Projekt

Marta Denisiuk, Nicole Kahlau

Zbiór danych i biblioteki

Źródło danych: https://www.kaggle.com/residentmario/ramen-ratings

```
library(plyr)
library(dplyr)
library(countrycode)
library(ggplot2)
library(viridis)
library(ggthemes)
library(gridExtra)
library(ggpubr)
library(grid)
library(stringr)
library(car)
library(nortest)
library(stats)
library(userfriendlyscience)
library(RVAideMemoire)
library(corrplot)
library(ggridges)
```

```
read.csv('ramen-ratings.csv') -> ramen
```

Zbiór danych zawiera informacje o ramenie. Większość zmiennych są zmiennymi kategorycznymi, jedna zmienna jest ilościowa - ocena danego ramenu. Celami badawczymi były:

- Na którym koentynencie najłatwiej znaleźć dobry ramen? Zależność oceny od kontynentu.
- Czy miejsce geograficzne tworzy upodobanie względem rodzaju podawania ramenu?
- Wpływ najczęstszych cech ramenu na ocenę -> czyli jak bardzo ramen z kurczakiem, ostry lub z wołowiną smakuje?

Zmienne

zmienna	opis
`Review`	Numer recenzji dań
`Brand`	Firma
`Variety`	Nazwa dania
`Style`	rodzaj opakowania w którym podawane jest danie
`Country`	Kraj/Stan/Miasto w którym jest danie
`Stars` O	cena
`Top.Ten`	Top 10 dań w latach 2012 - 2016

Pierwszym krokiem jest czyszczenie danych oraz tworzenie nowych zmiennych. Aby wyciągnąć wnioski z naszych danych tworzymy zmienną Continent oraz StarsInterval odpowiadającą za przedziały ocen. Połączyłyśmy obie Ameryki w jedność, ze względu na małą liczebność Ameryki Południowej.

```
unique(ramen$Country)
```

```
##  [1] "Japan"         "Taiwan"        "USA"           "India"
##  [5] "South Korea"   "Singapore"     "Thailand"      "Hong Kong"
##  [9] "Vietnam"       "Ghana"         "Malaysia"      "Indonesia"
## [13] "China"         "Nigeria"       "Germany"       "Hungary"
## [17] "Mexico"        "Fiji"          "Australia"     "Pakistan"
## [21] "Bangladesh"    "Canada"        "Nepal"         "Brazil"
## [25] "UK"            "Myanmar"       "Netherlands"   "United States"
## [29] "Cambodia"      "Finland"       "Sarawak"       "Philippines"
## [33] "Sweden"        "Colombia"      "Estonia"       "Holland"
## [37] "Poland"        "Dubai"
```

```
ramen$Country[ramen$Country == 'United States'] <- 'USA'
ramen$Continent <- countrycode(sourcevar = ramen[, "Country"],
                               origin = "country.name",
                               destination = "continent")
ramen$Continent[ramen$Country == 'Dubai'] <- 'Asia'
ramen$Continent[ramen$Country == 'Holland'] <- 'Europe'
ramen$Continent[ramen$Country == 'Sarawak'] <- 'Asia'
unique(ramen$Continent)
```

```
## [1] "Asia"     "Americas" "Africa"   "Europe"   "Oceania"
```

```
ramen$StarsInterval <- c("0-1", "1-2", "2-3", "3-4",
                         "4-5")[findInterval(as.numeric(as.character(ramen$Stars)) ,
                                             c(0, 1, 2, 3, 4, Inf) )]
```

Usuwamy puste wiersze.

```
unique(ramen$Style)
```

```
## [1] "Cup"  "Pack" "Tray" "Bowl" "Box"  "Can"  "Bar"  ""
```

```
ramen <- ramen[ramen$Style != "",]
sort(unique(ramen$Stars))
```

```
##  [1] "0"       "0.1"     "0.25"    "0.5"     "0.75"    "0.9"     "1"
##  [8] "1.1"     "1.25"    "1.5"     "1.75"    "1.8"     "2"       "2.1"
## [15] "2.125"   "2.25"    "2.3"     "2.5"     "2.75"    "2.8"     "2.85"
## [22] "2.9"     "3"       "3.0"     "3.00"    "3.1"     "3.125"   "3.2"
## [29] "3.25"    "3.3"     "3.4"     "3.5"     "3.50"    "3.6"     "3.65"
## [36] "3.7"     "3.75"    "3.8"     "4"       "4.0"     "4.00"    "4.125"
## [43] "4.25"    "4.3"     "4.5"     "4.50"    "4.75"    "5"       "5.0"
## [50] "5.00"    "Unrated"
```

```
ramen <- ramen[ramen$Stars != 'Unrated',]
round(as.numeric(ramen$Stars),2) -> ramen$Stars
revalue(ramen$Brand, c("A-One"="A1")) -> ok
```

W oryginalnym zbiorze mamy zmienną Variety która zawiera nazwy potraw. W poniższym kodzie na podstawie tej zmiennej tworzymy zbiór słów określających ramen i ich liczebność.

```
#Budujemy listę składników z nazw dań
strsplit(ramen$Variety, ' ') -> variety # rozdzielamy wyrazy w każdej nazwie
```

```
mark <- c() # lista składników
for (i in 1:length(variety)){
  for (j in 1:length(variety[[i]])){    # zmiana list w liście na jedną listę
    mark <- c(mark, variety[[i]][[j]])
  }
}
mark <- gsub("\\(|\\)|\\[|\\]","",mark) # usuwanie znaków
mark <- sub(' ','',mark)
specified <- c('noodles','noodle','flavour','artificial','ramen','instant','flavor',
               'sauce','cup','bowl','rice', 'with',
               'a','the','soup','men','la','i','the','-',
               'y','in','ly','de','&','mi','no') #lista zbędnych słów
tolower(mark) -> mark
mark[! mark %in% specified] -> mark # usuwanie słów
table(mark) -> mark.table # zliczenie każdego słowa
as.data.frame(mark.table) -> mark.table # ramka danych składników
arrange(mark.table,-Freq) -> mark.table # sortowanie po ilości składników
which(is.na(str_detect(tolower(ramen$Variety[i]),
                       as.character(mark.table$mark)))) -> error
mark.table$mark[error]
```

```
## [1]
## 1442 Levels:  'kua-chap' -dry "a" "phnom "tom / 1 10 100 100% 1000 2 275 ... zzaldduck
```

```
mark.table <- mark.table[-c(error),]
head(mark.table, 10)
```

```
##       mark Freq
## 1  chicken  328
## 2    spicy  276
## 3     beef  233
## 4    curry  127
## 5      tom  127
## 6   shrimp  126
## 7      hot  119
## 8  seafood  109
## 9     pork  102
## 10   style   90
```

Dzięki temu, możemy teraz do naszego orginalnego zbioru danych dodać zmienną Mark która przypisze każdemu wierszowi cechę która jest najczęstsza ogółem. Możemy wnioskować również, że są to cechy najważniejsze, które wpływają również na smak, a co za tym idzie możliwe, że i na ocenę. Będziemy się temu przyglądać w dalszej części pracy.

```
for (i in 1:nrow(ramen)){
  which(str_detect(tolower(ramen$Variety[i]), as.character(mark.table$mark)))[1] -> first
  ramen$Mark[i] <- as.character(mark.table$mark[first])
}
```

Sprawdzamy jakiego typu są nasze zmienne.

```
glimpse(ramen)
```

```
## Rows: 2,575
## Columns: 10
## $ Review..      <int> 2580, 2579, 2578, 2577, 2576, 2575, 2574, 2573, 2572,...
## $ Brand         <chr> "New Touch", "Just Way", "Nissin", "Wei Lih", "Ching'...
```

```
## $ Variety       <chr> "T's Restaurant Tantanmen ", "Noodles Spicy Hot Sesam...
## $ Style         <chr> "Cup", "Pack", "Cup", "Pack", "Pack", "Pack", "Cup", ...
## $ Country       <chr> "Japan", "Taiwan", "USA", "Taiwan", "India", "South K...
## $ Stars         <dbl> 3.75, 1.00, 2.25, 2.75, 3.75, 4.75, 4.00, 3.75, 0.25,...
## $ Top.Ten       <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "...
## $ Continent     <chr> "Asia", "Asia", "Americas", "Asia", "Asia", "Asia", "...
## $ StarsInterval <chr> "3-4", "1-2", "2-3", "2-3", "3-4", "4-5", "4-5", "3-4...
## $ Mark          <chr> "tantanmen", "spicy", "chicken", "tom", "curry", "kim...
```

Widzimy, że parę zmiennych musimy ustawić na kategoryczne aby uzyskać nasze cele badawcze.

```
as.factor(ramen$Style) -> ramen$Style
as.factor(ramen$Country) -> ramen$Country
as.factor(ramen$StarsInterval) -> ramen$StarsInterval
as.factor(ramen$Brand) -> ramen$Brand
as.factor(ramen$Mark) -> ramen$Mark
glimpse(ramen)
```

```
## Rows: 2,575
## Columns: 10
## $ Review..      <int> 2580, 2579, 2578, 2577, 2576, 2575, 2574, 2573, 2572,...
## $ Brand         <fct> New Touch, Just Way, Nissin, Wei Lih, Ching's Secret,...
## $ Variety       <chr> "T's Restaurant Tantanmen ", "Noodles Spicy Hot Sesam...
## $ Style         <fct> Cup, Pack, Cup, Pack, Pack, Pack, Cup, Tray, Pack, Pa...
## $ Country       <fct> Japan, Taiwan, USA, Taiwan, India, South Korea, Japan...
## $ Stars         <dbl> 3.75, 1.00, 2.25, 2.75, 3.75, 4.75, 4.00, 3.75, 0.25,...
## $ Top.Ten       <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "...
## $ Continent     <chr> "Asia", "Asia", "Americas", "Asia", "Asia", "Asia", "...
## $ StarsInterval <fct> 3-4, 1-2, 2-3, 2-3, 3-4, 4-5, 4-5, 3-4, 0-1, 2-3, 4-5...
## $ Mark          <fct> tantanmen, spicy, chicken, tom, curry, kimchi, spice,...
```

Teraz wszystko jest poprawione. Okazało się, że musiałyśmy wprowadzić sporo poprawek, jak i utworzyć nowe zmienne aby mieć ciekawe badania. Teraz przejdziemy do II części projektu - eksploracji danych i wyciąganiu wniosków. Do każdego przedstawionego wcześniej podpunktu utworzyłyśmy wykresy które zawierają długi kod. Aby praca była bardziej przejrzysta i estetyczna postanowiłyśmy je ukryć w pliku PDF. Kody wykresów które nie były skomplikowane zostawiłyśmy.

Badania

Na początek sprawdzimy czy nasza zmienna Stars pochodzi z rozkładu normalnego. W tym celu, posłużymy się testem Lielleforsa'a który jest oparty na teście Kołmogorowa-Smirnova. Sprawdza hipotezę zerową wskazującą na rozkład zbliżony do rozkładu normalnego. Wartości $p > 0.05$ potwierdzają spełnienie założenia o rozkładzie normalnym.

```
lillie.test(ramen$Stars)
```

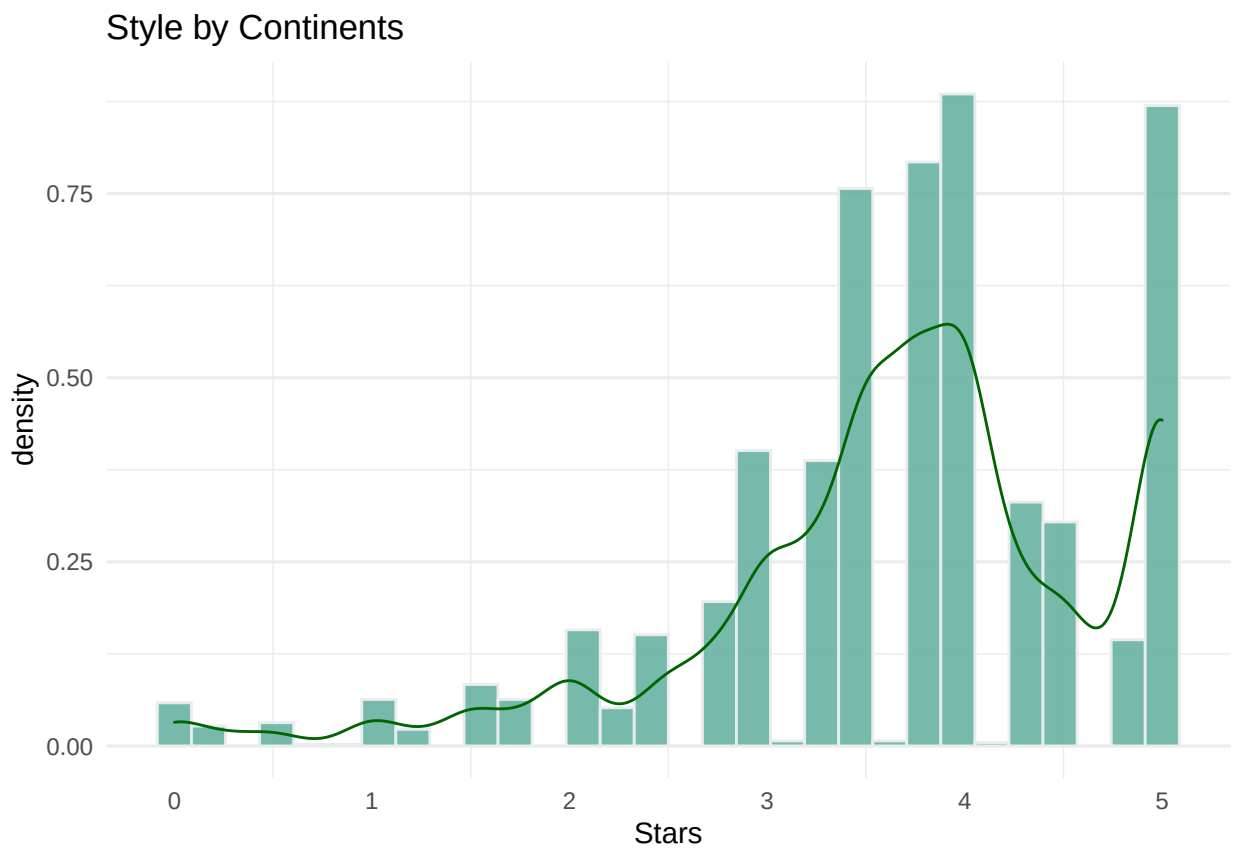
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ramen$Stars
## D = 0.14503, p-value < 2.2e-16
```

Wynik wyszedł $p = 0$ a więc odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej, czyli nie są.

W takim razie zobaczmy też jak układają się oceny na histogramie.

```
theme = theme_set(theme_minimal())
theme = theme_update(legend.position="none",legend.title=element_blank(),
                     panel.grid.major.x=element_blank())
ggplot(ramen, aes(x=Stars))+
  geom_histogram(aes(y= ..density..),fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  geom_density(alpha=.2,color="darkgreen") + ggtitle("Style by Continents")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Już na pierwszy rzut oka widać, że nie mamy do czynienia z rozkładem normalnym. Statystycy prowadzą spór, czy w różnych modelach np. analizy wariancji konieczne jest spełnienie tego założenia. Zobaczymy co przyniosą kolejne badania. Możemy podejrzewać, że średnia ocen oscyluje pomiędzy 3 a 4. Potwierdzimy to głównymi statystykami przy okazji badając też inne zmienne.

