

# WhaSAP\_vivino\_analysis

2025-01-25

Učitajmo podatke:

```
dataset_vina = read.csv("vivino dataset.csv")
dim(dataset_vina)
```

```
## [1] 12205    13
```

Iz kojih sve država nam dolaze podatci o vinima?

```
unique(dataset_vina$Country)
```

```
## [1] "Argentina"      "Australia"       "Chile"          "Germany"
## [5] "Spain"           "France"          "Italy"           "Portugal"
## [9] "United States"   "South Africa"
```

Koliko zapisa o vinima imamo iz svake od tih pojedinih država?

```
table(dataset_vina$Country)
```

```
##
##          Argentina      Australia       Chile        France       Germany
##             418              391            330          2022         1067
##          Italy          Portugal     South Africa     Spain United States
##            1890             1784            1485          2019          799
```

Kratki pregled vrijednosti podataka:

```
summary(dataset_vina)
```

```
##      Winery          Year          Wine_ID          Wine
##  Length:12205  Length:12205  Min.   : 531  Length:12205
##  Class  :character  Class  :character  1st Qu.:1135203  Class  :character
##  Mode   :character  Mode   :character  Median  :1425545  Mode   :character
##                               Mean   :2122684
##                               3rd Qu.:2486838
##                               Max.  :10205770
##      Rating          Reviews          Price          Region
##  Min.   :1.90  Min.   : 25.0  Min.   : 2.07  Length:12205
##  1st Qu.:3.70  1st Qu.: 55.0  1st Qu.: 8.95  Class  :character
##  Median :3.90  Median :122.0  Median :17.90  Mode   :character
##  Mean   :3.92  Mean   :498.5  Mean   :42.61
##  3rd Qu.:4.10  3rd Qu.:330.0  3rd Qu.:37.00
##  Max.   :4.90  Max.   :114425.0 Max.   :6511.31
##      Primary_Grape      Natural          Country          Style
##  Length:12205  Length:12205  Length:12205  Length:12205
##  Class  :character  Class  :character  Class  :character  Class  :character
##  Mode   :character  Mode   :character  Mode   :character  Mode   :character
##      
```

```
##  
## Country_Code  
## Length:12205  
## Class :character  
## Mode :character  
##  
##  
##
```

Svaki redak predstavlja značajke o vinu: Winery - Naziv vinarije koja proizvodi vino. Year - Godina berbe ili proizvodnje vina. WineID - Jedinstveni identifikacijski broj vina u bazi podataka. Wine - Naziv vina. Rating - Prosječna ocjena vina na temelju recenzija, na ljestvici od 1.0 do 5.0. Reviews - Broj recenzija koje je vino primilo. Price - Cijena vina (valuta nije navedena). Region - Regija u kojoj je vino proizvedeno. Primary\_Grape - Glavna sorta grožđa korištena u vinu. Natural - Označava je li vino prirodno. *Country - Država u kojoj je vino proizvedeno*. *Style - Vrsta vina (npr. crveno, bijelo, pjenušavo)*. *Country\_Code - Kod države prema međunarodnom standardu*. prirodno vino podrazumijeva proizvodnju uz minimalno intervenciju, tj. bez kemikalija i pesticida

Koje sve vrste vina (Style) imamo među danim podatcima?

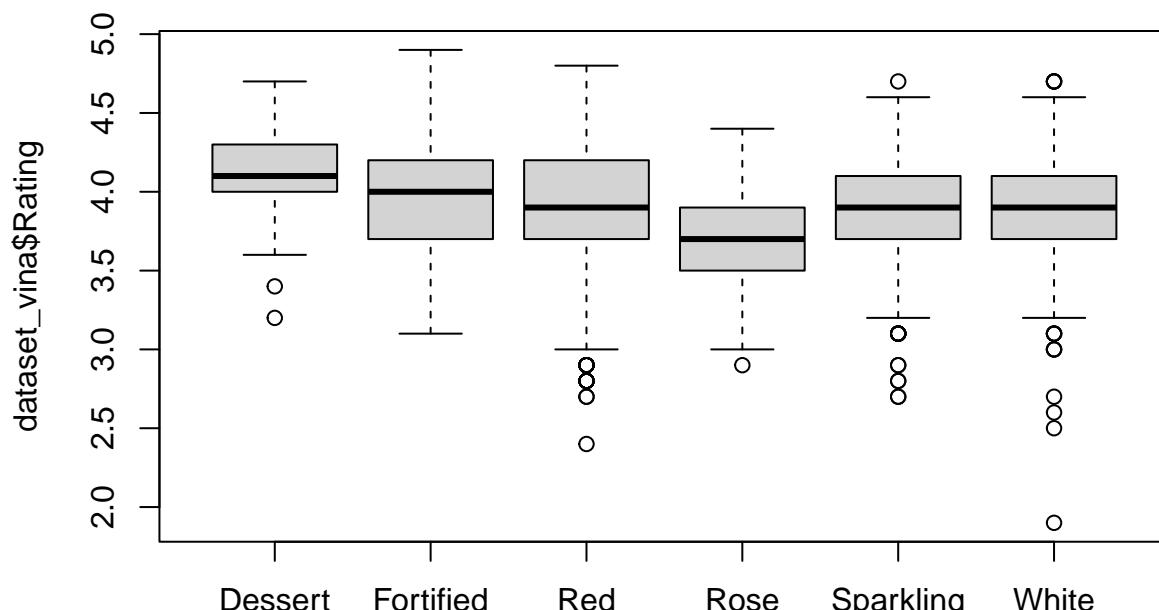
```
unique(dataset_vina$Style)      # jedinstveni nazivi vrste vina  
  
## [1] "Rose"      "White"     "Red"       "Sparkling" "Dessert"   "Fortified"
```

Neka od ključnih pitanja koja nas zanimaju su:

- Postoji li razlike u ocjenama vina među različitim vrstama vina?

Provjerimo vizualno postoje li razlike u ocjenama za različite vrste vina.

```
# Graficki prikaz podataka  
boxplot(dataset_vina$Rating ~ dataset_vina$Style)
```



Na boxplot-u možemo vidjeti da određene srednje vrijednosti (poput one za npr. Rose) znatno odskaču od ostalih, ali postoje i srednje vrijednosti koje djeluju blisko (npr. za Dessert, Fortified i Sparkling).

# ANOVA

ANOVA (engl. *ANalysis Of VAriance*) je metoda kojom testiramo sredine više populacija. U analizi varijance pretpostavlja se:da je ukupna varijabilnost u podatcima posljedica varijabilnosti podataka unutar svake pojedine grupe (populacije) i varijabilnosti između različitih grupa. Varijabilnost unutar pojedinog uzorka je rezultat slučajnosti, a ako postoje razlike u sredinama populacija, one će biti odražene u varijabilnosti među grupama. Jedan od glavnih ciljeva analize varijance je ustanoviti jesu li upravo te razlike između grupa samo posljedica slučajnosti ili su statistički značajne.

## Jednofaktorska ANOVA

Kod jednofaktorske ANOVA-e proučavamo k različitim populacijama koje se razlikuju na temelju jednog kriterija. Postoji k slučajnih uzoraka (po jedan iz svake populacije) i svaki uzorak je veličine n. Želimo testirati:  $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ ,  $H_1$ : barem dvije sredine nisu jednake

Pretpostavka koju podatci moraju poštivati kako bismo mogli provesti ANOVA test jest: - populacije su nezavisne - populacije su normalno distribuirane s očekivanjima  $\mu_1, \mu_2, \dots, \mu_k$  - populacije imaju jednake varijance  $\sigma^2$ .

Budući da su naši podatci jednoznačno podijeljeni u grupe na temelju vrste vina, te nema nikakvih preklapanja među grupama, zadovoljeno je svojstvo nezavisnosti populacija.

Sada slijedi provjera normalne distribucije populacije i jednakosti varijance među populacijama.

Provjera normalnosti može se za svaku pojedinu grupu napraviti Kolmogorov-Smirnovljevim testom ili Lillieforsovom inaćicom Kolmogorov-Smirnovljevog testa. Lillieforsovom inaćica KS testa koristi se kada želimo testirati da li podaci dolaze iz normalne distribucije, a ne poznaju se očekivanje i varijanca populacije. U ovom slučaju razmatrat ćemo vrstu vina kao varijablu koja određuje grupe (populacije) i razliku u ocjenama kao zavisnu varijablu.

$H_0$ : podatci dolaze iz normalne distribucije  $H_1$ : podatci ne dolaze iz normalne distribucije

Provjera normalnosti zavisne varijable ocjena na temelju svih ocjena. Provjera normalnosti zavisne varijable ocjena unutar grupe kreirane na temelju varijable vrste vina.

```
require(nortest)
```

```
## Loading required package: nortest
lillie.test(dataset_vina$Rating)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dataset_vina$Rating
## D = 0.080853, p-value < 2.2e-16
lillie.test(dataset_vina$Rating[dataset_vina$Style == 'Dessert'])

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Dessert"]
## D = 0.10375, p-value = 0.0006551
lillie.test(dataset_vina$Rating[dataset_vina$Style == 'Fortified'])

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```

## data: dataset_vina$Rating[dataset_vina$Style == "Fortified"]
## D = 0.086086, p-value = 5.411e-11
lillie.test(dataset_vina$Rating[dataset_vina$Style=='Red'])

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Red"]
## D = 0.085468, p-value < 2.2e-16
lillie.test(dataset_vina$Rating[dataset_vina$Style=='Rose'])

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Rose"]
## D = 0.082224, p-value = 7.92e-06
lillie.test(dataset_vina$Rating[dataset_vina$Style=='Sparkling'])

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Sparkling"]
## D = 0.089246, p-value < 2.2e-16
lillie.test(dataset_vina$Rating[dataset_vina$Style=='White'])

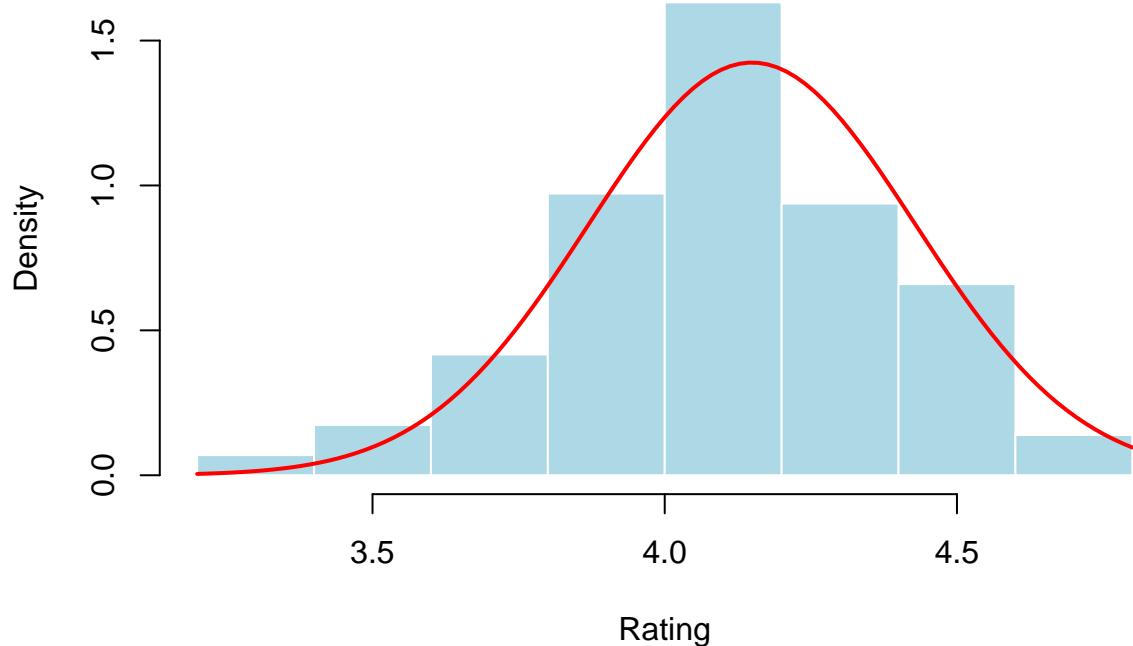
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dataset_vina$Rating[dataset_vina$Style == "White"]
## D = 0.081579, p-value < 2.2e-16

plot_histogram_with_normal_dist <- function(data, style_name) {
  hist(data, main = paste(style_name, "Ratings"), xlab = "Rating", prob = TRUE,
        col = "lightblue", border = "white")
  curve(dnorm(x, mean = mean(data), sd = sd(data)), add = TRUE, col = "red", lwd = 2)
}

# Prikaz podataka podijeljenih u grupe pomoću histograma
plot_histogram_with_normal_dist(dataset_vina$Rating[dataset_vina$Style == 'Dessert'], 'Dessert')

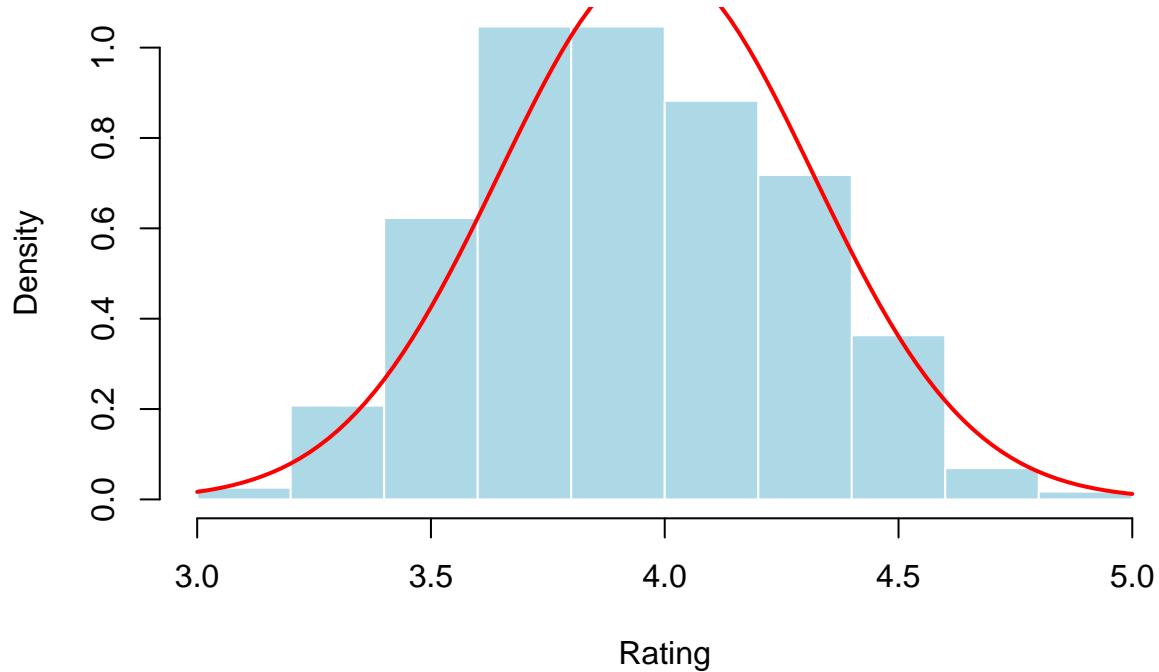
```

## Dessert Ratings



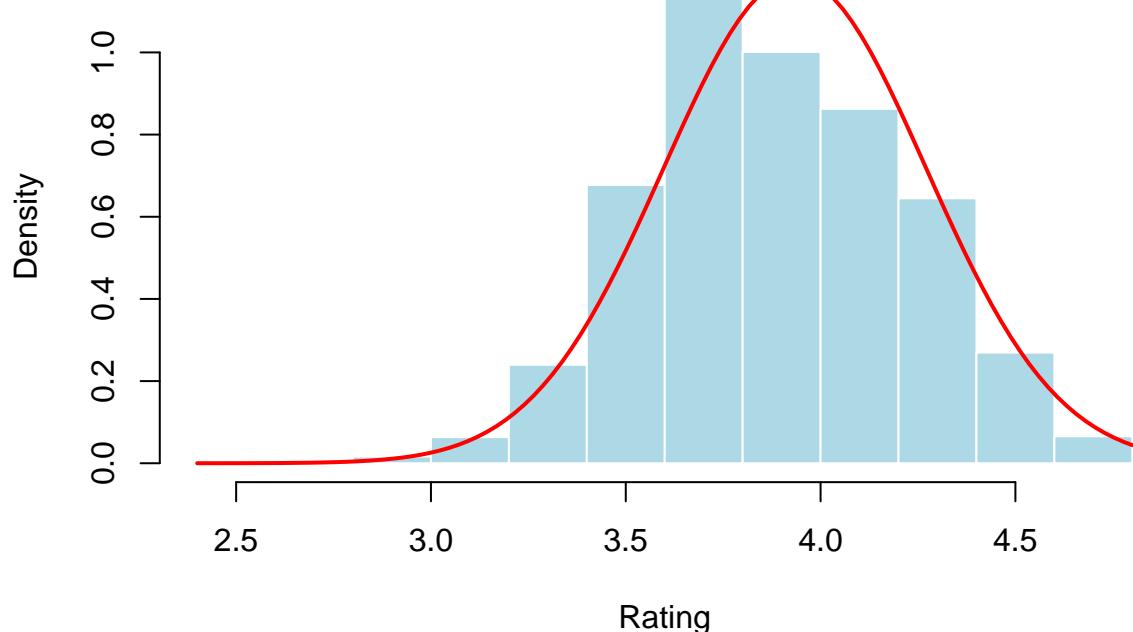
```
plot_histogram_with_normal_dist(dataset_vina$Rating[dataset_vina$Style == 'Fortified'], 'Fortified')
```

## Fortified Ratings



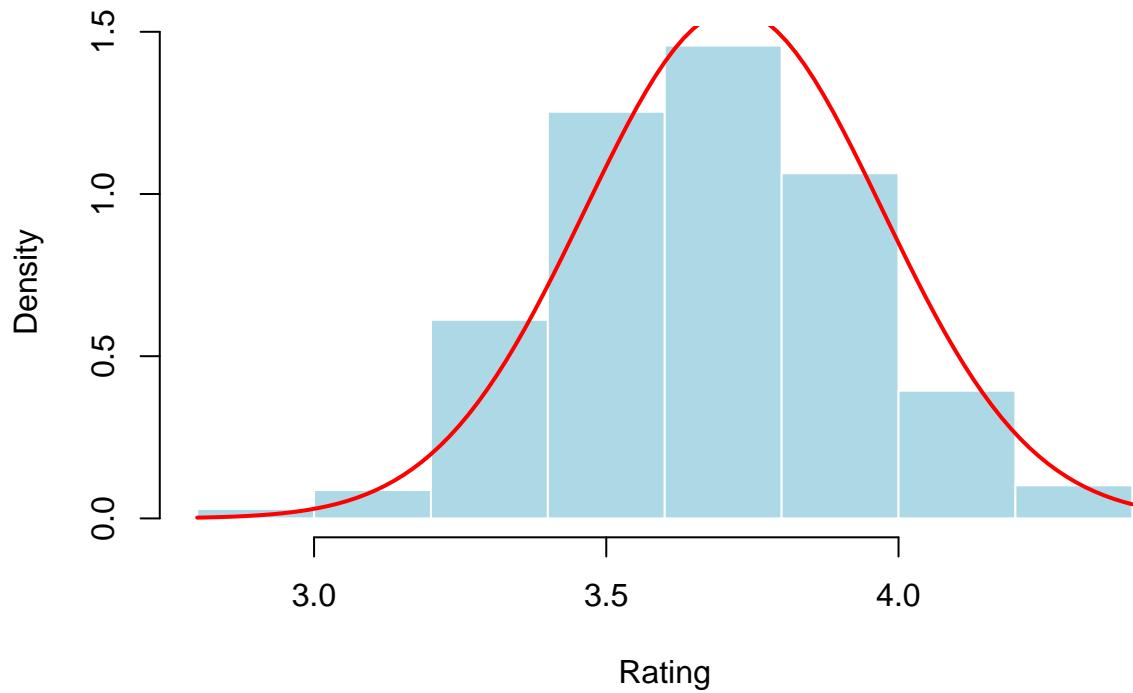
```
plot_histogram_with_normal_dist(dataset_vina$Rating[dataset_vina$Style == 'Red'], 'Red')
```

