Moce testow + ANOVA

Jakub Ignatik

15 listopada 2018

Rozdział I: Moc testów normalności

Wprowadzenie

Zaleca się odtwarzanie pliku w formacie HTML

Celem tej części projektu jest sprawdzenie mocy testów normalności dla danych z trzech rozkładów: t-Studenta, jednostajnego, i wykładniczego. Wykorzystam do tego pięć testów: test Shapiro-Wilka, Jarque-Bera, Kołmogorowa, chi-kwadrat i Andersona Darlinga. Projekt będzie opierać się na napisaniu funkcji obliczającej moce testów dla rozkładów o zadanej długości, a następnie zwizualizowaniu powyższego. Każdy z testów opiera się na innej własności rozkłądu normalnego, więc ciekawie będzie wyglądało ich starcie z podanymi wyżej rozkładami.
Hipotezy: Z pewnością wraz ze zwiększaniem się próby moc testu będzie wzrastać. Co do rozkładów, z których będą brane dane, spodziewam się najsłabszej mocy przy testach dla rozkładu t-Studenta, gdyż spośród wszystkich trzech przypomina on najbardziej rozkład normalny. Największą z kolei moc osiągnie zapewne rozkład wykładniczy. Jeśli chodzi o testy, które wykazywać się będą największą mocą, nie mam zdecydowanych faworytów, domyślam się jednak, że we wszystkich przypadkach bardzo dobrze wypadnie test Kołmogorowa, którego moc znana mi jest z wcześniejszych projektów.
Dane: Rozważam dane o długości 5, 10, 25 oraz 50. Poziom istotności ustawiłem na 0,05, a liczbę symulacji na 1000.

Wizualizacja

Na początku załaduję biblioteki, wprowadzę podane we wprowadzeniu wartości oraz stworzę ramkę danych, w której na obecną chwilę znajdą się tylko długości prób.

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(tseries)
library(nortest)
library(reshape2)
library(stringr)
library(car)

set.seed(111)
#ilość symulacji
N <- 1000
#poziom istotności
alpha <- .05
#długości próby
sample_l <- c(5,10,25,50)

params <- expand.grid(sample_l)
names(params) <- "length"
```

Następny krok to wprowadzenie funkcji, która za argumenty bierze długość próby, rozkład dla danych oraz test normalności rozkładu.

```
moc_testu <- function(length, distribution, normtest){
  #Przypisanie nazwy użytego rozkładu do zmiennej służącej do generowania obserwacji
  eval(parse(text = paste0("used_distribution <- r", distribution)))
  #Przypisanie nazwy użytego testu do zmiennej służącej do testowania normalności
  eval(parse(text = paste0("used_normtest <- ", normtest, ".test")))

  #Jeśli użyty zostanie test Andersona-Darlinga, wektor długości jest wydłużany, aby spełnić wymóg doty
  if (normtest == "ad"){
    params$length = params$length + 3
  }

  #Obliczanie mocy testów
  powers <- sapply(1:nrow(params), function(p){
    length <- params[p, 1]
    p_sim <-sapply(rep(length, N), function(x){
      if (distribution == "t"){
        my_sample <- used_distribution(length, 2)
      }
      else{
        my_sample <- used_distribution(length)
      }
      if (normtest == "ks"){
        used_normtest(my_sample, pnorm)$p.value
      }
      else{
        used_normtest(my_sample)$p.value
      }
    })
    mean(p_sim < alpha)
  })

}
```

Wyniki funkcji dla każdej kombinacji rozkładu i testu umieszczam w ramce danych, którą to następnie transformuję do postaci wąskiej.

```
power <- bind_cols(params, t_shapiro = moc_testu(length, "t", "shapiro"),
                   unif_shapiro = moc_testu(length, "unif", "shapiro"),
                   exp_shapiro = moc_testu(length, "exp", "shapiro"),
                   t_jb = moc_testu(length, "t", "jarque.bera"),
                   unif_jb = moc_testu(length, "unif", "jarque.bera"),
                   exp_jb = moc_testu(length, "exp", "jarque.bera"),
                   t_ks = moc_testu(length, "t", "ks"),
                   unif_ks = moc_testu(length, "unif", "ks"),
                   exp_ks = moc_testu(length, "exp", "ks"),
                   t_chi = moc_testu(length, "t", "pearson"),
                   unif_chi = moc_testu(length, "unif", "pearson"),
                   exp_chi = moc_testu(length, "exp", "pearson"),
                   t_ad = moc_testu(length, "t", "ad"),
                   unif_ad = moc_testu(length, "unif", "ad"),
                   exp_ad = moc_testu(length, "exp", "ad"))
```

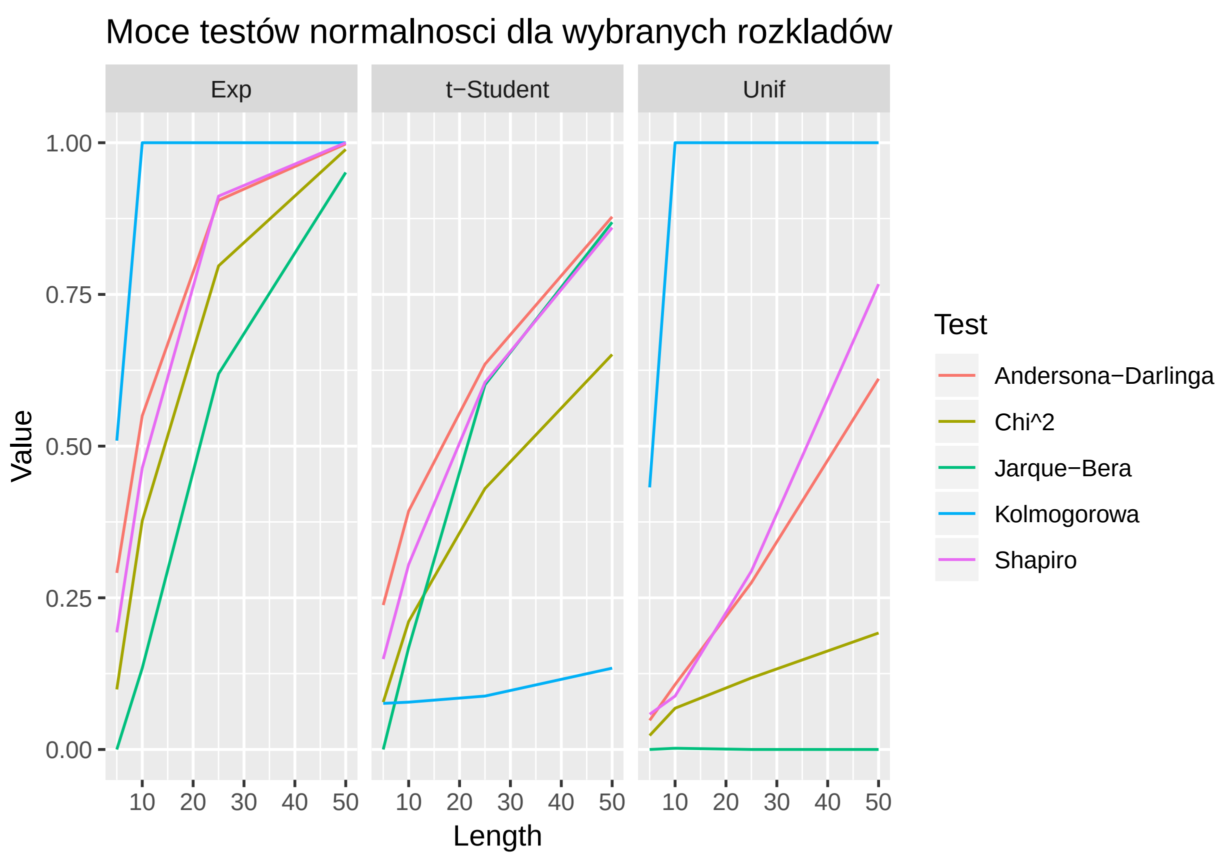
```
power2 <- melt(power, "length")
```

Stworzę teraz wykres, który przedstawi moce testów w rozdzieleniu na każdy z rozkładów. Najpierw jednak w ramce danych utworzę dwie kolumny, a w każdej z nich znajdzie się nazwa użytego testu i rozkładu dla każdego z wierszy, w celu polepszenia czytelności wykresu.

```
power2[,4:5] <- ""
colnames(power2) <- c("Length","Variable","Value","Rozklad","Test")
power2[str_detect(power2[,2], "t_") == TRUE,4] <- "t-Student"
power2[str_detect(power2[,2], "exp_") == TRUE,4] <- "Exp"
power2[str_detect(power2[,2], "unif_") == TRUE,4] <- "Unif"

power2[str_detect(power2[,2], "shapiro") == TRUE,5] <- "Shapiro"
power2[str_detect(power2[,2], "jb") == TRUE,5] <- "Jarque-Bera"
power2[str_detect(power2[,2], "ks") == TRUE,5] <- "Kołmogorowa"
power2[str_detect(power2[,2], "chi") == TRUE,5] <- "Chi^2"
power2[str_detect(power2[,2], "ad") == TRUE,5] <- "Andersona-Darlinga"