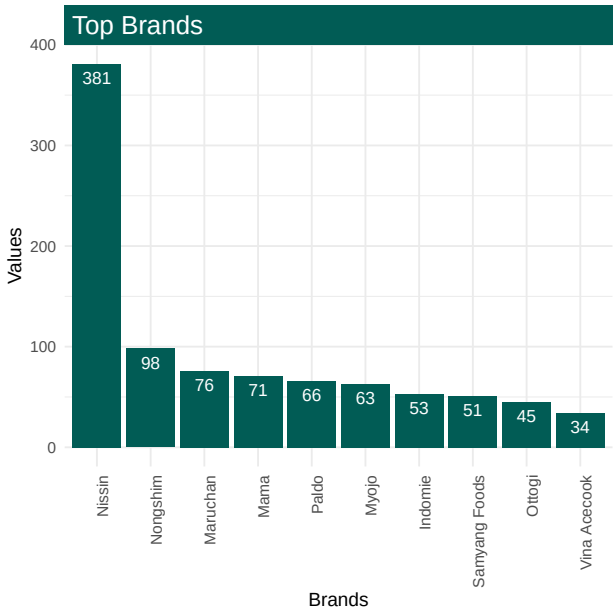
```
summary(ramen)
```

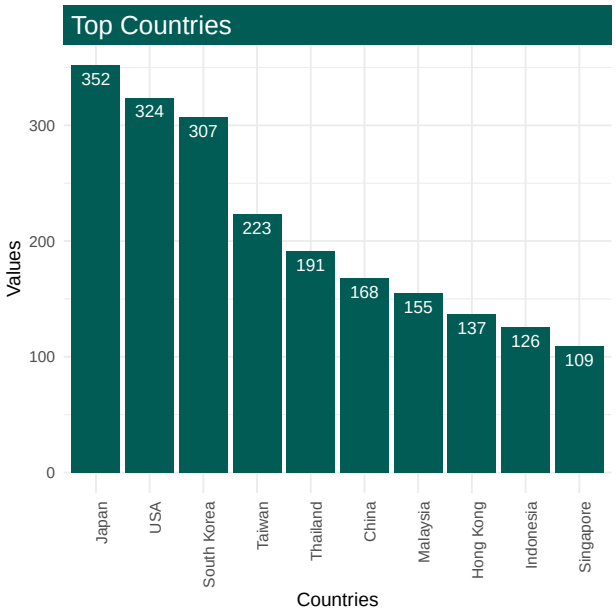
```
##     Review..          Brand        Variety           Style
##  Min.   :   1.0   Nissin  : 381   Length:2575        Bar :   1
##  1st Qu.: 646.5   Nongshim:  98   Class :character   Bowl: 481
##  Median :1290.0   Maruchan:  76   Mode  :character   Box :   6
##  Mean   :1290.2   Mama    :  71                      Can :   1
##  3rd Qu.:1934.5   Paldo   :  66                      Cup : 450
##  Max.   :2580.0   Myojo   :  63                      Pack:1528
##                   (Other) :1820                      Tray: 108
##         Country         Stars          Top.Ten          Continent
##  Japan      : 352   Min.   :0.000   Length:2575        Length:2575
##  USA        : 324   1st Qu.:3.250   Class :character   Class :character
```

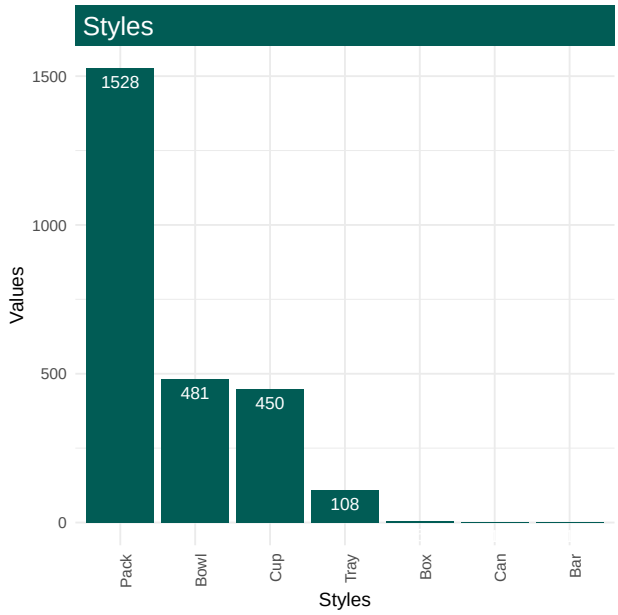
```
##  South Korea: 307   Median :3.750   Mode  :character   Mode  :character
##  Taiwan     : 223   Mean   :3.655
##  Thailand   : 191   3rd Qu.:4.250
##  China      : 168   Max.   :5.000
##  (Other)    :1010
##  StarsInterval      Mark
##  0-1:  54      chicken: 327
##  1-2: 103      spicy  : 233
##  2-3: 250      beef   : 187
##  3-4:1041      tom    : 150
##  4-5:1127      curry  : 106
##                pork   :  86
##                (Other):1486
```

Średnia gwiazd wynosi 3.65. Mimo, że oceny przeważnie są średnie/pozytywne to możemy spotkać negatywne czy nawet z punktacją 0. Widzimy też, że mamy 2580 ocen różnych ramenów w różnych częściach świata. Jeśli chodzi o zmienne porządkowe, to summary przedstawiło nam wartości wszystkie lub te najczęściej występujące. Dzięki temu widzimy które marki są dużymi korporacjami, Azja (jak można było się spodziewać z powodu pochodzenia ramenu) króluje nad kontynentami, kurczak jest najczęstszym słowem w nazwie ramenu, a w opakowaniach występuje duża nierówność w ilości. W krajach widzimy przewagę państw azjatyckich, ale na podium znalazło się również USA.

Możemy również przedstawić dane na wykresach. Z powodu estetyki, oraz z tego, że w dalszych badaniach będziemy skupiać się i tak na najliczniejszyc grupach, przedstawione zostały tylko top z każdej zmiennej.





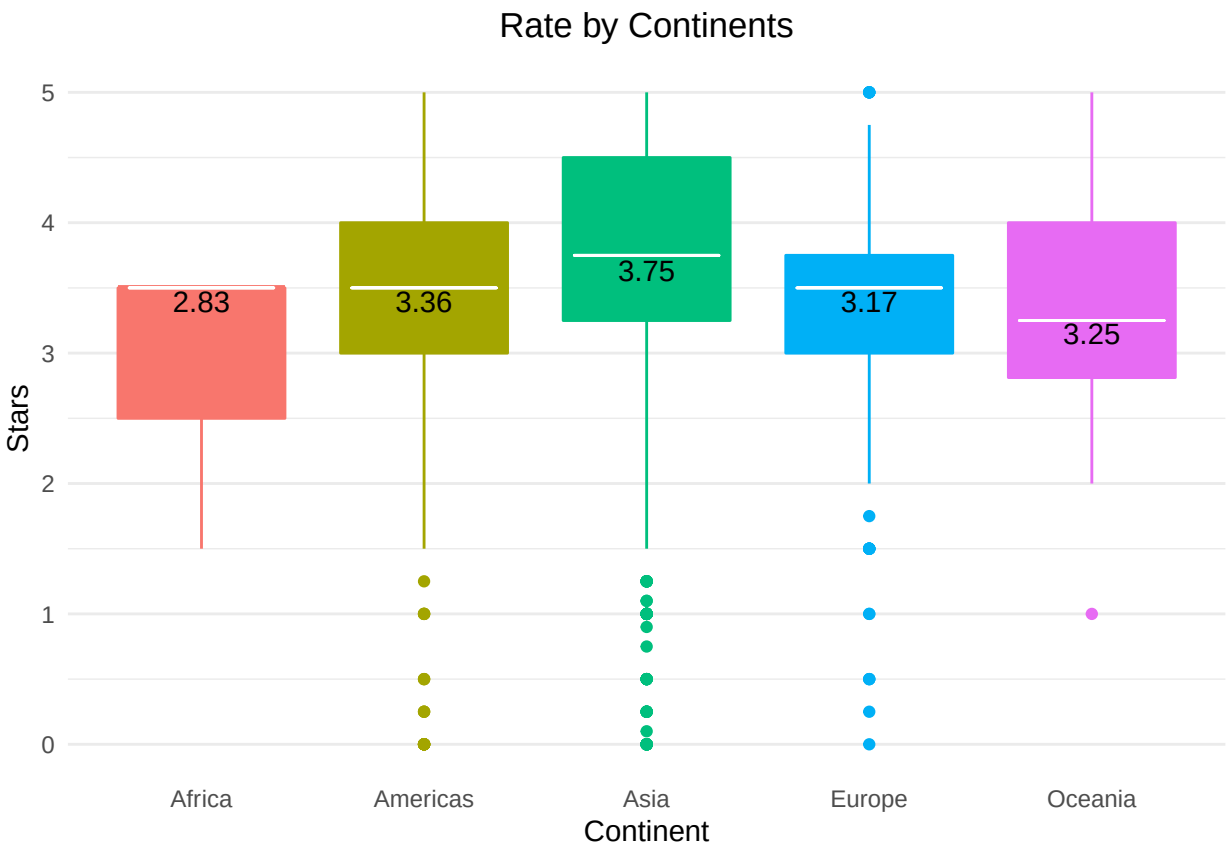


Czy na dobry Ramen najłatwiej natrafimy w Azji?

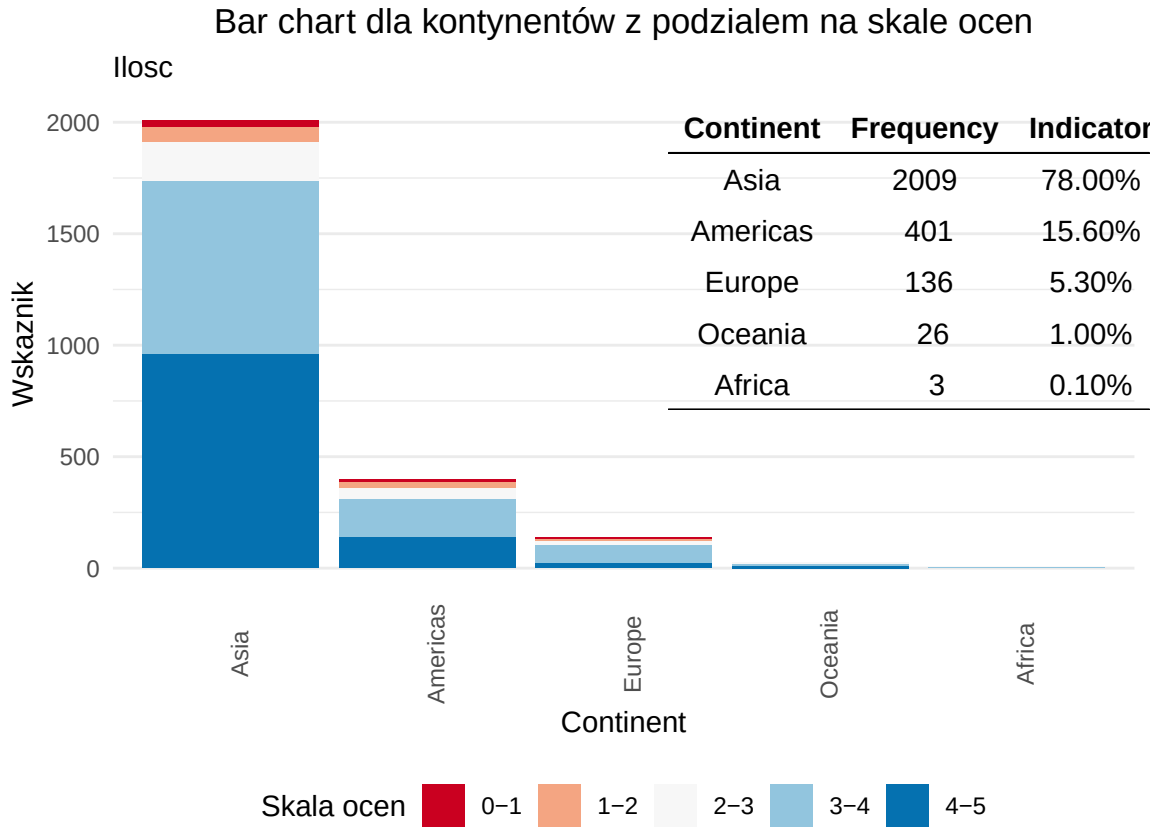
Hipoteza H_0: Ocena Ramenu nie zależy od kontynentu

Hipoteza H_1: Ocena Ramenu zależy od kontynentu

```
title = "Rate by Continents"
theme = theme_set(theme_minimal())
theme = theme_update(legend.position="none",legend.title=element_blank(),
                     panel.grid.major.x=element_blank())
boxplot = ggplot(ramen, mapping=aes_string(y = 'Stars', x = 'Continent')) +
  ggtitle(title) + theme(plot.title = element_text(hjust = 0.5))
boxplot = boxplot + geom_boxplot(outlier.colour = NULL,
                                 aes_string(colour="Continent", fill="Continent")) +
  stat_summary(geom = "crossbar", width=0.65, fatten=0, color="white",
               fun.data = function(x){ return(c(y=median(x), ymin=median(x),
                                                ymax=median(x))) })
mean.n <- function(x){ return(c(y = median(x)*0.97, label = round(mean(x),2))) }
boxplot = boxplot + stat_summary(fun.data = mean.n, geom = "text",
                                 fun = mean, colour = "black")
boxplot
```



Na pierwszy rzut oka widzimy, ze wariancje w poszczególnych grupach wydają się być zbliżone. Problem może stanowić Afryka ze względu na małą liczebność i wariancja może nie być zbyt dobrze odzwierciedlona na wykresie. Poniższy wykres przedstawia liczebność każdej grupy.



Continent	Frequency	Indicator
Asia	2009	78.00%
Americas	401	15.60%
Europe	136	5.30%
Oceania	26	1.00%
Africa	3	0.10%

Grupami które będą mają możliwość uzyskania wiarygodnych wyników jest Azja, Ameryka i ewentualnie Europa. Niestety różnice pomiędzy grupami są tak duże, że nie możemy mówić tutaj o równoliczności grup. Mamy też do czynienia z obserwacjami odstającymi (wykresy pudełkowe). Z tego względu podejrzewamy, że w naszym przypadku lepiej sprawdzą się testy nieparametryczne, które nie muszą spełniać tych wymagań. Przydatne są również, kiedy naszą zmienne nasze są porządkowe, a własnie Continent jest taką zmienną. Jednak zanim wykluczymy całkowicie testy parametryczne, spróbujmy przeprowadzić analizę wariancji.

