

Проект

Дисциплина: Статистика и емпирични методи

- Практикум

Тема:

Предпочитано удоволствие ли са филмите?

Изпълнител:

Деница Стоянова

ФМИ, курс 2, гр. 2

Ф.№ 71904

Анкетирах 80 човека с цел да разбера дали филмите са предпочитано удоволствие за тях. Прилагам линк към проведената анкета:

<https://docs.google.com/forms/d/e/1FAIpQLSeFqWGttr9c-H5n2Uil6mvCdX6cNG848DCnpJY5eloJlqX1aA/viewform>

Правя описателен анализ върху получените резултати - анализ на едномерна променлива и анализ на взаимодействието между две променливи. Като за всеки въпрос въвеждам данните във вектори и след това ги обединявам в един data frame.

1. Въвеждане на данните и анализ на едномерна променлива

Въпрос 1: Гледаш ли филми?

❖ Въвеждане на данните:

```
watching <- c("Да, постоянно", "Рядко", "Рядко", "Само когато имам време", "Рядко", "Само когато имам време", "Да, постоянно", "Да, постоянно", "Рядко", "Да, постоянно", "Само когато имам време", "Рядко", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Да, постоянно", "Рядко", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Рядко", "Само когато имам време", "Само когато имам време", "Да, постоянно", "Рядко", "Само когато имам време", "Рядко", "Да, постоянно", "Да, постоянно", "Рядко", "Само когато имам време", "Да, постоянно", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Да, постоянно", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Рядко", "Само когато имам време", "Само когато имам време", "Да, постоянно", "Да, постоянно", "Само когато имам време", "Само когато имам време", "Рядко", "Само когато имам време", "Само когато имам време", "Рядко", "Да, постоянно", "Само когато имам време", "Да, постоянно", "Рядко", "Само когато имам време", "Само когато имам време", "Само когато имам време", "Рядко", "Да, постоянно", "Да, постоянно")
```

❖ Анализ:

При категориите променливи най-добре честотата се вижда посредством таблици, затова използвам функцията `table()`. С `prop.table()` взимам процентното разпределение.

```
table_watching <- table(watching)
```

```
watching
```

Да, постоянно	Рядко	Само когато имам време
20	18	42

```
prop_table_watching <- prop.table(table_watching)
```

watching	Да, постоянно	Рядко	Само когато имам време
	0.250	0.225	0.525

Използвам barplot за да представя графично честотното разпределение на категориите променливи.

```
barplot(height = prop_table_watching, col = rainbow(5), main = "Гледаш ли филми?")
```



На графиката ясно се вижда, че преобладават хората, които гледат филми само когато имат време. Техният брой е 42. От анкетираните няма такива, които не гледат филми.

Използвам piechart за да изобразя процентното разпределение на получените данни.

```
piepercent_watching <- round(100*table_watching/sum(table_watching), 1)
```

```
pie(table_watching, labels = piepercent_watching, main = " Гледаш ли филми?", col = rainbow(n = length(table_watching)*2))
```

```
legend(x = "bottomleft", legend = c("Да, постоянно", "Само когато имам време", "Рядко", "Не"), cex = 0.8, fill = rainbow(length(table_watching) * 2))
```

Гледаш ли филми?



🚦 Въпрос 2: По колко часа на ден в интервала 0 - 24 отделяш за гледане на филми?

❖ Въвеждане на данните:

```
hours_watching <-  
c(3,6,6,5,5,1,5,5,2,5,2,1,2,3,4,2,4,1,3,3,2,3,3,2,5,1,3,1,6,3,3,4,2,4,4,2,3,6,3,4,3,2,2,4,2,2,4,2,2,2,2,4,5,2,3,2,3,1,4,3,2,6,  
3,2,2,1,8,2,2,2,1,2,1,2,3,3,2,8,5)
```

❖ Анализ:

Намирам модата - най-често срещаната стойност във вектора.

```
modeFunction <- function(x) {  
  res_table <- table(x)  
  return(names(res_table)[res_table == max(res_table)])  
}  
  
modeFunction(hours_watching)
```

```
[1] "2"
```

След прилагане на modeFunction установявам, че най-често срещаната стойност е 2. Следователно повечето от анкетираните отделят 2 часа на ден от времето си за гледане на филми.

Прилагам summary - описателна статистика за центъра на разпределението. Тя дава информация за минималната стойност, 1-вия квартил, медиана (2-рия квартил), 3-тия квартил и максималната стойност.

```
summary(hours_watching)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	3.062	4.000	8.000

Намирам вариацията (дисперсията):

```
var(hours_watching)
```

```
[1] 2.540348
```

След това намирам стандартното отклонение. То е оценка на вариацията, която показва колко далече са наблюденията от очакването. За разлика от обхвата, взема под внимание всички наблюдения. Стандартното отклонение е производно на вариацията и представлява корен квадратен от дисперсията.

```
sd(hours_watching)
```

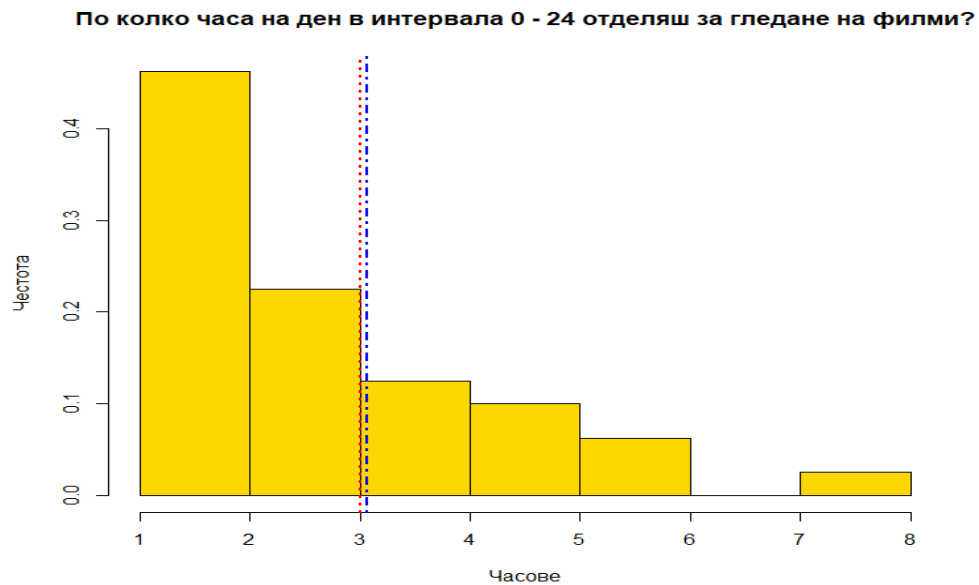
```
[1] 1.593847
```

Използвам хистограма, за да представя разпределението на непрекъснатите променливи. Синята вертикална прекъсната линия показва къде се намира средната стойност, а червената – медианата. Както се вижда, техните стойности са много близки.

```
hist(hours_watching, main = "По колко часа на ден в интервала 0 - 24 отделяш за гледане на филми?", xlab = "Часове", ylab = "Честота", col = "gold1", prob = T)
```

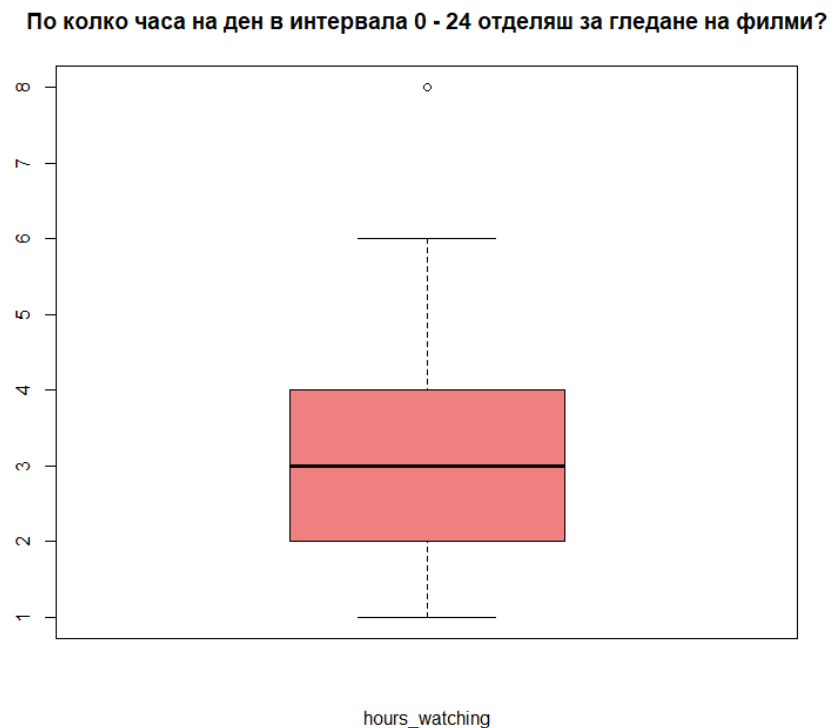
```
abline(v = mean(hours_watching), lwd = 2, lty = 4, col = "blue")
```

```
abline(v = median(hours_watching), lwd = 2, lty = 3, col = "red")
```



Използвам boxplot за откриването на потенциални outlier-и.