power2 %>% ggplot(aes(x = Length, y = Value, color = Test)) + geom_line() +
  ggtitle("Moce testów normalności dla wybranych rozkładów") +
  facet_wrap(~Rozklad)
```



Na powyższym wykresie można na pierwszy rzut można zaobserwować dwa zjawiska:
-Moc wszystkich testów rośnie wraz ze zwiększeniem się długości próby (potwierdza to moją hipotezę ze wstępu). Uwagę z pewnością przykuwa test Kołmogorowa, który w pierwszym i ostatnim przypadku charakteryzuje się niemal pionowym wzrostem mocy, ale w przypadku rozkładu t-Studenta przyrost mocy jest niewielki. Zarówno dla rozkładu wykładniczego jak i t-Studenta podobnie wygląda przyrost mocy testów,

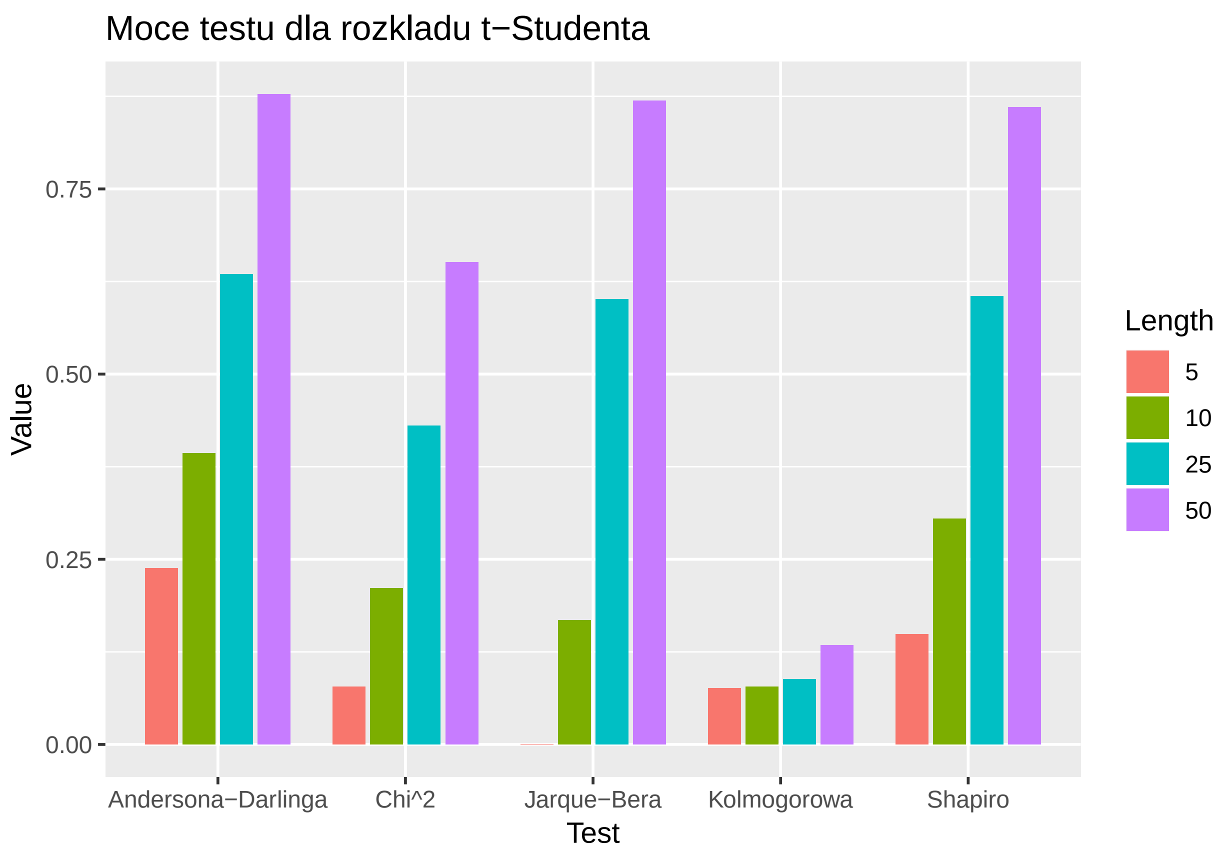
jednak to w pierwszym przypadku zostaje szybciej osiągnięta moc zbliżona do jedności, gdyż to dla rozkładu wykładniczego moce testów mają początkowo wyższe wartości. Moce dla ostatniego z rozkładóW odstają od reszty, osiągając wysoką moc znacznie później w porównaniu do swoich poprzedników.
-Testy osiągają największą moc przy rozkładzie wykładniczym (zgodnie z moją prognozą), jednak najniższa moc testu jest osiągana nie dla rozkładu t-Studenta, a dla rozkładu jednostajnego.

W kolejnej części wizualizacji zajmę się omówieniem z osobna każdego z rozkładów.

```
#Przekształcenie długości na typ "factor", aby nadać kolejność słupkom na wykresie
ord <- c("5","10","25","50")
power2$Length <- factor(power2$Length, ord)

power2 %>% filter(Rozklad == "t-Student") %>%
  ggplot(aes(x = Test, y = Value, fill = Length)) +
  geom_bar(stat = "identity", position = position_dodge(0.8), width = 0.7) +
  ggtitle("Moce testu dla rozkładu t-Studenta")
```

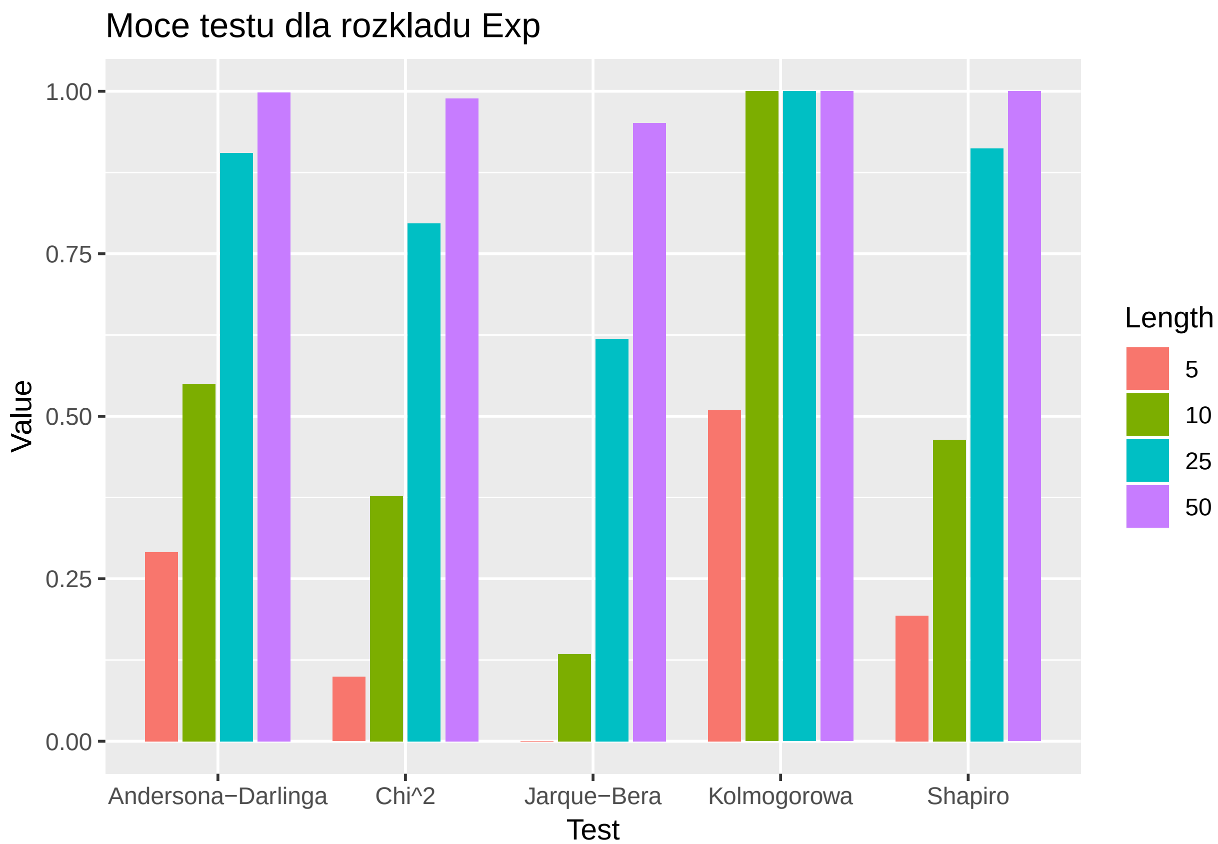
```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```



Jak wynika z wykresu, największą moc wobec danych z rozkładu t-Studenta wykazują testy Andersona-Darlinga oraz Shapiro-Wilka (należy brać poprawkę na to, że test Andersona-Darlinga jest wykonywany w rzeczywistości dla prób o długości Length+3). Test Andersona-Darlinga wykazał tak dużą moc, gdyż porównując ogony (na czym to opiera się test) rozkładu normalnego i t-Studenta z dwoma stopniami swobody (taki rozkład został wzięty do testu) można zauważyć niemałe różnice.
Najsłabiej wypadł test Kołmogorowa, gdyż opiera się on na różnicach średnich, które to są zbliżone dla rozkładu normalnego i t-Studenta.
Ciekawie prezentuje się test Jarque-Bera, który nie wykazywał się wysoką mocą przy małych próbach (odpowiednio 0 i 0,17 dla 5 i 10 obserwacji), ale już przy liczbie obserwacji równej 25 dorównał najmocniejszym