Testem statystycznym jednorodności wariancji jest test Flignera-Killeena oraz test Levene'a.

```
leveneTest(Stars ~ Continent, ramen)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    4  1.5563 0.1833
##       2570
```

```
fligner.test(Stars ~ Continent, ramen)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Stars by Continent
## Fligner-Killeen:med chi-squared = 8.3185, df = 4, p-value = 0.08058
```

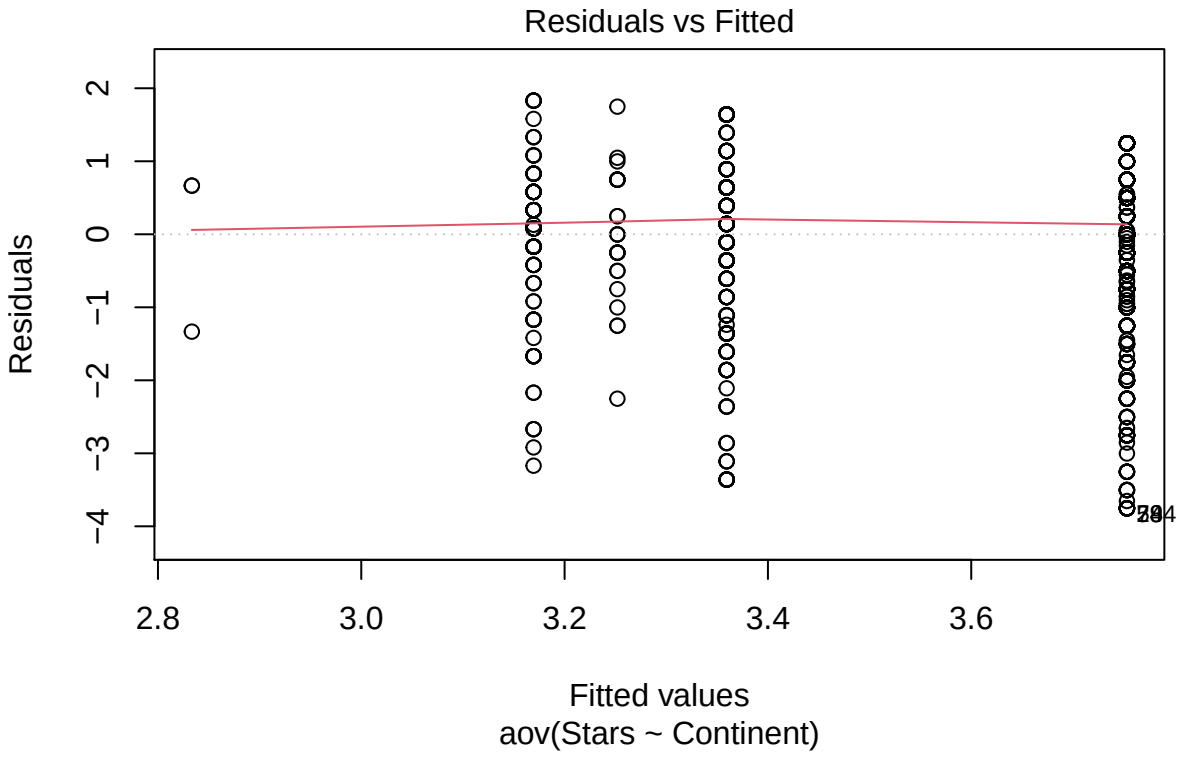
Widzimy, że w obu przypadkach p.value > 0.05, jednak test Flignera wykazał niewiele większą wartość, a że nasze dane są zaszumione, to w takich przypadkach p.value określa się na poziomie 0.1. Nie mamy wystarczającego dowodu by powiedzieć, że nasze wariancje są jednorodne. A co się stanie jeśli dopasujemy model ANOVA? Sprawdźmy.

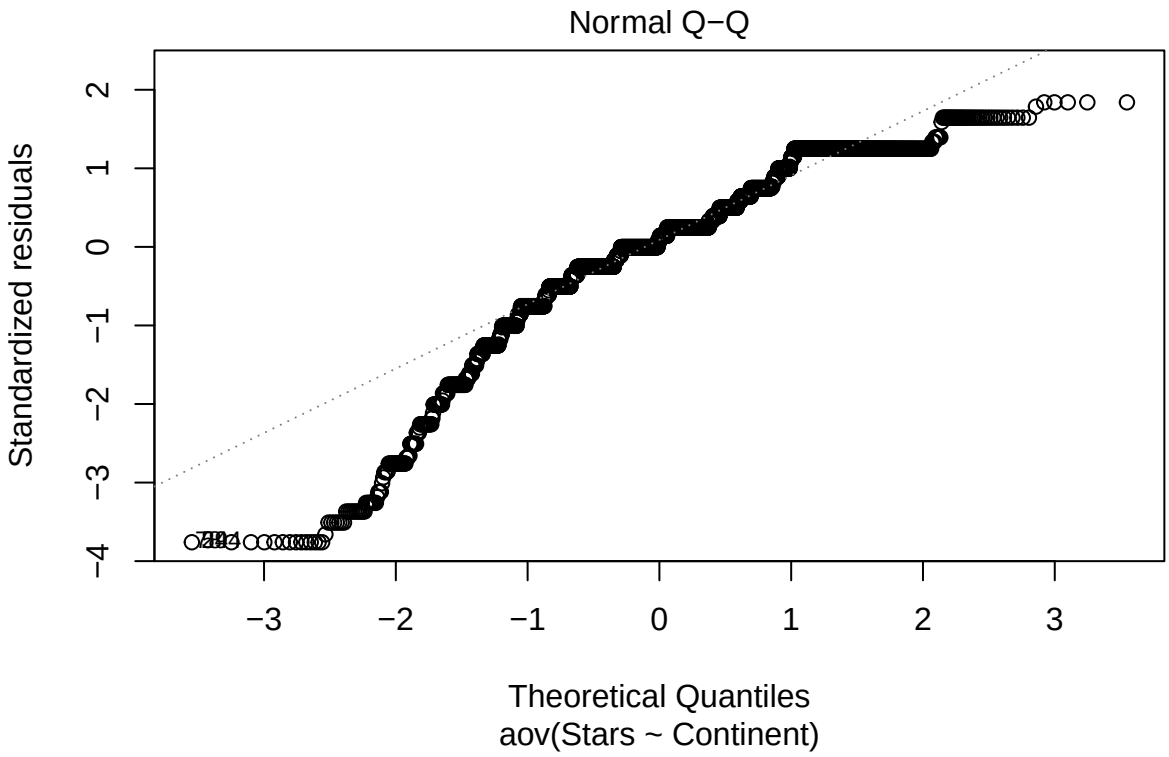
```
aov.continent <- aov(Stars ~ Continent, ramen)
summary(aov.continent)
```

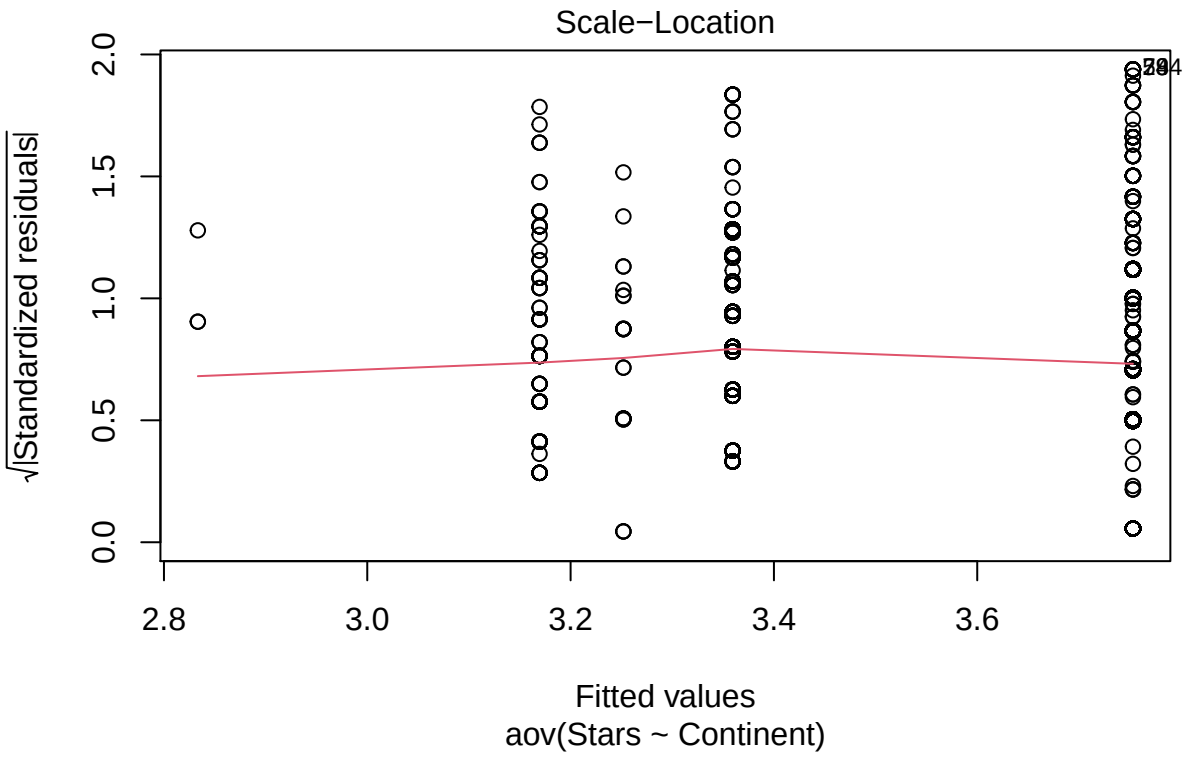
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Continent      4   92.7  23.177   23.25 <2e-16 ***
## Residuals   2570 2562.5   0.997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

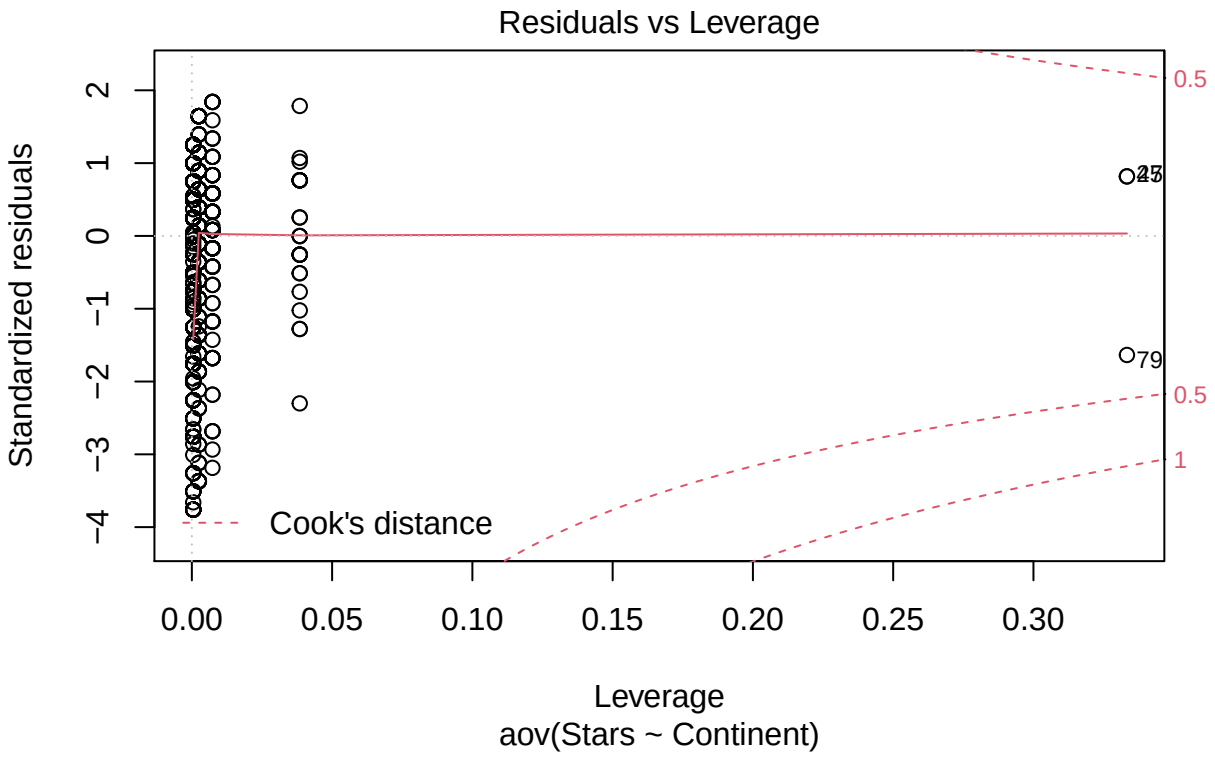
Test F wykazał istotną zależność pomiędzy średnimi w grupach określonych zmienną Continent (tzn. co najmniej jedna ze średnich istotnie odstaje od reszty).

```
plot(aov.continent)
```




Normal Q−Q
Standardized residuals
−4
−3
−2
−1
0
1
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(Stars ~ Continent)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.8
3.0
3.2
3.4
3.6
Fitted values
aov(Stars ~ Continent)



Pierwszy wykres diagnostyczny pokazuje, iż wariancje w grupach nie są do końca równe. Drugi wykres sprawdza warunek normalności i jak mogliśmy się spodziewać, mamy duże odchyłki - ma ciężkie ogony. Taka wielkość odchyleń może negatywnie wpłynąć na jakość analizy. Sprawdźmy jeszcze podsumowanie tej analizy.

```
summary.lm(aov.continent)
```

```
##
## Call:
## aov(formula = Stars ~ Continent, data = ramen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7532 -0.4607  0.0805  0.6406  1.8305
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.75317    0.02228 168.471  < 2e-16 ***
## ContinentAmericas -0.39376    0.05461  -7.210 7.33e-13 ***
## ContinentEurope   -0.58368    0.08847  -6.597 5.07e-11 ***
## ContinentOceania  -0.50124    0.19709  -2.543    0.011 *
## ContinentAfrica   -0.91983    0.57693  -1.594    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9985 on 2570 degrees of freedom
## Multiple R-squared:  0.03492,    Adjusted R-squared:  0.03341
## F-statistic: 23.24 on 4 and 2570 DF,  p-value: < 2.2e-16
```

Wszystkie 4 kontynenty wpływają negatywnie na ocenę w stosunku do Azji. Widzimy, że Afryka (prawdopodobnie ze względu na małą liczebność) nie jest istotna statystycznie. Jednak widzieliśmy na wykresach, że model nie został dobrze dopasowany, więc rozsądne będzie odrzucenie go. W takim wypadku, jak już wcześniej wspomniałyśmy, zajmiemy się testami nieparametrycznymi.

Odpowiednikiem jednoczynnikowej analizy wariancji będzie test Kruskala-Wallisa. Stosujemy go, gdy chcemy porównać co najmniej trzy grupy pod względem jakiejś zmiennej ilościowej, dokładnie tak jak w przypadku analizy wariancji.

```
kruskal.test(Stars ~ Continent, ramen)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Stars by Continent
## Kruskal-Wallis chi-squared = 100.48, df = 4, p-value < 2.2e-16
```

Drugim testem nieparametrycznym jest test mediany.

```
mood.medtest(Stars~Continent, data=ramen, exact = FALSE)
```

```
##
##  Mood's median test
##
## data:  Stars by Continent
## X-squared = 74.509, df = 4, p-value = 2.531e-15
```

Czyli mediany są różne na kontynentach.