```
boxplot(hours_watching, col = "lightcoral", main = "По колко часа на ден в интервала 0 - 24  
отделяш за гледане на филми?", xlab = "hours_watching")
```



Открит е потенциален outlier.

Въпрос 3: Какво предпочиташ?

❖ Въвеждане на данните:

```
movies_or_serials <- c("Сериали","Сериали","Филми","Сериали","Филми","Сериали","Сериали","И  
двете","Филми","И двете","И двете","Сериали","Филми","И двете","И двете","Филми","И двете","И  
двете","Сериали","И двете","И двете","Сериали","И двете","И двете","И двете","Филми","И  
двете","Филми","Филми","И двете","И двете","Филми","Филми","Филми","Филми","И двете","И  
двете","Филми","Сериали","Филми","Филми","Филми","Филми","Филми","Филми","Филми","И двете","И  
двете","И двете","Филми","И двете","И двете","Филми","Филми","Филми","И двете","И  
двете","Филми","Сериали","Сериали","И двете","Филми","И двете","Филми","И двете","И двете","И двете","И  
двете","Филми","Филми","Филми","Филми","Сериали","Филми")
```

❖ Анализ:

Тук отново има категорийни променливи, затова използвам функцията `table()`, а с `prop.table()` взимам процентното разпределение.

```
table_movies_or_serials <- table(movies_or_serials)
```

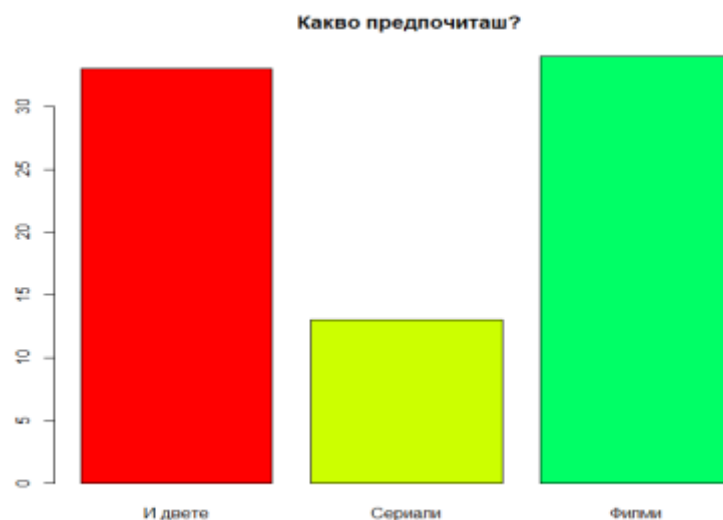
```
movies_or_serials  
И двете    Сериали    Филми  
      33         13         34
```

```
prop_table_movies_or_serials <- prop.table(table_movies_or_serials)
```

```
movies_or_serials  
И двете    Сериали    Филми  
  0.4125   0.1625   0.4250
```

Използвам `barplot` за да представя графично честотното разпределение

```
barplot(height = table_movies_or_serials, col = rainbow(5), main = "Какво предпочиташ?")
```



От графиката установявам, че броя на хората, посочили, че предпочитат филми е почти равен на броя хора, отговорили, че предпочитат и двете. Двете категории се различават с единица.

Използвам piechart за да изобразя процентното разпределение на получените данни.

```
piepercent_movies_or_serials <- round(100*table_movies_or_serials/sum(table_movies_or_serials), 1)
```

```
pie(table_movies_or_serials, labels = piepercent_movies_or_serials, main = "Какво предпочиташ?", col = rainbow(n = length(table_movies_or_serials)*2))
```

```
legend(x = "bottomleft", legend = c("И двете", "Сериали", "Филми"), cex = 0.8, fill = rainbow(length(table_movies_or_serials)*2))
```



🚦 Въпрос 4: Какво е за теб гледането на филми?

❖ Въвеждане на данните:

Тъй като на този въпрос всеки анкетиран има право да избере повече от един отговор, имам наличие на повече отговори от броя на анкетираните.

Поради тази причина не включвам данните в data frame-а и ги въвеждам по следния начин:

```
what_are_movies <- c(rep("Време за релакс",64),rep("Възможност да науча нещо ново",22),rep("Загуба на време",8))
```

❖ Анализ:

Тук отново има категорийни променливи, затова използвам функцията table(), а с prop.table() взимам процентното разпределение.

```
table_what_are_movies <- table(what_are_movies)
```

```
what_are_movies
Време за релакс    Възможност да науча нещо ново    Загуба на време
        64                22                        8
```

```
prop_table_what_are_movies <- prop.table(table_what_are_movies)
```

```
what_are_movies
Време за релакс    Възможност да науча нещо ново    Загуба на време
0.68085106         0.23404255                 0.08510638
```

Използвам barplot за да представя графично честотното разпределение

```
barplot(height = table_what_are_movies, col = rainbow(5), main = "Какво е за теб гледането на филми?")
```



От графиката става ясно, че за хората гледането на филми е предимно време за релакс. Малка част от тях са отговорили че е

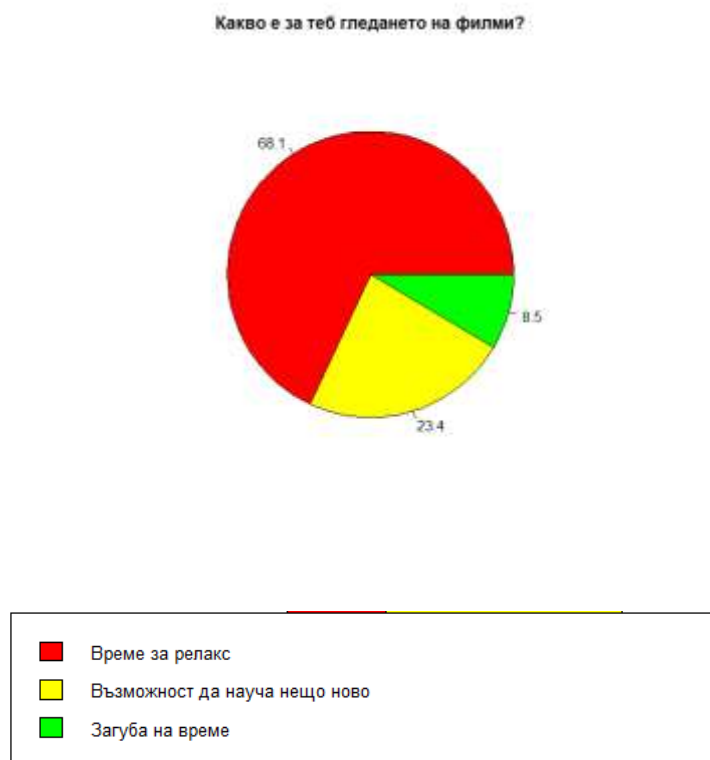
възможност да научат нещо ново. А за 8 от анкетираните филмите са само загуба на време.

Използвам piechart за да изобразя процентното разпределение на получените данни.

```
piepercent_what_are_movies <- round(100*table_what_are_movies/sum(table_what_are_movies), 1)
```

```
pie(table_what_are_movies, labels = piepercent_what_are_movies, main = "Какво е за теб гледането на филми?", col = rainbow(n = length(piepercent_what_are_movies)*2))
```

```
legend(x = "bottomleft", legend = c("Време за релакс", "Възможност да науча нещо ново", "Загуба на време"), cex = 0.8, fill = rainbow(length(table_watching)*2))
```



❖ Въпрос 5: Обикновено колко филми (в интервала 1- 20) гледаш на седмица?

❖ Въвеждане на данните:

```
movies_count <-  
c(4,9,9,7,6,2,6,5,2,20,2,1,2,3,4,2,6,1,2,6,1,4,4,5,9,1,2,2,11,5,2,5,3,5,5,3,3,10,5,5,7,5,3,8,2,2,10,3,3,2,3,2,7,10,5,7,1,5,2,6,  
5,3,12,5,3,10,3,10,2,2,4,1,7,3,4,4,3,4,12,9)
```

❖ Анализ:

Намирам модата - най-често срещаната стойност във вектора.

```
modeFunction <- function(x) {  
  res_table <- table(x)  
  return(names(res_table)[res_table == max(res_table)])  
}  
  
modeFunction(movies_count)
```

```
[1] "2"
```

След прилагане на modeFunction установявам, че най-често срещаната стойност е 2. Следователно повечето от анкетираните гледат по 2 филма на седмица.

Прилагам summary - описателна статистика за центъра на разпределението.