## Red Ratings



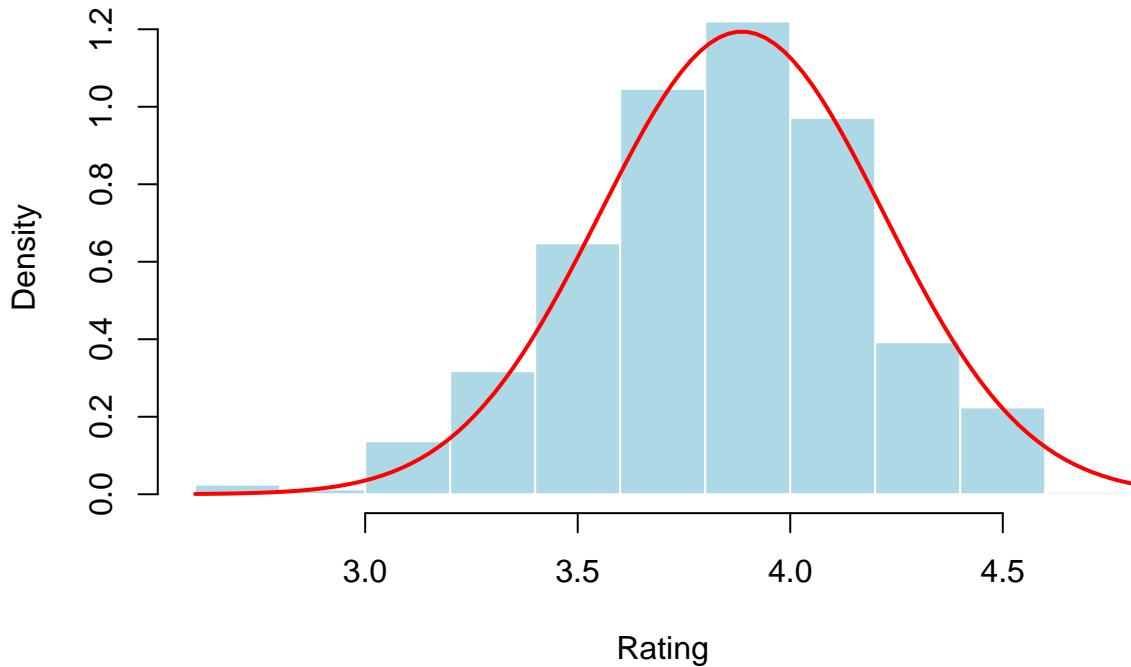
```
plot_histogram_with_normal_dist(dataset_vina$Rating[dataset_vina$Style == 'Rose'], 'Rose')
```

## Rose Ratings



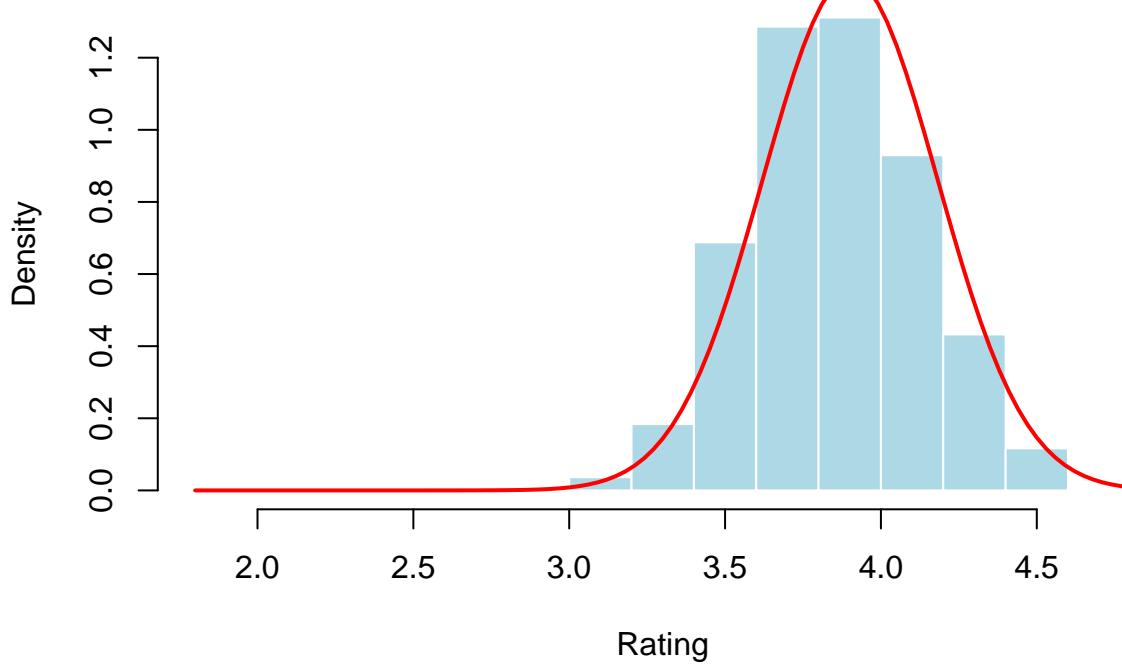
```
plot_histogram_with_normal_dist(dataset_vina$Rating[dataset_vina$Style == 'Sparkling'], 'Sparkling')
```

## Sparkling Ratings



```
plot_histogram_with_normal_dist(dataset_vina$Rating[dataset_vina$Style == 'White'], 'White')
```

## White Ratings



Vidimo da su p-vrijednosti provedenog testa vrlo male (<5%). Razlog zbog kojeg p-vrijednosti mogu biti ovako male je i osjetljivost testa na outliere koje smo vidjeli u početnom boxplot-u. Također, jedan od razloga zbog kojeg testovi odbacuju nul-hipotezu mogao bi biti i zato što imamo diskretizirane (4.0, 4.1, ...) umjesto

kontinuiranih podataka

Provodimo KS test za provjeru normalnosti:

```
ks.test(dataset_vina$Rating[dataset_vina$Style == 'Dessert'], "pnorm",
        mean = mean(dataset_vina$Rating[dataset_vina$Style == 'Dessert'], na.rm = TRUE),
        #na.rm = True odbacuje vrijednosti koje nedostaju
        sd = sd(dataset_vina$Rating[dataset_vina$Style == 'Dessert'], na.rm = TRUE))

## Warning in ks.test.default(dataset_vina$Rating[dataset_vina$Style ==
## "Dessert"], : ties should not be present for the one-sample Kolmogorov-Smirnov
## test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Dessert"]
## D = 0.10375, p-value = 0.09009
## alternative hypothesis: two-sided

ks.test(dataset_vina$Rating[dataset_vina$Style == 'Fortified'], "pnorm",
        mean = mean(dataset_vina$Rating[dataset_vina$Style == 'Fortified'], na.rm = TRUE),
        sd = sd(dataset_vina$Rating[dataset_vina$Style == 'Fortified'], na.rm = TRUE))

## Warning in ks.test.default(dataset_vina$Rating[dataset_vina$Style ==
## "Fortified"], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Fortified"]
## D = 0.086086, p-value = 0.0003806
## alternative hypothesis: two-sided

ks.test(dataset_vina$Rating[dataset_vina$Style == 'Red'], "pnorm",
        mean = mean(dataset_vina$Rating[dataset_vina$Style == 'Red'], na.rm = TRUE),
        sd = sd(dataset_vina$Rating[dataset_vina$Style == 'Red'], na.rm = TRUE))

## Warning in ks.test.default(dataset_vina$Rating[dataset_vina$Style == "Red"], :
## ties should not be present for the one-sample Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Red"]
## D = 0.085468, p-value < 2.2e-16
## alternative hypothesis: two-sided

ks.test(dataset_vina$Rating[dataset_vina$Style == 'Rose'], "pnorm",
        mean = mean(dataset_vina$Rating[dataset_vina$Style == 'Rose'], na.rm = TRUE),
        sd = sd(dataset_vina$Rating[dataset_vina$Style == 'Rose'], na.rm = TRUE))

## Warning in ks.test.default(dataset_vina$Rating[dataset_vina$Style == "Rose"], :
## ties should not be present for the one-sample Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
```

```

## data: dataset_vina$Rating[dataset_vina$Style == "Rose"]
## D = 0.082224, p-value = 0.01936
## alternative hypothesis: two-sided

ks.test(dataset_vina$Rating[dataset_vina$Style == 'Sparkling'], "pnorm",
        mean = mean(dataset_vina$Rating[dataset_vina$Style == 'Sparkling'], na.rm = TRUE),
        sd = sd(dataset_vina$Rating[dataset_vina$Style == 'Sparkling'], na.rm = TRUE))

## Warning in ks.test.default(dataset_vina$Rating[dataset_vina$Style ==
## "Sparkling"], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: dataset_vina$Rating[dataset_vina$Style == "Sparkling"]
## D = 0.089246, p-value = 5.568e-06
## alternative hypothesis: two-sided

ks.test(dataset_vina$Rating[dataset_vina$Style == 'White'], "pnorm",
        mean = mean(dataset_vina$Rating[dataset_vina$Style == 'White'], na.rm = TRUE),
        sd = sd(dataset_vina$Rating[dataset_vina$Style == 'White'], na.rm = TRUE))

## Warning in ks.test.default(dataset_vina$Rating[dataset_vina$Style == "White"],
## : ties should not be present for the one-sample Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: dataset_vina$Rating[dataset_vina$Style == "White"]
## D = 0.081579, p-value < 2.2e-16
## alternative hypothesis: two-sided

```

Provedbom KS-testa dobivamo veće vjerojatnosti, te na razinučajnosti od 1% ne bismo mogli odbaciti hipoteze da Dessert, Fortified, Rose i Sparkling ne prate normalu razdiobu. Ali slično kao i u provedbi Lillieforsove inačice KS testa, na preciznost testa utječu outlieri i diskretizirane vrijednosti.

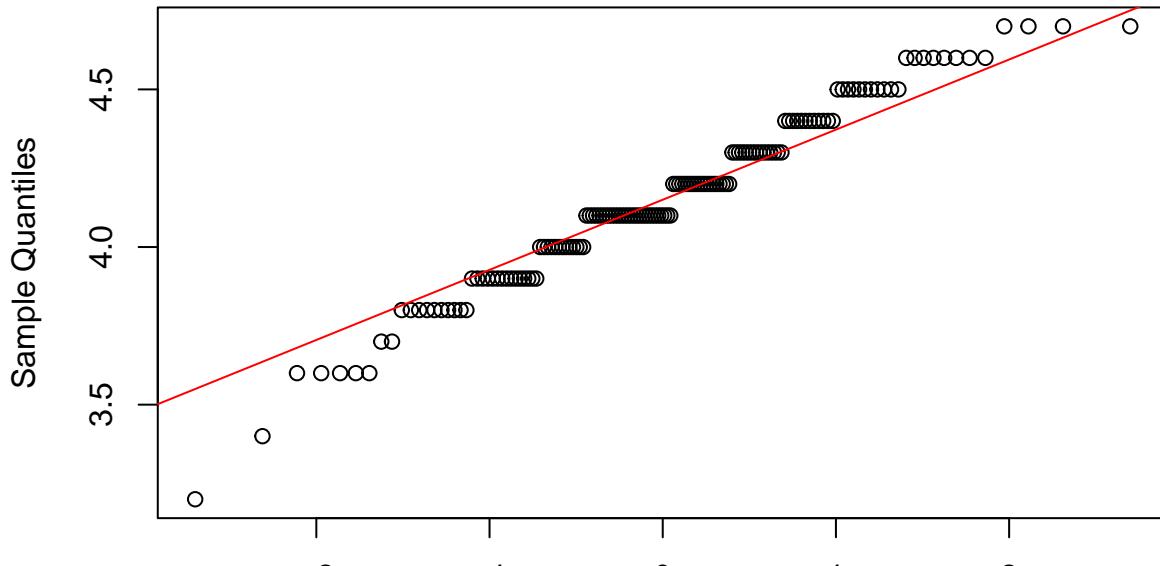
Prikaz podataka podijeljenih u grupe na Q-Q plot-u:

```

qqnorm(dataset_vina$Rating[dataset_vina$Style == "Dessert"], main = "Q-Q Plot: Dessert Ratings")
qqline(dataset_vina$Rating[dataset_vina$Style == "Dessert"], col = "red")

```

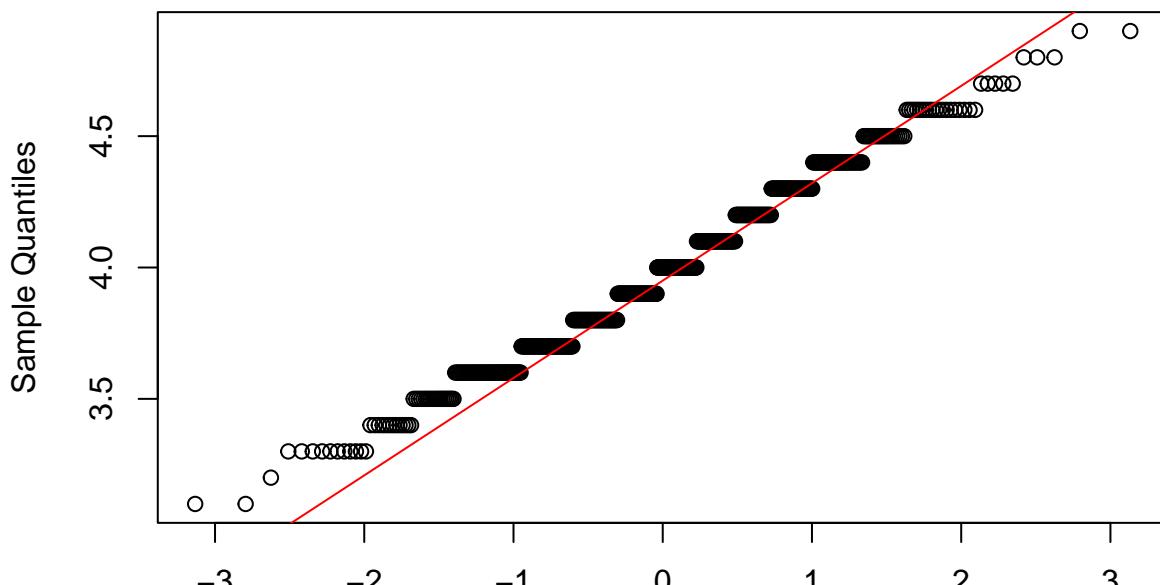
## Q-Q Plot: Dessert Ratings



Theoretical Quantiles

```
qqnorm(dataset_vina$Rating[dataset_vina$Style == "Fortified"], main = "Q-Q Plot: Fortified Ratings")
qqline(dataset_vina$Rating[dataset_vina$Style == "Fortified"], col = "red")
```

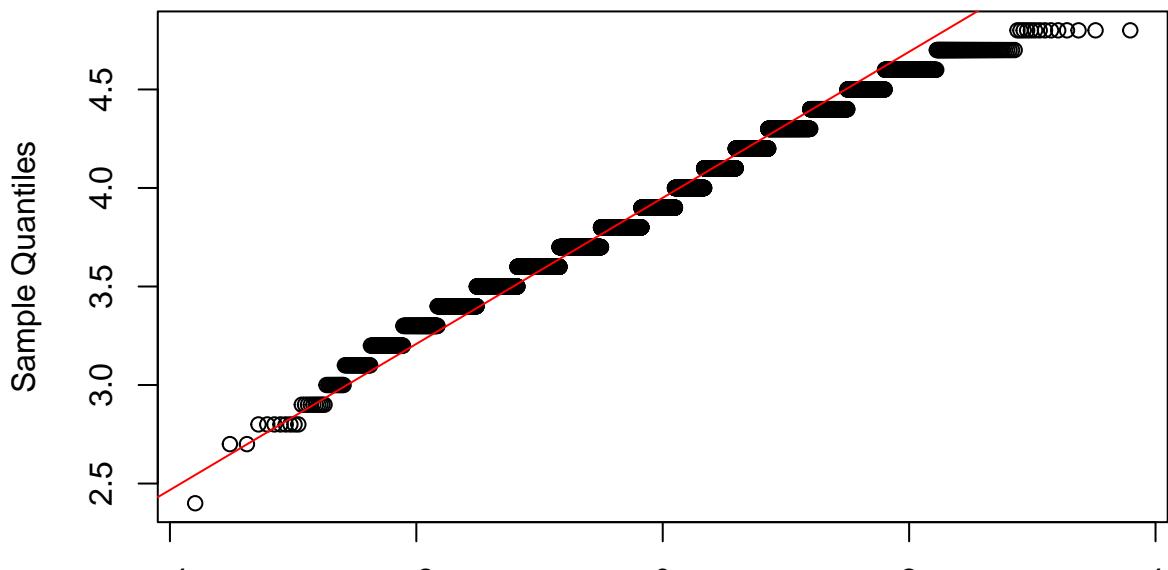
## Q-Q Plot: Fortified Ratings



Theoretical Quantiles

```
qqnorm(dataset_vina$Rating[dataset_vina$Style == "Red"], main = "Q-Q Plot: Red Ratings")
qqline(dataset_vina$Rating[dataset_vina$Style == "Red"], col = "red")
```

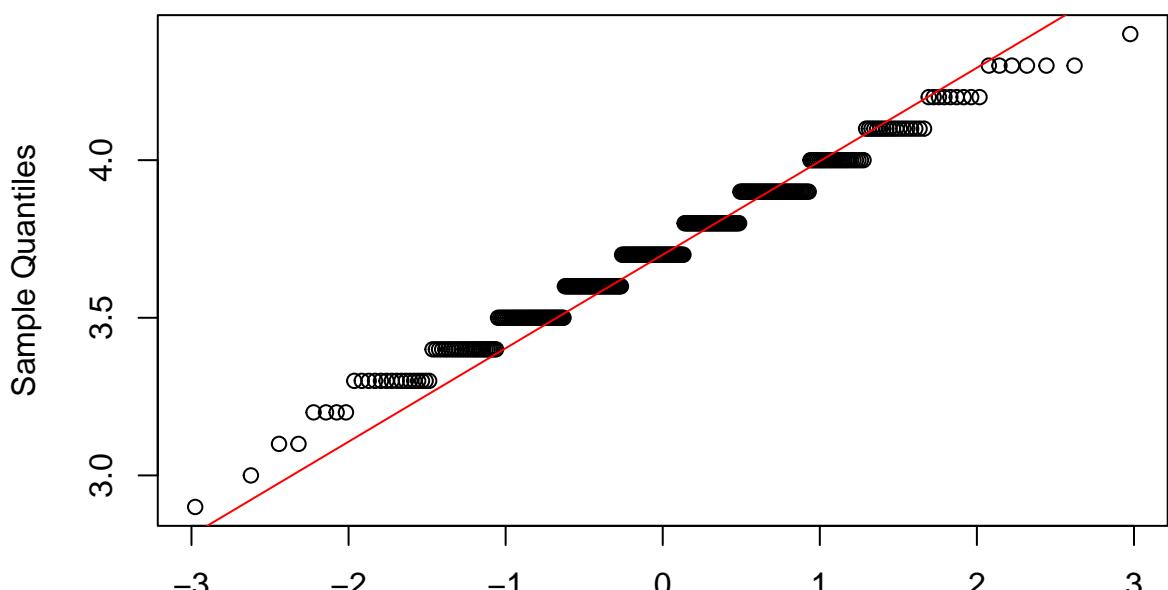
### Q-Q Plot: Red Ratings



Theoretical Quantiles

```
qqnorm(dataset_vina$Rating[dataset_vina$Style == "Rose"], main = "Q-Q Plot: Rose Ratings")
qqline(dataset_vina$Rating[dataset_vina$Style == "Rose"], col = "red")
```

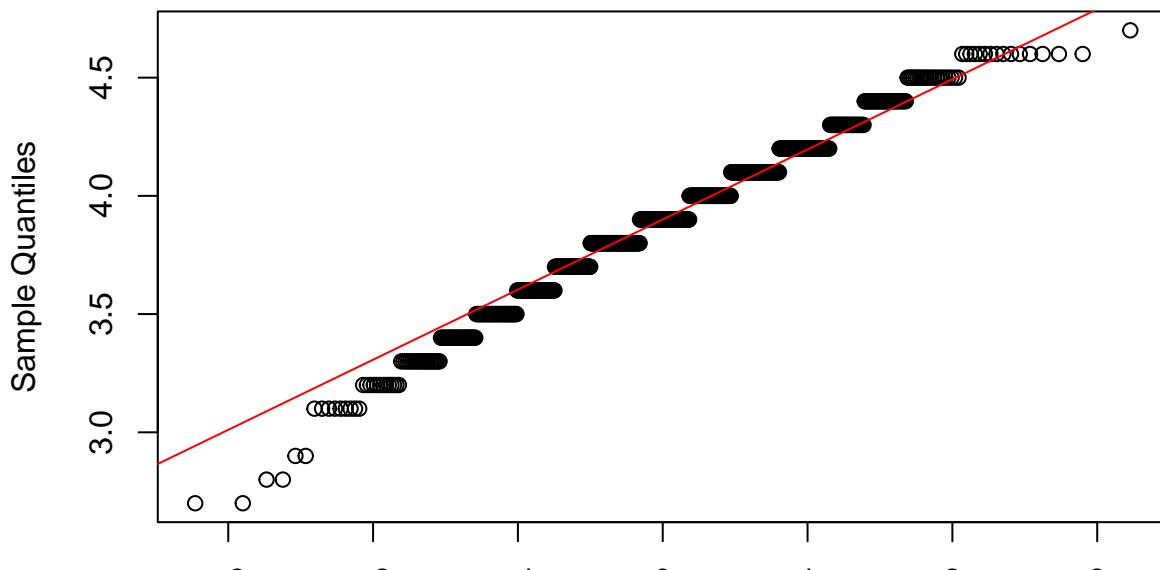
### Q-Q Plot: Rose Ratings



Theoretical Quantiles

```
qqnorm(dataset_vina$Rating[dataset_vina$Style == "Sparkling"], main = "Q-Q Plot: Sparkling Ratings")
qqline(dataset_vina$Rating[dataset_vina$Style == "Sparkling"], col = "red")
```