Wyszedł nam wynik istotny statystycznie, a więc możemy przypuszczać, że conajmniej jedna grupa różni się od innej grupy. Aby dowiedzieć się więcej szczegółów, wykonamy test post-hoc Gamesa-Howella, dla nierównych wariancji, ze względu na to, że nie uzyskaliśmy pełnej zgodności co do tego.

```
posthocTGH(ramen$Stars,ramen$Continent,method = "games-howell")
```

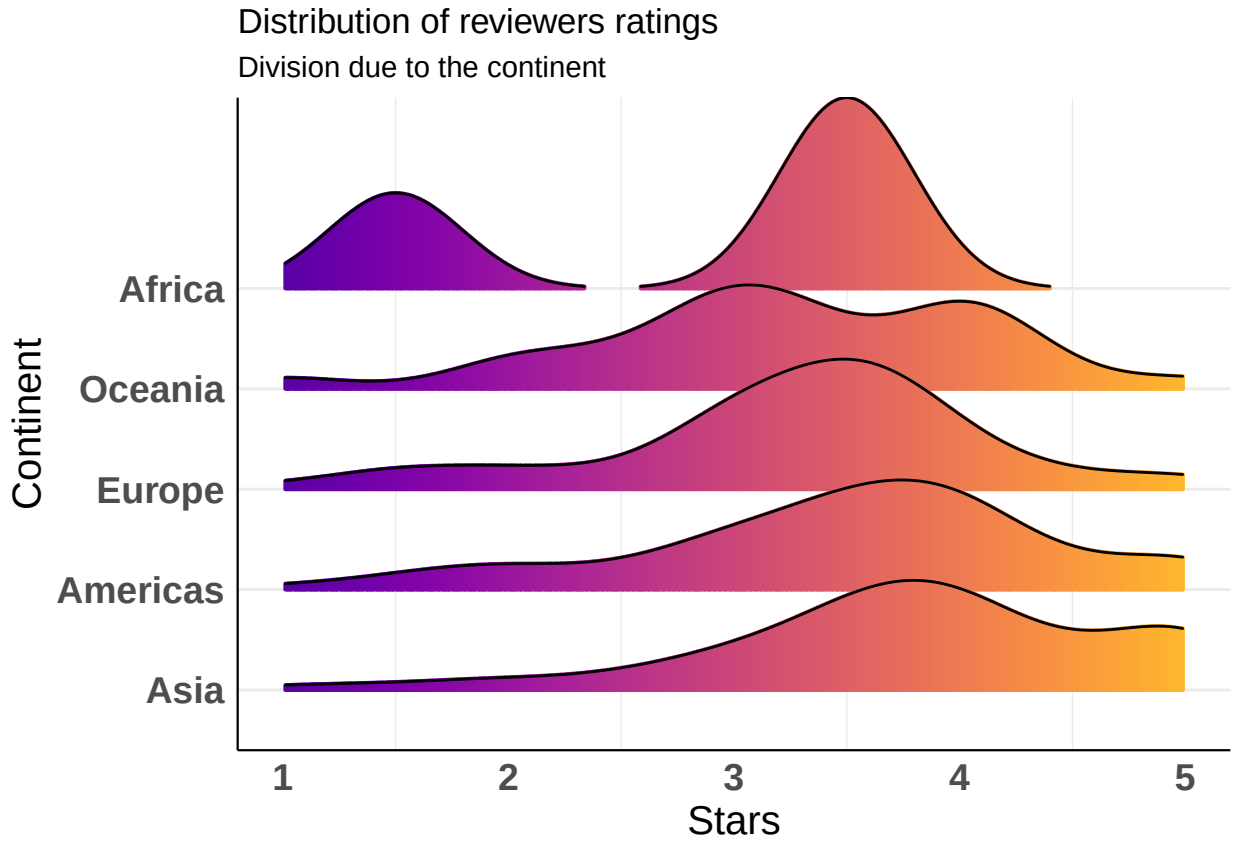
```
##             n means variances
## Asia     2009   3.8      0.95
## Americas  401   3.4      1.24
## Europe    136   3.2      0.95
## Oceania    26   3.3      0.77
## Africa      3   2.8      1.33
```

```
## Registered S3 methods overwritten by 'ufs':
##   method                     from
##   grid.draw.ggProportionPlot userfriendlyscience
##   pander.associationMatrix   userfriendlyscience
##   pander.dataShape           userfriendlyscience
##   pander.descr               userfriendlyscience
##   pander.normalityAssessment userfriendlyscience
##   print.CramersV             userfriendlyscience
##   print.associationMatrix    userfriendlyscience
##   print.confIntOmegaSq       userfriendlyscience
##   print.confIntV             userfriendlyscience
##   print.dataShape            userfriendlyscience
##   print.descr                userfriendlyscience
##   print.ggProportionPlot     userfriendlyscience
##   print.meanConfInt          userfriendlyscience
##   print.multiVarFreq         userfriendlyscience
##   print.normalityAssessment  userfriendlyscience
```

```
##   print.regrInfluential      userfriendlyscience
##   print.scaleDiagnosis       userfriendlyscience
##   print.scaleStructure       userfriendlyscience
##   print.scatterMatrix        userfriendlyscience
```

```
##                    diff ci.lo   ci.hi    t    df    p
## Americas-Asia    -0.394 -0.56 -0.2303 6.59 529.9 <.01
## Europe-Asia      -0.584 -0.82 -0.3447 6.74 153.8 <.01
## Oceania-Asia     -0.501 -1.01  0.0067 2.89  25.8  .05
## Africa-Asia      -0.920 -6.04  4.1997 1.38   2.0  .69
## Europe-Americas  -0.190 -0.47  0.0863 1.89 262.9  .33
## Oceania-Americas -0.107 -0.63  0.4163 0.59  30.5  .97
## Africa-Americas  -0.526 -5.59  4.5404 0.79   2.0  .92
## Oceania-Europe    0.082 -0.47  0.6303 0.43  37.9  .99
## Africa-Europe    -0.336 -5.33  4.6539 0.50   2.1  .98
## Africa-Oceania   -0.419 -5.04  4.1994 0.61   2.3  .96
```

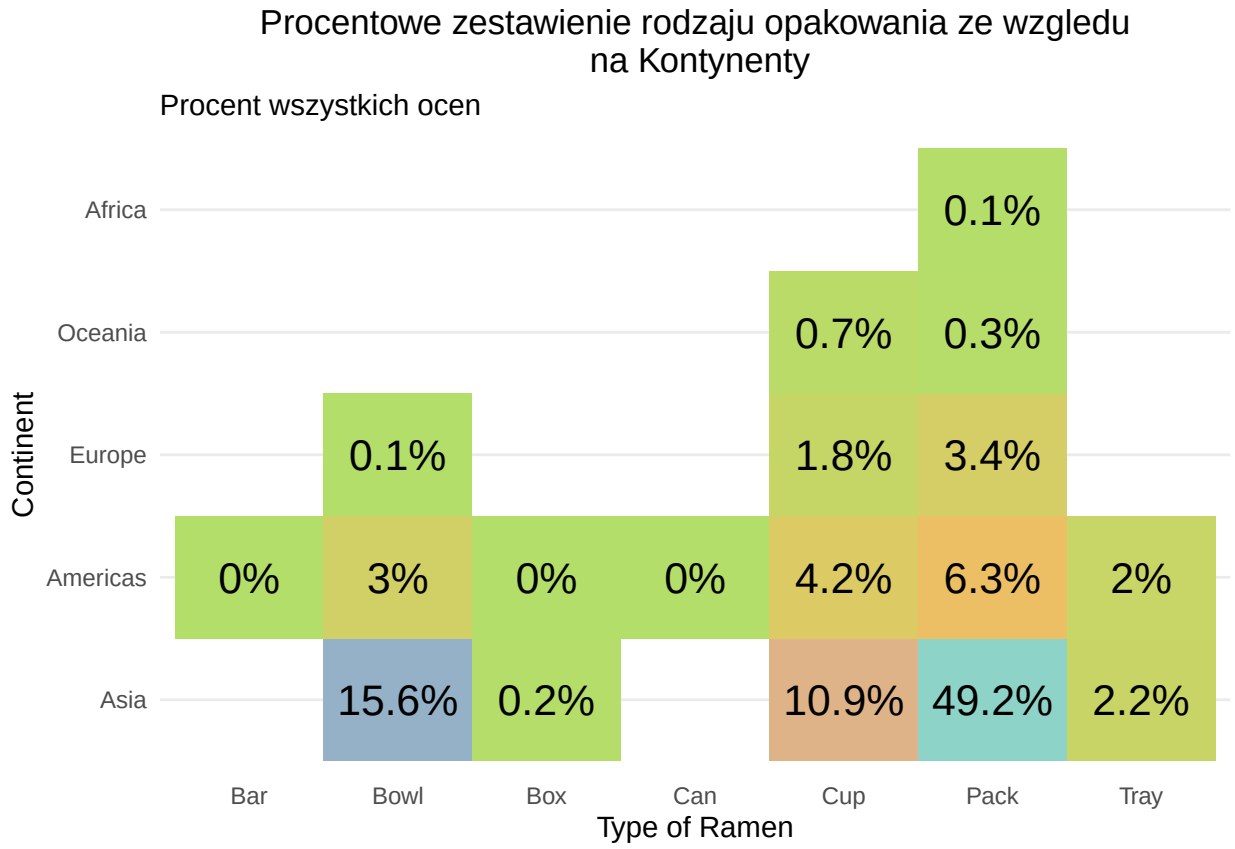
Widzimy, że statystycznie istotne są tylko różnice pomiędzy Azją a Ameryką i Azją a Europą. Afryka i Ocenia nie różnią się od siebie i są grupami o bardzo małej liczebności. Dlatego możemy twierdzić, że w Azji po pierwsze jest największa różnorodność co do wyboru ramenu oraz istnieje spore prawdopodobieństwo, że szybko natrafimy na dobry ramen, chociaż testy wykazały jedynie zależność oceny od kontynentu Azja, Europa, Ameryka. Biorąc pod uwagę średnie widzimy, że Azja osiąga najlepszy wynik. Dodałyśmy poniższy wykres składający się z histogramów poszczególnych kontynentów i widzimy, że znowu wygrywa Azja.



Czy kultura, miejsce geograficzne ma wpływ na upodobania "pojdemnikowe"? Czyli zależność rodzaju pojemnika od kontynentu

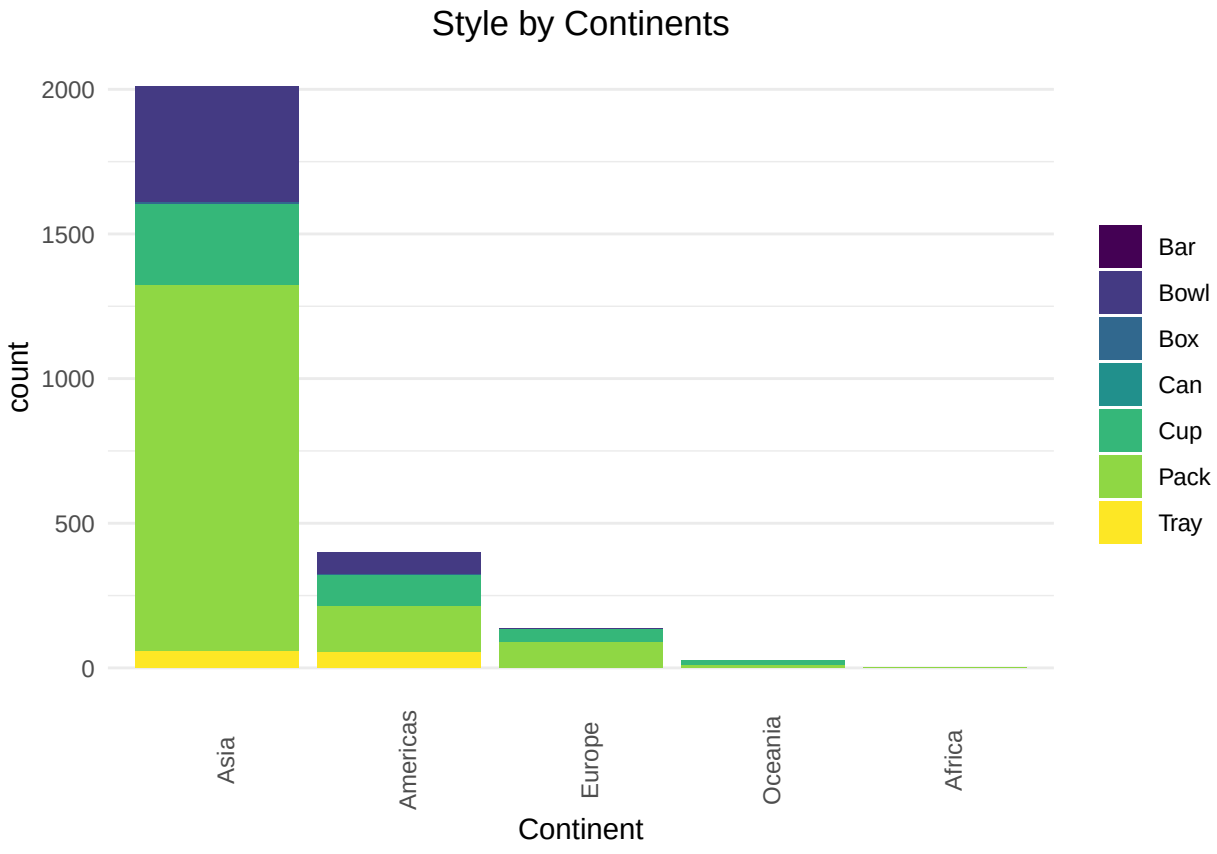
Hipoteza H_0: Rodzaj opakowania nie zależy od kontynentu

Hipoteza H_1: Rodzaj opakowania zależy od kontynentu




```
## List of 4
##  $ axis.title     :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : num 15
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text      :List of 11
##   ..$ family       : NULL
##   ..$ face         : chr "bold"
##   ..$ colour       : NULL
##   ..$ size         : num 14
##   ..$ hjust        : NULL
```

```
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.line      :List of 6
##   ..$ colour       : NULL
##   ..$ size         : num 0.4
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ legend.position: chr "none"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```



Na pierwszy rzut oka, widzimy, że “Pack”, “Cup” i “Bowl” są głównymi rodzajami opakowań.

W tej hipotezie posłużymy się testem niezależności chi-kwadrat. Test ten pozwala ocenić czy zaobserwowany rozkład zależy od drugiej zmiennej.

```
chisq <- chisq.test(ramen$Style,ramen$Continent,simulate.p.value = TRUE)
chisq
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  ramen$Style and ramen$Continent
## X-squared = 253.2, df = NA, p-value = 0.001499
```

W naszym przypadku, rzędy(rodzaje pojemników) i kolumny(kontynenty) są statystycznie istotne (odrzucamy hipotezę zerową na rzecz alternatywnej), p.value < 0.05. Czyli możemy uznać, że są od siebie zależne, teraz pokażemy tylko które opakowanie z którym kontynentem jest najbardziej związane.