```
summary(movies_count)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	2.00	4.00	4.85	6.00	20.00

Намирам вариацията (дисперсията):

```
var(movies_count)
```

```
[1] 11.14177
```

След това намирам стандартното отклонение.

```
sd(movies_count)
```

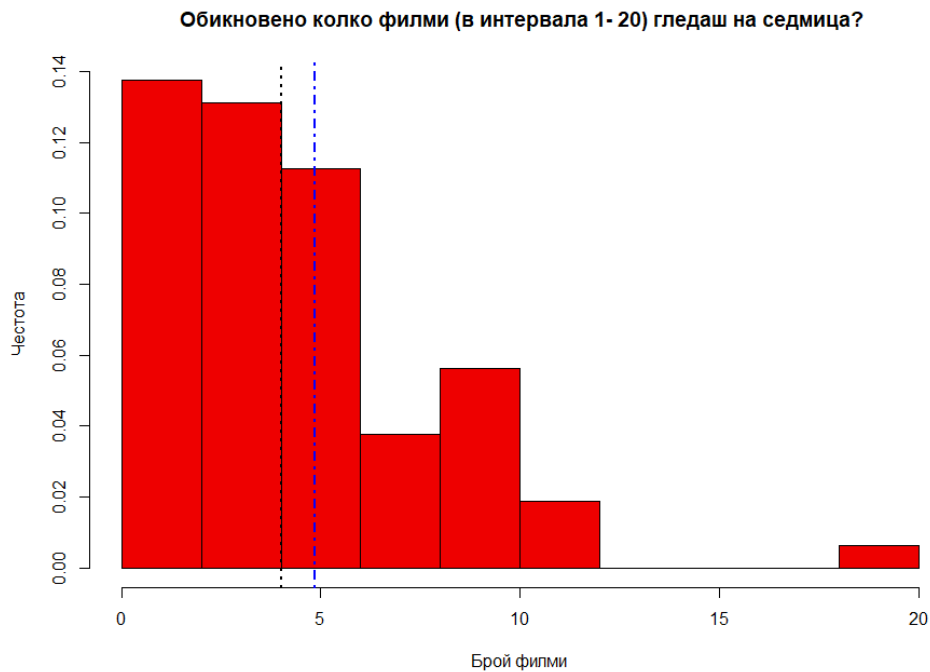
```
[1] 3.337929
```

Използвам хистограма, за да представя разпределението на непрекъснати те променливи. Синята вертикална прекъсната линия показва къде се намира средната стойност, а черната – медианата. Както се вижда, техните стойности са близки.

```
hist(movies_count, main = "Обикновено колко филми (в интервала 1- 20) гледаш на седмица?", xlab = "Брой филми", ylab = "Честота", col = "red2", prob = T)
```

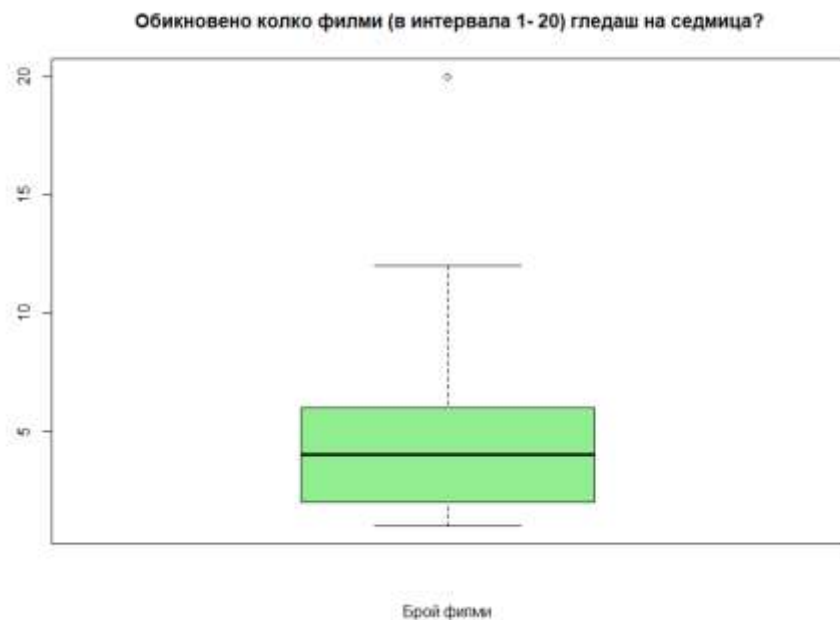
```
abline(v = mean(movies_count), lwd = 2, lty = 4, col = "blue")
```

```
abline(v = median(movies_count), lwd = 2, lty = 3, col = "black")
```



Използвам `boxplot` за откриването на потенциални outlier-и.

```
boxplot(movies_count, col = "lightgreen", main = " Обикновено колко филми (в интервала 1- 20)  
гледаш на седмица?", xlab = "Брой филми")
```



Открит е потенциален outlier.

Въпрос 6: Кой е любимият ти жанр филми?

❖ Въвеждане на данните:

Тъй като на този въпрос отново всеки анкетиран има право да избере повече от един отговор, имам наличие на повече отговори от броя на анкетираните. Поради тази причина не включвам данните в data frame-а и ги въвеждам по следния начин:

```
genre_movies <-  
c(rep("Комедия",50),rep("Екшън",28),rep("Трилър",19),rep("Ужаси",21),rep("Драма",23),rep("Роман-  
тичен",36),rep("Фентъзи",17),rep("Анимация",16),rep("Семеен",18),rep("Приключенски",29))
```

❖ Анализ:

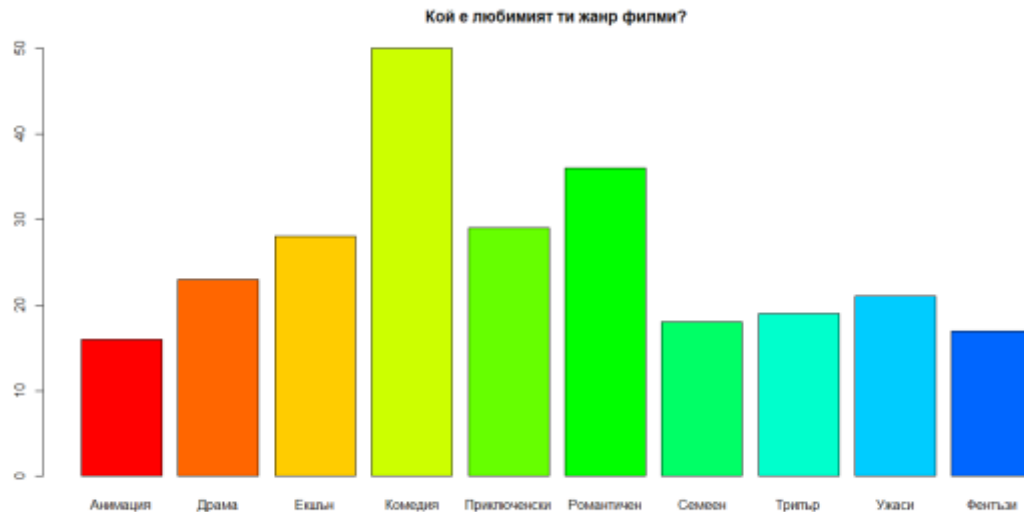
Тук отново има категорийни променливи, затова използвам функцията table(), а с prop.table() взимам процентното разпределение.

```
> table_genre_movies <- table(genre_movies)  
> table_genre_movies  
genre_movies  
Анимация    Драма    Екшън    Комедия    Приключенски    Роман-тичен    Семеен  
    16        23        28        50          29          36          18  
Трилър      Ужаси    Фентъзи  
    19        21        17  
> prop_table_genre_movies <- prop.table(table_genre_movies)  
> prop_table_genre_movies  
genre_movies  
Анимация    Драма    Екшън    Комедия    Приключенски    Роман-тичен    Семеен  
0.06225681  0.08949416  0.10894942  0.19455253  0.11284047  0.14007782  0.07003891  
Трилър      Ужаси    Фентъзи  
0.07392996  0.08171206  0.06614786
```

Използвам barplot за да представя графично честотното разпределение

```
barplot(height = table_genre_movies, col = rainbow(15), main = "Кой е любимият ти жанр  
филми?")
```

От графиката става ясно, че любимият жанр на най-много от анкетираните е комедия, следван от романтичен и приключенски. Едва за 16 от тях любимият жанр е анимация. Анимацията, наред със семеен, ужаси и фентъзи са най-малко предпочитаните филми.

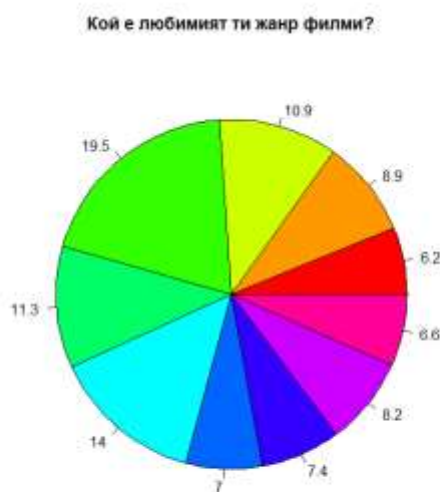


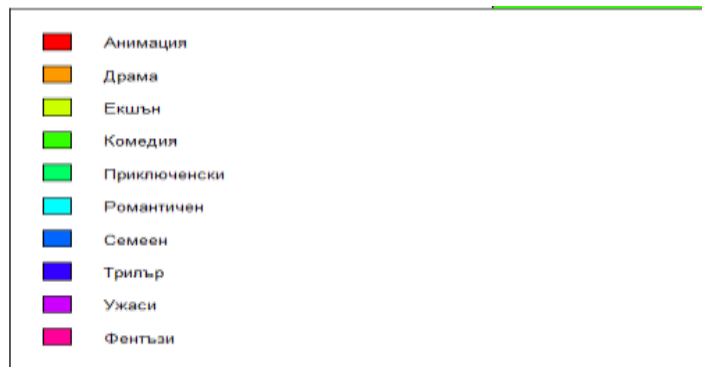
Използвам piechart за да изобразя процентното разпределение на получените данни.

```
piepercent_genre_movies <- round(100*table_genre_movies/sum(table_genre_movies), 1)
```

```
pie(table_genre_movies, labels = piepercent_genre_movies, main = "Кой е любимият ти жанр филми?", col = rainbow(n = length(table_genre_movies)))
```

```
legend(x = "bottomleft", legend =  
с("Анимация","Драма","Екшън","Комедия","Приключенски","Романтичен","Семейен","Трилър","  
Ужаси","Фентъзи"), cex = 0.8,fill = rainbow(length(table_genre_movies)))
```





❖ Въпрос 7: Колко дълъг (в интервала 20 - 180 минути) би бил идеалният филм за теб?