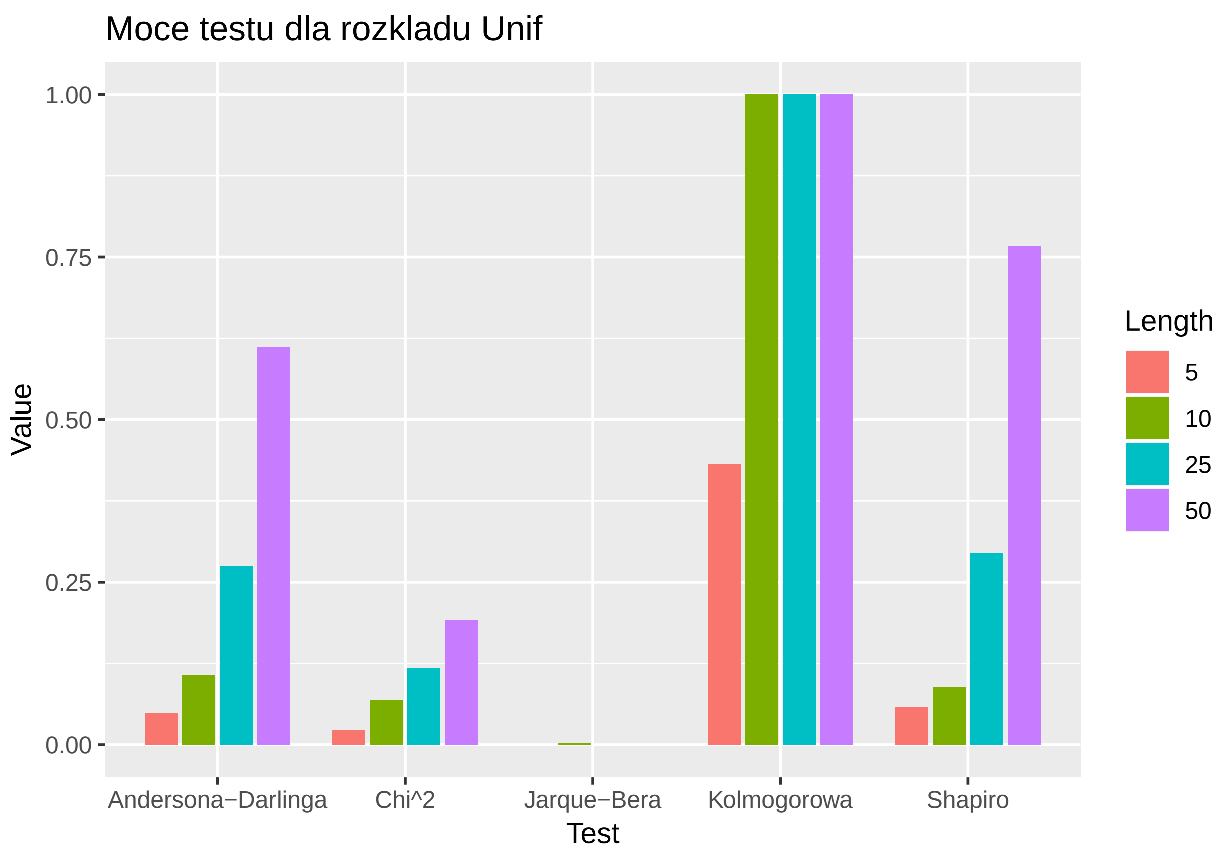
testom - Andersona-Darlinga i Shapiro-Wilka.

```
power2 %>% filter(Rozklad == "Exp") %>% ggplot(aes(x = Test, y = Value, fill = Length)) +
  geom_bar(stat = "identity", position = position_dodge(0.8), width = 0.7) +
  ggtitle("Moce testu dla rozkładu Exp")
```



Można zauważyć, że testy dla rozkładu wykładniczego wykazują się znacznie większą mocą niż testy dla rozkładu t-Studenta (co było także widoczne na pierwszym, ogólnym wykresie). Tym razem jest jeden zdecydowany faworyt, a jest to test Kołmogorowa. Średnia jest zatem tym, co zdecydowanie odróżnia oba rozkłady i to już przy niewielkiej próbie - 5 obserwacji pozwala na uzyskanie mocy równej 0,5, a zwiększenie tej liczby do 10 skutkuje uzyskaniem maksymalnej mocy.
Moce pozostałych testówW nie różnią się zbytnio między sobą, jednak wyraźnie odstaje test Jarque-Bera, który osiągnął bardzo zbliżone wyniki do testu dla rozkładu t-Studenta. Ponownie zaczyna od zerowej mocy, aby mocno przyśpieszyć przy liczbie obserwacji równej 25. Tym razem jednak okazało się to niewystarczające, aby znaleźć się w czołówce (chociaż przy liczbie obserwacji równej 50 niewiele brakuje do mocy równej 1, która to została osiągnięta przez inne testy).

```
power2 %>% filter(Rozklad == "Unif") %>% ggplot(aes(x = Test, y = Value, fill = Length)) +
  geom_bar(stat = "identity", position = position_dodge(0.8), width = 0.7) +
  ggtitle("Moce testu dla rozkładu Unif")
```



Rozkład jednostajny charakteryzuje się najmniejszą mocą testów spośród wszystkich trzech rozkładów. Przy liczbie obserwacji równej 50 większość testów miała moc równą 1, a tymczasem przy rozkładzie jednostajnym taka sytuacja występuje jedynie przy teście Kołmogorowa, który to ponownie okazał się najlepszy i to z jeszcze większym zapasem wobec innych testów.
Stawkę zamyka ponownie test Jarque-Bera, chociaż tym razem różnica jest znacznie bardziej widoczna, gdyż na wykresie możemy dostrzec tylko jeden, bliski zeru słupek dla długości równej 10. Rozkład jednostajny okazał się dla testu tak podobny do rozkładu normalnego, że moc testu oscyluje wokół zera. Oba rozkłady muszą posiadać bardzo podobną skośność i kurtozę.

Podsumowanie

Wyniki częściowo potwierdziły moje hipotezy:
Długość próby: We wszystkich przypadkach, dla każdego rozkładu i każdego testu, zwiększenie liczby obserwacji wpłynęło na zwiększenie mocy testu. Oznacza to, że zwiększenie długości próby skutkuje wzrostem mocy testu. Najbardziej jest to widoczne dla testu Jarque-Bera, gdzie dla małych prób test ma bardzo niską moc, a dla wysokich o wiele większą w dość krótkim czasie.
Rozkłady: Zgodnie z hipotezą, największą mocą charakteryzują się testy dla rozkładu wykładniczego, zatem rozkład wykładniczy jest tym, który najbardziej ze wszystkich różni się od rozkładu normalnego i najłatwiej wykryć różnice. Zaskoczyły mnie wyniki testów dla rozkładu t-Studenta, gdyż spodziewałem się dla niego najmniejszej mocy. Okazało się jednak, że różnice nie są tak trudne do wykrycia jak przy rozkładzie jednostajnym.
Testy normalności:
-Test Kołmogorowa okazał się rzeczywiście mocny, jednak nie we wszystkich przypadkach: przy rozkładzie t-Studenta okazał się tym, który wypadł najsłabiej
-Testy Shapiro-Wilka i Andersona-Darlinga okazały się być mocnymi testami, które miały zbliżoną do siebie

moc. Oba jednak wypadły gorzej w starciu z rozkładem jednostajnym. -Test Chi^2 nie został wspomniany przeze mnie ani razu podczas analizowania wykresów, gdyż nie był ani razu na wysokiej pozycji, ale nie zamykał też stawki na ostatnim miejscu. Nie jest to mocny test, jest raczej słabym testem, zwłaszcza przy małych ilościach obserwacji
-Test Jarque-Bera wypadł w tym zestawieniu najgorzej. Dla wszystkich rozkładóW wypadał słabo przy małych próbach i przyśpieszał przy większej ilości obserwacji, doganiając resztę testów, jednak przy rozkładzie jednostajnym nie wykazał żadnej mocy, jaka to sytuacja nie wystąpiła dla żadnego innego testu

Rozdział II: ANOVA

Wprowadzenie

Celem drugiej części projektu jest wykonanie badania ANOVĄ. W projekcie wykorzystam ANOVĘ jedno oraz dwuczynnikową, a także MANOVĘ.
Dane: Dane opierają się na dwóch źródłach:
-*IMDB:* Baza danych z tej strony została przeze mnie pobrana ze strony Kaggle (https://www.kaggle.com/PromptCloudHQ/imdb-data/version/1). Obejmuje ona 1000 najpopularniejszych filmów na portalu IMDB w latach 2008-2016
-*Filmweb:* Przygotowałem własnoręcznie trzy ramki danych dotyczące aktorów, aktorek i reżyserów. Za pomocą webscrappingu pobrałem z Filmwebu rankingi wyżej wymienionych osób. Każdą ramkę po kolei łączyłem z bazą IMDB na podstawie klucza w postaci imienia i nazwiska aktora/aktorki/reżysera. Po dokonanych operacjach w bazie zostały 83 kompletne rekordy. Ostatnim krokiem było własnoręczne dopisywanie do każdego z filmów zmiennej 0-1 oznaczającej zdobycie przez filmu Oscara.
Metodologia: Na potrzeby projektu zdecydowałem się na pozostawienie jednej zmiennej zależnej (Ocena) oraz trzech czynników grupujących: gatunku filmu, jego długości oraz tego, czy dany film zdobył Oscara. Uznałem za interesujące sprawdzenie, czy te właśnie parametry mają wpływ na ocenę filmu. W tym celu przeprowadzę następujące badania:
-ANOVA jednoczynnikowa (Ocena ~ Gatunek, Ocena ~ Dlugosc, Ocena ~ Oscary)
-ANOVA wieloczynnikowa (Ocena ~ Dlugosc + Oscary)
-MANOVA (Ocena + Dlugosc ~ Oscary)
Hipotezy: Jeśli chodzi o ANOVĘ jednoczynnikową, spodziewam się, że tylko liczba zdobytych Oscarów będzie miała wpływ na ocenę filmu. Zdobycie Oscara oznacza, że dany film był szczególnie doceniony przez jury, a więc mógł podobać się także szerszej publiczności. Nie sądzę, żeby długość filmu oraz jego gatunek miały taki wpływ. Udany film nie musi być długi, a gatunek nie gra roli, gdy chodzi o ocenę tego, czy film jest dobry.
Co do ANOVY dwuczynnikowej, myślę, że wykaże ona wpływ jedynie Oscarów, a model będzie pozbawiony interakcji pomiędzy zmiennymi kategorycznymi.
Myślę, że MANOVA wykaże wpływ Oscarów na oba te czynniki jednocześnie, gdyż liczba Oscarów wpływa wg mnie zarówno na ocenę filmu jak i jego długość (najdłuższe filmy często mają Oscara).