```
chisq$observed # nasze wartośći zaobserwowane
```

```
##            ramen$Continent
## ramen$Style Asia Americas Europe Oceania Africa
##        Bar     0        1      0       0      0
##        Bowl  401       78      2       0      0
##        Box     5        1      0       0      0
##        Can     0        1      0       0      0
##        Cup   280      107     46      17      0
##        Pack 1267      161     88       9      3
##        Tray   56       52      0       0      0
```

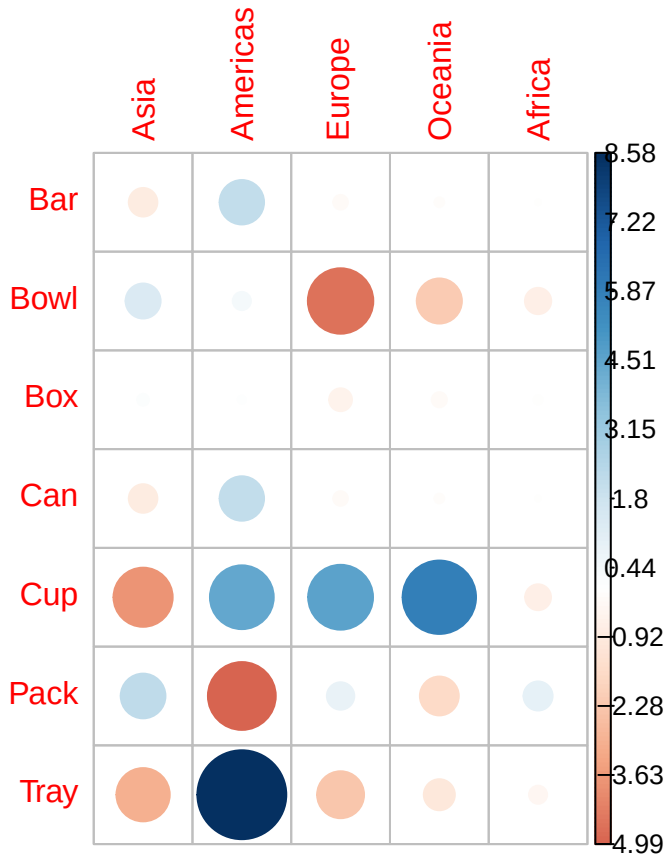
```
round(chisq$expected,2)  # średnie wartości
```

```
##            ramen$Continent
## ramen$Style    Asia Americas Europe Oceania Africa
##        Bar     0.78     0.16   0.05    0.01   0.00
##        Bowl  375.27    74.91  25.40    4.86   0.56
##        Box     4.68     0.93   0.32    0.06   0.01
##        Can     0.78     0.16   0.05    0.01   0.00
##        Cup   351.09    70.08  23.77    4.54   0.52
##        Pack 1192.14   237.95  80.70   15.43   1.78
##        Tray   84.26    16.82   5.70    1.09   0.13
```

```
round(chisq$residuals, 3) # reszty
```

```
##            ramen$Continent
## ramen$Style   Asia Americas Europe Oceania Africa
##        Bar  -0.883    2.139 -0.230  -0.100 -0.034
##        Bowl  1.328    0.358 -4.643  -2.204 -0.749
##        Box   0.147    0.068 -0.563  -0.246 -0.084
##        Can  -0.883    2.139 -0.230  -0.100 -0.034
##        Cup  -3.794    4.411  4.560   5.844 -0.724
##        Pack  2.168   -4.989  0.812  -1.637  0.914
##        Tray -3.079    8.579 -2.388  -1.044 -0.355
```

```
corrplot(chisq$residuals, is.cor = FALSE)
```

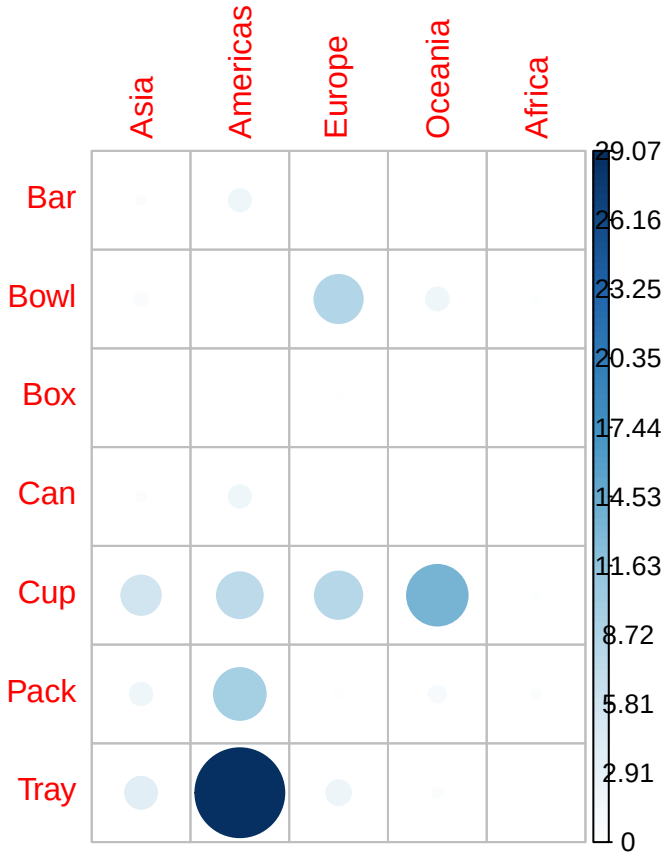


Zwizualizowaliśmy reszty testu chi kwadrat na wykresie. Na niebiesko jest zaznaczony pozytywny związek pomiędzy zmiennymi. Przykładowo taki związek mamy pomiędzy Afryką a "Tray", czy np "Cup" a Oceanią. Natomiast negatywny związek, czyli nie ma korelacji to pomiędzy "Box" a Europą, albo "Pack" a Ameryką. Negatywny związek w prostym języku to odpychanie się tych zmiennych od siebie. Czyli jak jedna będzie wzrastać to to nie ma wpływu na wzrost drugiej. Procentowy udział zmiennych:

```
contrib <- 100*chisq$residuals^2/chisq$statistic
round(contrib, 3)
```

```
##            ramen$Continent
## ramen$Style   Asia Americas Europe Oceania Africa
##        Bar   0.308    1.808  0.021   0.004  0.000
##        Bowl  0.697    0.050  8.516   1.918  0.221
##        Box   0.009    0.002  0.125   0.024  0.003
##        Can   0.308    1.808  0.021   0.004  0.000
##        Cup   5.685    7.683  8.214  13.487  0.207
##        Pack  1.857    9.829  0.261   1.058  0.330
##        Tray  3.744   29.066  2.253   0.431  0.050
```

```
corrplot(contrib, is.cor = FALSE)
```



Względny udział każdej komórki w całkowitym wyniku Chi-kwadrat daje pewne wskazanie charakteru zależności między wierszami i kolumnami tabeli kontyngentów. Możemy zauważyć, że:

- Kolumna "America" jest silnie związana z kategoriami "Tray", "Cup","Pack"
- Kolumna "Europa" jest silnie powiązana z "Bowl" i "Cup"
- Wiersz "Cup" jest połączony z każdym kontynentem prócz Afryka

Czy cecha ramenu ma wpływ na jego ocenę?

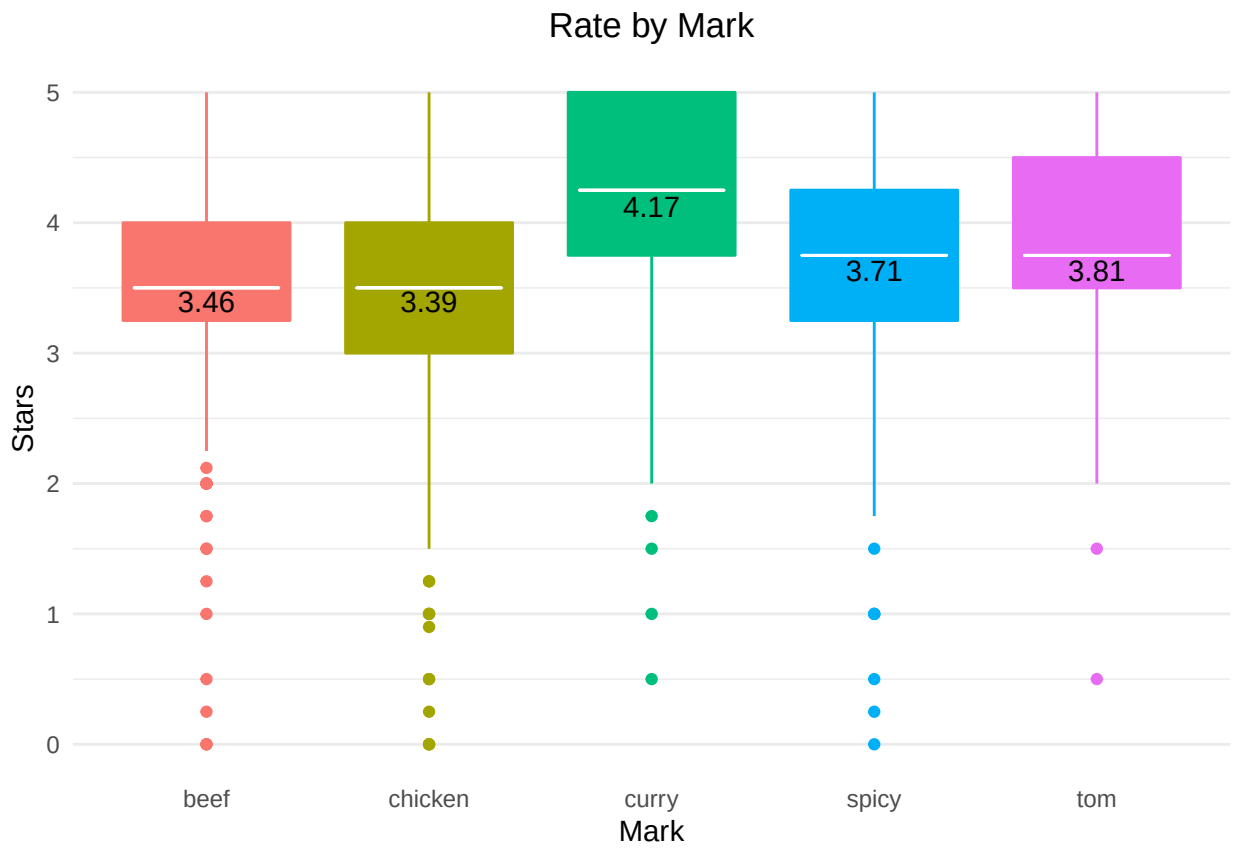
Hipoteza H_0: Cecha nie wpływa na ocenę

Hipoteza H_1: Cecha wpływa na ocenę

Wybraliśmy dane z top 5 cechami, aby zminimalizować ilość grup, a jednocześnie mając dużą ilość danych.

```
ramen.top.mark <- ramen %>% filter_at(vars(Mark),
                                      any_vars(. %in% as.character(mark.table[1:5,1])))
```

Rysujemy wykresy pudełkowe




```
leveneTest(Stars ~ Mark, ramen.top.mark)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.3367 0.8533
##       998
```

```
fligner.test(Stars ~ Mark, ramen.top.mark)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Stars by Mark
## Fligner-Killeen:med chi-squared = 2.2199, df = 4, p-value = 0.6954
```

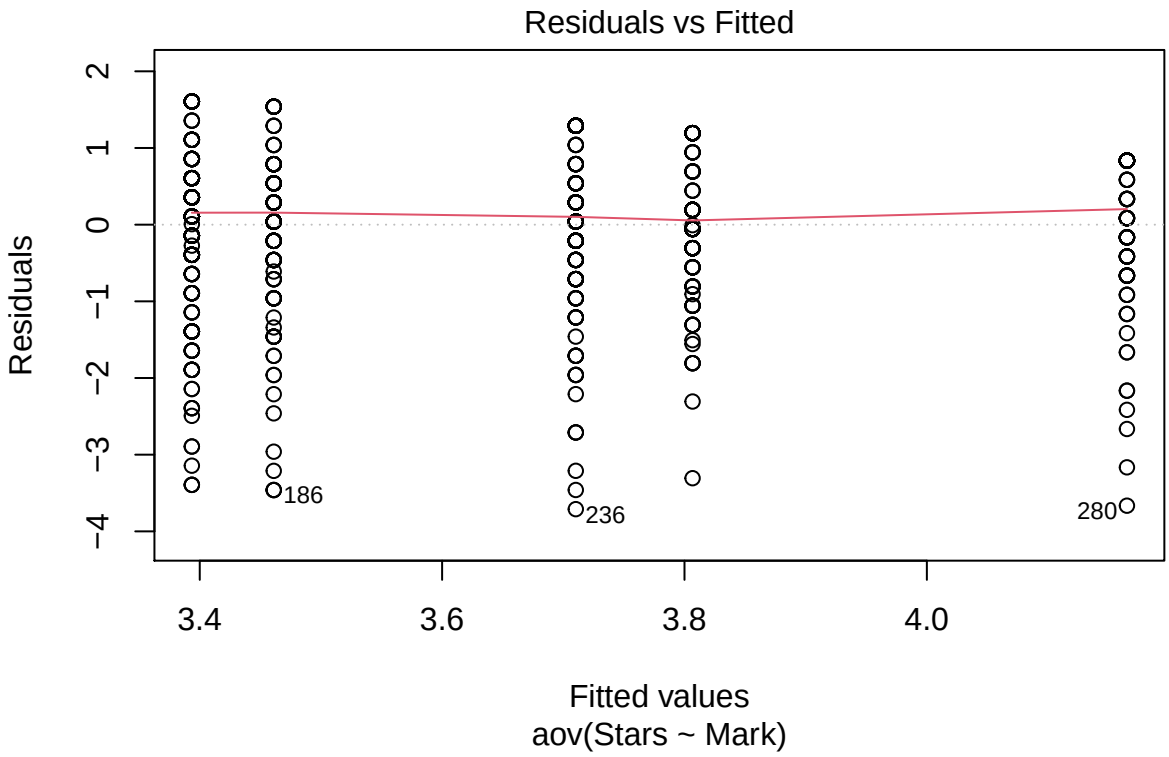
Widzimy, że p-wartość dla obu testów wyszła duża, a zatem nie odrzucamy hipotezy iż wariancje są jednorodne. Standardowo kolejny nasz krok to dopasowanie modelu ANOVA:

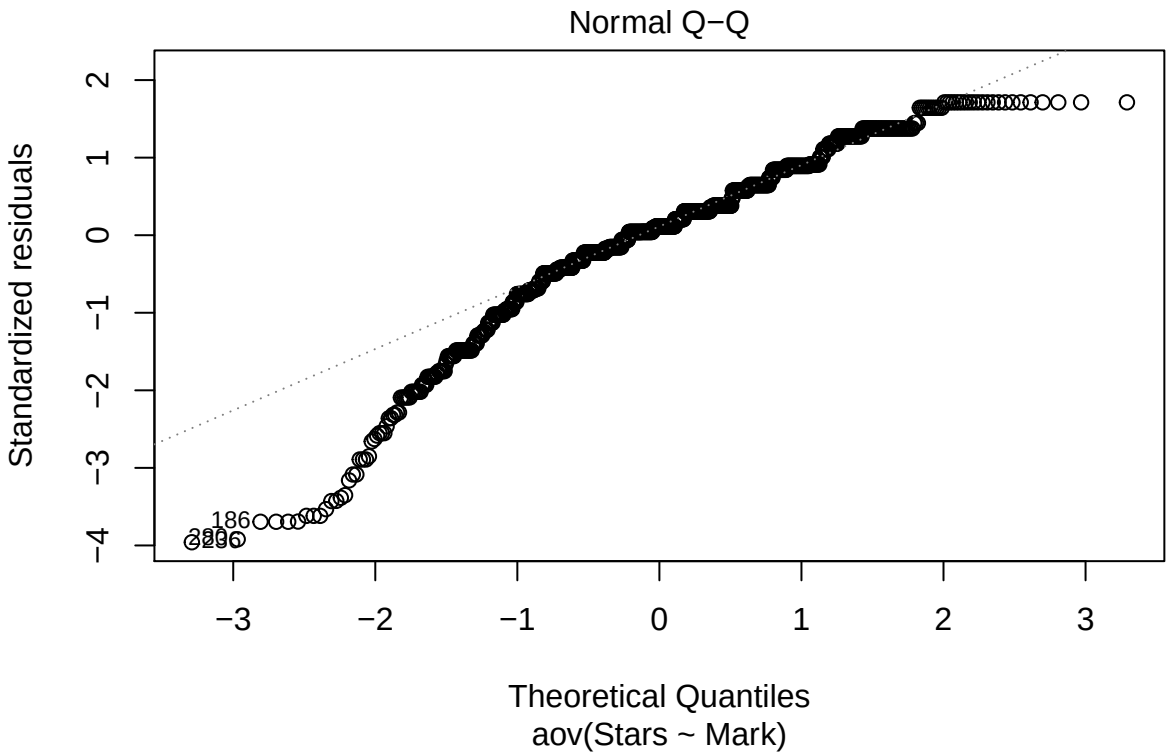
```
aov.mark <- aov(Stars~Mark, ramen.top.mark)
summary(aov.mark)
```

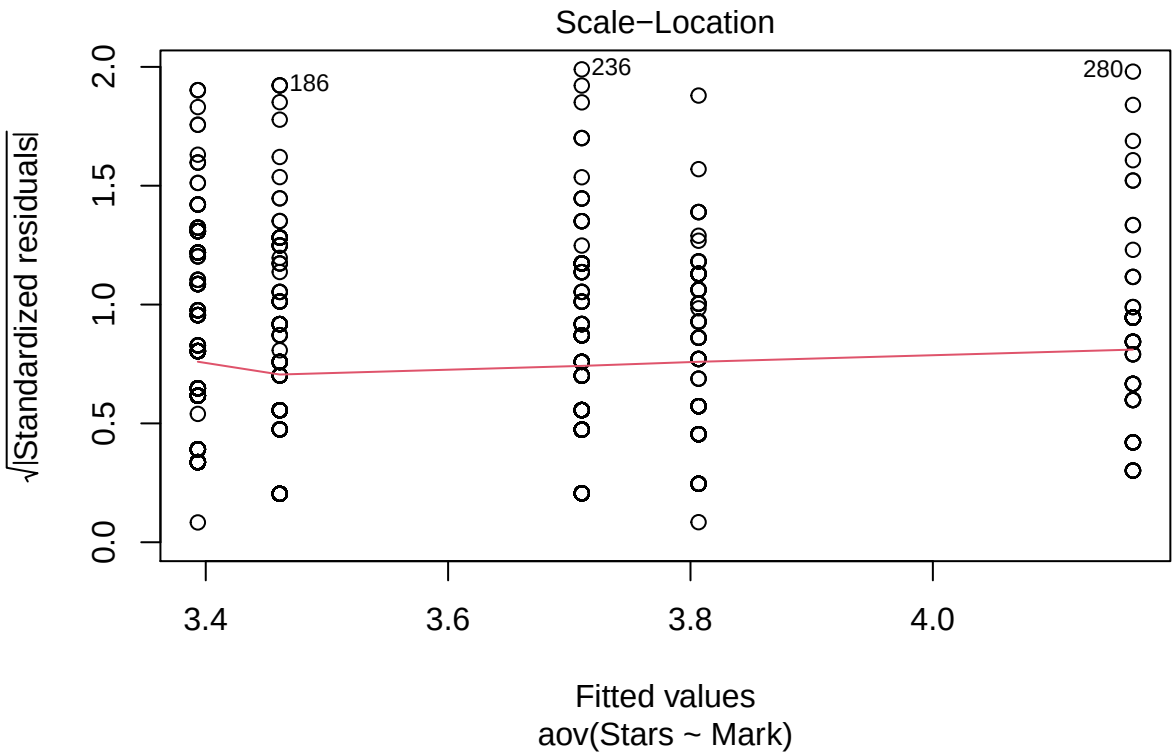
```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Mark          4   60.1  15.029   17.04 1.6e-13 ***
## Residuals   998  880.5   0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