❖ Въвеждане на данните:

```
ideal_movie_length <-
c(20,30,120,100,45,20,20,180,120,180,90,120,150,150,120,120,90,100,90,180,160,120,180,90,135,180,180,130,90,100,160
,169,120,105,105,160,80,20,80,120,180,60,100,120,90,120,105,120,120,110,80,120,150,180,180,120,82,30,90,150,90,70,1
60,90,90,90,90,90,95,150,120,160,120,110,110,120,30,20,180,40)
```

❖ Анализ:

Намирам модата - най-често срещаната стойност във вектора.

```
modeFunction <- function(x) {
  res_table <- table(x)
  return(names(res_table)[res_table == max(res_table)])
}
modeFunction(ideal_movie_length)
[1] "120"
```

След прилагане на modeFunction установявам, че най-често срещаната стойност е 120. Следователно за повечето от анкетираните идеалният филм би бил дълъг 120 минути.

Прилагам summary - описателна статистика за центъра на разпределението.

```
summary(ideal_movie_length)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
20.0   90.0   115.0   110.8  150.0   180.0
```

Намирам вариацията (дисперсията):

```
var(ideal_movie_length)
```

```
[1] 1999.601
```

След това намирам стандартното отклонение.

```
sd(ideal_movie_length)
```

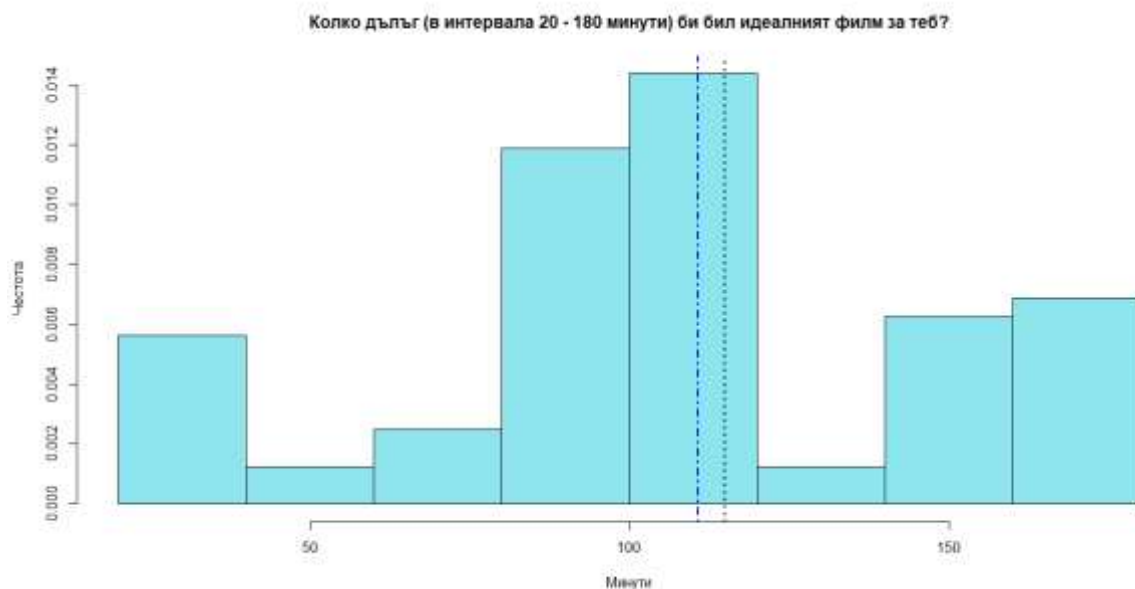
```
[1] 44.7169
```

Използвам хистограма, за да представя разпределението на непрекъснати те променливи. Синята вертикална прекъсната линия показва къде се намира средната стойност, а черната – медианата.

```
hist(ideal_movie_length, main = "Колко дълъг (в интервала 20 - 180 минути) би бил идеалният филм за теб?", xlab = "Минути", ylab = "Честота", col = "cadetblue2", prob = T)
```

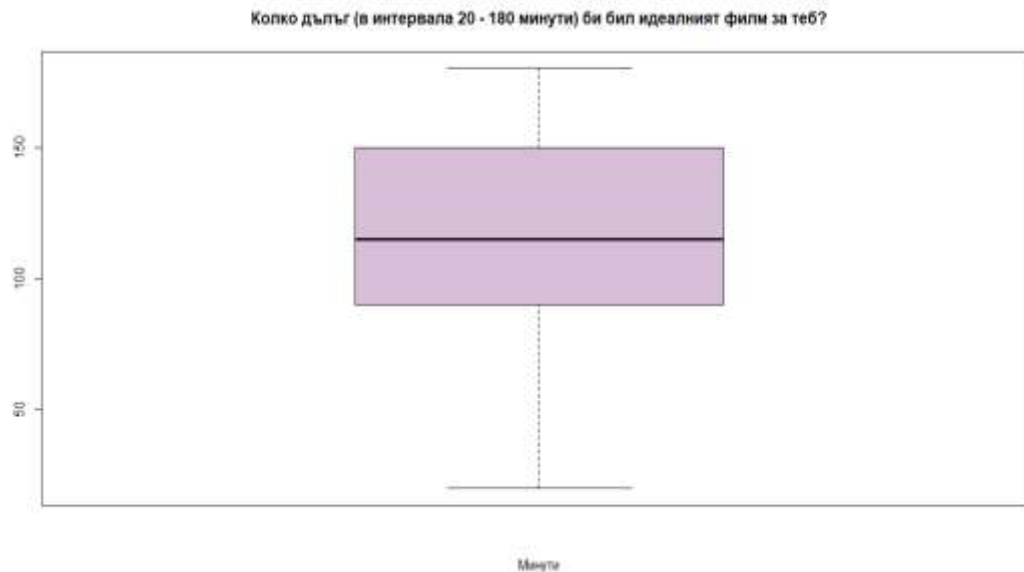
```
abline(v = mean(ideal_movie_length), lwd = 2, lty = 4, col = "blue")
```

```
abline(v = median(ideal_movie_length), lwd = 2, lty = 3, col = "black")
```



Използвам boxplot за откриването на потенциални outlier-и.

```
boxplot(ideal_movie_length, col = "thistle", main = "Колко дълъг (в интервала 20 - 180 минути) би бил идеалният филм за теб?", xlab = "Минути")
```

Няма открити потенциални outlier-и.

Въпрос 8: На кое лично устройство гледаш филми най-често?

❖ Въвеждане на данните:

Тъй като на този въпрос отново всеки анкетирани има право да избере повече от един отговор, имам наличие на повече отговори от броя на анкетирани. Поради тази причина не включвам данните в data frame-а и ги въвеждам по следния начин:

```
personal_device <-  
c(rep("Телевизор",50),rep("Компютър",37),rep("Таблет",5),rep("Смартфон",22))
```

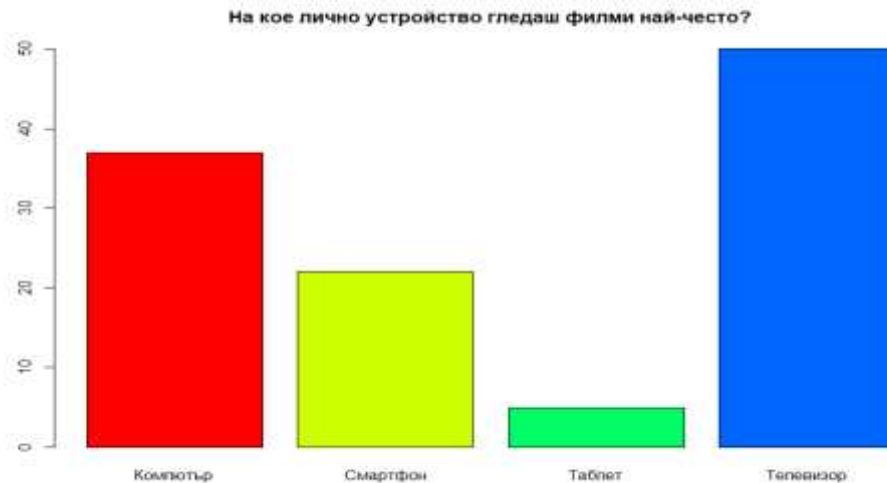
❖ Анализ:

Тук отново има категорични променливи, затова използвам функцията table(), а с prop.table() взимам процентното разпределение.

```
> table_personal_device <- table(personal_device)  
> table_personal_device  
personal_device  
  Компютър  Смартфон   Таблет Телевизор  
        37        22         5         50  
> prop_table_personal_device <- prop.table(table_personal_device)  
> prop_table_personal_device  
personal_device  
  Компютър  Смартфон   Таблет Телевизор  
0.32456140 0.19298246 0.04385965 0.43859649
```

Използвам barplot за да представя графично честотното разпределение

```
barplot(height = table_personal_device, col = rainbow(5), main = "На кое лично устройство гледаш филми най-често?")
```



Както се вижда от графиката, телевизорът е най-предпочитаното устройство за гледане на филми. Най-малка част от хората гледат филми от таблет. Едва 5 човека са го посочили като отговор.