Badanie ANOVĄ

Zasady randomizacji i niezależność danych

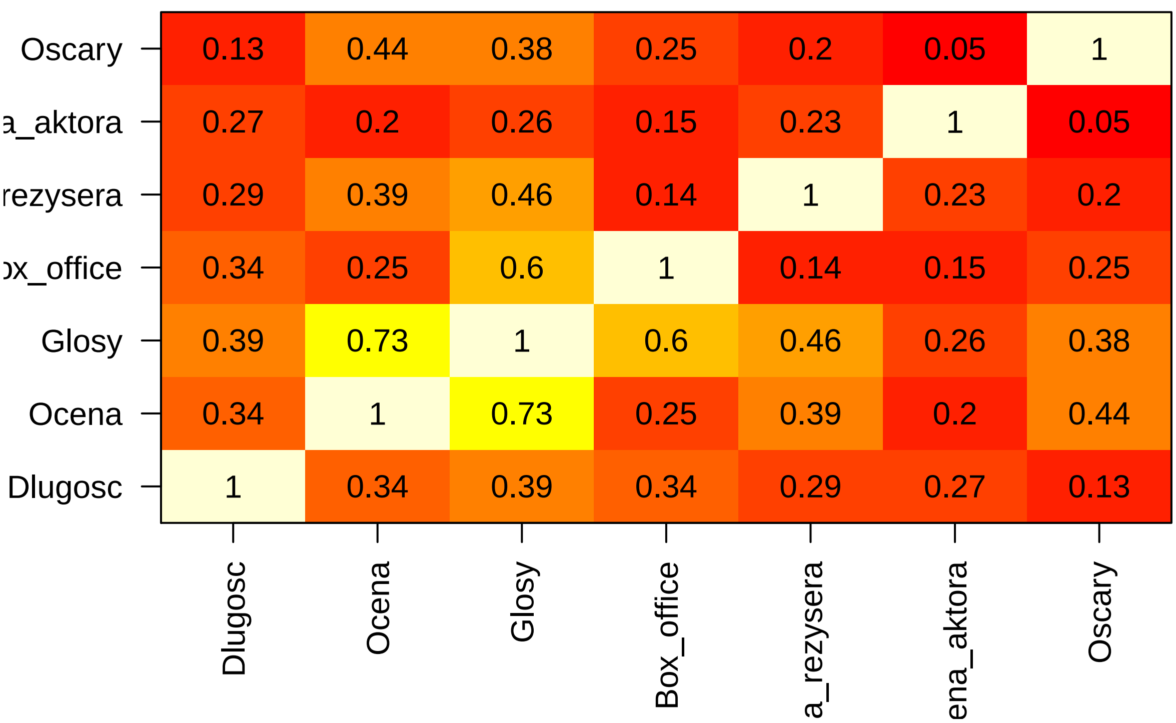
Wszystkie z modeli opisanych we wstępie spełniają obie zasady randomizacji: próbka została wybrana z populacji w sposób losowy(baza zawierała 1000 najlepszych filmów, jednak losowo wybrałem filmy do badania poprzez wykorzystanie bazy aktorów serwisu Filmweb), a elementy zostały przypisane do próbek również w sposób losowy, nie dzieliłem ich względem podobieństwa pozostałych parametrów.
Zmienna zależna Ocena oraz Długość posiadają wartości na skali przedziałowej (Ocena - od 1 do 10, Długość - od 89 do 191 minut).

Jeśli chodzi o niezależność danych, najlepiej przedstawi to macierz korelacji. Zanim jednak ją przedstawię, wgram dane potrzebne do projektu:

```
dane <- read.csv("C:/Users/Jan/Desktop/filmy2.csv", sep = ";", dec = ".")
#Większość filmów ma przypisane kilka gatunków filmowych, ja wybiorę tylko ten, który jest na pierwszym
dane$Gatunek <- gsub(",.*", "", dane$Gatunek)

#Do macierzy korelacji biorę tylko dane liczbowe
dane2 <- dane[,c(5:11)]

COR <- cor(dane2[,1:7])
image(x=seq(nrow(COR)), y=seq(ncol(COR)), z=cor(dane2[,1:7]), axes=F, xlab="", ylab="")
text(expand.grid(x=seq(dim(COR)[1]), y=seq(dim(COR)[2])), labels=round(c(COR),2))
box()
axis(1, at=seq(nrow(COR)), labels = rownames(COR), las=2)
axis(2, at=seq(ncol(COR)), labels = colnames(COR), las=1)
```



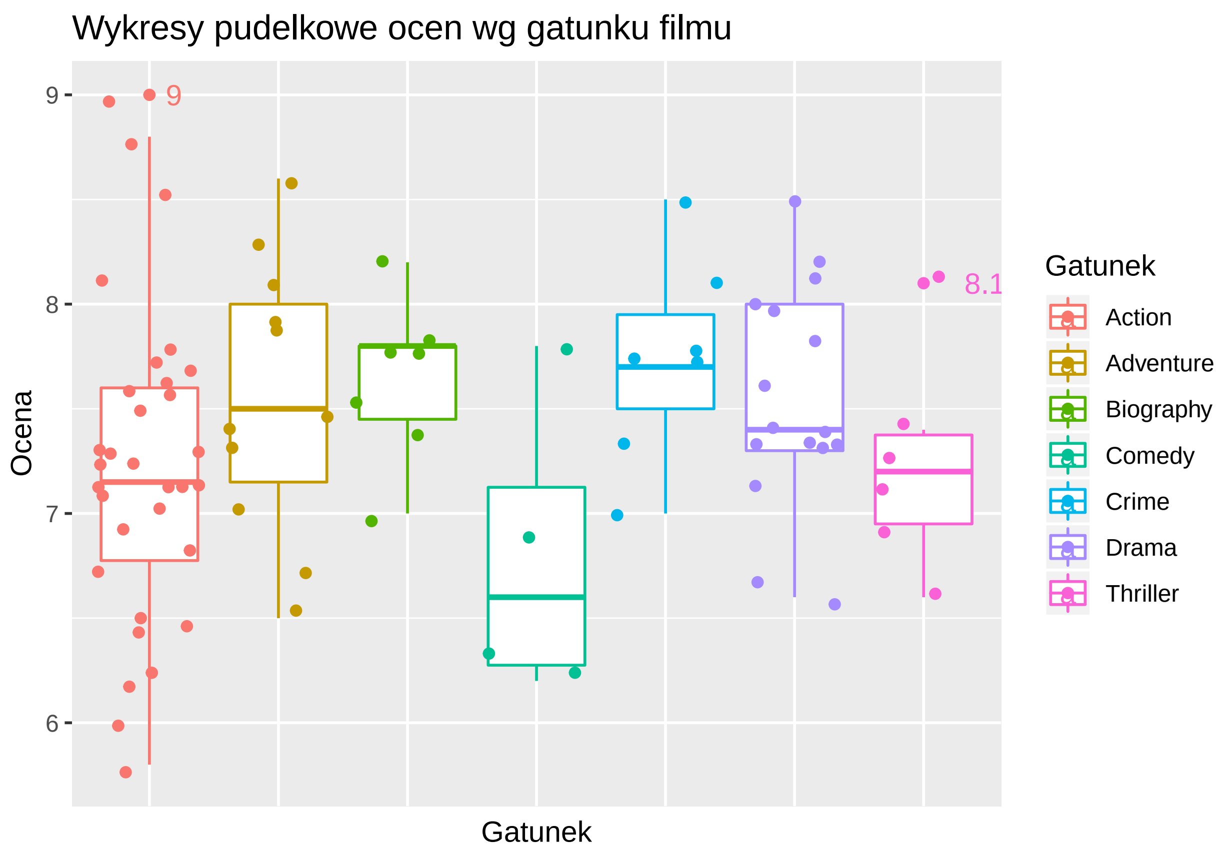
Jak wynika z macierzy korelacji, dane nie są ze sobą mocno skorelowane, korelacja jest dość słaba.

ANOVA jednoczynnikowa

Pierwszy model jaki wykonam to *Ocena ~ Gatunek*. Zbada on, czy średnie ocen różnią się w zależności od gatunku filmu. Na początku zamienię zmienną Gatunek na zmienną kategoryczną oraz zbadam różnorodność grup.

```
dane$Gatunek <- as.factor(dane$Gatunek)
```

```r
#Funkcja pozwoli na pokazanie na wykresie outlierów, gdyż nie będą one widoczne wśród innych obserwacji
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}
dane %>% group_by(Gatunek) %>% mutate(outlier = ifelse(is_outlier(Ocena), Ocena, as.numeric(NA))) %>%
  ggplot(aes(x = Gatunek, y = Ocena, color = Gatunek)) + geom_boxplot() +
  ggtitle("Wykresy pudełkowe ocen wg gatunku filmu") + geom_jitter() +
  geom_text(aes(label = outlier), na.rm = TRUE, hjust = -1) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



Z wykresu można pokusić się o wysnucie wniosku, że gatunek ma wpływ na ocenę filmu. Trzeba jednak mieć na uwadze, że wszystkie gaunki filmowe poza komedią mieszczą się w zakresie ocen 7-8, więc nie jest to drastyczna różnica. Wykres zwraca eż uwagę na to, że ilość ocen dla poszczególnych gatunków dość mocno się od siebie różni - od około 30 dla filmów akcji po 4 dla komedii. Na pewno więc nie będą stosowane testy wariancji zakładające równość podgrup.

Aby zostało spełnione założenie ANOVY o normalności zmiennej zależnej w podgrupach, przeprowadzę test Shapiro-Wilka.