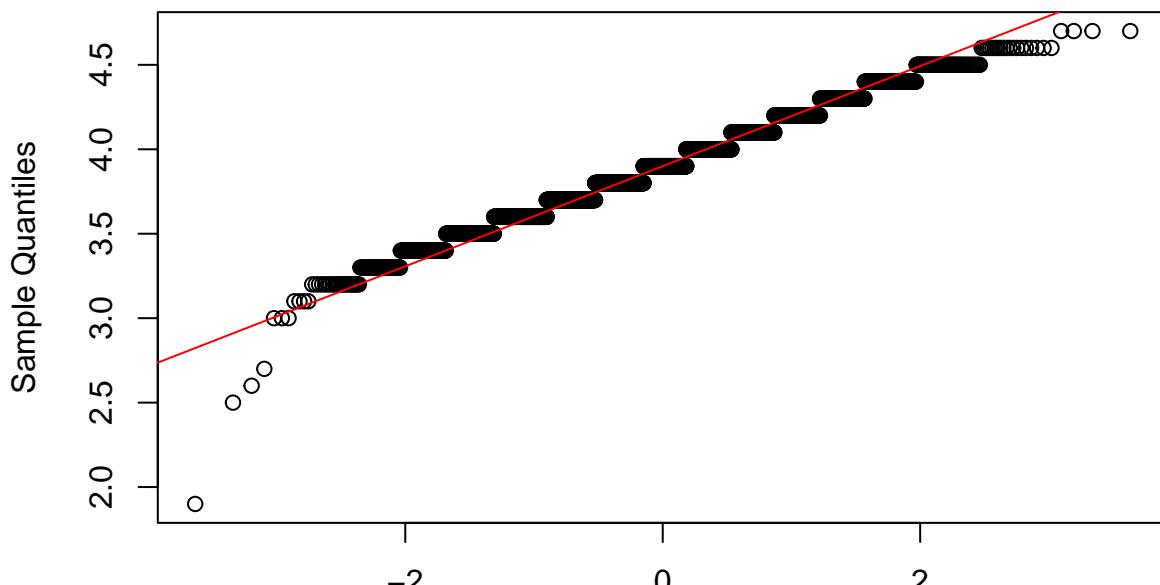
## Q-Q Plot: Sparkling Ratings



Theoretical Quantiles

```
qqnorm(dataset_vina$Rating[dataset_vina$Style == "White"], main = "Q-Q Plot: White Ratings")
qqline(dataset_vina$Rating[dataset_vina$Style == "White"], col = "red")
```

## Q-Q Plot: White Ratings



Theoretical Quantiles

Prikazom podataka na Q-Q Plot-u možemo vidjeti da vrijednosti vrlo blisko prate normalu razdiobu i na temelju njega zaključujemo da grafovi izgledaju u redu za primjenu ANOVE, tako da možemo koristiti parametarsku inačicu.

Kad su veličine grupa podjednake, ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti i homogenosti varijanci.

```
table(dataset_vina$Style)      # broj ocjena za svaku pojedinu vrstu vina

##
##   Dessert Fortified      Red      Rose Sparkling      White
##       144        578     6776       343       803      3561

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
# sortitani prikaz broja ocjena za svaku pojedinu vrstu vina, sortirano silazno
# po broju ocjena
dataset_vina %>%
  count(Style, sort = TRUE)

##      Style    n
## 1      Red 6776
## 2    White 3561
## 3 Sparkling  803
## 4 Fortified  578
## 5     Rose  343
## 6  Dessert  144
```

U našem slučaju veličine grupa znatno odskaču jedne od drugih, npr. vrsta vina "Red" sadrži 6776 ocjena, dok vrsta vina "Dessert" sadrži samo 144 ocjene.

Što se tiče homogenosti varijanci različitih populacija, potrebno je testirati:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \text{barem dvije varijance nisu iste.}$$

Navedenu hipotezu možemo testirati Bartlettovim testom. Bartlettov test u R-u implementiran je naredbom `bartlett.test()`.

```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom
bartlett.test(dataset_vina$Rating ~ dataset_vina$Style)

##
##  Bartlett test of homogeneity of variances
##
##  data: dataset_vina$Rating by dataset_vina$Style
##  Bartlett's K-squared = 192.28, df = 5, p-value < 2.2e-16
var((dataset_vina$Rating[dataset_vina$Style=='Dessert']))
```

```
## [1] 0.07846105
```

```

var((dataset_vina$Rating[dataset_vina$Style=='Fortified']))

## [1] 0.1128803

var((dataset_vina$Rating[dataset_vina$Style=='Red']))

## [1] 0.1139744

var((dataset_vina$Rating[dataset_vina$Style=='Rose']))

## [1] 0.06487699

var((dataset_vina$Rating[dataset_vina$Style=='Sparkling']))

## [1] 0.1116326

var((dataset_vina$Rating[dataset_vina$Style=='White']))

## [1] 0.07890023

```

Vidimo da je p-vrijednost Bartlettovog testa vrlo mala, pa na razini značajnosti 5% možemo odbaciti nultu hipotezu, odnosno zaključujemo da barem dvije varijance nisu iste, a ova tvrdnja je vidljiva i kod ispisa varijanci.

Bartlettov test varijance pokazuje nam da se varijance svih 6 skupina znatno različite. Možemo vidjeti i da su varijance vrsta vina: Fortified, Red i Sparkling relativno slične, te da su varijance vrsta vina: Dessert, Rose i White također međusobno slične. Zbog toga, vrste vina dijelimo u dva podskupa:

```

group_high_variance <- subset(dataset_vina, Style %in% c("Fortified", "Red", "Sparkling"))
group_low_variance <- subset(dataset_vina, Style %in% c("Dessert", "Rose", "White"))

# Bartlettov test homogenosti varijanci na podskupovima
bartlett_high <- bartlett.test(Rating ~ Style, data = group_high_variance)
bartlett_low <- bartlett.test(Rating ~ Style, data = group_low_variance)

print(bartlett_high)

##
##  Bartlett test of homogeneity of variances
##
##  data:  Rating by Style
##  Bartlett's K-squared = 0.16854, df = 2, p-value = 0.9192

print(bartlett_low)

##
##  Bartlett test of homogeneity of variances
##
##  data:  Rating by Style
##  Bartlett's K-squared = 5.6607, df = 2, p-value = 0.05899

```

Vidimo da je p-vrijednost Barlettovog testa kod podgrupe koja sadrži stilove: "Fortified", "Red", "Sparkling", vrlo visoka, zbog čega ne možemo odbaciti nultu hipotezu, te ima smisla provesti ANOVA test na ovoj skupini podataka.

Vidimo da je p-vrijednost Barlettovog testa kod podgrupe koja sadrži stilove: "Dessert", "Rose", "White", vrlo niska, blizu vrijednosti odbacivanja nulte hipoteze, zbog čega ne odbacivanju nulte hipoteze moramo pristupiti vrlo oprezno.

```

group_low_variance_2 <- subset(dataset_vina, Style %in% c("Dessert", "White"))

# Bartlettov test homogenosti varijanci na novom podskupu
bartlett_low <- bartlett.test(Rating ~ Style, data = group_low_variance_2)

print(bartlett_low)

```

```

##
##  Bartlett test of homogeneity of variances
##
## data: Rating by Style
## Bartlett's K-squared = 0.002133, df = 1, p-value = 0.9632

```

Kada odstranimo vrijednosti za "Rose" iz druge podskupine, možemo vidjeti da p-vrijednost Bartlettovog testa naraste na 0.9632 te u ovoj podskupini ne možemo odbaciti nultu hipotezu, te ima smisla provesti ANOVA test.

```

anova_high <- aov(Rating ~ Style, data = group_high_variance)
summary(anova_high)

```

```

##          Df Sum Sq Mean Sq F value    Pr(>F)
## Style      2     3.1   1.5576   13.7 1.14e-06 ***
## Residuals 8154  926.8   0.1137
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

anova_low <- aov(Rating ~ Style, data = group_low_variance)
summary(anova_low)

```

```

##          Df Sum Sq Mean Sq F value    Pr(>F)
## Style      2   20.19  10.093   129.9 <2e-16 ***
## Residuals 4045 314.29   0.078
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

anova_low_2 <- aov(Rating ~ Style, data = group_low_variance_2)
summary(anova_low_2)

```

```

##          Df Sum Sq Mean Sq F value    Pr(>F)
## Style      1    8.52   8.519    108 <2e-16 ***
## Residuals 3703 292.10   0.079
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Vrlo mala p-vrijednost ( $<0.001$ ) kod provođenja ANOVA testa implicira da postoji statistički značajna razlika u ocjenama između vrsta vina Fortified, Red i Sparkling. Vrlo mala p-vrijednost ( $<0.001$ ) kod provođenja ANOVA testa implicira da postoji statistički značajna razlika u ocjenama između vrsta vina Dessert, Rose i White. Vrlo mala p-vrijednost ( $<0.001$ ) kod provođenja ANOVA testa implicira da postoji statistički značajna razlika u ocjenama između vrsta vina Dessert i White

Zaključak: Svi ANOVA testovi pokazali su da postoji statistički značajna razlika među različitim podgrupama vrsta vina, odnosno možemo odgovoriti na početno pitanje i zaključiti da postoji razlika u ocjenama među različitim vrstama vina.

Grafički prikaz (boxplot) sa početka ovog dokumenta sugerira da postoji jasna razlika između grupa, što potvrđuje i ANOVA. Sada kada smo proučili zasebno svaku od podgrupa, kako bismo procijenili model koji pomoću varijable o vrsti vina objašnjava ocjene?

```

# Linearni model
model = lm(Rating ~ Style, data = dataset_vina)
summary(model)

##
## Call:
## lm(formula = Rating ~ Style, data = dataset_vina)
##
## Residuals:
##       Min     1Q Median     3Q    Max 
## -2.00121 -0.23396 -0.01808  0.21370  0.91851 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.14931   0.02658 156.103 < 2e-16 ***
## StyleFortified -0.16782   0.02971 -5.649 1.65e-08 ***
## StyleRed      -0.21535   0.02686 -8.017 1.18e-15 ***
## StyleRose      -0.43123   0.03167 -13.615 < 2e-16 ***
## StyleSparkling -0.26300   0.02887 -9.111 < 2e-16 ***
## StyleWhite     -0.24810   0.02711 -9.151 < 2e-16 ***  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 0.319 on 12199 degrees of freedom
## Multiple R-squared:  0.02147, Adjusted R-squared:  0.02107 
## F-statistic: 53.53 on 5 and 12199 DF, p-value: < 2.2e-16

```

Svi koeficijenti su statistički značajni ( $p < 0.001$ ), što znači da je mala šansa da su razlike između referencirane kategorije vrste vina (Dessert) i ostalih vrsta vina slučajne.

```

# Provjera ANOVA za usporedbu ocjena ovisno o vrsti vina
anova_result <- aov(Rating ~ Style, data = dataset_vina)
summary(anova_result)

```

```

##              Df Sum Sq Mean Sq F value Pr(>F)    
## Style          5  27.2  5.446   53.53 <2e-16 ***
## Residuals  12199 1241.1   0.102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

```

P-vrijednost ( $< 2.2e-16$ ) pokazuje da varijabla vrste vina (Style) ima statistički značajan utjecaj na ocjene (Rating).

Kruskal-Wallisov test je neparametarska alternativa (jednofaktorskoj) analizi varijance - "neparametarska ANOVA". Budući da je on robusniji na odstupanja pretpostavki ANOVA-e, možemo provesti i njega kako bismo potvrdili svoj zaključak o razlici ocjene ovisno o vrsti vina. Njegove hipoteze su:  $H_0$  : medijani distribucija svih uzoraka su jednaki  $H_1$  : barem dva medijana nisu jednaka

```

# Provjera Kruskal-Wallis testa za usporedbu ocjena ovisno o vrsti vina
kruskal_test_result <- kruskal.test(Rating ~ Style, data = dataset_vina)
print(kruskal_test_result)

```

```

## 
## Kruskal-Wallis rank sum test
## 
## data: Rating by Style
## Kruskal-Wallis chi-squared = 254.94, df = 5, p-value < 2.2e-16

```

Budući da je p-vrijednost vrlo mala, možemo odbaciti nullu hipotezu. To znači da postoji značajan dokaz da se ocjene razlikuju u barem jednom paru vrsta vina. Kruskal-Wallisov test nam ne govori koje se grupe razlikuju.

hi-kvadrat test za nezavisnost je test kojime možemo izračunati jesu li dvije kategoriske vrijednosti nezavisne ili povezane. To se odrađuje uspoređivanjem očekivanja i očitanog. Pošto se u zadatku od nas traži da testiramo povezanost vrste vina i prirodnosti vina, moramo koristiti hi-kvadrat test da odredimo povezanost.

```
contingency_table <- table(dataset_vina$Style, dataset_vina$Natural)
```

```
#podatci hi-kvadrat testa prikazuju se tablicom
print(contingency_table)
```

```
##
##          False True
##  Dessert     141   3
##  Fortified    573   5
##  Red         6542  234
##  Rose        333   10
##  Sparkling   759   44
##  White       3372  189
```

```
chi_sq_res <- chisq.test(contingency_table)

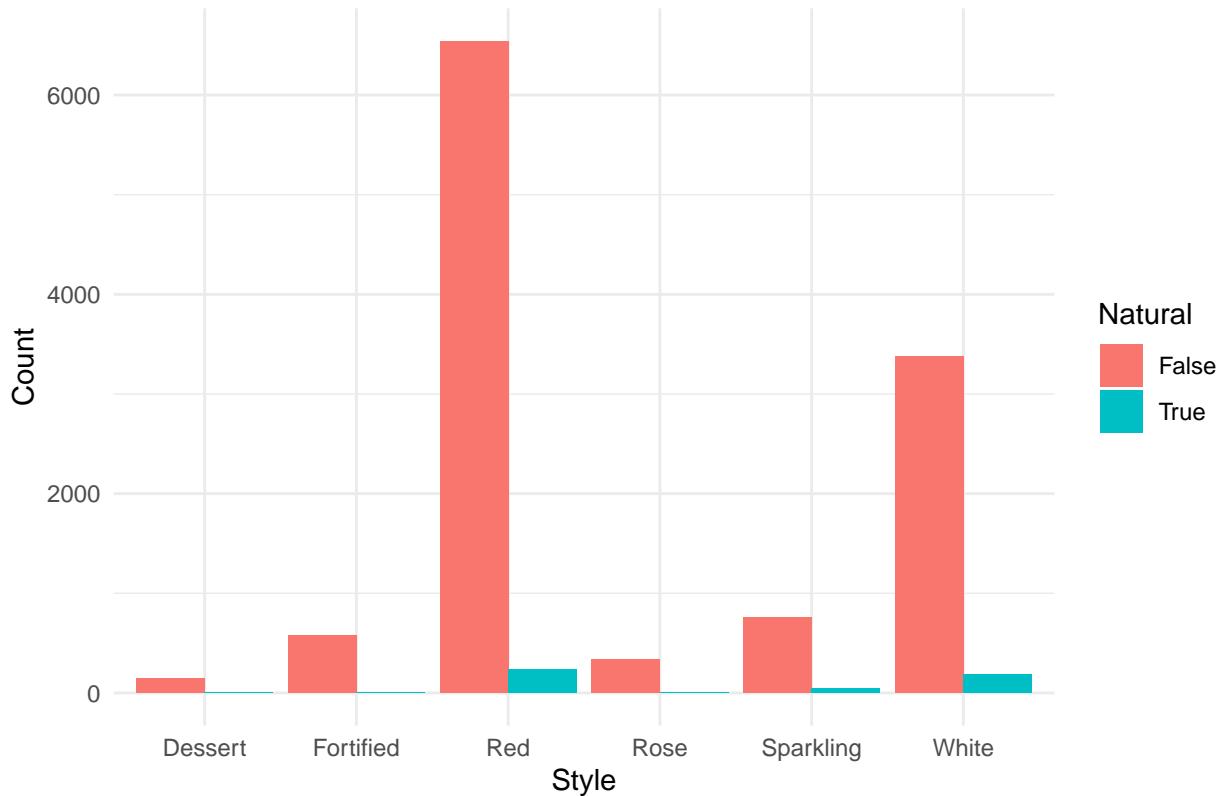
print(chi_sq_res)
```

```
##
##  Pearson's Chi-squared test
##
##  data: contingency_table
##  X-squared = 43.174, df = 5, p-value = 3.407e-08
```

```
contingency_df <- as.data.frame(contingency_table)
```

```
ggplot(contingency_df, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Style", y = "Count", fill = "Natural") +
  theme_minimal() +
  ggtitle("Contingency Table Visualization")
```

## Contingency Table Visualization



rezultat:  $\chi^2 = 43.174$ ,  $df = 5$ ,  $p\text{-value} = 3.407e-08$  pošto je  $p < 0.05$  možemo odbaciti hipotezu da su vrste vina povezane s da li su prirodne i prihvaćamo hipotezu da je to nepovezano kao točnu

PITANJE: Jesu li vina iz Francuske popularnija (imaju više recenzija) od onih iz Italije?

Hipoteze:  $H_0: \bar{y}_{FR} = \bar{y}_{IT}$   $H_1: \bar{y}_{FR} > \bar{y}_{IT}$

Prikaz podataka o broju recenzija vina

```
francuskaVina = dataset_vina[dataset_vina$Country == "France",]
talijanskaVina = dataset_vina[dataset_vina$Country == "Italy",]
cat("Prosječan broj recenzija Francuskih vina:", mean(francuskaVina$Reviews), "\n")
```

```
## Prosječan broj recenzija Francuskih vina: 906.7928
```

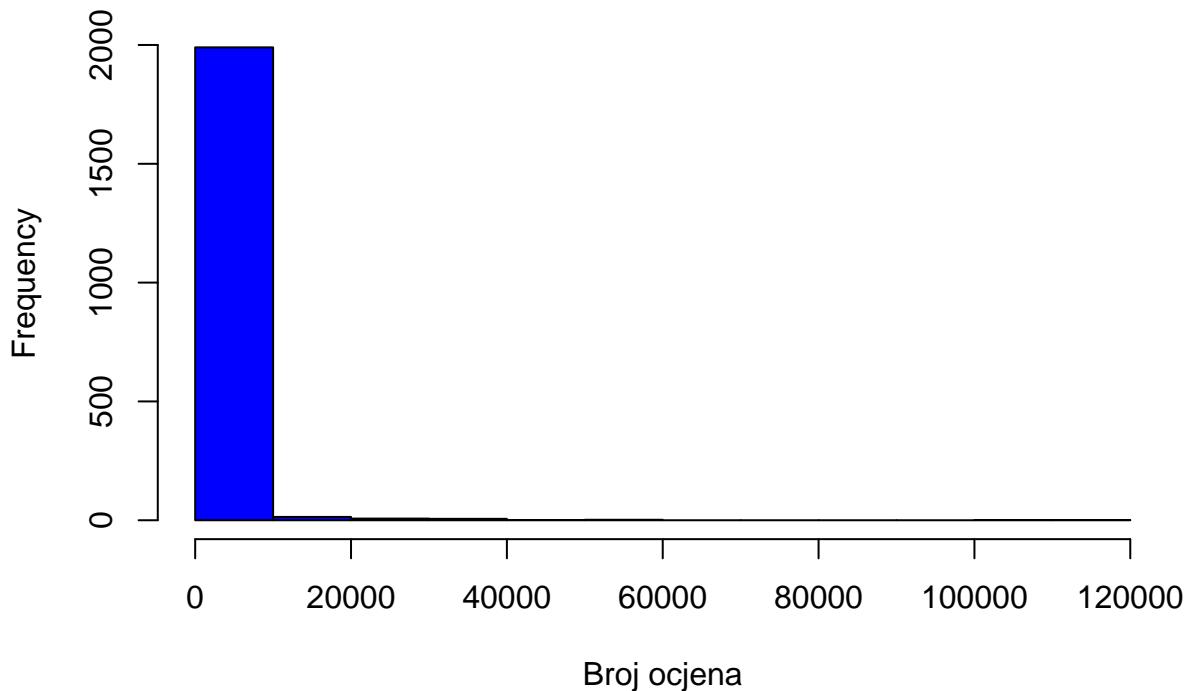
```
cat("Prosječan broj recenzija Talijanskih vina:", mean(talijanskaVina$Reviews), "\n")
```