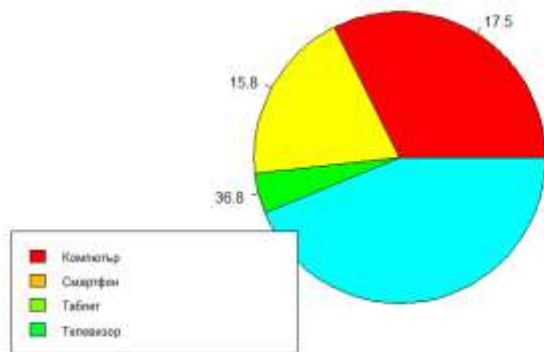
Използвам piechart за да изобразя процентното разпределение на получените данни.

```
piepercent_personal_device <- round(100*table_watching/sum(table_personal_device), 1)
```

```
pie(table_personal_device, labels = piepercent_personal_device, main = "На кое лично устройство гледаш филми най-често?", col = rainbow(n = length(table_watching)*2))
```

```
legend(x = "bottomleft", legend = c("Компютър", "Смартфон", "Таблет", "Телевизор"), cex = 0.8, fill = rainbow(length(table_personal_device)*2))
```

На кое лично устройство гледаш филми най-често?



Въпрос 9: Ходиш ли на кино?

❖ Въвеждане на данните:

```
cinema <-
```

c("Да","Понякога","Понякога ","Понякога","Да","Не","Да","Понякога","Понякога","Не","Да","Понякога","Понякога","П
онякога","Понякога","Понякога","Понякога","Понякога","Понякога","Понякога","Понякога","Понякога","Понякога","П
онякога","Не","Понякога","Да","Понякога","Понякога","Да","Не","Понякога","Понякога","Да","Да","Понякога","Поняк
ога","Понякога","Да","Понякога","Не","Понякога","Понякога","Понякога","Понякога","Понякога","Да","Понякога","По
някога","Понякога","Да","Не","Понякога","Не","Понякога","Да","Понякога","Понякога","Понякога","Понякога","Не","
Понякога","Понякога","Не","Понякога","Понякога","Не","Понякога","Понякога","Понякога","Понякога","Да","Да","Не","Не","Да","
Да","Да","Да","Понякога","Да")

❖ Анализ:

Тук отново има категорийни променливи, затова използвам функцията `table()`, а с `prop.table()` взимам процентното разпределение.

```
table cinema <- table(cinema)
```

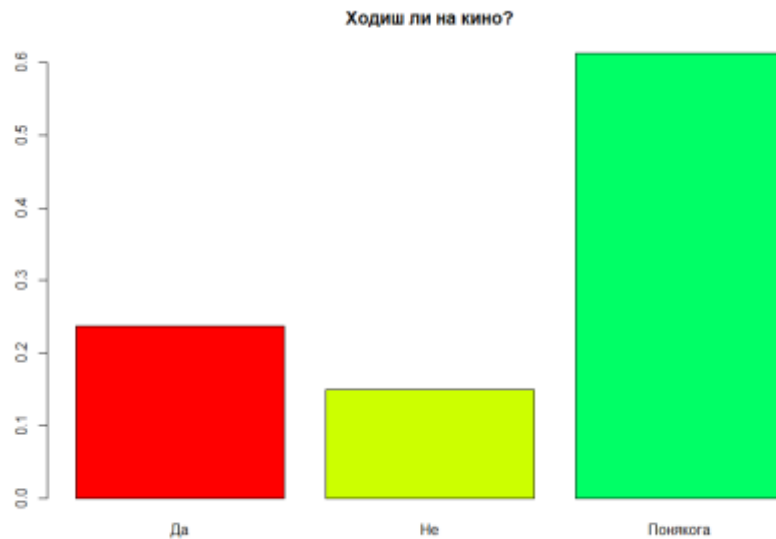
сінема		
Да	Не	Понякога
19	12	49

```
prop_table_cinema <- prop.table(table_cinema)
```

сінема		
Да	Не	Понякога
0.2375	0.1500	0.6125

Използвам `barplot` за да представя графично честотното разпределение

```
barplot(height = prop_table_cinema, col = rainbow(5), main = "Ходиш ли на кино?")
```



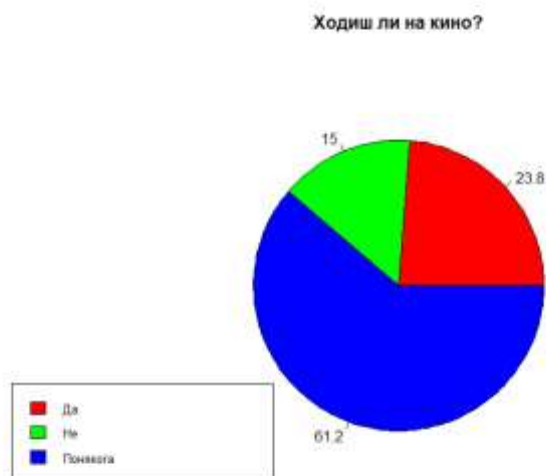
Както се вижда, по-голямата част от анкетираните са отговорили, че понякога ходят на кино. Само 12 човека от тях не посещават киното.

Използвам piechart за да изобразя процентното разпределение на получените данни.

```
piepercent_cinema <- round(100*table_cinema/sum(table_cinema), 1)
```

```
pie(table_cinema, labels = piepercent_cinema, main = "Ходиш ли на кино?", col = rainbow(n = length(table_cinema)))
```

```
legend(x = "bottomleft", legend = c("Да", "Не", "Понякога"), cex = 0.8, fill = rainbow(length(table_cinema)))
```



Въпрос 10: Каква сума обикновено заплащаш в киното?

❖ Въвеждане на данните:

Тъй като на този въпрос не са отговорили всички от анкетираните, то той няма да бъде включен в data frame-а.

```
pay_cinema <-  
c(15,10,25,6,20,15,15,10,10,15,10,10,20,10,15,20,6,15,20,15,35,30,10,12,40,15,40,40,20,12,15,20,12,15,10,15,10,40,15,15,  
13,10,10,15,13,13,9,25,20,10,20,10,10,22,15,10,10,14,10,20,5,8,20,30)
```

❖ Анализ:

Намирам модата - най-често срещаната стойност във вектора.

```
modeFunction <- function(x) {  
  
  res_table <- table(x)  
  
  return(names(res_table)[res_table == max(res_table)])  
  
}  
  
modeFunction(pay_cinema)  
  
[1] "10"
```

След прилагане на modeFunction установявам, че най-често срещаната стойност е 10.

Прилагам summary - описателна статистика за центъра на разпределението. Тя дава информация за минималната стойност, 1-вия квартил, медиана (2-рия квартил), 3-тия квартил и максималната стойност.

```
summary(pay_cinema)  
  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
  5.00   10.00   15.00   16.33   20.00   40.00
```

Намирам вариацията (дисперсията):

```
var(pay_cinema)  
  
[1] 72.19221
```

След това намирам стандартното отклонение. То е оценка на вариацията, която показва колко далече са наблюденията от очакването. За разлика от

обхвата, взема под внимаванеи всички наблюдения. Стандартното отклонение е производно на вариацията и представлява корен квадратен от дисперсията.

```
sd(pay_cinema)
```

