
2],table(kol_2017)[2]),
                           sum_of_duration=c(sum(films[films$released_year==2019
,4]), sum(films[films$released_year==2018,4]), sum(films[films$released_year==20
17,4])))
year_statistic

##   year number_of_films sum_of_duration
## 1 2019             147           15833
## 2 2018             284           30540
## 3 2017              70            7198
```

Заключение

В ходе практической работы я получила новые знания и умения в сфере программирования на языке R. Их я применила при анализе таблицы films.

С помощью функций базового R, а также функций из пакетов tidyr и dplyr

я смогла считать данные из .csv файла и преобразовать их в дата фрейм в RStudio.

Используя, различные возможности R, я получила интересные сведения об изучаемых фильмах. Мне удалось разбить все фильмы на группы в соответствии с их жанрами и годами выпуска, также у меня получилось отследить различия между пользовательской оценкой фильмов и оценкой критиков.

С помощью функций для построения графиков в R я определила наиболее и наименее популярные жанры фильмов.

Пройденный мною курс оказался довольно интересным. Информация преподносилась в нём доступно и понятно, было много увлекательных практических заданий.

Источники

1.) Hadley Wickham “Advanced R”, 2015 by Taylor & Francis Group, LLC
2.) Richard Cotton “ Learning R”, 2013, Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
3.) Курс «Основы программирования на R», Антон Антонов, платформа Stepik.
4.) https://stackoverflow.com

stepik

BIOINFORMATICS INSTITUTE

Результат

100%

С отличием

Настоящий сертификат подтверждает, что

Елизавета Дмитроченко

успешно завершил/а курс

Основы программирования на R

Антон Антонов

https://stepik.org/course/497 https://stepik.org/cert/1835313

07.12.2022