```
## Prosječan broj recenzija Talijanskih vina: 373.7037
```

Prikaz podataka i provjera normalnosti:

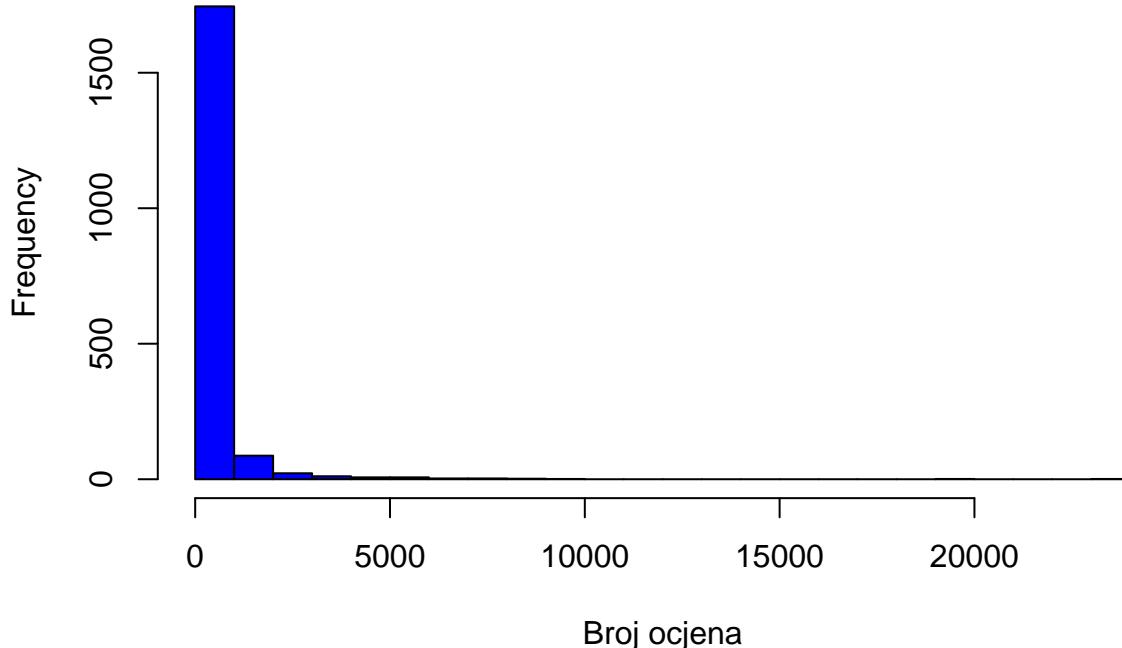
```
hist(francuskaVina$Reviews,
     main = "Broj ocjena vina iz Francuske",
     xlab = "Broj ocjena",
     col = "blue",
     breaks = 10)
```

## Broj ocjena vina iz Francuske



```
hist(talijanskaVina$Reviews,  
      main = "Broj ocjena vina iz Italije",  
      xlab = "Broj ocjena",  
      col = "blue",  
      breaks = 20)
```

## Broj ocjena vina iz Italije



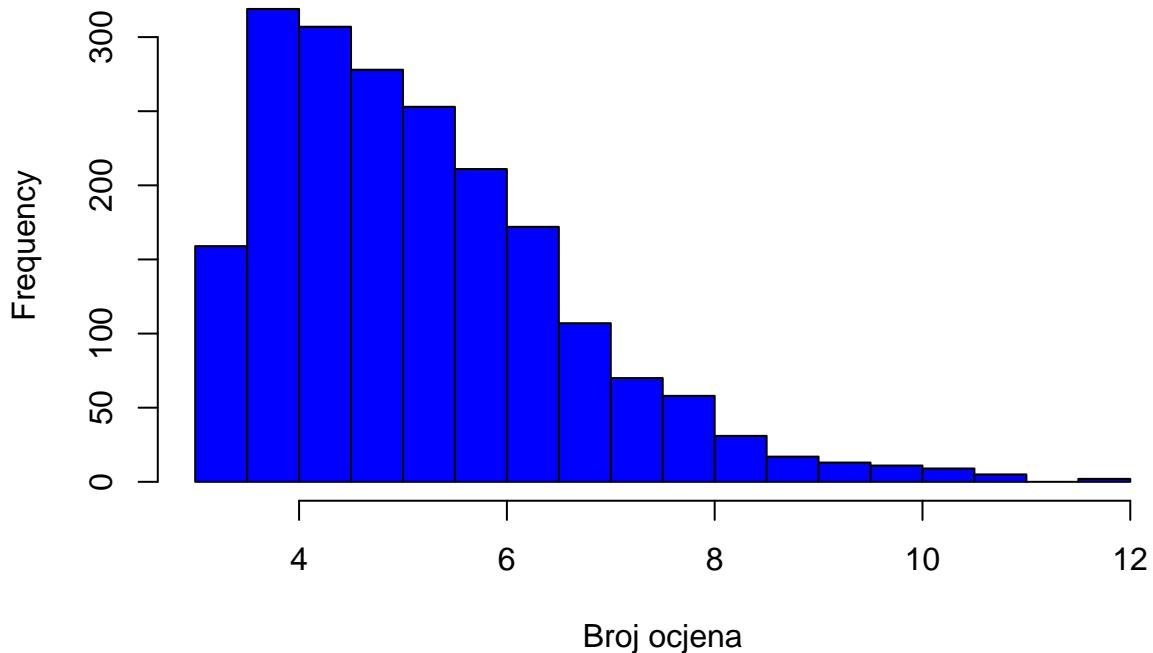
Iz histograma je vidljivo kako podatci nisu normalno distribuirani, zbog čega se primjenjuje logaritamska funkcija u pokušaju normalizacije podataka.

Prikaz podataka i provjera normalnosti logaritma broja recenzija

```
francuskaVina$LogReviews = log(franuskaVina$Reviews + 1)  
talijanskaVina$LogReviews = log(talijanskaVina$Reviews + 1)
```

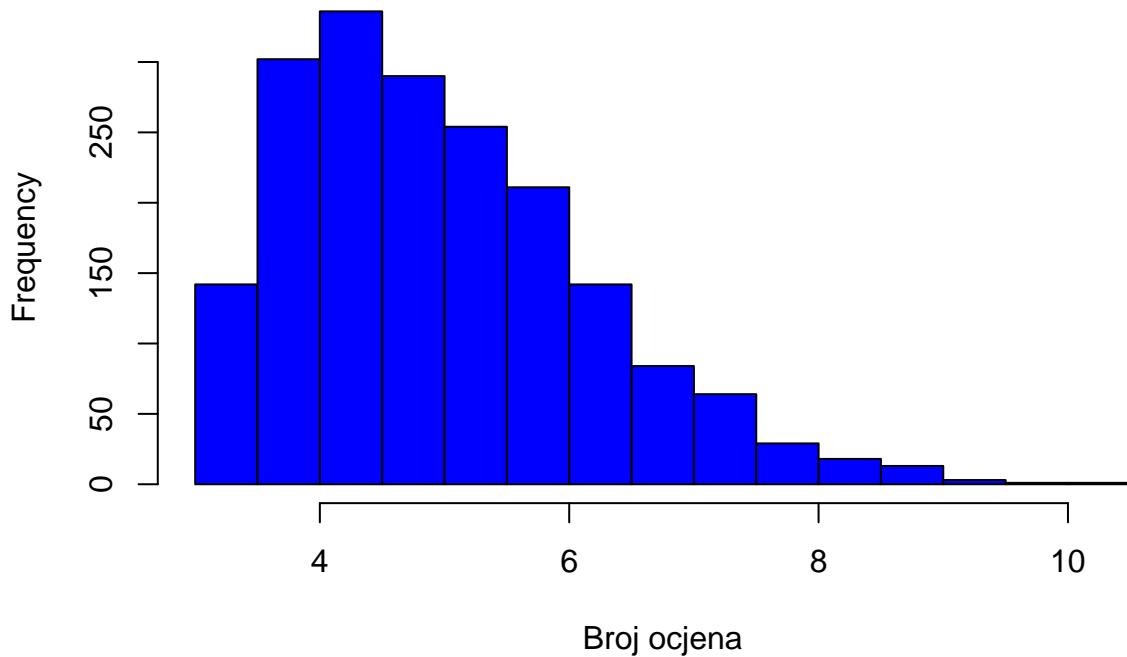
```
hist(franuskaVina$LogReviews,  
     main = "Broj ocjena vina iz Francuske",  
     xlab = "Broj ocjena",  
     col = "blue",  
     breaks = 20)
```

**Broj ocjena vina iz Francuske**



```
hist(talijanskaVina$LogReviews,  
     main = "Broj ocjena vina iz Italije",  
     xlab = "Broj ocjena",  
     col = "blue",  
     breaks = 20)
```

## Broj ocjena vina iz Italije



Provjera dobivenih rezultata Kolmogorov-Smirnovov testom Hipoteze: H0: podatci su iz normalne razdiobe  
H1: podatci nisu iz normalne razdiobe

```
france_data = francuskaVina$LogReviews
italy_data = talijanskaVina$LogReviews

ks.test(france_data, "pnorm", mean = mean(france_data), sd = sd(france_data))

## Warning in ks.test.default(france_data, "pnorm", mean = mean(france_data), :
## ties should not be present for the one-sample Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: france_data
## D = 0.091879, p-value = 2.986e-15
## alternative hypothesis: two-sided

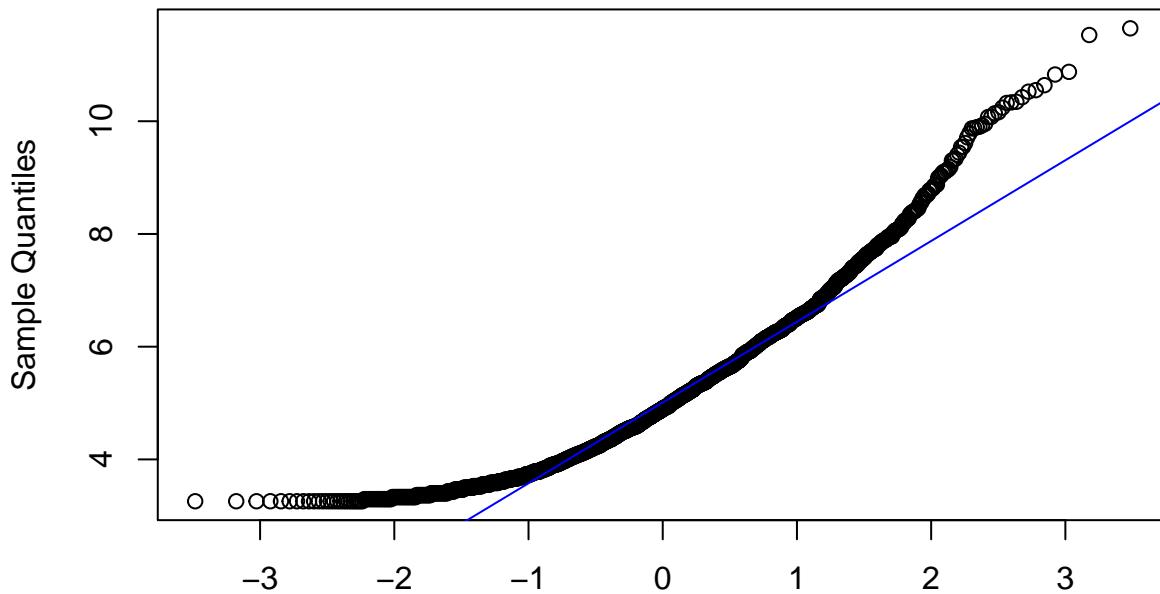
ks.test(italy_data, "pnorm", mean = mean(italy_data), sd = sd(italy_data))

## Warning in ks.test.default(italy_data, "pnorm", mean = mean(italy_data), :
## ties should not be present for the one-sample Kolmogorov-Smirnov test

##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: italy_data
## D = 0.074961, p-value = 1.193e-09
## alternative hypothesis: two-sided

qqnorm(france_data, main = "Francuska Q-Q grafikon")
qqline(france_data, col = "blue")
```

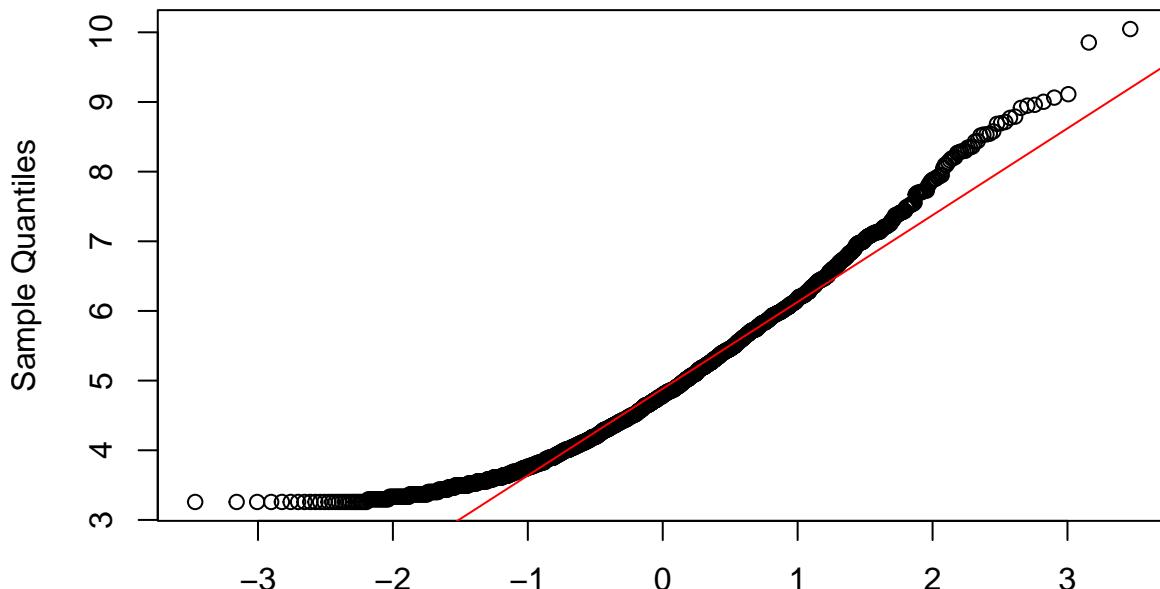
## Francuska Q-Q grafikon



Theoretical Quantiles

```
qqnorm(italy_data, main = "Italija Q-Q grafikon")
qqline(italy_data, col = "red")
```

## Italija Q-Q grafikon



Theoretical Quantiles

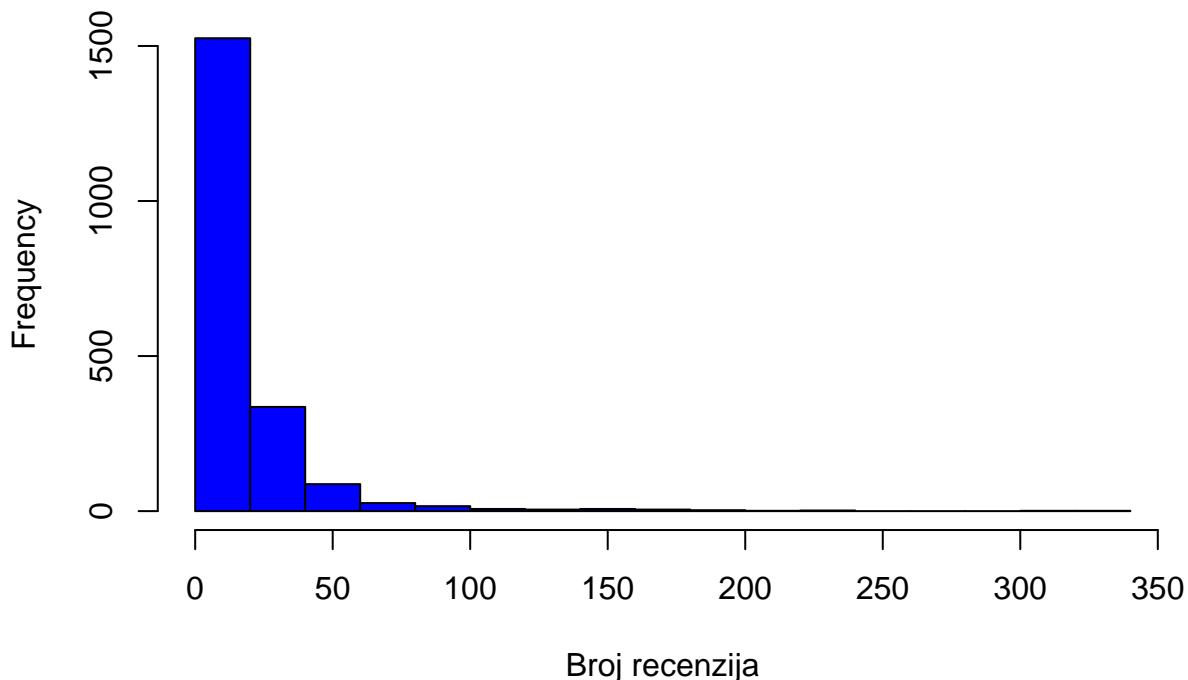
Iz KS testa vidimo da je p-vrijednost vrlo mala, što znači da možemo odbaciti hipotezu H<sub>0</sub>, tj. logaritam broja recenzija ne zadovoljava normalnu distribuciju.

Druga transformacija podataka dobivena je iz korijena broja recenzija:

```
francuskaVina$SqrtReviews = sqrt(franckuskaVina$Reviews)
talijanskaVina$SqrtReviews = sqrt(talijanskaVina$Reviews)

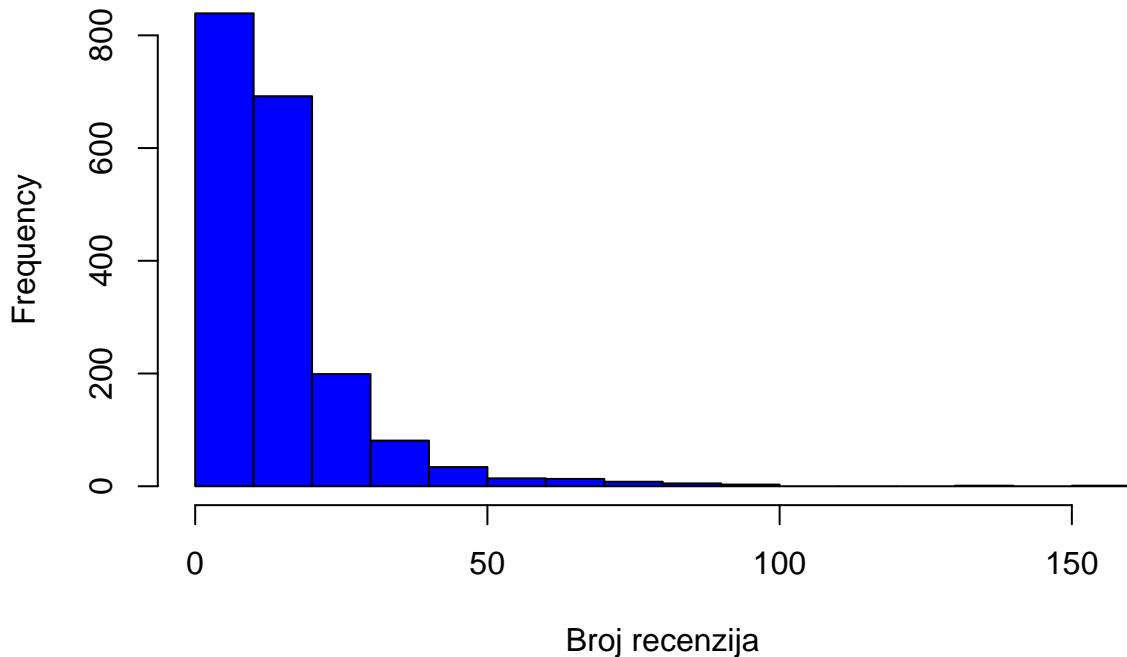
hist(franckuskaVina$SqrtReviews,
      main = "Broj recenzija Francuskog vina",
      xlab = "Broj recenzija",
      col = "blue",
      breaks = 20)
```

**Broj recenzija Francuskog vina**