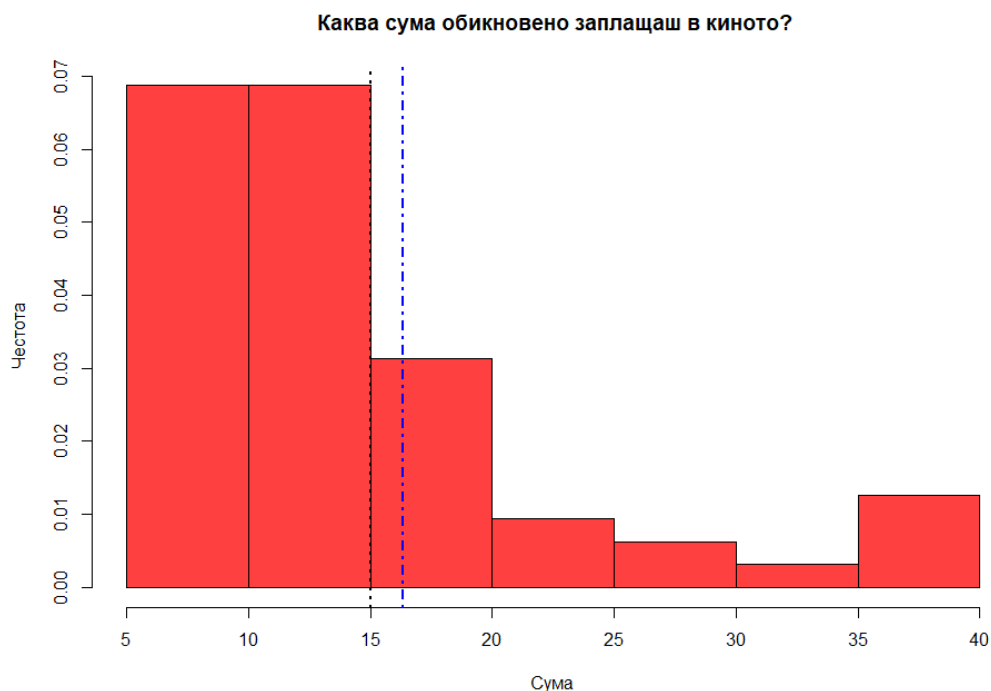
```
[1] 8.4966
```

Използвам хистограма, за да представя разпределението на непрекъснати те променливи. Синята вертикална прекъсната линия показва къде се намира средната стойност, а черната – медианата.

```
hist(pay_cinema, main = " Каква сума обикновено заплащаш в киното?",  
xlab = "Сума", ylab = "Честота", col = "brown1", prob = T)
```

```
abline(v = mean(pay_cinema), lwd = 2, lty = 4, col = "blue")
```

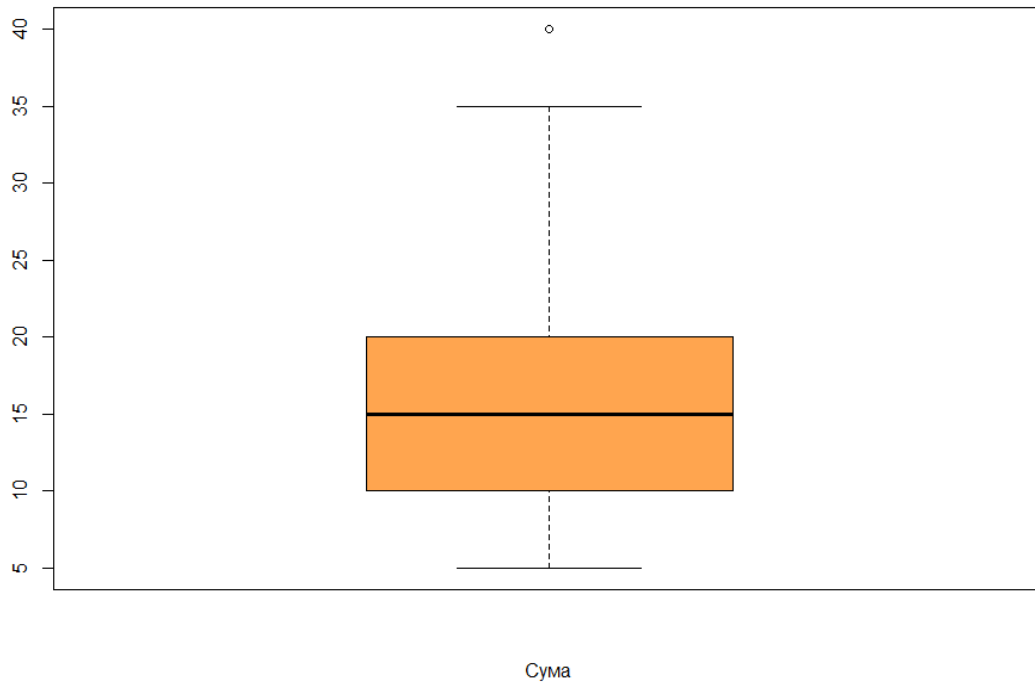
```
abline(v = median(pay_cinema), lwd = 2, lty = 3, col = "black")
```



Използвам boxplot за откриването на потенциални outlier-и.

```
boxplot(pay_cinema, col = "tan1", main = "Каква сума обикновено заплащаш в киното?", xlab =  
"Сума")
```

Каква сума обикновено заплащаш в киното?



Открит е потенциален outlier.

🚩 Въпрос 11: Би ли заменил гледането на филми за четене на книги?

❖ Въвеждане на данните:

```
books_or_movies <-
c("Не", "Не", "Да", "Не", "Да", "Да", "Да", "Да", "Да", "Не", "Не", "Да", "Да", "Не", "Не", "Да", "Не", "Да", "Не", "Да", "Не", "Да",
"Да", "Да", "Да", "Да", "Да", "Не", "Да", "Не", "Да", "Да", "Не", "Не", "Не", "Да", "Да", "Да", "Да", "Да", "Да", "Не", "Да", "Не",
"Не", "Да", "Не", "Да", "Не", "Да", "Не", "Не", "Да", "Не", "Да", "Да", "Да", "Да", "Да", "Да", "Не", "Да", "Не", "Не", "Да", "Да", "Да",
"Не", "Да", "Да", "Не", "Да", "Да", "Не", "Да", "Да", "Не", "Да")
```

❖ Анализ:

Тук отново има категорийни променливи, затова използвам функцията `table()`, а с `prop.table()` взимам процентното разпределение.

```
table_books_or_movies <- table(books_or_movies)
```

```
books_or_movies
да не
50 30
```

```
prop_table_books_or_movies <- prop.table(table_books_or_movies)
```

```
books_or_movies
  да    не
0.625 0.375
```

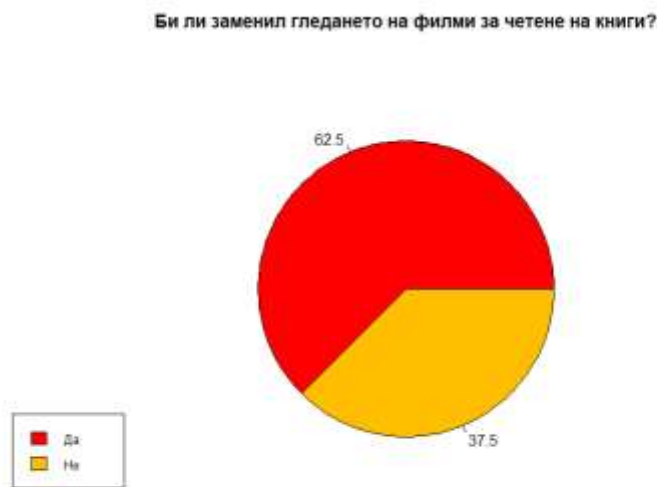
Използвам piechart за да изобразя процентното разпределение на получените данни.


```
piepercent_books_or_movies<- round(100*table_books_or_movies/sum(table_books_or_movies),
1)
```

```
pie(table_books_or_movies, labels = piepercent_books_or_movies, main = "Би ли заменил
гледането на филми за четене на книги?", col = rainbow(n = length(table_books_or_movies)*4))
```

```
legend(x = "bottomleft", legend = c("Да", "Не"), cex = 0.8, fill =
rainbow(length(table_books_or_movies)*4))
```

Както се вижда, по-голямата част от анкетираните биха заменили гледането на филми за четене на книги. Това води до извода, че независимо от бързо развиващите се технологии, хората не са обърнали гръб на книгите.



 Обединявам анализираните досега данни в един data frame.

```
moviesDF <-
data.frame(watching, hours_watching, movies_or_serials, movies_count, ideal_movie_length, cinema, books_or_movies)
```


2. Анализ на взаимодействието между две променливи

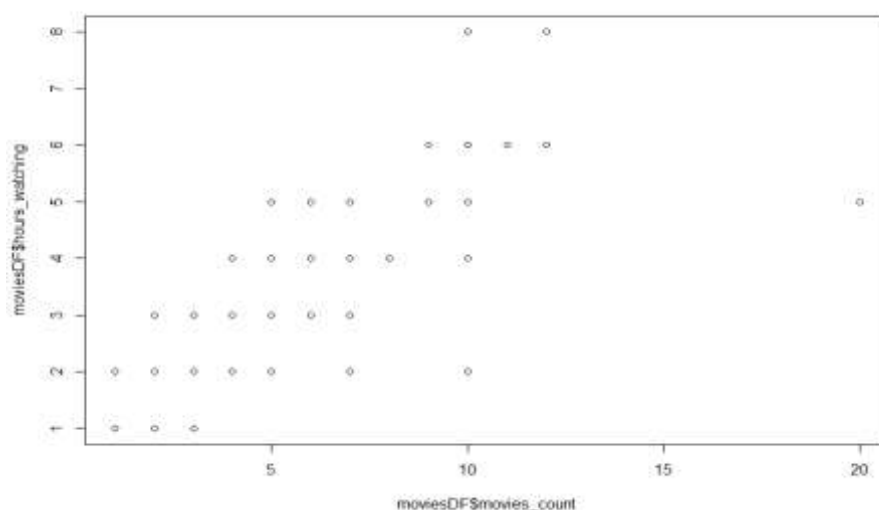
Числова VS числова

Ще анализирам връзката между променливите `movies_count` (брой филми, които гледат хората на седмица) и `hours_watching` (брой часове на ден, които отделят хората за гледане на филми).