```r
tapply(dane$Ocena, dane$Gatunek, shapiro.test)
```

```
## $Action
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96838, p-value = 0.4559
##
```

```
##
## $Adventure
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97379, p-value = 0.922
##
##
## $Biography
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.94673, p-value = 0.6999
##
##
## $Comedy
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.88804, p-value = 0.3741
##
##
## $Crime
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97569, p-value = 0.9361
##
##
## $Drama
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95937, p-value = 0.6505
##
##
## $Thriller
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95933, p-value = 0.8146
```

Jak widać, żaden z gatunków nie ma problemu z normalnością rozkładu zmiennej zależnej. Ostatnie założenie do zbadania dotyczy równości wariancji. ponieważ liczebności grup mocno się od siebie różnią, przeprowadzę test Bartletta.

```
bartlett.test(Ocena ~ Gatunek, data = dane)
```

```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  Ocena by Gatunek
## Bartlett's K-squared = 5.8929, df = 6, p-value = 0.4353
```

Również i to założenie zostało spełnione. Można zatem przejść do badania ANOVĄ.

```
summary(aov(Ocena ~ Gatunek, data = dane))
```

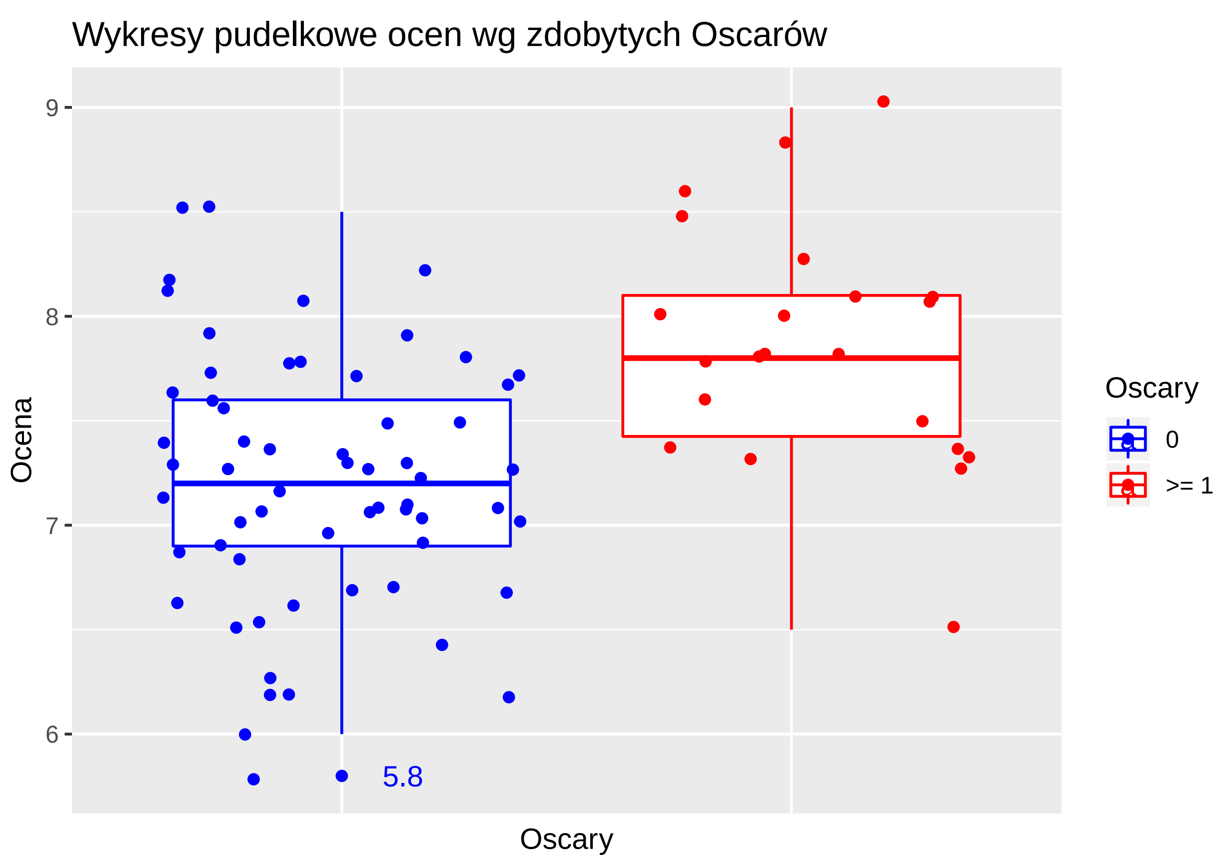
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Gatunek      6  4.508  0.7513   1.826  0.105
## Residuals   76 31.275  0.4115
```

Jak wynika z badania, pomimo tego, co widoczne było na wykresie, gatunek filmu nie ma wpływu na jego ocenę, co potwierdza moją hipotezę.

Kolejnym modelem będzie *Ocena ~ Oscary*. Podobnie jak w poprzednim przypadku, zamienię zmienną Oscary na typ factor oraz zbadam podstawowe statystyki.

```
dane$Oscary <- as.factor(dane$Oscary)

dane %>% group_by(Oscary) %>% mutate(outlier = ifelse(is_outlier(Ocena), Ocena, as.numeric(NA))) %>%
  ggplot(aes(x = Oscary, y = Ocena, color = Oscary)) + geom_boxplot() +
  ggtitle("Wykresy pudełkowe ocen wg zdobytych Oscarów") +
  scale_color_manual(labels = c("0", ">= 1"), values = c("blue", "red")) +
  guides(color=guide_legend("Oscary")) + geom_jitter() +
  geom_text(aes(label = outlier), na.rm = TRUE, hjust = -1) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



Również i w tym przypadku wykres wskazuje na różnice w ocenach filmów dla filmów oscarowych (mniej licznych, około 20) oraz tych, które Oscara nie dostały (około 60). Patrząc na wykres można wysnuć wniosek, że filmy z Oscarami są lepiej oceniane przez widzów. Też i tym razem potrzebny będzie test Bartletta.
Aby zostało spełnione założenie ANOVY o normalności zmiennej zależnej w podgrupach, przeprowadzę test Shapiro-Wilka.

```
tapply(dane$Ocena, dane$Oscary, shapiro.test)
```

```
## $`0`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.9891, p-value = 0.8645
##
##
## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97204, p-value = 0.7578
```

Jak widać, żaden z gatunków nie ma problemu z normalnością rozkładu zmiennej zależnej. Ostatnie założenie do zbadania dotyczy równości wariancji. ponieważ liczebności grup mocno się od siebie różnią, przeprowadzę test Bartletta.

```
bartlett.test(Ocena ~ Oscary, data = dane)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Ocena by Oscary
## Bartlett's K-squared = 0.065634, df = 1, p-value = 0.7978
```

Również i w tym przypadku założenie zostało spełnione. Można przejść zatem do badania ANOVĄ.

```
summary(aov(Ocena ~ Oscary, data = dane))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Oscary       1  6.946   6.946   19.51 3.07e-05 ***
## Residuals   81 28.837   0.356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tym razem wynik różni się od poprzedniego. Okazuje się, że średnie ocen dla tych filmów, które mają Oscara oraz dla tych, które go nie mają, różnią się między sobą, co tym razem znajduje odzwierciedlenie na wykresie. Ponownie zgadza się to z moją hipotezą, ale pozostaje pytanie, na ile mocny jest ten efekt.
Analiza post-hoc byłaby tutaj zbędna, gdyż zmienna kategoryczna ma tylko 2 poziomy i z góry wiadomo, które czynniki odpowiadają za różne średnie dla grup. Aby sprawdzić moc tego efektu, przeprowadzę badanie wielkości efektów eksperymentalnych. Najpierw sporządzę funkcje, które to policzą, a następnie obliczę wielkość tych efektów.

```
eta_sq <- function(aovm){
  sum_stats <- summary(aovm)[[1]]
  SSm <- sum_stats[["Sum Sq"]][1]
  SSr <- sum_stats[["Sum Sq"]][2]
  SSm/(SSm+SSr)
```

```r
}

omega_sq <- function(aovm){
  sum_stats <- summary(aovm)[[1]]
  SSm <- sum_stats[["Sum Sq"]][1]
  SSr <- sum_stats[["Sum Sq"]][2]
  DFm <- sum_stats[["Df"]][1]
  MSr <- sum_stats[["Mean Sq"]][2]
  (SSm-DFm*MSr)/(SSm+SSr+MSr)
}

eta_sq(aov(Ocena ~ Oscary, data = dane))
```

```
## [1] 0.1941174
```

```r
omega_sq(aov(Ocena ~ Oscary, data = dane))
```

```
## [1] 0.1823539
```