```
hist(talijanskaVina$SqrtReviews,
      main = "Broj recenzija Talijanskog vina",
      xlab = "Broj recenzija",
      col = "blue",
      breaks = 20)
```

## Broj recenzija Talijanskog vina



Korijen broja recenzija također ne dolazi iz normalne distribucije.

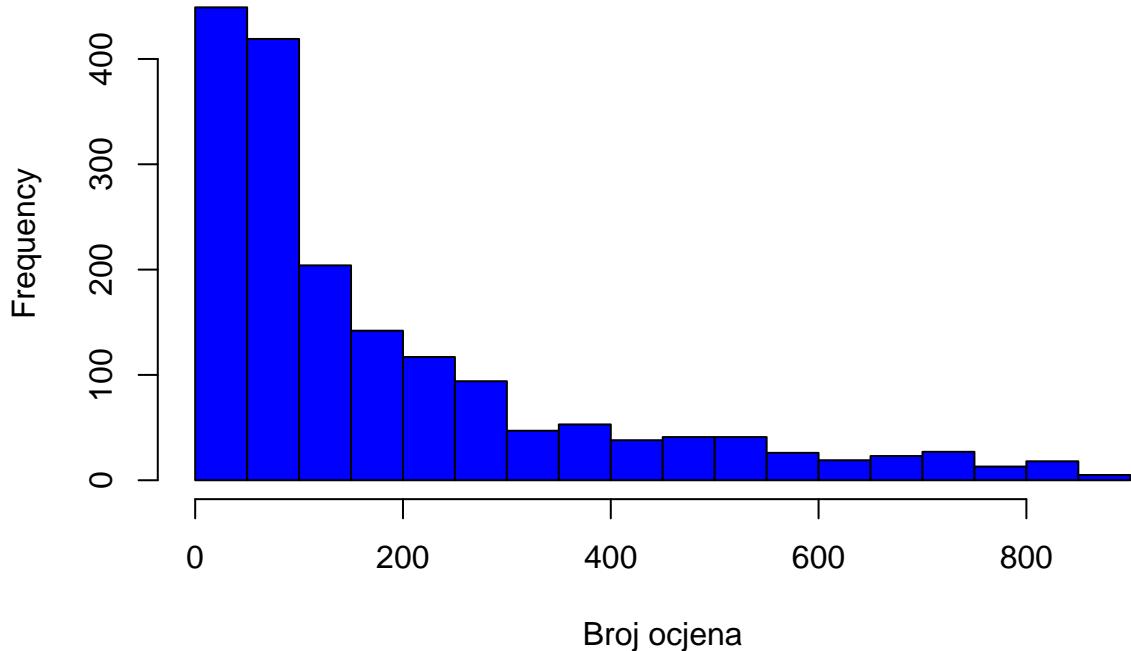
Izbacivanje outliera:

```
#Izbacivanje outliera za francuska vina
Q1_France = quantile(francuskaVina$Reviews, 0.25)
Q3_France = quantile(francuskaVina$Reviews, 0.75)
IQR_France = Q3_France - Q1_France
donjaGranicaFrance = Q1_France - 1.5 * IQR_France
gornjaGranicaFrance = Q3_France + 1.5 * IQR_France
francuskaVinaNoOutliers = francuskaVina[francuskaVina$Reviews >= donjaGranicaFrance
                                         & francuskaVina$Reviews <= gornjaGranicaFrance, ]

#Izbacivanje outliera za talijanska vina
Q1_Italy = quantile(talijanskaVina$Reviews, 0.25)
Q3_Italy = quantile(talijanskaVina$Reviews, 0.75)
IRQ_Italy = Q3_Italy - Q1_Italy
donjaGranicaItaly = Q1_Italy - 1.5*IRQ_Italy
gornjaGranicaItaly = Q3_Italy + 1.5*IRQ_Italy
talijanskaVinaNoOutliers = talijanskaVina[talijanskaVina$Reviews >= donjaGranicaItaly
                                         & talijanskaVina$Reviews <= gornjaGranicaItaly, ]

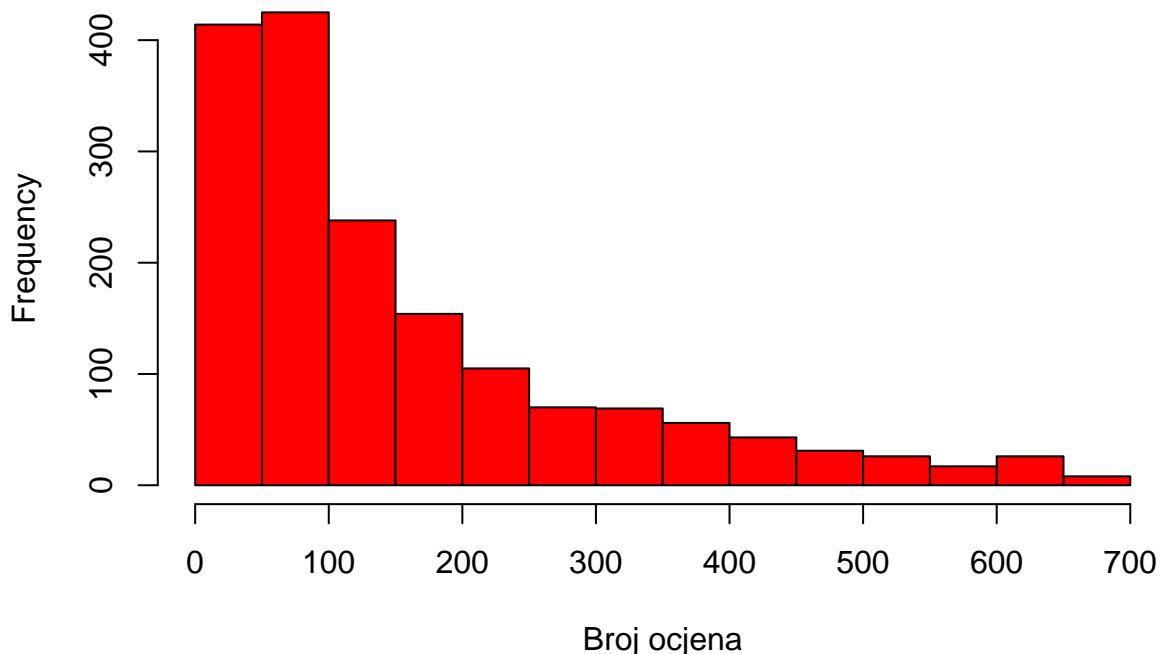
# Histogram
hist(francuskaVinaNoOutliers$Reviews,
      main = "Broj ocjena vina iz Francuske bez outliera",
      xlab = "Broj ocjena",
      col = "blue",
      breaks = 20)
```

## Broj ocjena vina iz Francuske bez outliera



```
hist(talijanskaVinaNoOutliers$Reviews,  
      main = "Broj ocjena vina iz Italije bez outliera",  
      xlab = "Broj ocjena",  
      col = "red",  
      breaks = 20)
```

## Broj ocjena vina iz Italije bez outliera



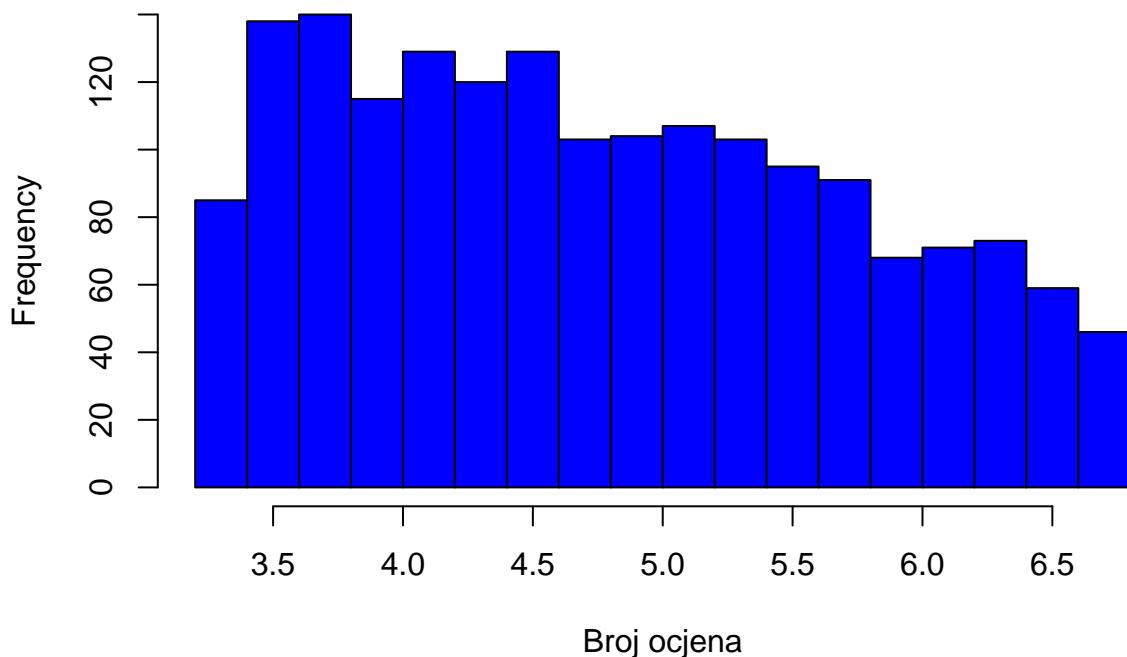
Izbacivanjem outliera bliže smo normalnoj razdiobi nego prije, ali podatci i dalje ne slijede normalnu razdiobu. Za pokušaj postizanja normalne razdiobe primjenjuje se logaritamska i korijen funkcija na podatke bez outliera:

```
francuskaVinaNoOutliers$LogReviews = log(franckuskaVinaNoOutliers$Reviews+1)
talijanskaVinaNoOutliers$LogReviews = log(talijanskaVinaNoOutliers$Reviews+1)

francuskaVinaNoOutliers$SqrtReviews = sqrt(franckuskaVinaNoOutliers$Reviews+1)
talijanskaVinaNoOutliers$SqrtReviews = sqrt(talijanskaVinaNoOutliers$Reviews+1)

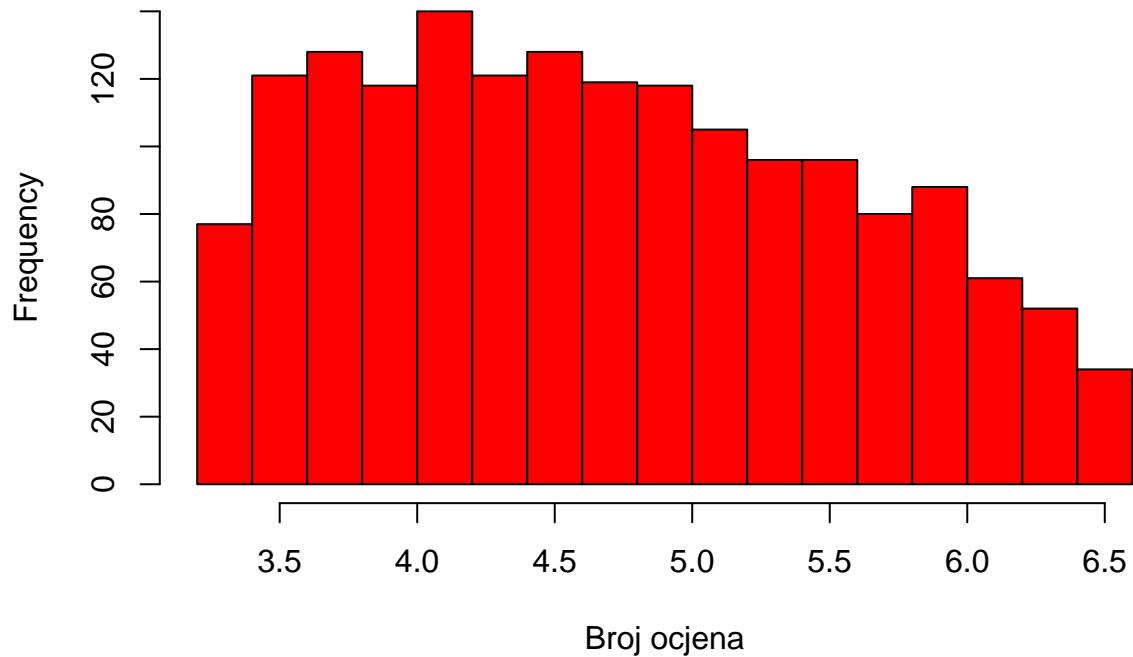
#Prikaz logaritamske promjene
hist(franckuskaVinaNoOutliers$LogReviews,
      main = "Broj ocjena vina iz Francuske bez outliera s logaritamskom izmjenom",
      xlab = "Broj ocjena",
      col = "blue",
      breaks = 20)
```

## Broj ocjena vina iz Francuske bez outliera s logaritamskom izmjenom



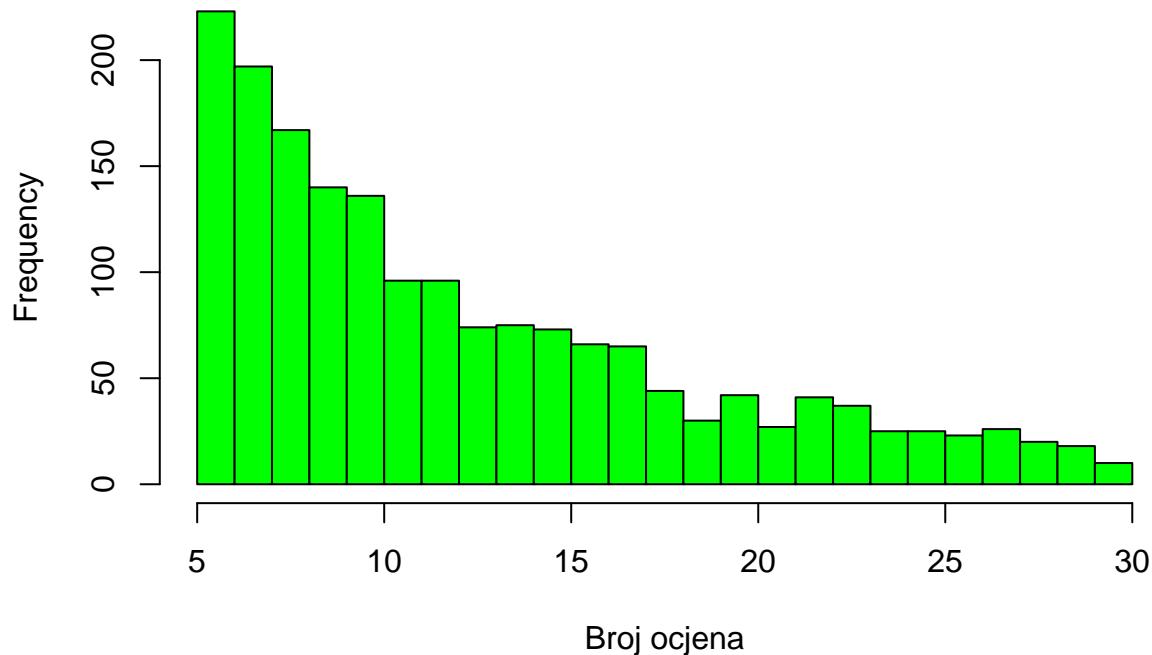
```
hist(talijanskaVinaNoOutliers$LogReviews,
      main = "Broj ocjena vina iz Italije bez outliera s logaritamskom izmjenom",
      xlab = "Broj ocjena",
      col = "red",
      breaks = 20)
```

## Broj ocjena vina iz Italije bez outliera s logaritamskom izmjenom



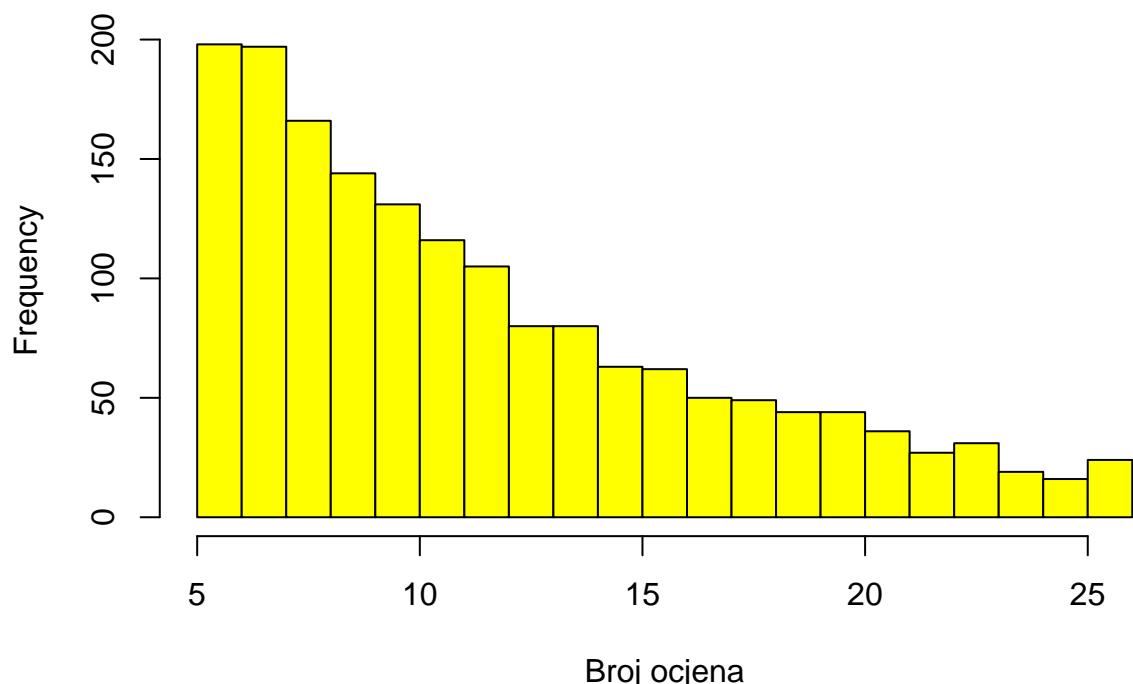
```
#Prikaz promjene korjenovanjem
hist(franckuskaVinaNoOutliers$SqrtReviews,
     main = "Broj ocjena vina iz Francuske bez outliera s korijenskom izmjenom",
     xlab = "Broj ocjena",
     col = "green",
     breaks = 20)
```

## Broj ocjena vina iz Francuske bez outliera s korijenskom izmjenom



```
hist(talijanskaVinaNoOutliers$SqrtReviews,
  main = "Broj ocjena vina iz Italije bez outliera s korijenskom izmjenom",
  xlab = "Broj ocjena",
  col = "yellow",
  breaks = 20)
```

## Broj ocjena vina iz Italije bez outliera s korijenskom izmjenom



Robusnost T-testa:

```
cat("Broj recenzija fr Vina:", length(franckuskaVina$Reviews), "\n")  
## Broj recenzija fr Vina: 2022  
cat("Broj recenzija ita Vina:", length(talijanskaVina$Reviews), "\n")  
## Broj recenzija ita Vina: 1890
```

Zbog velikog broja recenzija, n~1850, možemo koristiti t-test unatoč tome da podatci ne dolaze iz normalne razdiobe, također koristit će se logaritam broja recenzija jer je on bliže normalnoj razdiobi od broja recenzija

Također treba odrediti da li su varijance podataka jednake, to radimo F-testom o varijancama

```
var(franckuskaVina$LogReviews)  
## [1] 2.044662  
var(talijanskaVina$LogReviews)  
## [1] 1.429183  
var.test(franckuskaVina$LogReviews, talijanskaVina$LogReviews)
```

```
##  
## F test to compare two variances  
##  
## data: franckuskaVina$LogReviews and talijanskaVina$LogReviews  
## F = 1.4307, num df = 2021, denom df = 1889, p-value = 3.553e-15  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.309088 1.563293  
## sample estimates:  
## ratio of variances  
## 1.430651
```

Iz testa vidimo kako varijance nisu jednake tako da ne možemo koristiti t-test s nepoznatim ali jednakim varijancama, već moramo koristiti t-test s nepoznatim i različitim varijancama

```
t.test(franckuskaVina$LogReviews, talijanskaVina$LogReviews, alt="greater", var.equal = FALSE)
```

```
##  
## Welch Two Sample t-test  
##  
## data: franckuskaVina$LogReviews and talijanskaVina$LogReviews  
## t = 4.2694, df = 3862.8, p-value = 1.004e-05  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 0.1103185      Inf  
## sample estimates:  
## mean of x mean of y  
## 5.158850 4.979365
```

Iz izvedenog t-testa možemo vidjeti da je p-vrijednost vrlo mala, zbog čega možemo odbaciti našu hipotezu H0, tj. iz testa je vidljivo da su vina iz Francuske popularnija od vina iz Italije (imaju prosječno više recenzija)

Isti test možemo provesti i na podatcima bez outliera te koristimo podatke dobivene logaritamskom promjenom:

```
var(franckuskaVinaNoOutliers$LogReviews)
```

```
## [1] 0.9345072
```

```

var(talijanskaVinaNoOutliers$LogReviews)

## [1] 0.7654779

var.test(francuskaVinaNoOutliers$LogReviews, talijanskaVinaNoOutliers$LogReviews)

##
## F test to compare two variances
##
## data: francuskaVinaNoOutliers$LogReviews and talijanskaVinaNoOutliers$LogReviews
## F = 1.2208, num df = 1775, denom df = 1681, p-value = 3.516e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.110814 1.341546
## sample estimates:
## ratio of variances
## 1.220815

```

Iz izведенog testa vidimo kako su varijance različite i nepoznate te opet koristimo T-test s nepoznatim i različitim varijancama

```

t.test(francuskaVinaNoOutliers$LogReviews, talijanskaVinaNoOutliers$LogReviews,
       alt="greater", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: francuskaVinaNoOutliers$LogReviews and talijanskaVinaNoOutliers$LogReviews
## t = 2.6569, df = 3448.9, p-value = 0.003962
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 0.03168782      Inf
## sample estimates:
## mean of x mean of y
## 4.769199 4.685971

```

Iz provedenog testa vidimo kako je p-vrijednost 0.0004, iako je p-vrijednost veća kada smo izbacili outliere, i dalje je vrlo mala te ovdje također odbacujemo hipotezu H0 tj. prihvaćamo hipotezu H1 da su vina iz francuske popularnija.

Sljedeće što nas zanima vezano za ovaj skup podataka je: Može li se na temelju dostupnih podataka predvidjeti cijena vina?

Kako bih probali odgovoriti na ovo pitanje koristit ćemo linearnu regresiju.

Prvo izbacujemo nevaljanje vrijednosti iz seta podataka jer operacije koje su potrebne za izvođenje linearne regresije nije moguće izvoditi na njima.