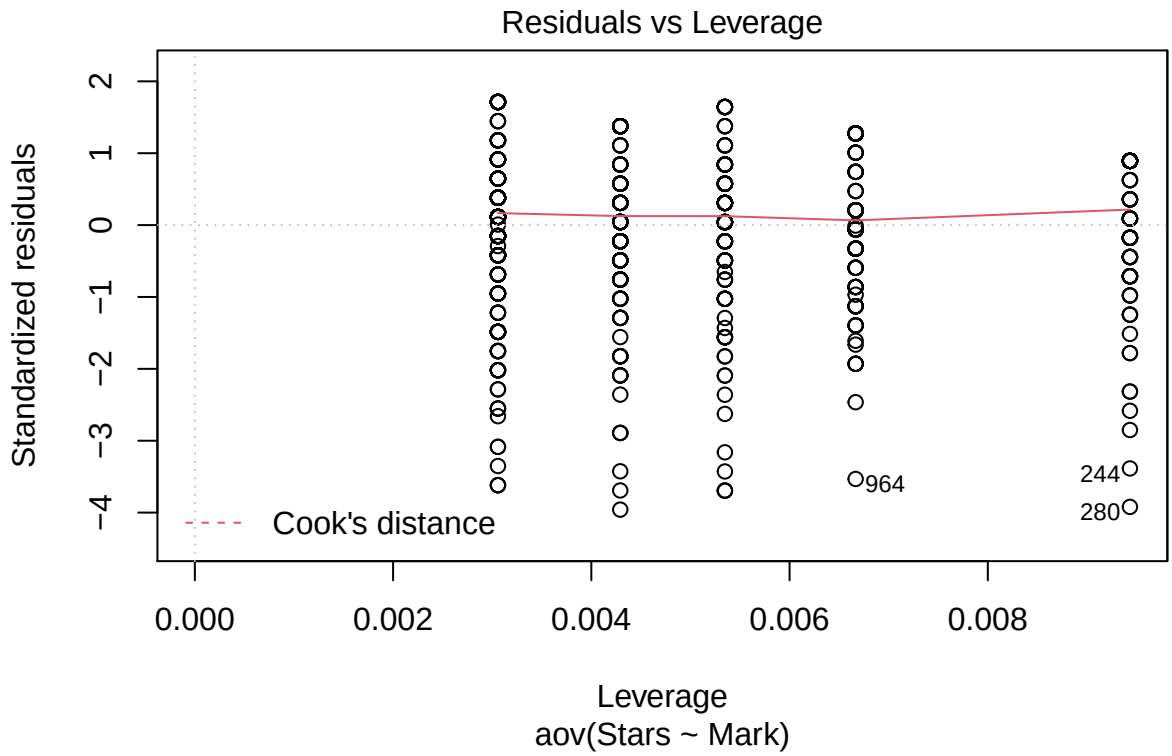
Test F wykazał istotną zależność pomiędzy średnimi w grupach określonych zmienną Mark.

```
plot(aov.mark)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(Stars ~ Mark)
186
−4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
186
236
280
3.4
3.6
3.8
4.0
Fitted values
aov(Stars ~ Mark)



I tak jak to było w przypadku wpływu kontynentu na ocenę, tak tutaj nie mamy do czynienia z rozkładem normalnym.

```
summary.lm(aov.mark)
```

```
##
## Call:
## aov(formula = Stars ~ Mark, data = ramen.top.mark)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7103 -0.3935  0.1065  0.6065  1.6065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.46107    0.06869  50.390  < 2e-16 ***
## Markchicken -0.06758    0.08611  -0.785 0.432754
## Markcurry    0.70402    0.11420   6.165 1.02e-09 ***
## Markspicy    0.24923    0.09222   2.703 0.006996 **
## Marktom      0.34560    0.10295   3.357 0.000818 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9393 on 998 degrees of freedom
## Multiple R-squared:  0.06391,    Adjusted R-squared:  0.06016
## F-statistic: 17.03 on 4 and 998 DF,  p-value: 1.6e-13
```

Ponownie używamy testów nieparametrycznych.

```
kruskal.test(Stars ~ Mark, ramen.top.mark)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Stars by Mark
## Kruskal-Wallis chi-squared = 74.097, df = 4, p-value = 3.093e-15
```

```
mood.medtest(Stars~Continent, data=ramen, exact = FALSE)
```

```
##
##  Mood's median test
##
## data:  Stars by Continent
## X-squared = 74.509, df = 4, p-value = 2.531e-15
```

mediany różnią się od siebie, a więc co najmniej jedna grupa różni się od innych.

```
as.matrix(as.data.frame(ramen.top.mark)) -> ramen.top.mark
posthocTGH(as.numeric(ramen.top.mark[,6]),as.factor(ramen.top.mark[,10]),
           method = "games-howell")
```

```
##           n means variances
## beef    187   3.5      0.89
## chicken 327   3.4      0.94
## curry   106   4.2      0.92
## spicy   233   3.7      0.87
## tom     150   3.8      0.74
##
##                 diff   ci.lo  ci.hi    t  df    p
## chicken-beef  -0.068 -0.3068  0.172 0.77 396  .94
## curry-beef     0.704  0.3856  1.022 6.08 215 <.01
## spicy-beef     0.249 -0.0035  0.502 2.70 397  .06
## tom-beef       0.346  0.0756  0.616 3.51 329 <.01
## curry-chicken  0.772  0.4759  1.067 7.19 180 <.01
## spicy-chicken  0.317  0.0941  0.540 3.89 510 <.01
## tom-chicken    0.413  0.1708  0.656 4.68 323 <.01
## spicy-curry   -0.455 -0.7613 -0.148 4.08 199 <.01
## tom-curry     -0.358 -0.6792 -0.038 3.08 211  .02
## tom-spicy      0.096 -0.1593  0.352 1.03 337  .84
```

Głównie chicken różni się od pozostałych, curry z beef też, tom-beef oraz spicy-curry. Możemy stwierdzić, że istnieje prawdopodbieństwo, że ocena zależy od cechy ramenu, chociaż w małym stopniu.

Czy duże korporacje osiągają te same wyniki recenzji na różnych kontynentach?

Hipoteza H_0: Ocena jest zależna od kontynentu

Hipoteza H_1: Ocena nie jest zależna od kontynentu

Jeżeli dowiedziemy, że są znaczące różnice w średnich miedzy kontynentami, to oznacza, że Nie osiagają tych samych wyników. Na początek wybór danych oraz wykres. Wybieramy dwie firmy z największych: Nissin i Maruchan.

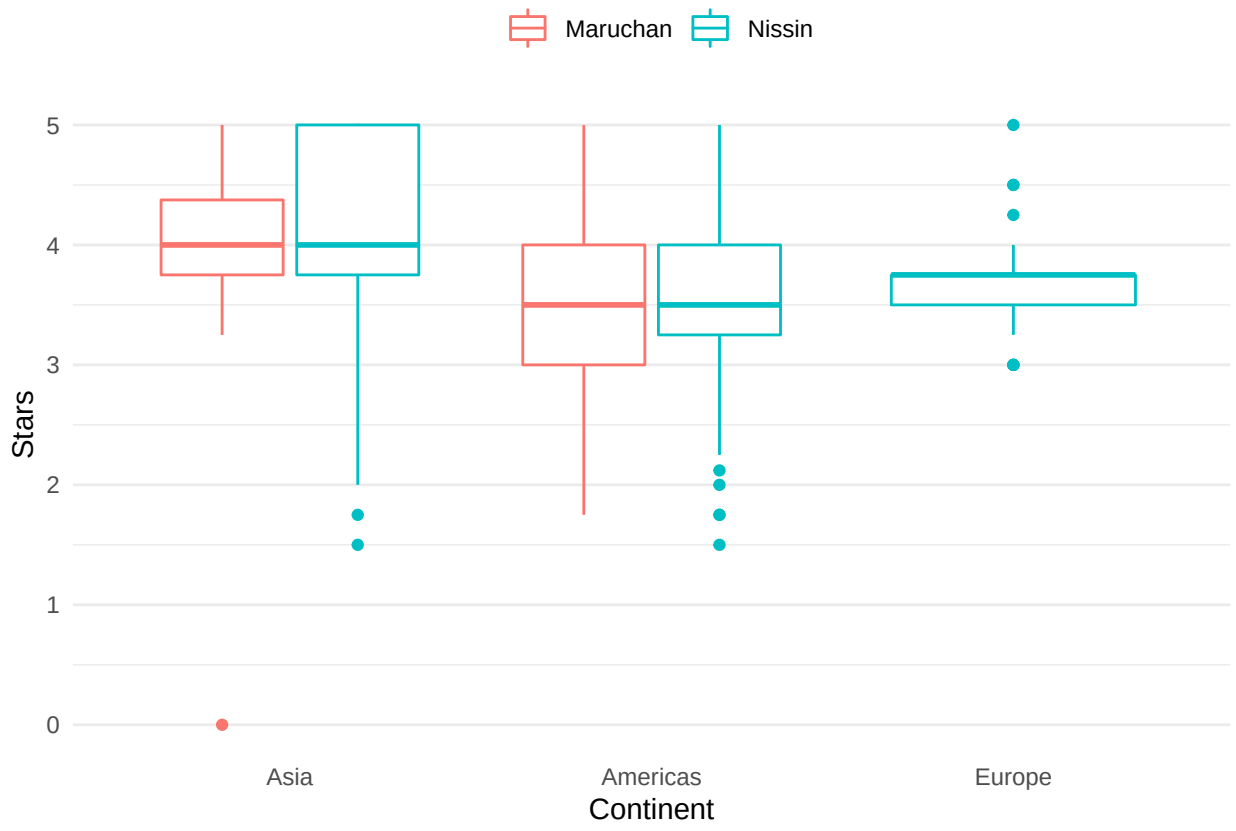
Nonghsim ze nieznanych problemóW technicznych nie chciało działać, więc wybrałyśmy trzecią firmę.

```
aggregate(Variety ~ Brand, ramen, length) -> ramen.values
ramen.values[order(ramen.values$Variety,decreasing = TRUE),]
```

```r
ramen[ramen$Brand == "Nissin",] -> nissin
ramen[ramen$Brand == "Maruchan",] -> nongshim
ramen.brand <- rbind(nissin, nongshim)
head(ramen.brand)
```

```
##    Review..  Brand                                                 Variety
## 3      2578 Nissin                                Cup Noodles Chicken Vegetable
## 14     2567 Nissin                                 Deka Buto Kimchi Pork Flavor
## 15     2566 Nissin Demae Ramen Bar Noodle Aka Tonkotsu Flavour Instant Noodle
## 21     2560 Nissin                                     Cup Noodles Laksa Flavour
## 28     2553 Nissin                               Hakata Ramen Noodle White Tonkotsu
## 32     2549 Nissin                             Gekikara Ramen Extra Hot Ramen Goreng
##    Style   Country Stars Top.Ten Continent StarsInterval     Mark
## 3    Cup       USA  2.25          Americas           2-3  chicken
## 14  Bowl     Japan  4.50              Asia           4-5     pork
## 15  Pack Hong Kong  5.00              Asia           4-5    demae
## 21   Cup Hong Kong  4.25              Asia           4-5    laksa
## 28  Bowl     Japan  4.75              Asia           4-5 tonkotsu
## 32  Pack Indonesia  4.50              Asia           4-5      hot
```

```r
title = "Rate by Continents"
theme = theme_set(theme_minimal())
theme = theme_update(legend.position="top",legend.title=element_blank(),
                     panel.grid.major.x=element_blank())
ggplot(ramen.brand, aes(x = Continent, y = Stars, color = Brand)) +  # ggplot function
  geom_boxplot()
```



Będziemy brać pod uwagę jedynie Azję i Ameryki, ze względu na to, że w Europie nie występuje Nonghsim.

```
ramen.brand <- ramen.brand[ramen.brand$Continent %in% c("Asia", "Americas"), ]
ramen.brand$Brand <- as.factor(as.character(ramen.brand$Brand))
ramen.brand$Continent <- as.factor(as.character(ramen.brand$Continent))
```

Przeprowadzamy test na jednolitość wariancji:

```
leveneTest(Stars~Continent, ramen.brand %>% filter(Brand == "Nissin"))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  2.2268 0.1365
##       356
```

```
leveneTest(Stars~Continent, ramen.brand %>% filter(Brand == "Maruchan"))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.3217 0.5723
##       74
```

W obu grupach p-value jest duże więc nie odrzucamy hipotezy zerowej o jednolitości wariancji.