`moviesDF$movies_count` VS `moviesDF$hours_watching`

Представям графично връзката използвайки `plot`:

```
plot(moviesDF$movies_count, moviesDF$hours_watching)
```



От графиката установявам, че съществува положителна линейна връзка. Следователно, ще използвам линейен модел за моделиране на връзката.

Прва корелационен анализ, който има за цел да измери силата на линейната връзка между двете променливи. Коефициентът на корелация (ρ) принадлежи на интервала $[-1, 1]$, а абсолютната му стойност определя силата на връзката.

Командата за корелация е `cor` (връща само едно число - корелацията между двете променливи) . С нея ще изследвам връзките между N-мерните числови данни.

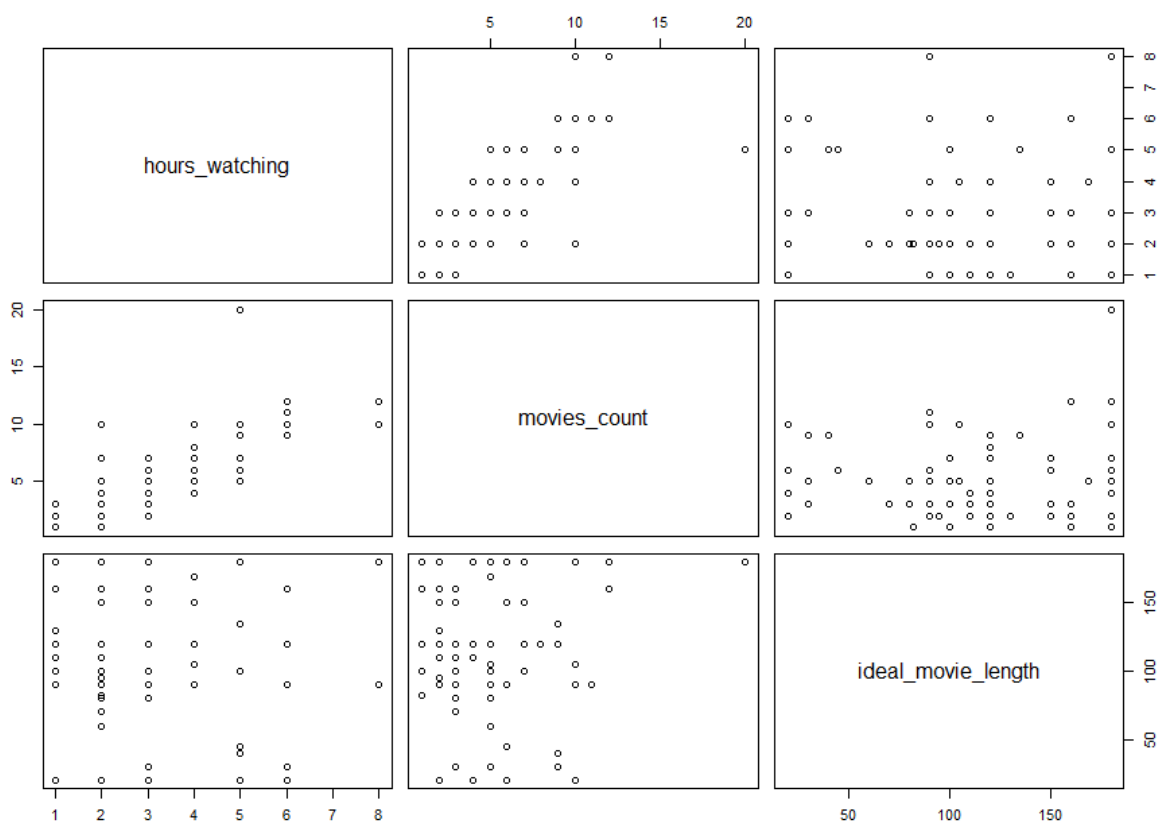
```
rho <- round(cor(moviesDF$movies_count, moviesDF$hours_watching), 3)
```

```
[1] 0.766
```

От получената стойност установявам, че има силна корелация между променливите `hours_watching` и `movies_count`.

Искам да изследвам корелациите между отделните променливи - `hours_watching`, `movies_count` и `ideal_movie_length`. Първо разглеждам графичното представяне между всеки две променливи.

```
pairs(moviesDF[, c("hours_watching", "movies_count", "ideal_movie_length")])
```



След това изследвам корелациите между тях:

```
cor(moviesDF[, c("hours_watching", "movies_count", "ideal_movie_length")])
```

	hours_watching	movies_count	ideal_movie_length
hours_watching	1.000000000	0.76553969	-0.001209932
movies_count	0.765539690	1.000000000	0.059037358
ideal_movie_length	-0.001209932	0.05903736	1.000000000

В получената симетрична матрица всички стойности по главния диагонал са единици което е следствие от формулата.

От тук ясно се вижда, че най-силна зависимост има между променливите `movies_count` и `hours_watching`. Разглеждам линейна регресия относно тези два въпроса, за да проуча и обобща връзките между посочените множества от непрекъснати променливи.

Прилагам функцията за линейна регресия `lm()`.

```
> model <- lm(movies_count ~ hours_watching, data=moviesDF)
> model

Call:
lm(formula = movies_count ~ hours_watching, data = moviesDF)

Coefficients:
(Intercept)  hours_watching
   -0.05992      1.60324
```

След като съм построила линеен модел, проверявам до колко този модел описва добре данните и какви са оценките на коефициенти му.

```
> summary(model)

Call:
lm(formula = movies_count ~ hours_watching, data = moviesDF)

Residuals:
    Min       1Q   Median       3Q      Max
-2.9563 -1.1466 -0.3530  0.8534 12.0437

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.05992    0.52602  -0.114    0.91
hours_watching  1.60324    0.15257  10.509 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.161 on 78 degrees of freedom
Multiple R-squared:  0.5861,    Adjusted R-squared:  0.5807
F-statistic: 110.4 on 1 and 78 DF,  p-value: < 2.2e-16
```

В резултат се показват 3 таблици. Първата от тях е `Residuals` - статистика за остатъците, `Coefficients` - коефициентите и `Residual standard error` - до колко линейната регресия работи добре.

Таблицата Residuals - тя дава информация за минималната стойност, 1-вия квантил, медиана (2-рия квантил), 3-тия квантил и максималната стойност.

Таблицата Coefficients по редове показва участващите коефициенти. Първата колона показва оценката, втората - стандартната грешка с която се построява доверителен интервал, третата колона представлява частното на оценката и стандартната грешка, а в последната колона се намерат стойностите на p-value.

Първо ще проверя дали коефициентите са статистически значими - дали е необходимо да участват в анализа. За всеки един коефициент проверявам хипотезата дали той е равен на 0. За да бъде значим трябва да бъде отхвърлена тази хипотеза. За да се отхвърли H_0 , то стойността на p-value трябва да бъде по-малка от 0.05.

Разглеждам оценките пред коефициента Intercept . Оценката е -0.05992. Той е статистически незначим, защото стойността на p-value е $0.91 > 0.05$. Тоест, този коефициент може да отпадне от анализа.

Чрез третата таблица проверявам до колко модела описва добре данните. Разглеждам статистиките Multiple R-squared или Adjusted R-squared. Тези стойности са в интервала [0-1].

Стойността на Adjusted R-squared е 0.5807, тоест не може да бъде направен изводът че моделът описва много добре данните.

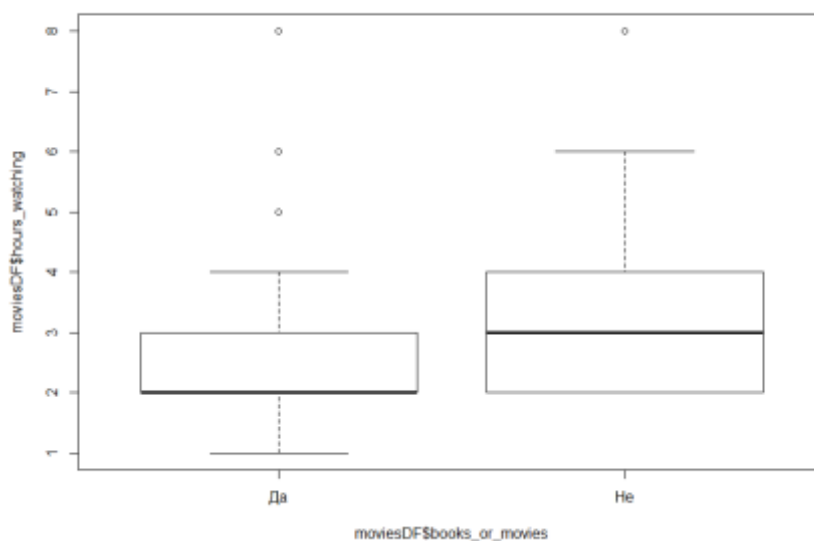
Категорийна VS числова / Числова VS категорична

Ще изследвам връзките между променливите hours_watching (брой часове на ден, които отделят хората за гледане на филми) и books_or_movies (дали хората биха заменили гледането на филми за четене на книги).

moviesDF\$hours_watching VS moviesDF\$books_or_movies

Изследвам данните графично, използвайки boxplot.

```
hours_watching_vs_books_or_movies <- boxplot(moviesDF$hours_watching ~  
moviesDF$books_or_movies)
```



Средната дебела линия във всеки един boxplot е медианата, страните на правоъгълника са 1 и 3-ти квантил, а дължината на опашките са минималната и максималните стойности. От графиката става ясно, че стойностите на втората група са по-големи, защото медианата, максималната стойност и третия квантил за втората група са по-големи от тези на първата.