```

dataset_vina <- dataset_vina[dataset_vina$Year != "N.V." & !is.na(dataset_vina$Price)
                           & dataset_vina$Year != '', ]

```

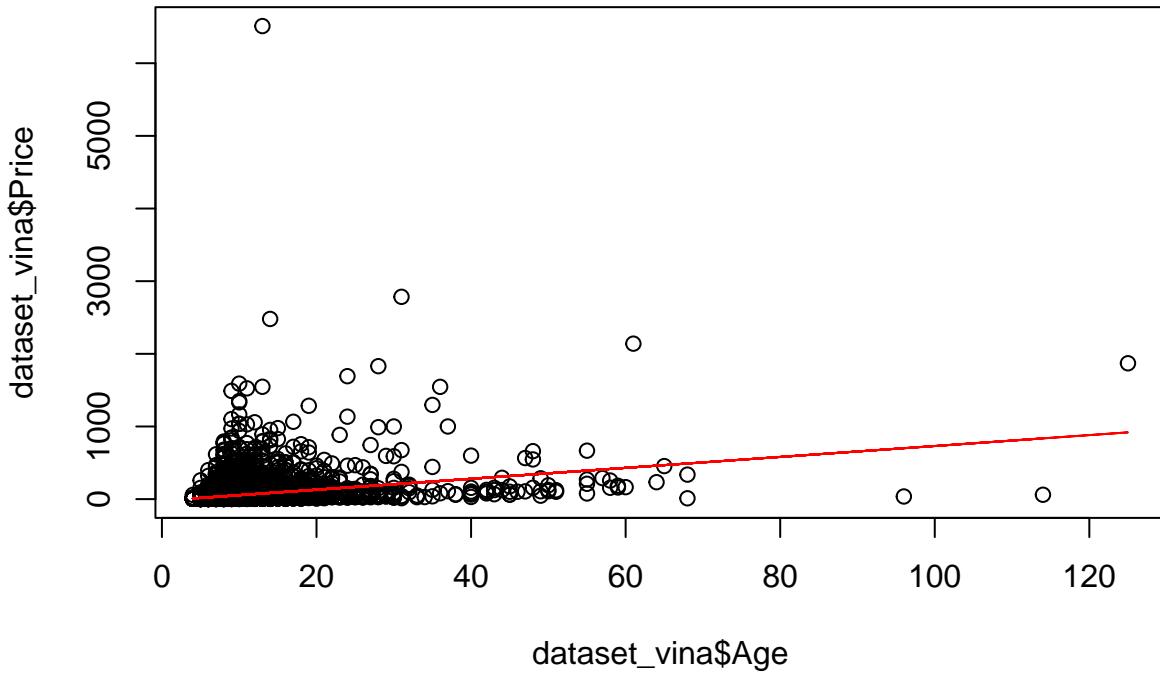
Kako kod LR pokušavamo pronaći vezu između ulaznih i izlaznih (cijena) varijabli dobra je praksa prvo promotriti utjecaj pojedinačnih varijabli na izlaznu. Različiti se prikazi koriste za numeričke i kategoriskske varijable. Već se iz samih grafova vidi da postoje neke zavisnosti među varijablama, npr. vidljivo je da će vjerojatno s rastom ocjene rasti i cijena, dok broj recenzija ne izgleda da uzrokuje toliku promjenu u cijeni.

```

dataset_vina$Age = 2025 - as.numeric(dataset_vina$Year)
fit.age = lm(Price~Age, data=dataset_vina)

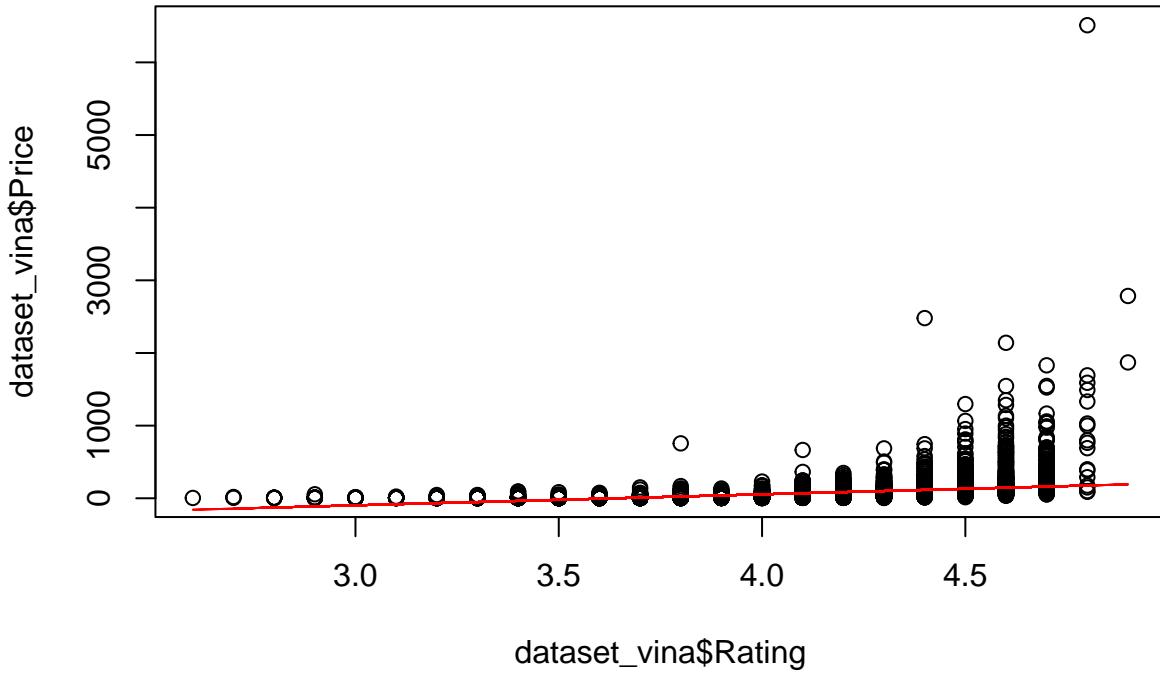
```

```
plot(dataset_vina$Age,dataset_vina$Price)
lines(dataset_vina$Age,fit.age$fitted.values,col='red')
```



```
fit.rating = lm(Price~Rating,data=dataset_vina)

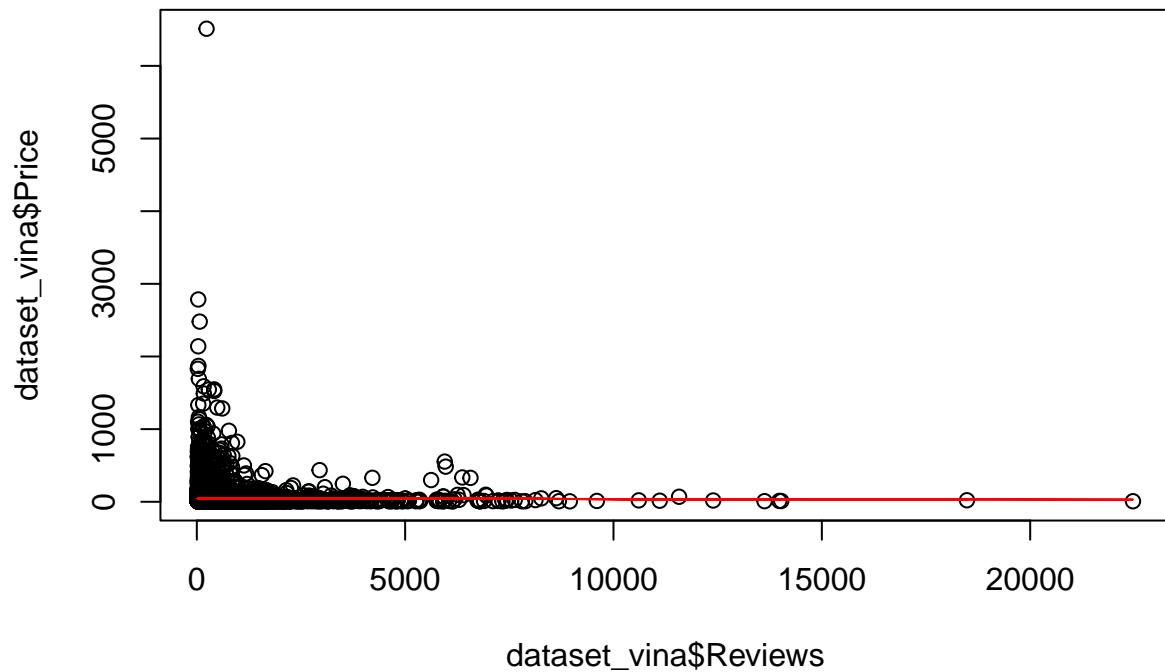
plot(dataset_vina$Rating,dataset_vina$Price)
lines(dataset_vina$Rating,fit.rating$fitted.values,col='red')
```



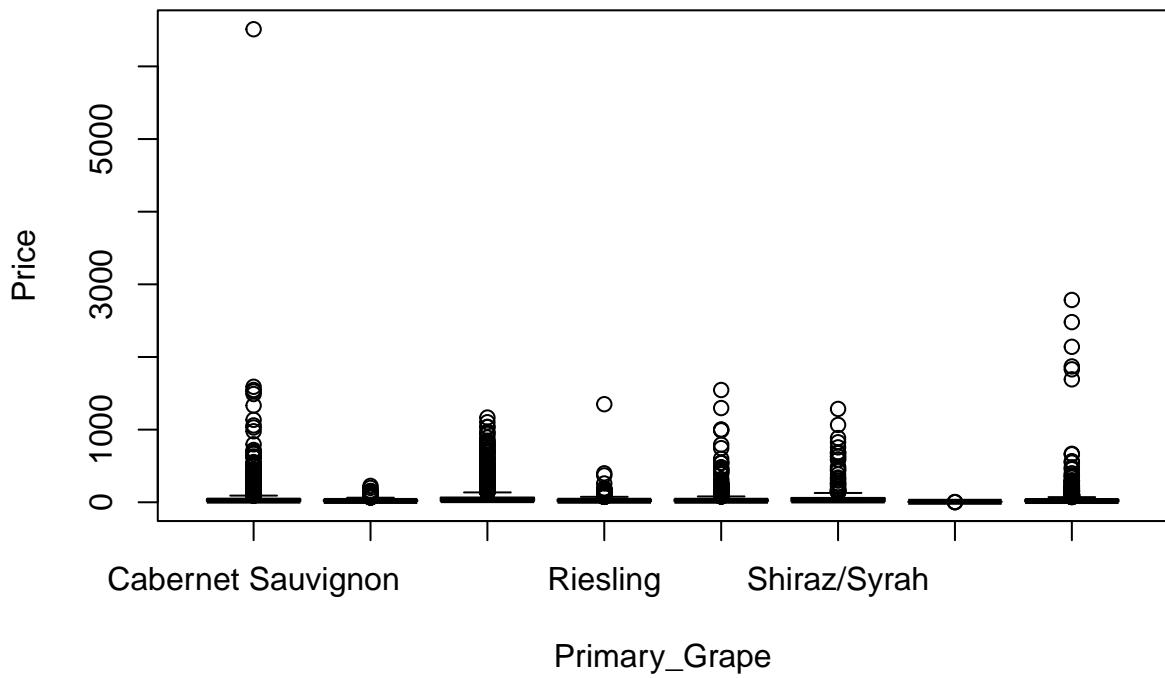
```
fit.reviews = lm(Price~Reviews,data=dataset_vina)

plot(dataset_vina$Reviews,dataset_vina$Price)
```

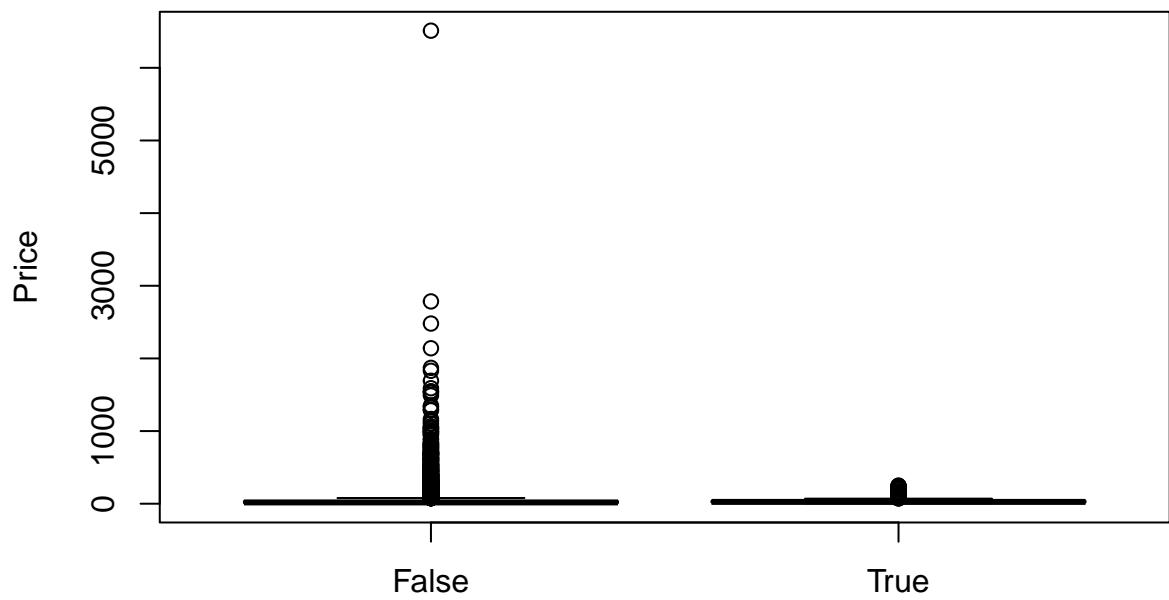
```
lines(dataset_vina$Reviews,fit.reviews$fitted.values,col='red')
```



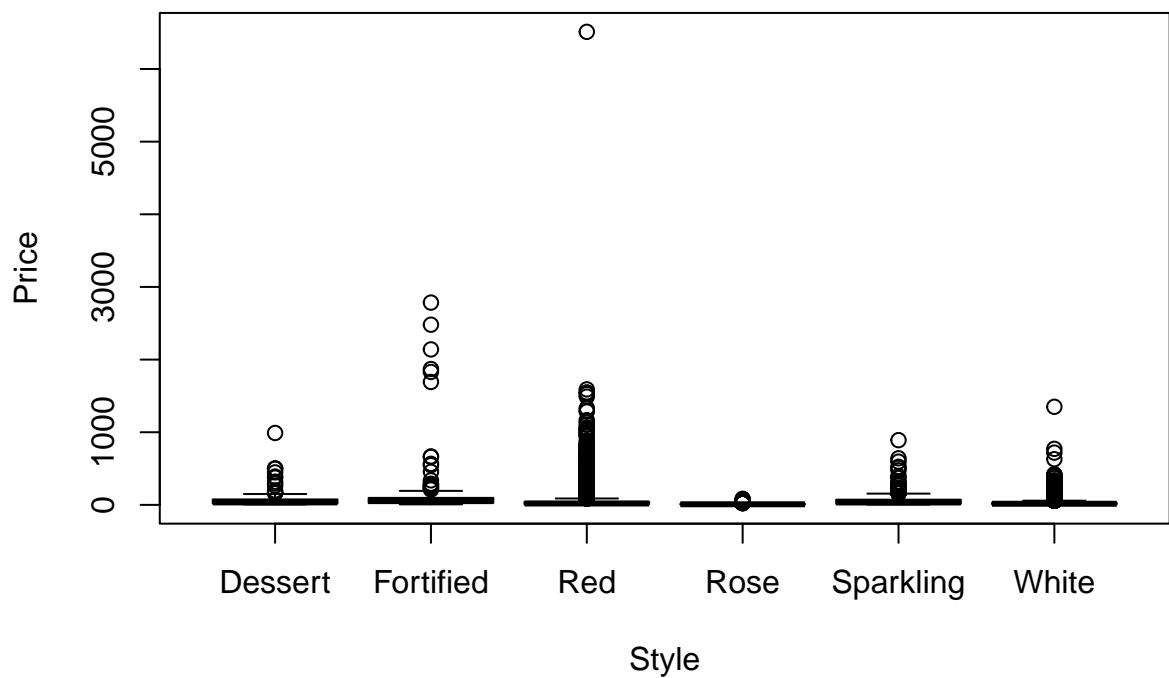
```
boxplot(Price~Primary_Grape,data=dataset_vina)
```



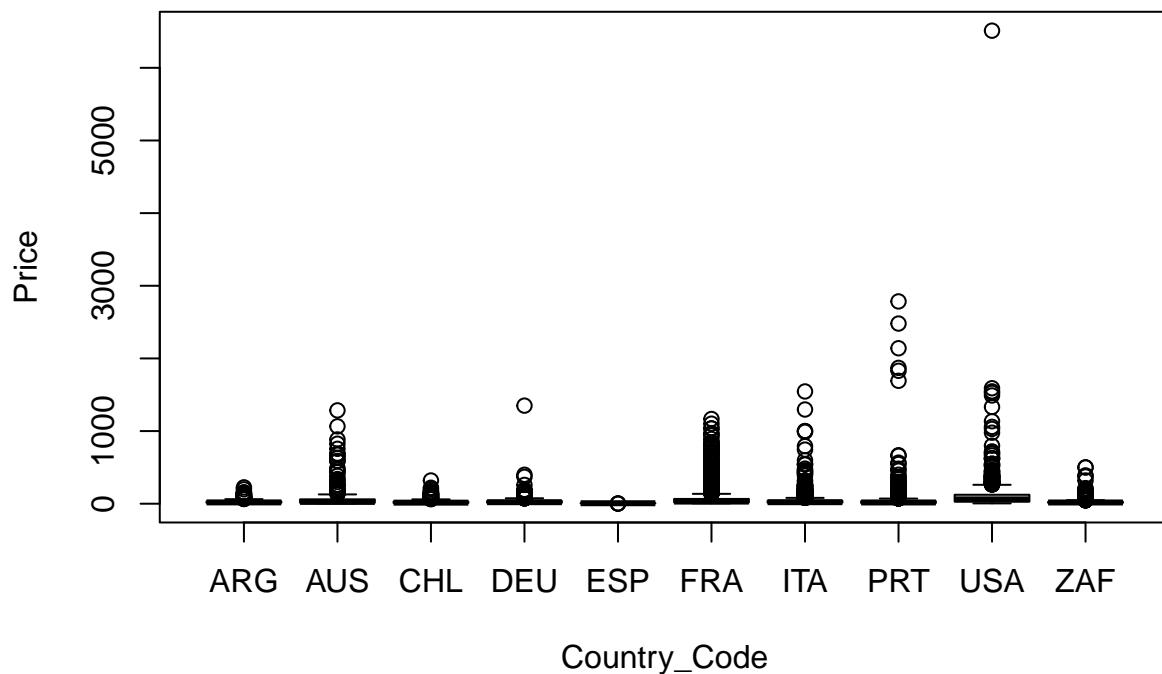
```
boxplot(Price~Natural,data=dataset_vina)
```



```
boxplot(Price~Style,data=dataset_vina)
```



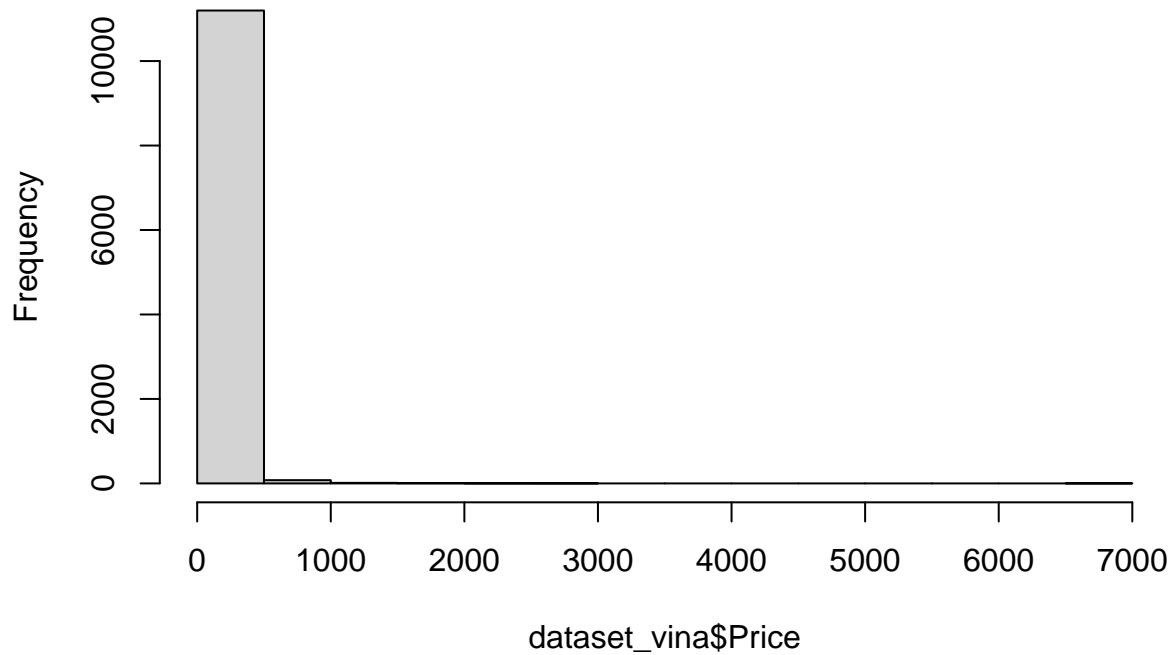
```
boxplot(Price~Country_Code,data=dataset_vina)
```



Linearna regresija je jako osjetljiva na stršeće vrijednosti pa ćemo varijablu Price provući kroz logaritamsku funkciju kako bi se umanjio njihov utjecaj. Pomoću histograma možemo vidjeti da provlačenjem cijene kroz logaritamsku funkciju distribucija više nalikuje normalnoj stoga ćemo u kasnijim modelima zasigurno koristiti  $\log(\text{Cijena})$ .