```
aov.nissin <- aov(Stars~Continent, ramen.brand %>% filter(Brand == "Nissin"))
aov.maruchan <- aov(Stars~Continent, ramen %>% filter(Brand == "Maruchan") %>%
                      filter(Continent == "Asia"|Continent =="Americas"))

summary(aov.nissin)
```

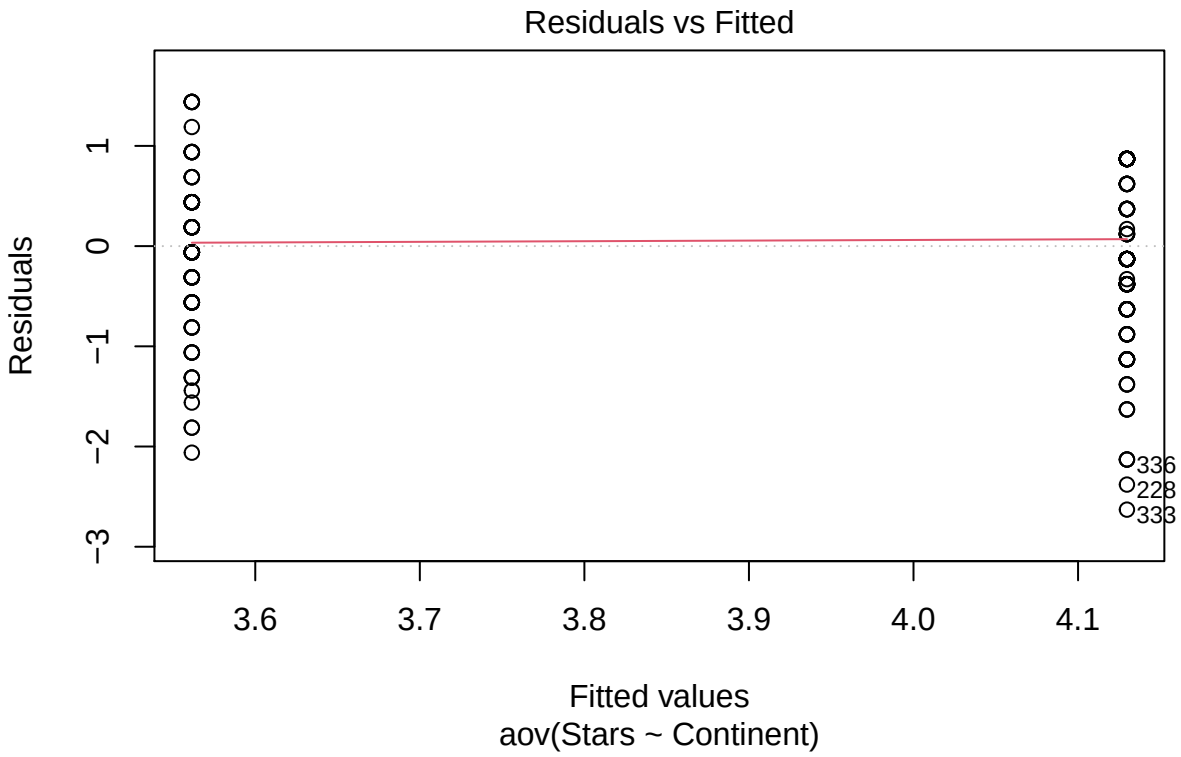
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Continent     1  26.17  26.172   48.54 1.57e-11 ***
## Residuals   356 191.94   0.539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

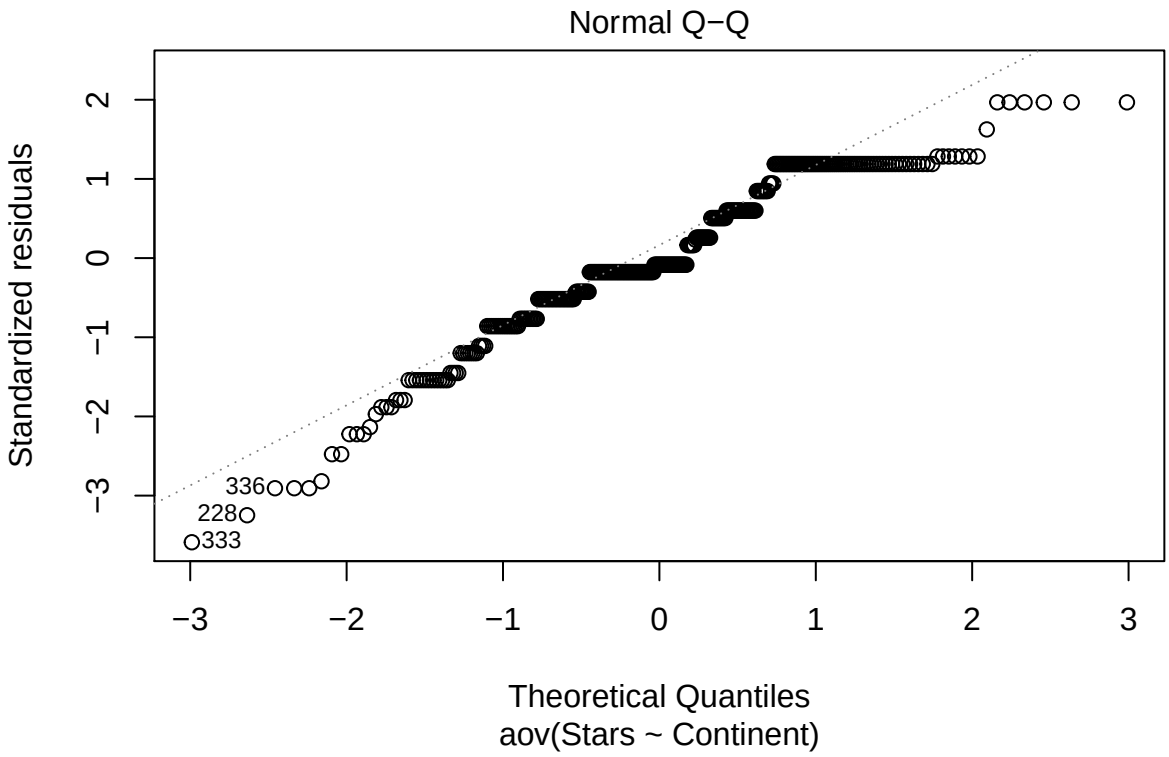
```
summary(aov.maruchan)
```

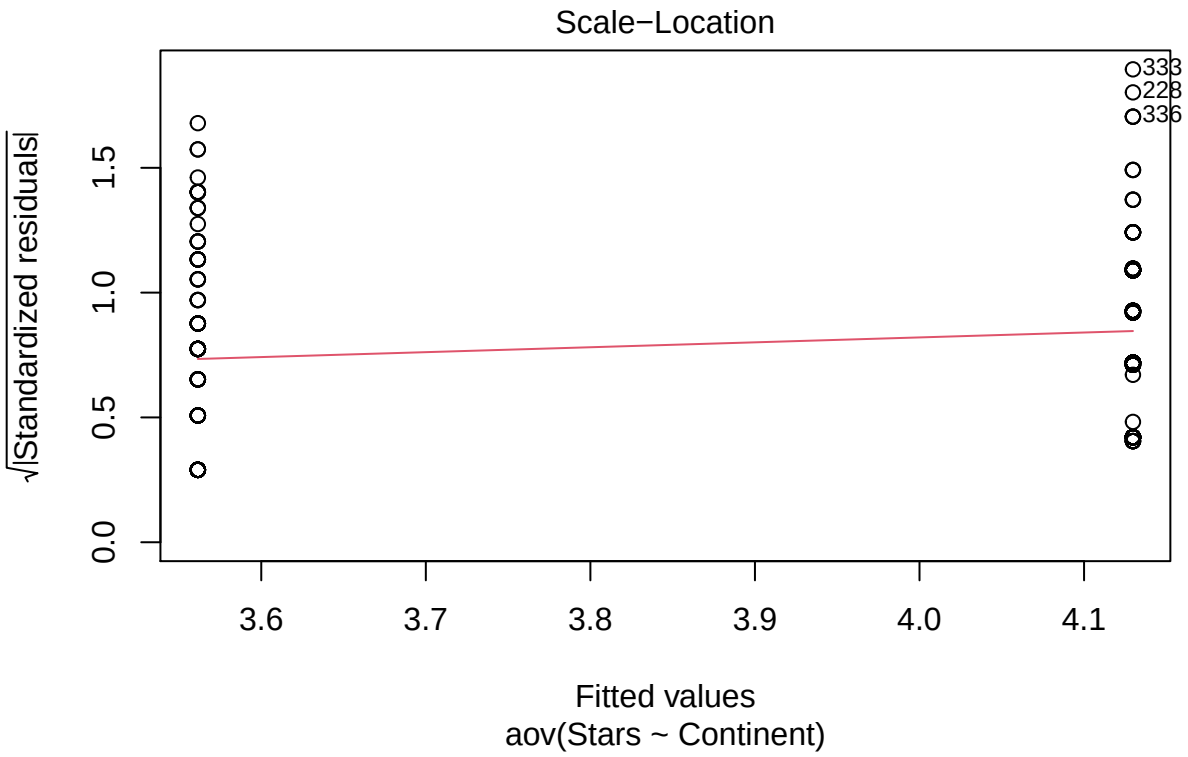
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Continent    1   4.91   4.909   6.853 0.0107 *
## Residuals   74  53.00   0.716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

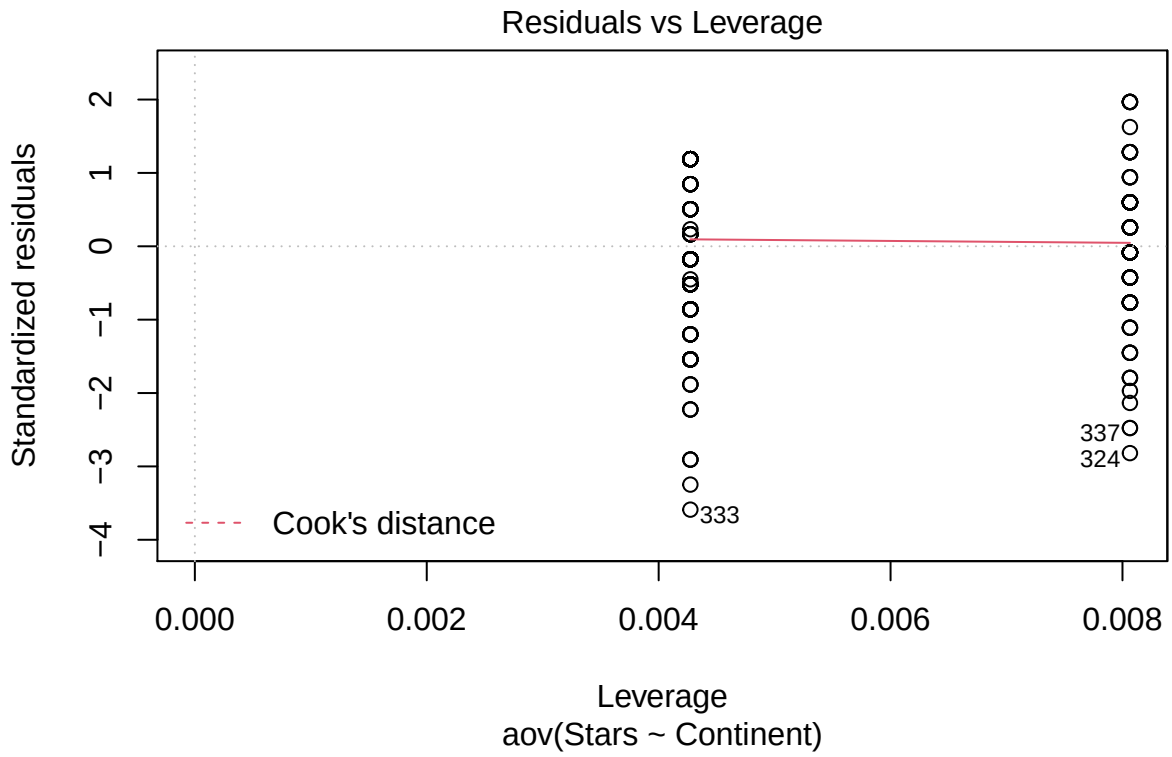
Widzimy, że w przypadku nissin jak i maruchan test F wykazał istotną zależność pomiędzy średnimi w grupach względem kontynentu.