Искам да изследвам дали има някаква значима разлика в средния брой часове на ден, които отделят хората за гледане на филми за различните предпочитания относно филмите и книгите. За тази цел ще създам два вектора, които ще пазят информация за двете групи, които искам да изследвам - един, който ще съдържа часовете за гледане на филми на всички, отговорили, че биха заменили гледането на филми за четене на книги и аналогично за тези, които не биха ги заменили.

```
said_yes <- moviesDF$hours_watching[moviesDF$books_or_movies == 'Да']
```

```
said_no <- moviesDF$hours_watching[moviesDF$books_or_movies == 'Не']
```

Първо ще започна с изследването дали двата вектора са нормално разпределени, за тази цел ще приложим `shapiro.test`. Нивото на съгласие е $\alpha = 0.05$. Нулевата хипотеза на теста (H_0) е, че разпределението е нормално, а алтернативната, че не е нормално разпределено. Ако p-value е по-малко от нивото на съгласие - тогава отхвърлям H_0 и приемам алтернативната хипотеза.

```
> shapiro.test(said_yes)

      shapiro-wilk normality test

data:  said_yes
W = 0.85213, p-value = 1.768e-05

> shapiro.test(said_no)

      shapiro-wilk normality test

data:  said_no
W = 0.87576, p-value = 0.002256
```

След прилагане на теста установявам, че и двата вектора не са нормално разпределени тъй като стойностите им за p-value са по-малки от нивото на съгласие и отхвърлям нулевата хипотеза. От тук следва, че трябва да бъде използван непараметричен тест. Тестът, който ще използвам е wilcoxon test.

Хипотези:

$H_0: E(\text{said_yes}) - E(\text{said_no}) = 0$ - няма разлика

$H_1: E(\text{said_yes}) - E(\text{said_no}) \neq 0$ - има разлика

```
> wilcox.test(hours_watching ~ books_or_movies, data = moviesDF, conf.int = TRUE, exact = FALSE)

      wilcoxon rank sum test with continuity correction

data:  hours_watching by books_or_movies
W = 488.5, p-value = 0.007505
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1.000025e+00 -7.898998e-05
sample estimates:
difference in location
 -0.9999517
```

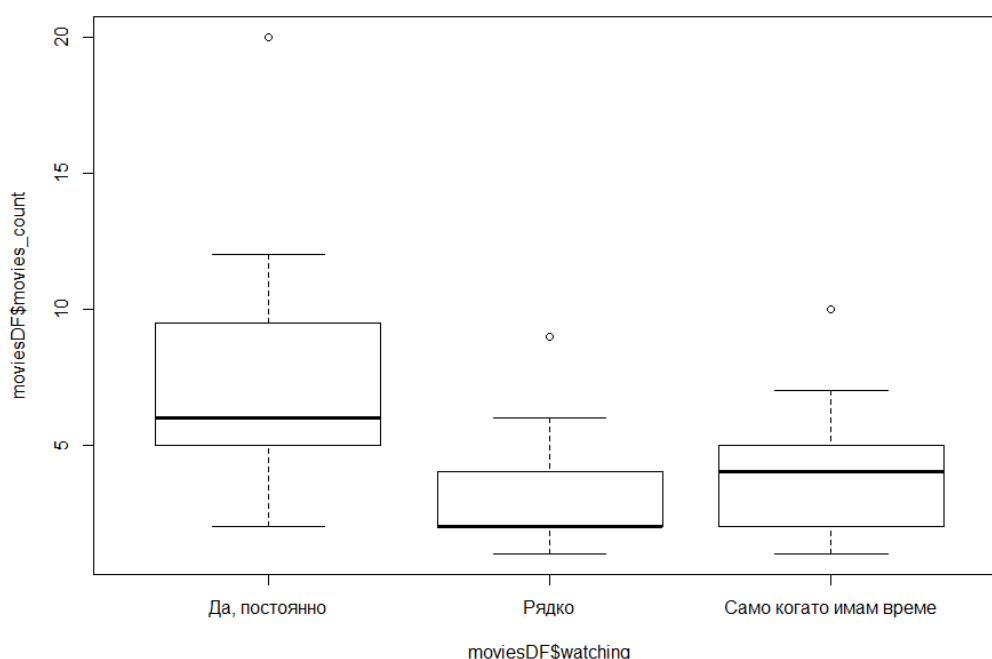
Стойността на p-value за теста е $0.007505 < 0.05 = \alpha$, следователно отхвърлям H_0 . Тоест съществува статистически значима разлика между очакваните стойности за часовете за гледане на филми, които отделят хората в двете различни категории. Анкетираните, които са отговорили, че не биха заменили гледането на филми за четене на книги отделят по-голяма част от времето си на ден за гледане на филми.

Ще изследвам локациите на разпределенията при повече от две групи. Искam да изследвам дали съществува значима разлика в очакването за броя гледани филми на седмица между различните категории за това колко често хората гледат филми.

moviesDF\$movies_count VS moviesDF\$watching

Изследвам данните графично, използвайки boxplot.

```
movies_count_vs_watching <- boxplot(moviesDF$movies_count ~ moviesDF$watching)
```



Средната дебела линия във всеки един boxplot е медианата, страните на правоъгълника са 1 и 3-ти квантил, а дължината на опашките са минималната и максималните стойности. От графиката става ясно, че стойностите на първата група са най-големи, защото медианата, максималната стойност, минималната стойност, първия и третия квантил за тази група са по-големи в сравнение с останалите две групи.

За изследването на разлика между локациите на повече от две групи ще използвам One-way ANOVA (параметричен тест) или Kruskal тест (непараметричния еквивалент на ANOVA).

One-way ANOVA има своите първоначални предположения, които ако бъдат нарушени трябва да бъде използван Kruskal тест.

Предположения

1. За всяка една група, разпределението на стойностите трябва да бъде нормално разпределена
2. Статистически еднаква дисперсия при всички групи (хомогенност на дисперсиите)

Ще започна с изследване на разпределението на данните по различните групи с функцията aggregate. Като функция за агрегация ще използвам теста на Shapiro-wilk.

```
aggregate(movies_count ~ watching, data = moviesDF, FUN = function(x) {shapiro.test(x)$p.value})
```

```
      watching movies_count
1      да, постоянно 0.0174779585
2      Рядко 0.0005776424
3 Само когато имам време 0.0003200209
```

Минималната стойност p-value за трите групи е $0.0003200209 < 0.05 = \alpha$ => отхвърлям H_0 => приемам, че и трите групи не са нормално разпределени.

Хомогенността на дисперсиите ще проверя с помощта на теста на bartlett, с нулева хипотеза за равенство на дисперсиите между различните групи.

```
bartlett.test(movies_count ~ watching, data = moviesDF)
```

```
      Bartlett test of homogeneity of variances
data:  movies_count by watching
Bartlett's K-squared = 8.5917, df = 2, p-value = 0.01363
```

Стойността на p-value е $0.01363 < 0.05 = \alpha$ => съществува статистически значима разлика между дисперсиите.

От получените резултати установявам, че трябва да използвам Kruskal тест.

```
kruskal.test(movies_count ~ watching, data = moviesDF)
```

```
      kruskal-wallis rank sum test
```



```
data: movies_count by watching  
kruskal-wallis chi-squared = 19.015, df = 2, p-value = 7.429e-05
```

Стойността на p-value за теста е $7.429e-05 \ll \alpha = 0.05 \Rightarrow$ съществува статистически значима разлика между групите.

Post-hoc анализ за Kruskal-Wallis тест:

```
> pairwise.wilcox.test(moviesDF$movies_count, moviesDF$watching,  
+                       p.adjust.method = "BH", exact = FALSE)  
  
Pairwise comparisons using wilcoxon rank sum test  
  
data: moviesDF$movies_count and moviesDF$watching  
  
      Да, постоянно Рядко  
Рядко      0.00047      -  
Само когато имам време 0.00137      0.02765  
  
P value adjustment method: BH
```

В получената таблица са записани стойностите на p-value при изследването на разликите между групите. Така статистически значима разлика се получава при категориите (Да, постоянно, Рядко), (Да, постоянно, Само когато имам време), (Рядко, Само когато имам време).

3. Заключение

След анализа на всички въпроси от анкетата, смея да твърдя, че филмите са силно предпочитано удоволствие сред хората. Всеки от анкетираните отделя част от времето си, за да изгледа някой интересен и завладяващ филм. Установих, че има връзка между броя филми, които гледат хората на седмица и броя часове на ден, които отделят за гледане на филми.