```
hist(dataset_vina$Price)
```

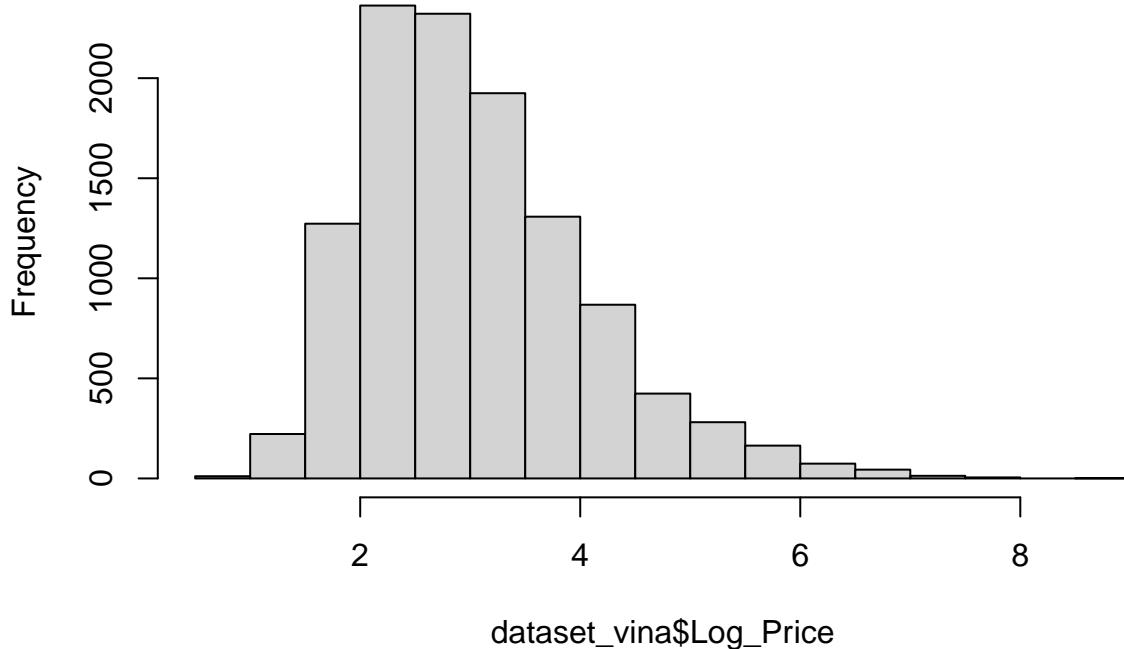
**Histogram of dataset\_vina\$Price**



```
dataset_vina$Log_Price <- log(dataset_vina$Price)
```

```
hist(dataset_vina$Log_Price)
```

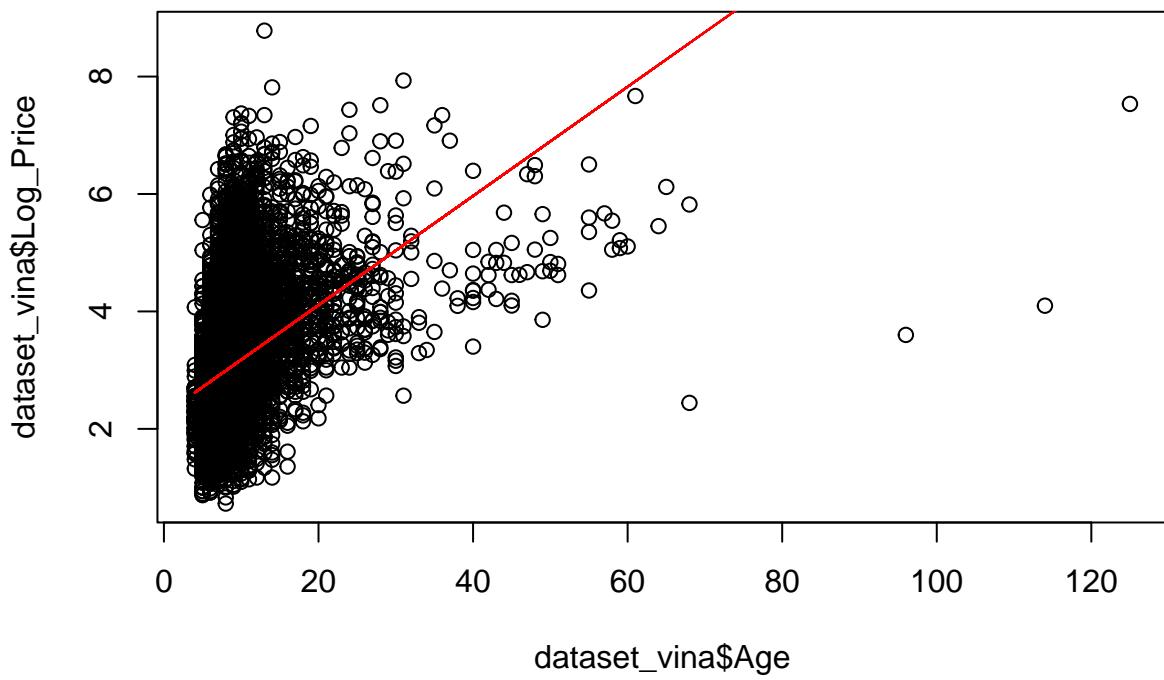
Histogram of dataset\_vina\$Log\_Price



Sada se bolje vidi kako numeričke varijable (Age, Reviews i Ratings) utjeću na cijenu vina. Uz to su dodani još grafovi s varijablama Age i Reviews provućenim kroz log.

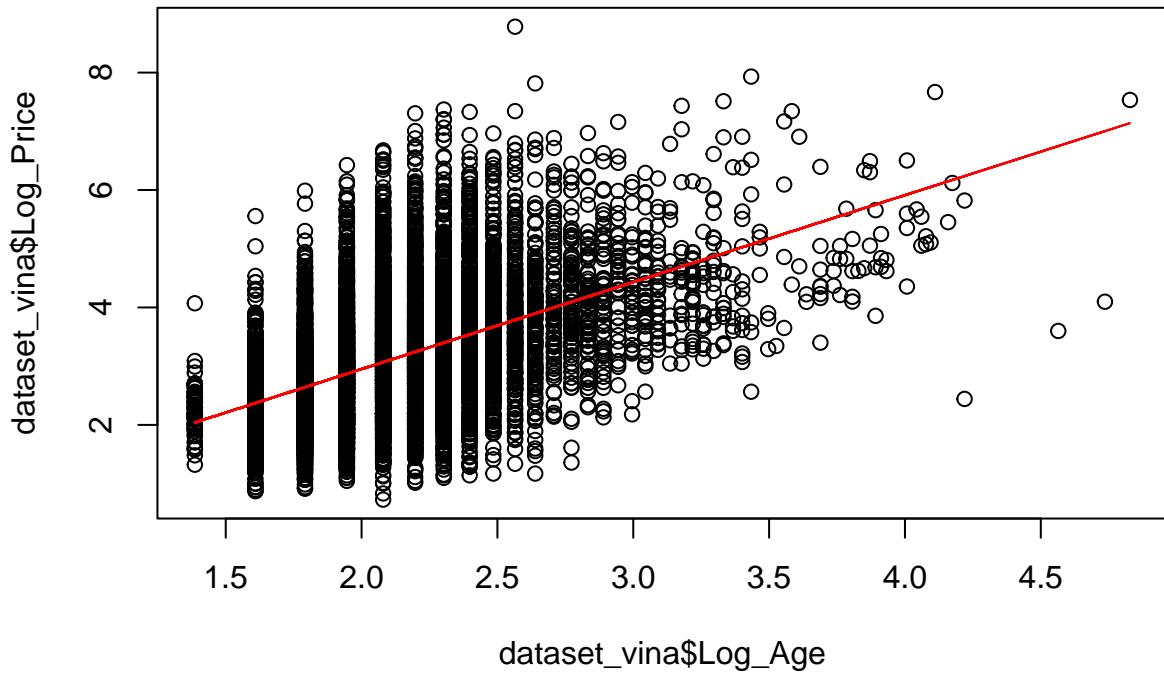
```
dataset_vina$Age = 2025 - as.numeric(dataset_vina$Year)
fit.age = lm(Log_Price~Age,data=dataset_vina)

plot(dataset_vina$Age,dataset_vina$Log_Price)
lines(dataset_vina$Age,fit.age$fitted.values,col='red')
```



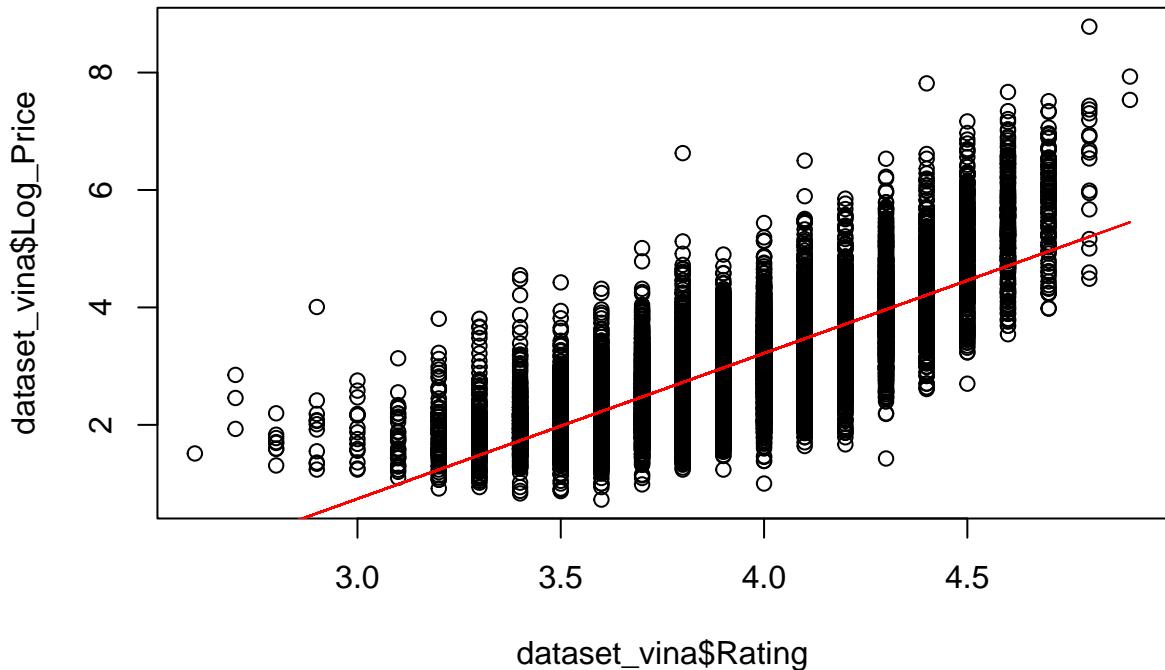
```
dataset_vina$Log_Age = log(dataset_vina$Age)
fit.age = lm(Log_Price~Log_Age,data=dataset_vina)

plot(dataset_vina$Log_Age,dataset_vina$Log_Price)
lines(dataset_vina$Log_Age,fit.age$fitted.values,col='red')
```



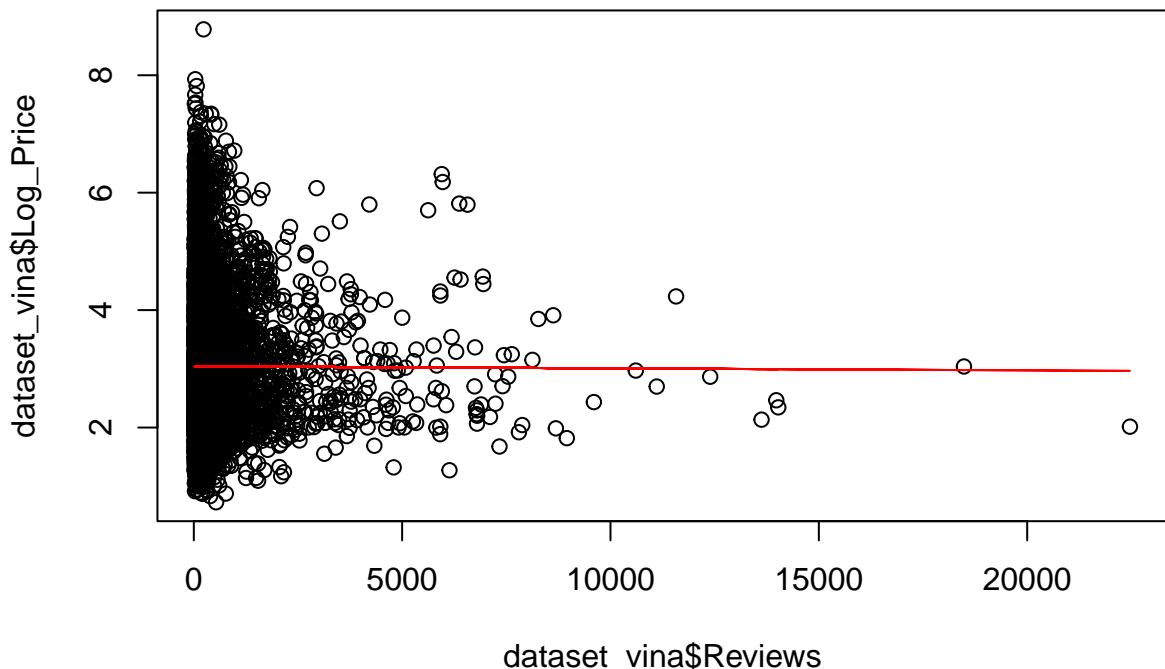
```
fit.rating = lm(Log_Price~Rating,data=dataset_vina)

plot(dataset_vina$Rating,dataset_vina$Log_Price)
lines(dataset_vina$Rating,fit.rating$fitted.values,col='red')
```



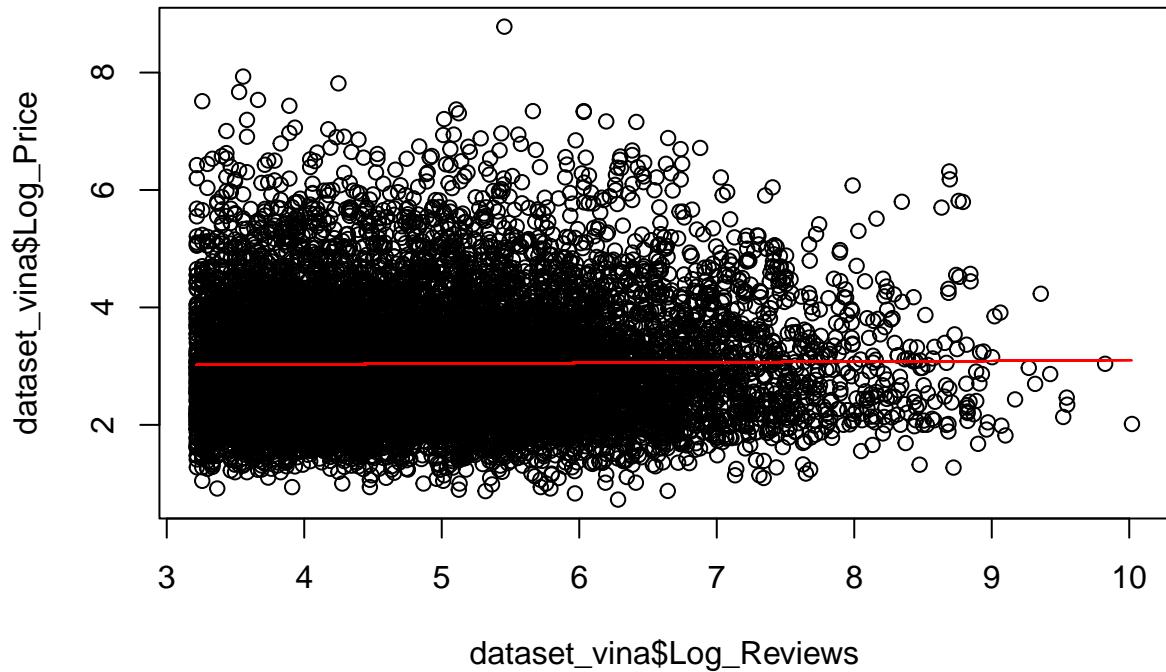
```
fit.reviews = lm(Log_Price~Reviews,data=dataset_vina)

plot(dataset_vina$Reviews,dataset_vina$Log_Price)
lines(dataset_vina$Reviews,fit.reviews$fitted.values,col='red')
```



```
dataset_vina$Log_Reviews = log(dataset_vina$Reviews)
fit.reviews = lm(Log_Price~Log_Reviews,data=dataset_vina)