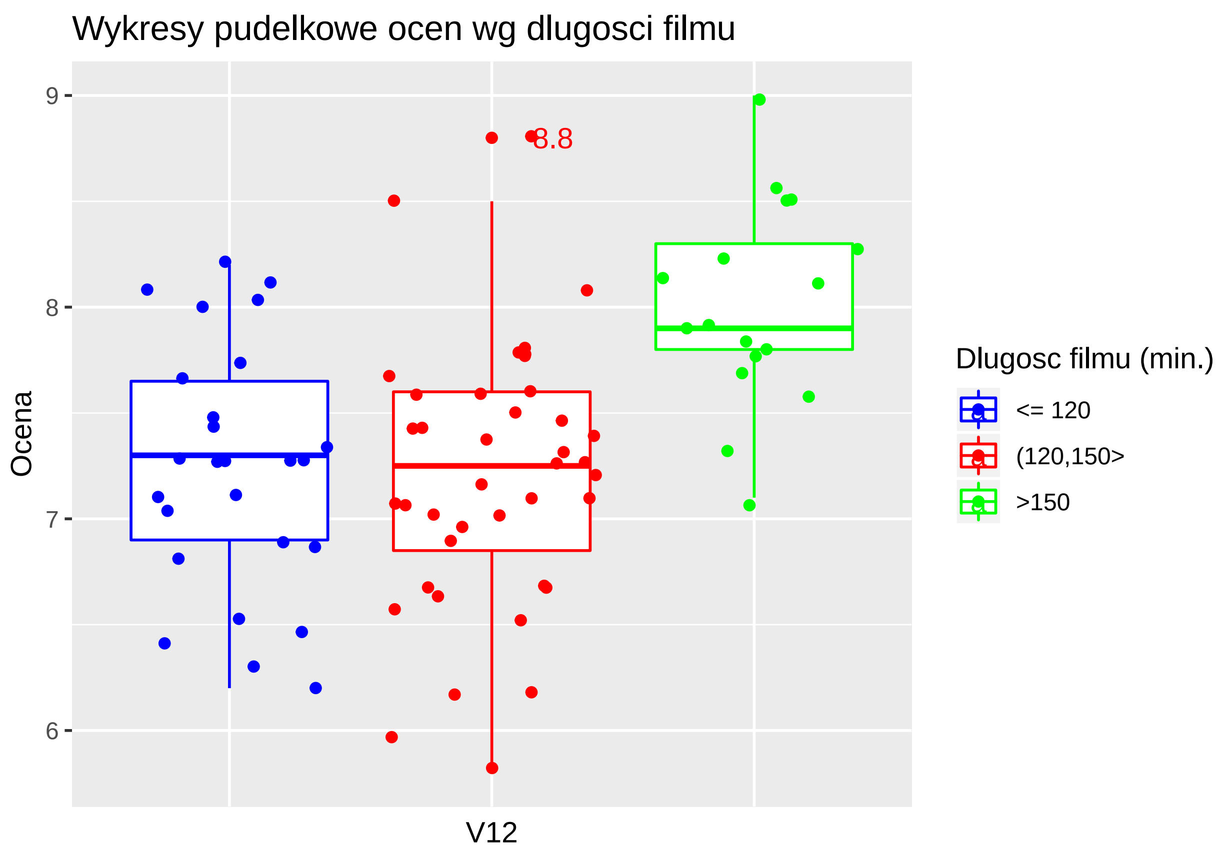
Pierwsza wielkość, eta^2 informuje o tym, jaki % zmienności oceny jest wyjaśniany przez to, czy film zdobył Oscara. W powyższym wypadku jest to 19%, więc efekt jest słaby.
Druga wielkość, omega^2 pokazuje to samo, lecz odnosi się nie do próby, a do całej populacji. W tym wypadku 18% zmienności oceny jest wyjaśniany przez oscarowość filmu dla wszystkich filmów, nie tylko tych w próbie. Efek jest również i w tym przypadku słaby.

Kolejnym modelem będzie *Ocena ~ Dlugosc*. Długość nie jest jednak zmienną kategoryczną, więc utworzę nową zmienną (V12), która podzieli filmy na 3 zbiory: <= 120 minut, (120, 150> minut oraz >150 minut. Po zainicjalizowaniu nowej zmiennej zamienię ją na typ factor oraz zbadam jej podstawowe statystyki, aby następnie utworzyć model właściwy, tj. *Ocena ~ V12*.

```r
dane[,12] <- NULL
for (i in 1:nrow(dane)){
  if (dane$Dlugosc[i] <= 120){
    dane[i,12] <- 1
  }
  else if (dane$Dlugosc[i] <= 150){
    dane[i,12] <- 2
  }
  else{
    dane[i,12] <- 3
  }
}

dane$V12 <- as.factor(dane$V12)

dane %>% group_by(V12) %>% mutate(outlier = ifelse(is_outlier(Ocena), Ocena, as.numeric(NA))) %>%
  ggplot(aes(x = V12, y = Ocena, color = V12)) + geom_boxplot() +
  ggtitle("Wykresy pudełkowe ocen wg długości filmu") +
  scale_color_manual(labels = c("<= 120", "(120,150>", ">150"), values = c("blue", "red", "green")) +
  guides(color=guide_legend("Długość filmu (min.)")) + geom_jitter() +
  geom_text(aes(label = outlier), na.rm = TRUE, hjust = -1) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



Jak widać na wykresie, najkrótsze filmy oraz te o średniej długości są do siebie bardzo zbliżone: średnia jest praktycznie identyczna, chociaż filmy z drugiej grupy mają większy rozrzut ocen. Zdecydowanie odstaje ostatnia z grup, sugerując, że najdłuższe filmy są tymi, które zgarniają najlepsze oceny. Wielkość grup jest do siebie dosyć zbliżona (26, 40, 17), jeśli porównać to z poprzednimi zmiennymi kategorycznymi.
Najpierw sprawdzę normalność zmiennej zależnej względem podgrup.

```
tapply(dane$Ocena, dane$V12, shapiro.test)
```

```
## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95569, p-value = 0.3139
##
##
## $`2`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98071, p-value = 0.716
##
##
## $`3`
##
##  Shapiro-Wilk normality test
```

```
##
## data:  X[[i]]
## W = 0.98301, p-value = 0.9795
```

Zgodnie z powyższym testem, żadna z podgrup nie ma problemóW z normalnością ocen. Czas na zbadanie kolejnego założenia ANOVY.

```
bartlett.test(Ocena ~ V12, data = dane)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Ocena by V12
## Bartlett's K-squared = 1.4112, df = 2, p-value = 0.4938
```

Założenie równości wariancji zostało spełnione. Pozostał model ANOVY.

```
summary(aov(Ocena ~ V12, data = dane))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## V12          2  8.496   4.248   12.45 1.96e-05 ***
## Residuals   80 27.287   0.341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Moja hipoteza okazała się fałszywa. Okazuje się, że wykres ponownie dał słuszną sugestię nierówności średniej w grupach. Ale czy na pewno? Możliwe jest, że są pewne różnice nie tylko względem trzeciej grupy. W tym celu przeprowadzę test post-hoc, test Tuckeya, który wskaże mi pary odpowiedzialne za powodzenie badania ANOVĄ.

```
TukeyHSD(aov(Ocena ~ V12, data = dane))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Ocena ~ V12, data = dane)
##
## $V12
##            diff        lwr       upr     p adj
## 2-1 -0.03096154 -0.3823123 0.3203892 0.9758889
## 3-1  0.77330317  0.3382837 1.2083226 0.0001713
## 3-2  0.80426471  0.4004626 1.2080668 0.0000254
```

Różnice występują dla par 2-1 oraz 3-1, zatem rzeczywiście grupa 1 i 2 nie odbiegają od siebie średnimi ocen. Jak będzie się przedstawiał efekt eksperymentalny, skoro nie ma różnic względem wszystkich podgrup?

```
eta_sq(aov(Ocena ~ V12, data = dane))
```

```
## [1] 0.2374307
```

```
omega_sq(aov(Ocena ~ V12, data = dane))
```

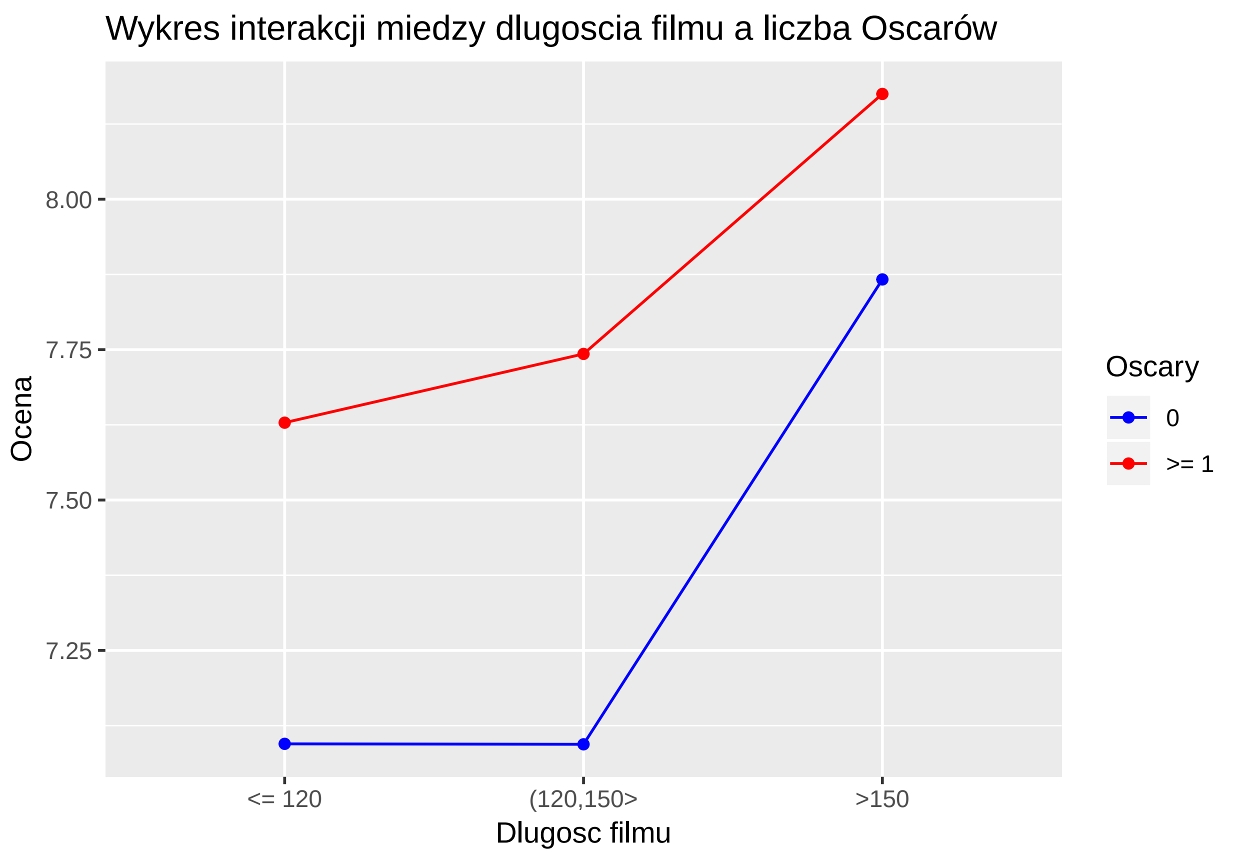
```
## [1] 0.2163046
```

Pomimo braku udziału wszystkich grup w nierówności średniej ocen, długość filmu objaśnia 24% zmienności Oceny w próbie. W populacji wynik jest niewiele gorszy i wynosi 22%. Jednak ponownie jest to efekt słaby.