```
plot(aov.nissin)
```

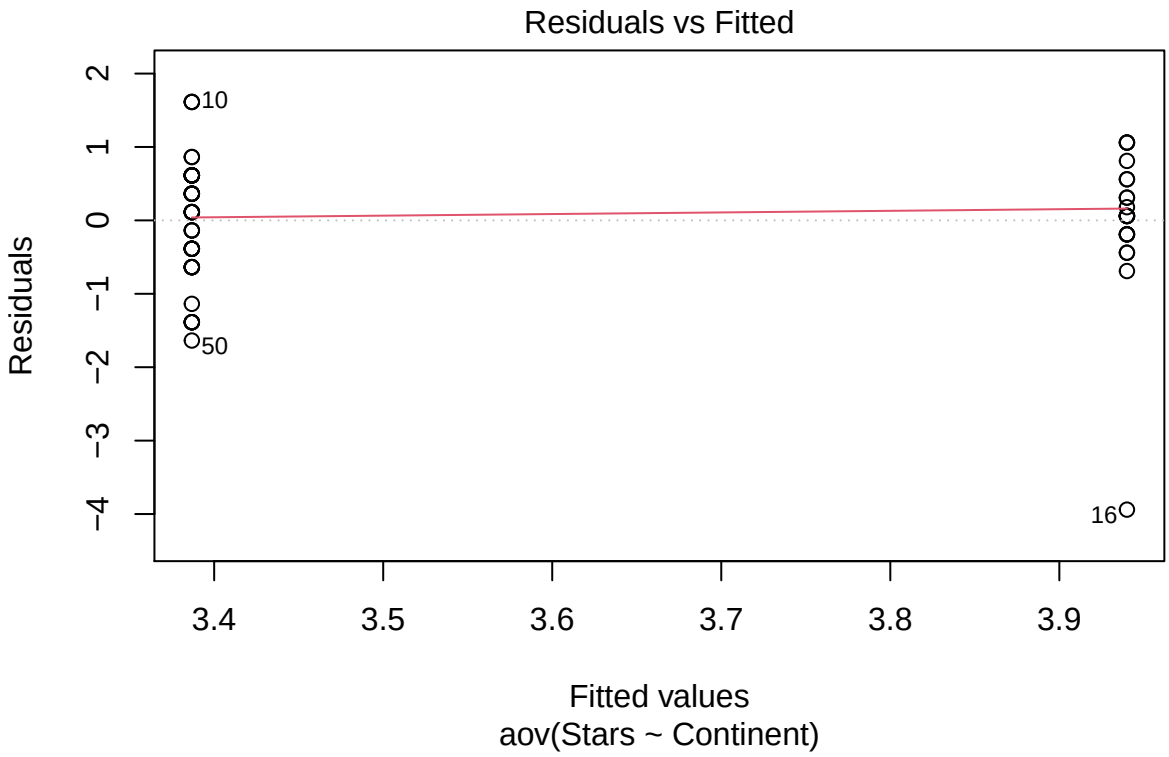

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
3.6
3.7
3.8
3.9
4.0
4.1
336
228
333
Fitted values
aov(Stars ~ Continent)

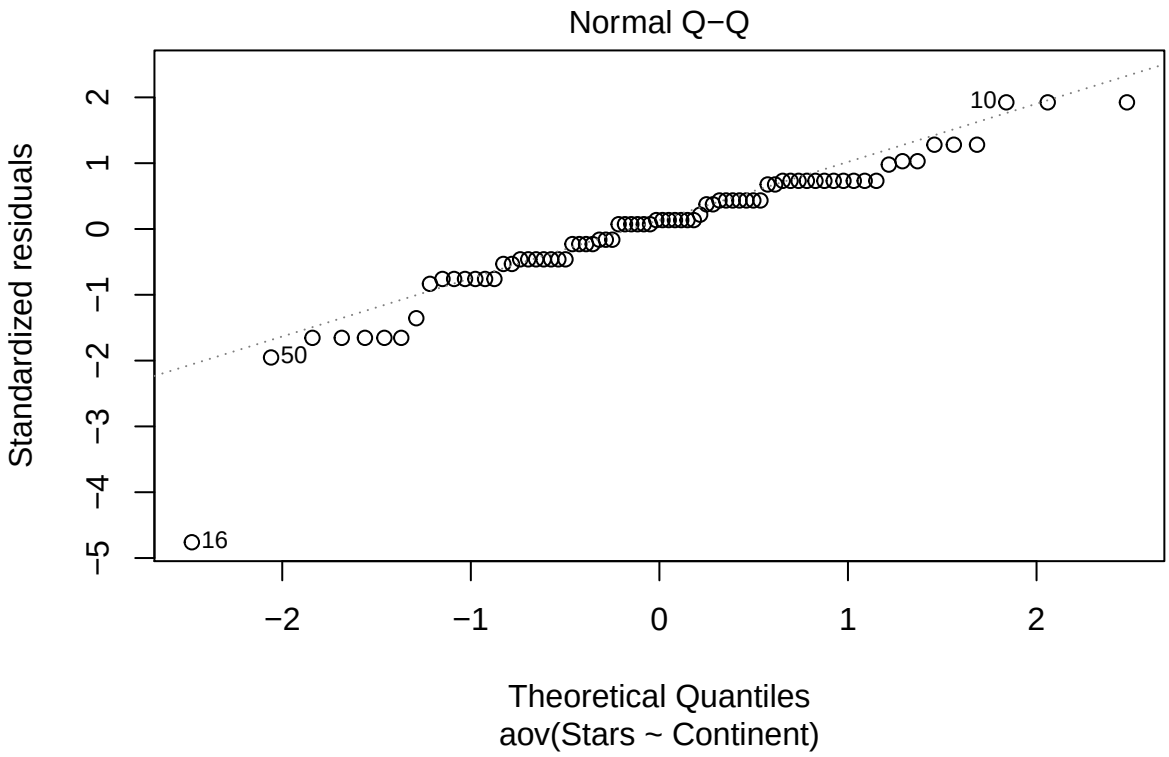

Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(Stars ~ Continent)
336
228
333
−3
−2
−1
0
1
2
3

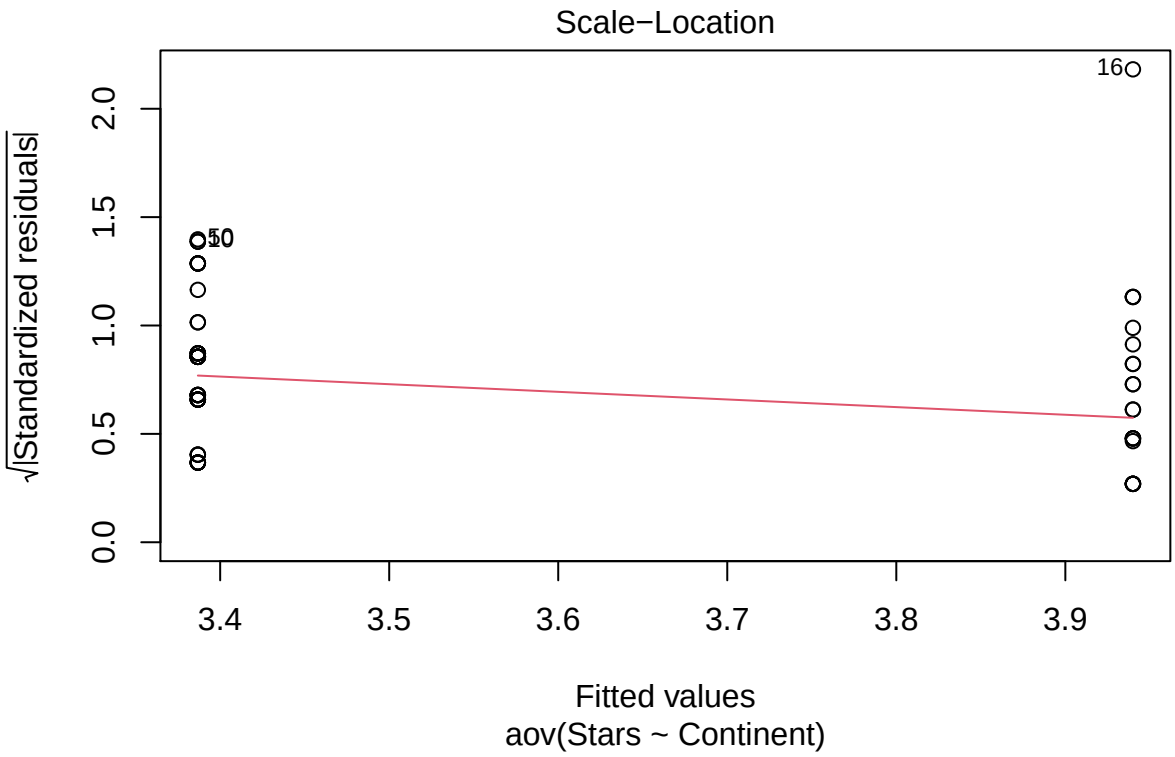

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3.6
3.7
3.8
3.9
4.0
4.1
333
228
336
Fitted values
aov(Stars ~ Continent)

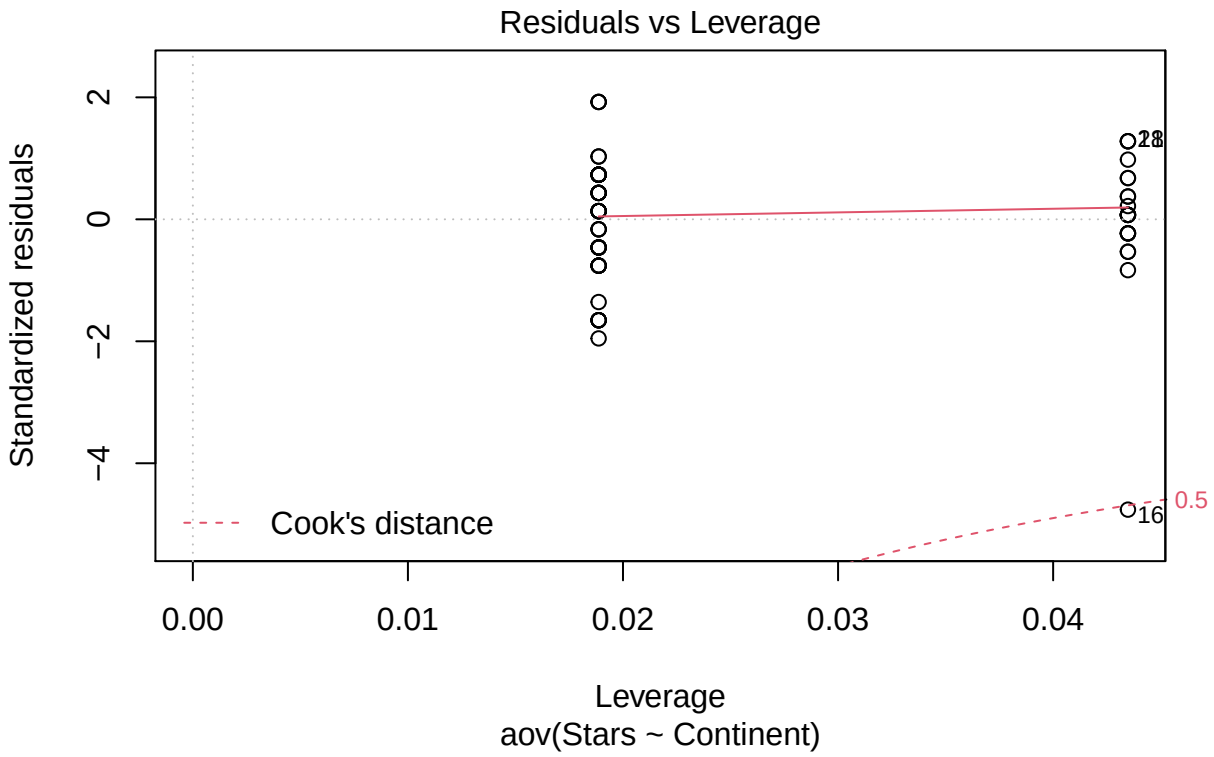



```
plot(aov.maruchan)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
−4
10
50
16
3.4
3.5
3.6
3.7
3.8
3.9
Fitted values
aov(Stars ~ Continent)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(Stars ~ Continent)
10
50
16
−5
−4
−3
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
16
50
3.4
3.5
3.6
3.7
3.8
3.9
Fitted values
aov(Stars ~ Continent)



Nie jesteśmy w stanie do końca odczytać z pierwszych wykresóW czy wariancje są sobie równe (ale wykazaliśmy to już w poprzednich testach). Zaskakujące może być dla nas to, że drugi wykres przedstawia w obu przypadkach całkiem dobrze dopasowany model liniowy. Odchyłki są niewielkie

```
summary.lm(aov.nissin)
```

```
##
## Call:
## aov(formula = Stars ~ Continent, data = ramen.brand %>% filter(Brand ==
##     "Nissin"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62970 -0.37970 -0.06145  0.62030  1.43855
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.56145    0.06594  54.010  < 2e-16 ***
## ContinentAsia  0.56825    0.08156   6.967 1.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7343 on 356 degrees of freedom
## Multiple R-squared:   0.12,  Adjusted R-squared:  0.1175
## F-statistic: 48.54 on 1 and 356 DF,  p-value: 1.572e-11
```

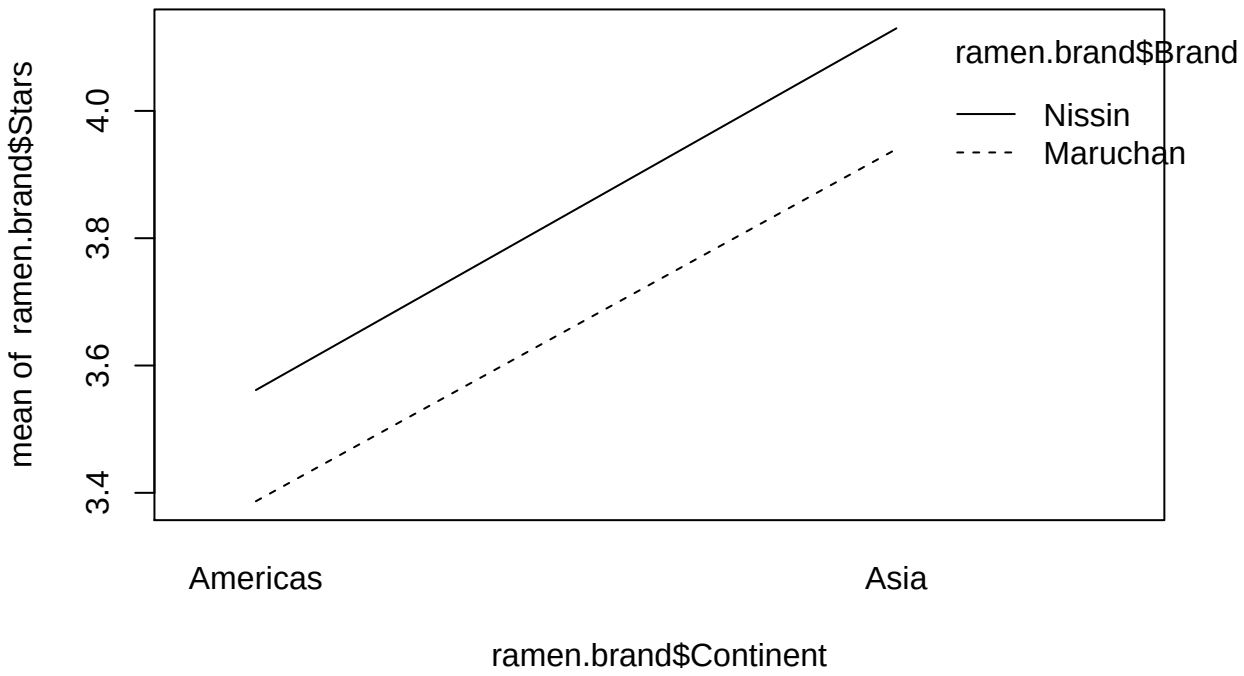
```
summary.lm(aov.maruchan)
```

```
##
## Call:
## aov(formula = Stars ~ Continent, data = ramen %>% filter(Brand ==
##     "Maruchan") %>% filter(Continent == "Asia" | Continent ==
##     "Americas"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9400 -0.3868  0.1132  0.6132  1.6132
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.9400     0.1765  22.327   <2e-16 ***
## ContinentAmericas  -0.5532     0.2113  -2.618   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8463 on 74 degrees of freedom
## Multiple R-squared:  0.08476,    Adjusted R-squared:  0.07239
## F-statistic: 6.853 on 1 and 74 DF,  p-value: 0.01072
```

W firmie nissin Ameryka powoduje spadek oceny o około 0.56. Podobnie dzieje się w firmie maruchan.

Narysujemy teraz wykres interakcji.

```
interaction.plot(ramen.brand$Continent, ramen.brand$Brand, ramen.brand$Stars)
```



Czyli w obu firmach kontynent ma wpływ na średnią ocenę ramenu. W Azji odnotowywujemy większe wartości.

Spróbujmy dopasować dwuczynnikowy model z wszystkimi interakcjami. Spodziewamy się, że Brand nie ma wpływu na ocenę jak to wcześniej było pokazane na wykresach.

```
twoway.full <- aov(Stars~Continent*Brand, ramen.brand)
summary(twoway.full)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## Continent         1  38.18   38.18  67.031 3.07e-15 ***
## Brand             1   1.88    1.88   3.306   0.0697 .
## Continent:Brand   1   0.00    0.00   0.005   0.9419
## Residuals       430 244.94    0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nasz test potwierdził, że Brand nie ma wpływu na ocenę ramenu na poszczególnych kontynentach. Nie jesteśmy w takim razie wysnuć wniosku, czy dane firmy dbają o jakość produktu na całym świecie, ponieważ nie ma pomiędzy temu związku.

```
summary.lm(twoway.full)
```

```
##
## Call:
## aov(formula = Stars ~ Continent * Brand, data = ramen.brand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.9400 -0.3797 -0.0615  0.6132  1.6132
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.38679    0.10367  32.668  < 2e-16 ***
## ContinentAsia               0.55321    0.18845   2.936  0.00351 **
## BrandNissin                 0.17466    0.12386   1.410  0.15923
## ContinentAsia:BrandNissin  0.01504    0.20626   0.073  0.94190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7547 on 430 degrees of freedom
## Multiple R-squared:  0.1406, Adjusted R-squared:  0.1346
## F-statistic: 23.45 on 3 and 430 DF,  p-value: 4.495e-14
```

Sprawdźmy jeszcze model addytywny. Podejrzewamy, że tak samo jak w poprzednim modelu, nie będziemy w stanie wyciągnąć istotnie statystycznych informacji.

```
twoway.add <- aov(Stars~Continent+Brand, ramen.brand)
summary(twoway.add)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Continent     1  38.18   38.18  67.186 2.85e-15 ***
## Brand         1   1.88    1.88   3.314   0.0694 .
## Residuals   431 244.95    0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nasze obawy się sprawdziły. Jedynie kontynent ma wpływ na ocenę ramenu. Na poprzednich wykresach pudełkowych widzieliśmy, że Azja istotnie wpływa na wzrost oceny ramenu.

```
anova(twoway.add, twoway.full)
```

```
## Analysis of Variance Table
##
## Model 1: Stars ~ Continent + Brand
## Model 2: Stars ~ Continent * Brand
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    431 244.95
## 2    430 244.94  1 0.0030295 0.0053 0.9419
```

Ostatecznie potwierdzenie, że żaden model nie jest dobrze dopasowany.