ANOVA dwuczynnikowa

Model, jakim zajmę się w tej części projektu, poświęconej ANOVIE dwuczynnikowej, to *Ocena ~ V12 + Oscary*. Dlaczego w modelu jest znak (+) zamiast (*), wyjaśni poniższy wykres.

```
dane %>% ggplot(aes(x = V12, y = Ocena, color = Oscary, group = Oscary)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line") +
  scale_color_manual(labels = c("0", ">= 1"), values = c("blue", "red")) +
  xlab("Długość filmu") + scale_x_discrete(labels = c("<= 120", "(120,150>", ">150")) + ggtitle("Wykres
```



Wykres interakcji to wykres, który pokazuje łączny wpływ działania kilku czynników (tzn., czy jeden czynnik wpływa na zmienną zależną zależnie od drugiego czynnika). Można zauważyć, że Oscary i V12 to czynniki addytywne, nieposiadające efektu interakcji. Dlatego właśnie w modelu pojawił się znak (+), który oznacza jej brak.

Zanim przejdę do badania, należy zbadać normalność w podgrupach, których jest 6 (3 poziomy długości * 2 poziomy Oscarów).

```
podgrupy <- split(dane,list(dane$Oscary,dane$V12))
#w szóstej kolumnie są podane oceny filmów
for (i in 1:6){
  print(shapiro.test(podgrupy[[i]][,6]))
}
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  podgrupy[[i]][, 6]
## W = 0.96666, p-value = 0.7083
##
##
##  Shapiro-Wilk normality test
##
## data:  podgrupy[[i]][, 6]
## W = 0.81379, p-value = 0.05593
##
##
##  Shapiro-Wilk normality test
##
## data:  podgrupy[[i]][, 6]
## W = 0.98753, p-value = 0.9621
##
##
##  Shapiro-Wilk normality test
##
## data:  podgrupy[[i]][, 6]
## W = 0.79495, p-value = 0.0365
##
##
##  Shapiro-Wilk normality test
##
## data:  podgrupy[[i]][, 6]
## W = 0.97341, p-value = 0.9223
##
##
##  Shapiro-Wilk normality test
##
## data:  podgrupy[[i]][, 6]
## W = 0.98462, p-value = 0.9821
```

Można zauważyć, że dla i=4 jest lekki problem z normalnością rozkładu. Oto grupa, która wywołała ten problem:

```
podgrupy[[4]]
```

```
##                Aktor           Rezyser
## 13 Leonardo DiCaprio Christopher Nolan
## 18    Bradley Cooper    Clint Eastwood
## 37    Eddie Redmayne       David Yates
## 45    Christian Bale        Adam McKay
## 49  Daniel Day-Lewis  Steven Spielberg
## 52    Bradley Cooper  David O. Russell
## 69         Tom Hanks  Steven Spielberg
##                                  Tytul   Gatunek Dlugosc Ocena   Glosy
## 13                           Inception    Action     148   8.8 1583625
## 18                      American Sniper    Action     133   7.3  353305
## 37 Fantastic Beasts and Where to Find Them Adventure     133   7.5  232072
## 45                        The Big Short Biography     130   7.8  246360
## 49                              Lincoln Biography     150   7.4  207497
## 52                Silver Linings Playbook    Comedy     122   7.8  564364
## 69                      Bridge of Spies     Drama     142   7.6  217938
##    Box_office Ocena_rezysera Ocena_aktora Oscary V12
```

```
## 13     292.57           8.77          8.96      1   2
## 18     350.12           8.27          8.22      1   2
## 37     234.02           6.29          8.76      1   2
## 45      70.24           7.62          8.61      1   2
## 49     182.20           8.21          8.60      1   2
## 52     132.09           7.26          8.22      1   2
## 69      72.31           8.21          8.81      1   2
```

Jest to podgrupa filmów posiadających Oscary, o średniej długości. Danych jest zaledwie 7, więc stąd mógł pojawić się problem z normalnością. Poza tym, wartość p-value wykazała wartość 0,0365, więc dla mniejszego współczynnika ufności problem nie istnieje. Uznałem zatem, że wszystko jest w porządku. Czas na zbadanie równości wariancji w podgrupach. Tym razem wykorzystam test Levene'a, który również nie potrzebuje założenia o równoliczności grup (test Bartletta nie chciał współpracować przy ANOVIE dwuczynnikowej).

```
leveneTest(Ocena ~ Oscary*V12, data = dane)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  0.4058 0.8434
##       77
```

Test potwierdził równość wariancji w podgrupach, więc wszystkie założenia ANOVY są spełnione. Można przejść do badania.

```
summary(aov(Ocena ~ Oscary+V12, data = dane))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Oscary       1  6.946   6.946  23.570 5.96e-06 ***
## V12          2  5.556   2.778   9.426 0.000213 ***
## Residuals   79 23.281   0.295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wynik nie powinien być zaskoczeniem, gdyż już wcześniej wykazałem, że oba te czynniki wpływają na ocenę filmu. Aby zobaczyć, czy różnice występują przy tych samych parach, wykonam test Tukeya.

```
TukeyHSD(aov(Ocena ~ Oscary+V12, data = dane))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Ocena ~ Oscary + V12, data = dane)
##
## $Oscary
##          diff      lwr       upr p adj
## 1-0 0.6554396 0.386719 0.9241603 6e-06
##
## $V12
##           diff        lwr       upr     p adj
## 2-1 0.03080104 -0.2958615 0.3574635 0.9724324
## 3-1 0.64132550  0.2368734 1.0457776 0.0008540
## 3-2 0.61052446  0.2350962 0.9859527 0.0006149
```

Również i w tym wypadku nie zaszły żadne radykalne zmiany. Moja dociekliwość zapytała mnie jednak, jak wyglądałby test Tukeya oraz model ANOVY, jeśli uwzględnić efekt interakcji. Wyniki okazały się interesujące.

```
summary(aov(Ocena ~ Oscary*V12, data = dane))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
```

```
## Oscary       1  6.946   6.946  23.260 6.99e-06 ***
## V12          2  5.556   2.778   9.302 0.000241 ***
## Oscary:V12   2  0.286   0.143   0.479 0.621045
## Residuals   77 22.995   0.299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wyniki ZARAZ będą interesujące, gdyż w ANOVIE nie ma nic odkrywczego. Efekt interakcji nie jest tu istotny statystycznie.

```r
TukeyHSD(aov(Ocena ~ Oscary*V12, data = dane))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Ocena ~ Oscary * V12, data = dane)
##
## $Oscary
##          diff       lwr       upr p adj
## 1-0 0.6554396 0.3848208 0.9260584 7e-06
##
## $V12
##           diff        lwr       upr     p adj
## 2-1 0.03080104 -0.2981995 0.3598015 0.9727908
## 3-1 0.64132550  0.2339787 1.0486724 0.0009426
## 3-2 0.61052446  0.2324092 0.9886397 0.0006820
##
## $`Oscary:V12`
##                  diff         lwr       upr     p adj
## 1:1-0:1  0.5338345865 -0.17234533 1.2400145 0.2454798
## 0:2-0:1 -0.0007974482 -0.46075993 0.4591650 1.0000000
## 1:2-0:1  0.6481203008 -0.05805961 1.3543002 0.0905915
## 0:3-0:1  0.7719298246  0.12562787 1.4182318 0.0100316
## 1:3-0:1  1.0802631579  0.40710890 1.7534174 0.0001646
## 0:2-1:1 -0.5346320346 -1.19925985 0.1299958 0.1868711
## 1:2-1:1  0.1142857143 -0.73944417 0.9680156 0.9987691
## 0:3-1:1  0.2380952381 -0.56680902 1.0429995 0.9537352
## 1:3-1:1  0.5464285714 -0.28019183 1.3730490 0.3909813
## 1:2-0:2  0.6489177489 -0.01571006 1.3135456 0.0596505
## 0:3-0:2  0.7727272727  0.17210570 1.3733488 0.0042982
## 1:3-0:2  1.0810606061  0.45163491 1.7104863 0.0000468
## 0:3-1:2  0.1238095238 -0.68109473 0.9287138 0.9976039
## 1:3-1:2  0.4321428571 -0.39447755 1.2587633 0.6475348
## 1:3-0:3  0.3083333333 -0.46775779 1.0844245 0.8536609
```

Okazuje się, że przy efekcie interakcji niektóre zestawienia mają istotnie różne średnie - efekt interakcji nie zachodzi jako całość, ale dla poszczególnych zestawień już tak:
-Brak Oscara:Długi film - Brak Oscara:Krótki film
-Brak Oscara:Długi film - Brak Oscara:Średni film
-Oscar:Długi film - Brak Oscara:Krótki film
-Oscar:Długi film - Brak Oscara:Średni film
Istotną rolę gra w tych interakcjach długi film, więc zapewne to on je powoduje.

Na koniec sprawdzę moc efektu. W tym celu utworzę zaktualizowane funkcje, które służą do liczenia powyższego przy ANOVIE wieloczynnikowej.

```r
eta_sq <- function(aovm){
  sum_stats <- summary(aovm)[[1]]
  SSm1 <- sum_stats[["Sum Sq"]][1]
  SSm2 <- sum_stats[["Sum Sq"]][2]
  SSr <- sum_stats[["Sum Sq"]][3]
  print("Pierwszy efekt: ")
  print(SSm1/(SSm1+SSm2+SSr))
  print("Drugi efekt: ")
  print(SSm2/(SSm1+SSm2+SSr))
  print("Suma: ")
  print(SSm1/(SSm1+SSm2+SSr) + SSm2/(SSm1+SSm2+SSr))
}

omega_sq <- function(aovm){
  sum_stats <- summary(aovm)[[1]]
  SSA <- sum_stats[["Sum Sq"]][1]
  SSB <- sum_stats[["Sum Sq"]][2]
  SSr <- sum_stats[["Sum Sq"]][3]
  DFA <- sum_stats[["Df"]][1]
  DFB <- sum_stats[["Df"]][2]
  MSr <- sum_stats[["Mean Sq"]][3]
  print("Pierwszy efekt: ")
  print((SSA-DFA*MSr)/(SSA+SSB+SSr+MSr))
  print("Drugi efekt: ")
  print((SSB-DFB*MSr)/(SSA+SSB+SSr+MSr))
  print("Suma: ")
  print((SSA-DFA*MSr)/(SSA+SSB+SSr+MSr) + (SSB-DFB*MSr)/(SSA+SSB+SSr+MSr))
}

eta_sq(aov(Ocena ~ Oscary+V12, data = dane))
```

```
## [1] "Pierwszy efekt: "
## [1] 0.1941174
## [1] "Drugi efekt: "
## [1] 0.1552654
## [1] "Suma: "
## [1] 0.3493828
```

```r
omega_sq(aov(Ocena ~ Oscary+V12, data = dane))
```

```
## [1] "Pierwszy efekt: "
## [1] 0.1843634
## [1] "Drugi efekt: "
## [1] 0.1376604
## [1] "Suma: "
## [1] 0.3220237
```

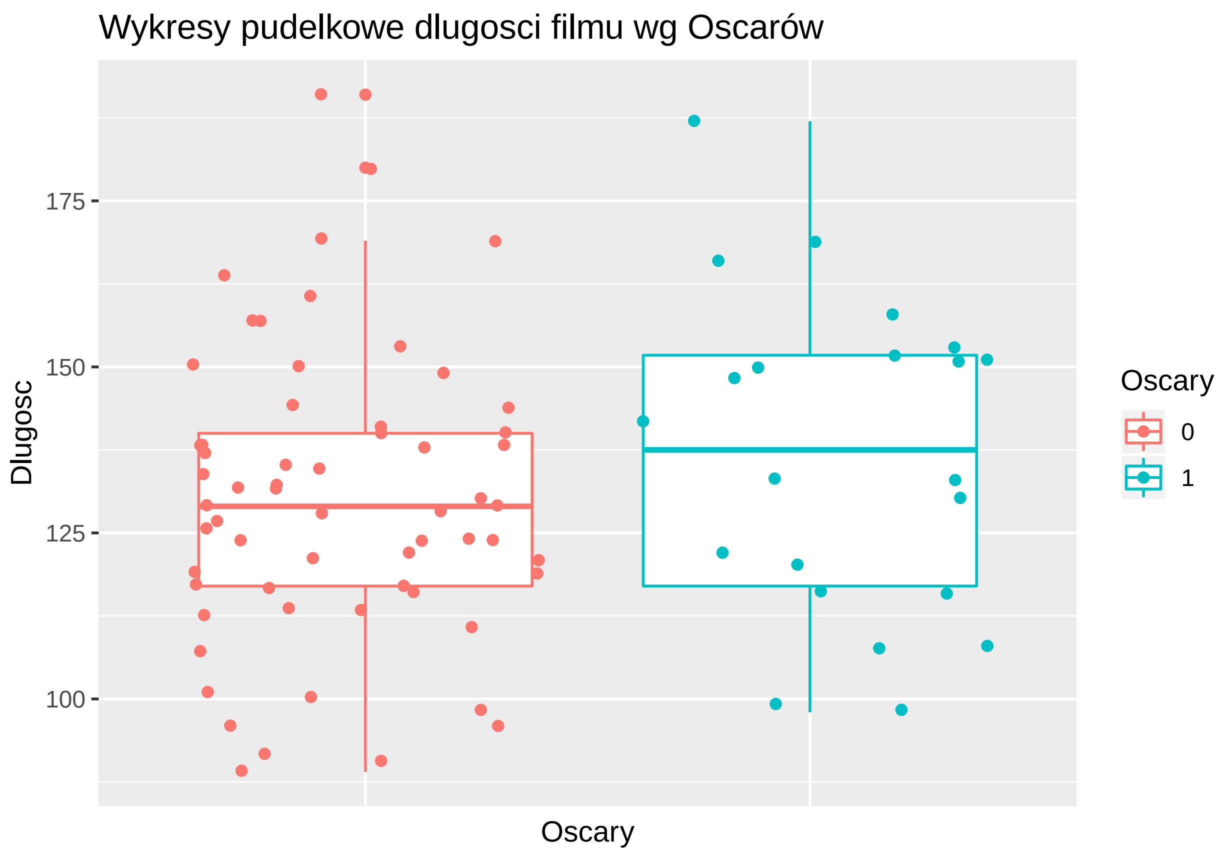
Jak widać, obie zmienne wyjaśniają łącznie około 35% zmienności próby losowej, natomiast około 32% zmienności w całej populacji. Jest to jednak wciąż efekt słaby.

MANOVA

W ostatniej części badania zmierzę się wielowymiarową ANOVĄ. Osoba czytająca moją pracę mogła zauważyć, że we wstępie zmienna Dlugosc występowała w dwóch rolach: jako zmienna kategoryczna oraz zależna. Jak

to okazało się już przy ANOVIE jednoczynnikowej, było to lekkie nagięcie rzeczywistości, gdyż z DLugosci została utworzona zmienna kategoryczna, ona sama zaś w sobie jest zmienną zależną. W tym punkcie właśnie tak ją potraktuję: tworząc model *Ocena + Dlugosc ~ Oscary*.
Na początku sprawdzę na wykresie relację między długością filmu a liczbą zdobytych Oscarów (zależność między oceną a liczbą zdobytych Oscarów zbadałem wcześniej).

```
dane %>% ggplot(aes(x = Oscary, y = Dlugosc, col = Oscary)) + geom_boxplot() +
  ggtitle("Wykresy pudełkowe długości filmu wg Oscarów") + geom_jitter() +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



Jak wynika z wykresu, średnia długości nie różni się zbytnio między filmami z Oscarami i bez nich. Dla przypomnienia, w przypadku wykresu Ocena-Oscary wykres wskazywał większe różnice.
Wykonam teraz test normalności, ale wyłącznie wg zmiennej Dlugosc, gdyż wg zmiennej Oscary testy zostały przeprowadzone przy ANOVIE jednoczynnikowej.

```
tapply(dane$Dlugosc, dane$Oscary, shapiro.test)
```

```
## $`0`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97989, p-value = 0.4126
##
##
## $`1`
##
##  Shapiro-Wilk normality test
```

```
##
## data:  X[[i]]
## W = 0.96645, p-value = 0.6291
```

Założenia o normalności zmiennych zależnych zostały spełnione, pora więc na test równości wariancji, ale ponownie wyłącznie dla zmiennej Dlugosc.

```r
bartlett.test(Dlugosc ~ Oscary, data = dane)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Dlugosc by Oscary
## Bartlett's K-squared = 0.30695, df = 1, p-value = 0.5796
```

Również ten warunek został spełniony, więc można przystąpić do badania MANOVĄ.

```r
summary(manova(cbind(Ocena, Dlugosc) ~ Oscary, data = dane))
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## Oscary     1 0.19446   9.6561      2     80 0.0001752 ***
## Residuals 81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Jak widać, średnie ocen i długości filmów różnią się między filmami, jednak nie jest to do końca prawdziwe stwierdzenie. Możliwe, że średnie tylko jednej z tych cech (a konkretnie konkretnie zmiennej Ocena, gdyż ANOVA jednoczynnikowa to pokazała) różnią się względem zdobytych Oscarów.

```r
summary(aov(cbind(Ocena, Dlugosc) ~ Oscary, data = dane))
```

```
##  Response Ocena :
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Oscary       1  6.9461  6.9461  19.511 3.067e-05 ***
## Residuals   81 28.8368  0.3560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Dlugosc :
##             Df Sum Sq Mean Sq F value Pr(>F)
## Oscary       1    748  748.03  1.4947  0.225
## Residuals   81  40538  500.47
```

Rzeczywiście tylko średnie ocen różnią się względem Oscarów, tym samym nie potwierdziłą się moja hipoteza ze wstępu o wpływie Oscarów na oba czynniki jednocześnie.

Podsumowanie

Wynki częściowo potwierdziły moje hipotezy:
ANOVA jednoczynnikowa:
- *Ocena ~ Gatunek*: Potwierdziła się moja hipoteza, ANOVA nie wykazała różnic dla średnich ocen względem gatunków filmów
- *Ocena ~ Oscary*: Tutaj też potwierdziła się moja hipoteza. Oscary wpływają na różnice dla średnich ocen, ale ze słabym efektem.
- *Ocena ~ Dlugosc(V12)*: Zaskoczenie, nie spodziewałem się, że średnie ocen będą się różnić względem długośći filmów. Jak wykazał test Tuckeya, wszystko to za sprawą filmów powyżej 150 minut. Efekt jednak był słaby.
ANOVA wieloczynnikowa:

-*Ocena ~ Oscary + Dlugosc(V12)*: Prawidłowo przewidziałem, że jedynie Oscary będą miały wpływ na ocenę, a model będzie pozbawiony interakcji. Zaskoczeniem jednak była obecność interakcji dla niekórych tylko zbiorów.
MANOVA:
-*Ocena + Dlugosc ~ Oscary*: Nie sądziłem, że Oscary nie będą rozróżniać długości filmów. Moja hipoeza była nietrafna.