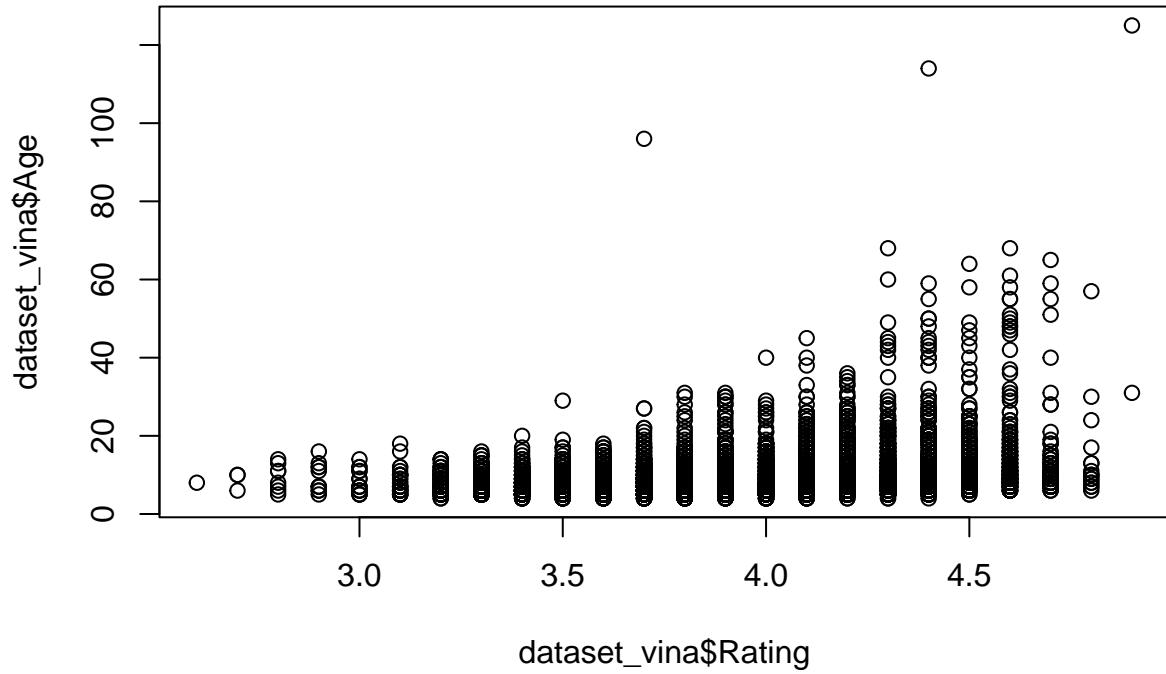
plot(dataset_vina$Log_Reviews,dataset_vina$Log_Price)
lines(dataset_vina$Log_Reviews,fit.reviews$fitted.values,col='red')
```



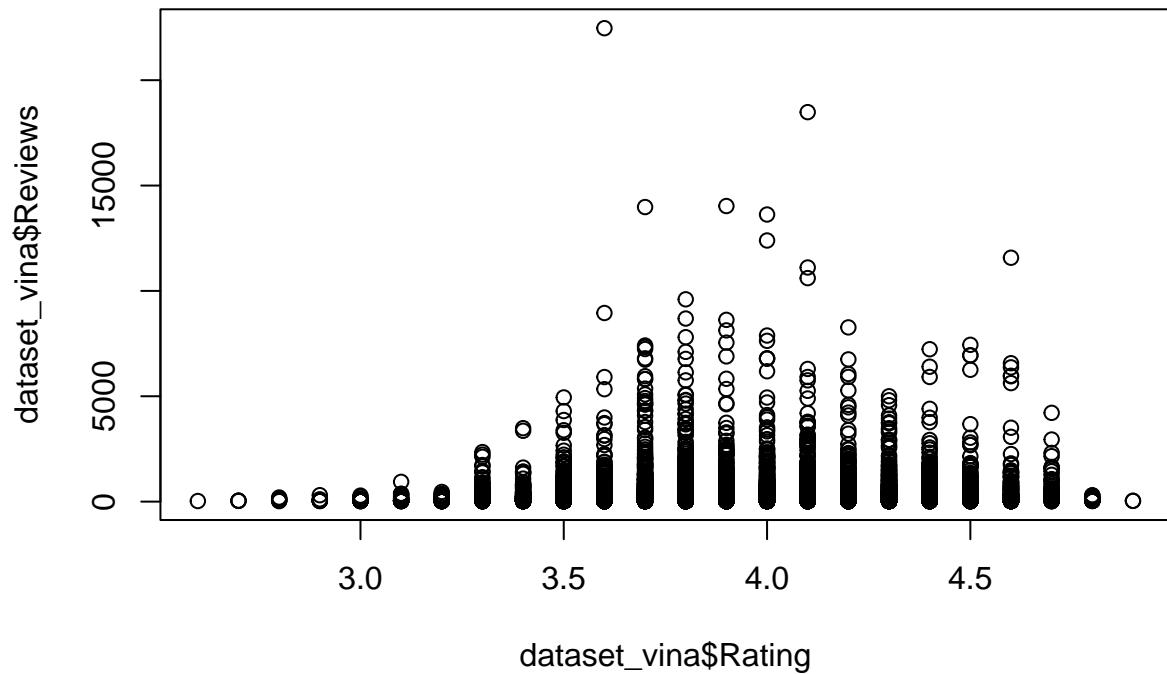
Korelacijska matrica prikazuje povezanost između nezavisnih varijabli, što nam je bitno jer linearna regresija loše funkcioniра kada su nezavisne varijable jako korelirane, stoga je u tom slučaju dobra praksa izbaciti one koje su višak.

Iako se ovdje vidi neka korelacija između starosti vina i ocjene smatramo da nije dovoljna za izbacivanje i jedne od te dvije varijable.

```
plot(dataset_vina$Rating, dataset_vina$Age)
```

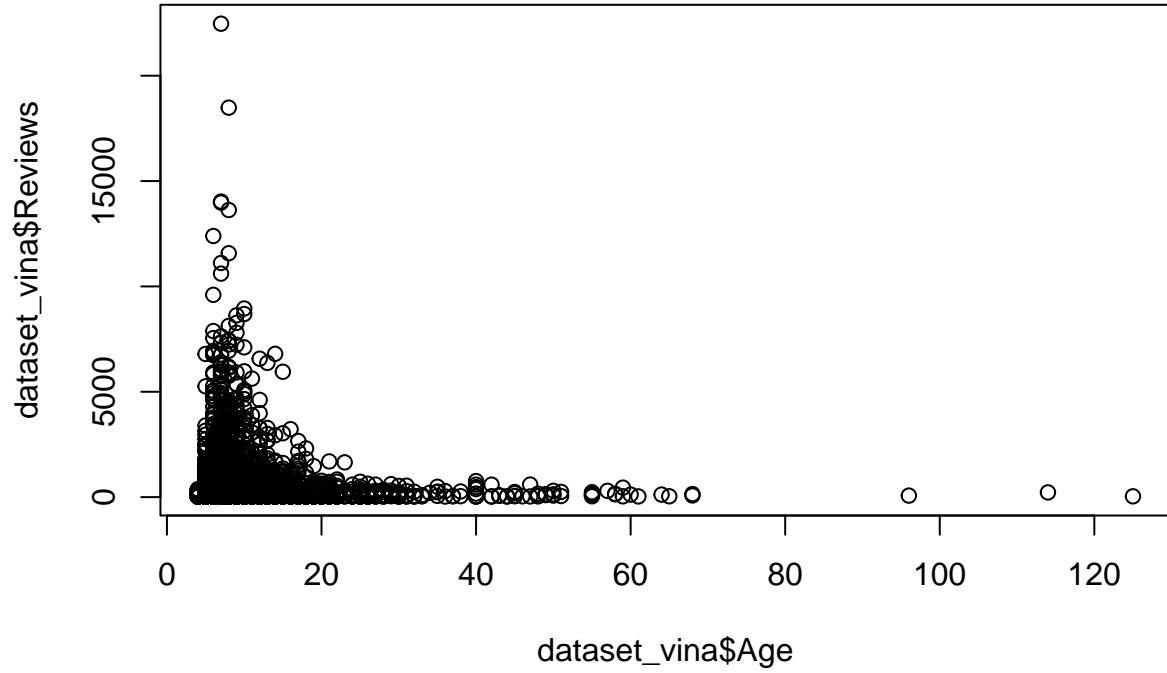


```
plot(dataset_vina$Rating, dataset_vina$Reviews)
```



dataset\_vina\$Rating

```
plot(dataset_vina$Age, dataset_vina$Reviews)
```



dataset\_vina\$Age

```
cor(cbind(dataset_vina$Age,dataset_vina$Rating,dataset_vina$Reviews))
```

```
##          [,1]      [,2]      [,3]
## [1,] 1.0000000 0.2977615 -0.01124750
## [2,] 0.2977615 1.0000000  0.04826539
## [3,] -0.0112475 0.04826539  1.00000000
```

Katgorijske varijable se predstavljaju dummy varijablama jer linearna regresija radi samo s numeričkim

podacima. Svaka kategorija predstavljena je svojom vlastitom indikatorском varijablu koja poprima vrijednost 1 u slučaju da originalna kategoriska varijabla poprima vrijednost te kategorije, a 0 inače.

```
require(fastDummies)
```

```
## Loading required package: fastDummies
dataset_vina.d = dummy_cols(dataset_vina, select_columns=c('Primary_Grape',
    'Natural', 'Country_Code', 'Style'))
```

U prvom modelu korištene su varijable: Age, Rating, Reviews, Natural, Style, Country\_Code, Primary\_Grape. Vidi se da model dosta loše procjenjuje po samoj prosječnoj grešci koja iznosi 105.3 eura te R-squared metrikom 0.2287. Također je vidljivo da su neke od kategoriskih varijabla visoko korelirane pa ćemo te varijable izbaciti iz modela, kao i one varijable koje nisu značajne za model tj. imaju visoku p vrijednost. Bitno je i da ovakav model za neke cijene predviđa negativne vrijednosti što nije prihvatljivo u ovom slučaju.

```
model <- lm(Price ~ Age + Rating + Reviews + Natural + Style + Country_Code +
    Primary_Grape, data = dataset_vina)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Age + Rating + Reviews + Natural + Style +
##     Country_Code + Primary_Grape, data = dataset_vina)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -624.2  -31.0   -7.8   15.0  6283.9
##
## Coefficients: (7 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.174e+02  1.816e+01 -28.487 < 2e-16 ***
## Age          5.460e+00  2.469e-01  22.117 < 2e-16 ***
## Rating       1.244e+02  3.582e+00  34.724 < 2e-16 ***
## Reviews      -5.108e-03  1.302e-03 -3.922 8.82e-05 ***
## NaturalTrue -9.844e+00  5.116e+00 -1.924 0.054365 .
## StyleFortified 9.083e+00  1.203e+01  0.755 0.450342
## StyleRed     1.809e+01  9.516e+00  1.901 0.057324 .
## StyleRose    2.185e+01  1.111e+01  1.967 0.049264 *
## StyleSparkling 7.482e+00  1.143e+01  0.655 0.512671
## StyleWhite   1.443e+01  9.595e+00  1.504 0.132502
## Country_CodeAUS 3.691e+01  7.477e+00  4.936 8.08e-07 ***
## Country_CodeCHL 8.997e+00  7.818e+00  1.151 0.249820
## Country_CodeDEU -1.751e+00  6.370e+00 -0.275 0.783451
## Country_CodeESP 1.996e+01  5.931e+00  3.365 0.000768 ***
## Country_CodeFRA 2.567e+01  5.907e+00  4.346 1.40e-05 ***
## Country_CodeITA 3.497e+00  5.850e+00  0.598 0.549957
## Country_CodePRT -1.150e+01  6.015e+00 -1.911 0.056015 .
## Country_CodeUSA 5.983e+01  6.467e+00  9.252 < 2e-16 ***
## Country_CodeZAF -1.881e+00  5.957e+00 -0.316 0.752252
## Primary_GrapeMalbec NA      NA      NA      NA
## Primary_GrapePinot Noir NA      NA      NA      NA
## Primary_GrapeRiesling NA      NA      NA      NA
## Primary_GrapeSangiovese NA      NA      NA      NA
## Primary_GrapeShiraz/Syrah NA      NA      NA      NA
```

```

## Primary_GrapeTempranillo          NA        NA        NA        NA
## Primary_GrapeTouriga Nacional     NA        NA        NA        NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105.3 on 11279 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.2275
## F-statistic: 185.8 on 18 and 11279 DF,  p-value: < 2.2e-16

```

Sljedeći model očekivano nije ništa bolji jer smo samo makli nepotrebne varijable i ništa nismo promjenili osim toga, ali on će biti referentni model s kojim ćemo uspoređivati ostale.

```

model_original <- lm(Price ~ Age + Rating + Reviews + Country_Code_AUS +
                      Country_Code_FRA + Country_Code_ESP + Country_Code_USA,
                      data = dataset_vina.d)

summary(model_original)

##
## Call:
## lm(formula = Price ~ Age + Rating + Reviews + Country_Code_AUS +
##     Country_Code_FRA + Country_Code_ESP + Country_Code_USA, data = dataset_vina.d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -608.2   -30.6    -8.3   14.6  6286.6 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.979e+02  1.375e+01 -36.201 < 2e-16 ***
## Age          5.240e+00  2.138e-01  24.516 < 2e-16 ***
## Rating       1.234e+02  3.562e+00  34.652 < 2e-16 ***
## Reviews      -4.585e-03  1.276e-03 -3.592 0.00033 ***
## Country_Code_AUS 4.002e+01  5.521e+00   7.250 4.45e-13 ***
## Country_Code_FRA 2.692e+01  2.868e+00   9.386 < 2e-16 ***
## Country_Code_ESP 2.229e+01  3.006e+00   7.417 1.28e-13 ***
## Country_Code_USA 6.311e+01  4.072e+00  15.498 < 2e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105.5 on 11290 degrees of freedom
## Multiple R-squared:  0.2263, Adjusted R-squared:  0.2258
## F-statistic: 471.8 on 7 and 11290 DF,  p-value: < 2.2e-16

```

Kada se koristi log od Cijene u modelu već se vidi znatno poboljšanje po R-squared i F testovima. No teško je bez računanja reći koliko je prosječna greška bolja.

```

model_log <- lm(Log_Price ~ Age + Rating + Reviews + Country_Code_AUS +
                      Country_Code_FRA + Country_Code_ESP + Country_Code_USA,
                      data = dataset_vina.d)

summary(model_log)

##
## Call:
## lm(formula = Log_Price ~ Age + Rating + Reviews + Country_Code_AUS +
##     Country_Code_FRA + Country_Code_ESP + Country_Code_USA, data = dataset_vina.d)

```

```

## 
## Residuals:
##   Min     1Q Median     3Q    Max
## -5.2375 -0.3293 -0.0258  0.2907  3.6595
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -4.751e+00 7.057e-02 -67.32 <2e-16 ***
## Age                   4.807e-02 1.097e-03  43.83 <2e-16 ***
## Rating                1.873e+00 1.828e-02 102.47 <2e-16 ***
## Reviews               -6.891e-05 6.549e-06 -10.52 <2e-16 ***
## Country_Code_AUS      4.464e-01 2.833e-02  15.76 <2e-16 ***
## Country_Code_FRA      3.825e-01 1.472e-02  25.99 <2e-16 ***
## Country_Code_ESP      -4.753e-01 1.542e-02 -30.81 <2e-16 ***
## Country_Code_USA      7.179e-01 2.090e-02  34.36 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5411 on 11290 degrees of freedom
## Multiple R-squared:  0.7242, Adjusted R-squared:  0.7241
## F-statistic:  4236 on 7 and 11290 DF,  p-value: < 2.2e-16

```

U sljedećem isječku koda isprobavamo modele s interakcijskim varijablama. Njih koristimo jer se ponekad učinci varijabli zajedno razlikuju od njihovog pojedinačnog doprinosa. To je očito slučaj i s našim setom podataka, jer se metrika R-sqr još dodatno poboljšava uvođenjem tih značajki.

```

model_log <- lm(Log_Price ~ Age + Rating + exp(Rating) + I(Age * Reviews * Rating)
+ log(Reviews) + Country_Code_AUS + Country_Code_FRA +
Country_Code_ESP + Country_Code_USA, data = dataset_vina.d)
summary(model_log)

```

```

## 
## Call:
## lm(formula = Log_Price ~ Age + Rating + exp(Rating) + I(Age *
##   Reviews * Rating) + log(Reviews) + Country_Code_AUS + Country_Code_FRA +
##   Country_Code_ESP + Country_Code_USA, data = dataset_vina.d)
## 
## Residuals:
##   Min     1Q Median     3Q    Max
## -4.4951 -0.2998 -0.0045  0.2746  3.5039
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.161e+00 2.238e-01 18.587 <2e-16 ***
## Age                   3.925e-02 1.053e-03 37.293 <2e-16 ***
## Rating                -1.046e+00 7.453e-02 -14.041 <2e-16 ***
## exp(Rating)          5.492e-02 1.364e-03 40.270 <2e-16 ***
## I(Age * Reviews * Rating) 4.835e-07 2.443e-07  1.979 0.0478 *
## log(Reviews)         -6.338e-02 5.422e-03 -11.689 <2e-16 ***
## Country_Code_AUS      3.901e-01 2.641e-02 14.771 <2e-16 ***
## Country_Code_FRA      3.768e-01 1.368e-02 27.541 <2e-16 ***
## Country_Code_ESP      -5.670e-01 1.451e-02 -39.085 <2e-16 ***
## Country_Code_USA       6.081e-01 1.969e-02 30.890 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## 
## Residual standard error: 0.5031 on 11288 degrees of freedom
## Multiple R-squared:  0.7616, Adjusted R-squared:  0.7615
## F-statistic: 4008 on 9 and 11288 DF, p-value: < 2.2e-16

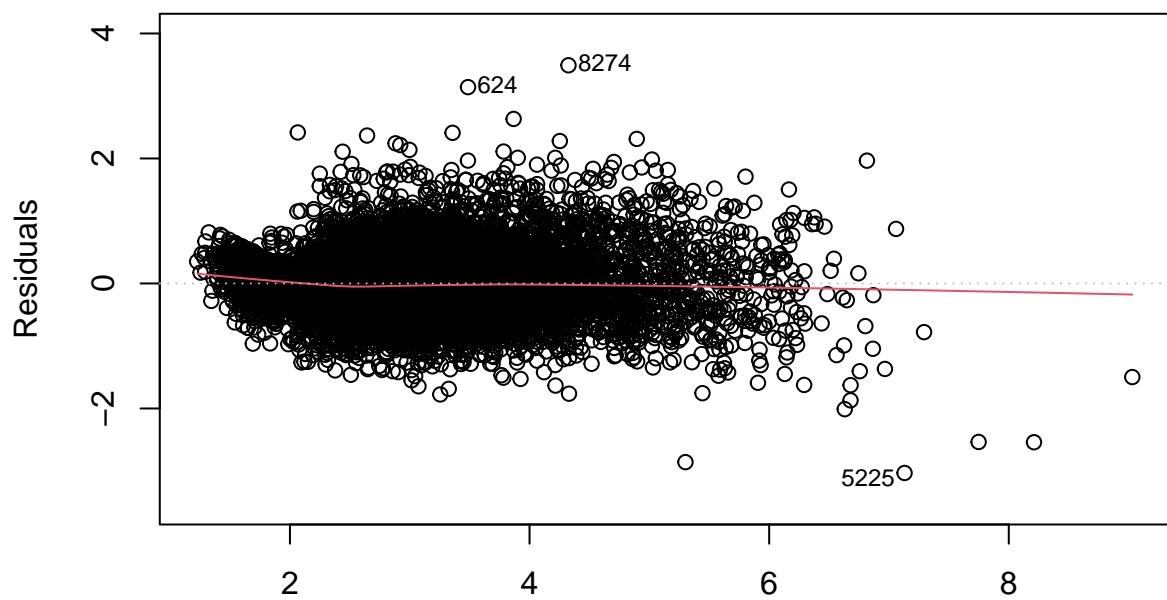
model_log <- lm(Log_Price ~ Age + Rating + exp(Rating) +
  I(log(Age) * log(Reviews) * exp(Rating)) + log(Reviews) +
  Country_Code_AUS + Country_Code_FRA + Country_Code_ESP +
  Country_Code_USA, data = dataset_vina.d)
summary(model_log)

##
## Call:
## lm(formula = Log_Price ~ Age + Rating + exp(Rating) + I(log(Age) *
##   log(Reviews) * exp(Rating)) + log(Reviews) + Country_Code_AUS +
##   Country_Code_FRA + Country_Code_ESP + Country_Code_USA, data = dataset_vina.d)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -3.0323 -0.2982 -0.0135  0.2713  3.4910
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.097e+00  2.184e-01 18.756 < 2e-16
## Age                      9.532e-03  1.628e-03  5.857 4.85e-09
## Rating                   -6.414e-01  7.475e-02 -8.580 < 2e-16
## exp(Rating)              3.114e-02  1.672e-03 18.627 < 2e-16
## I(log(Age) * log(Reviews) * exp(Rating)) 1.496e-03  6.333e-05 23.626 < 2e-16
## log(Reviews)             -2.321e-01  8.445e-03 -27.482 < 2e-16
## Country_Code_AUS          3.963e-01  2.579e-02 15.369 < 2e-16
## Country_Code_FRA          3.648e-01  1.337e-02 27.286 < 2e-16
## Country_Code_ESP          -5.711e-01  1.416e-02 -40.323 < 2e-16
## Country_Code_USA           6.124e-01  1.922e-02 31.864 < 2e-16
##
## (Intercept) *** 
## Age          ***
## Rating       ***
## exp(Rating) ***
## I(log(Age) * log(Reviews) * exp(Rating)) ***
## log(Reviews) ***
## Country_Code_AUS ***
## Country_Code_FRA ***
## Country_Code_ESP ***
## Country_Code_USA ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4912 on 11288 degrees of freedom
## Multiple R-squared:  0.7728, Adjusted R-squared:  0.7726
## F-statistic: 4266 on 9 and 11288 DF, p-value: < 2.2e-16

plot(model_log)

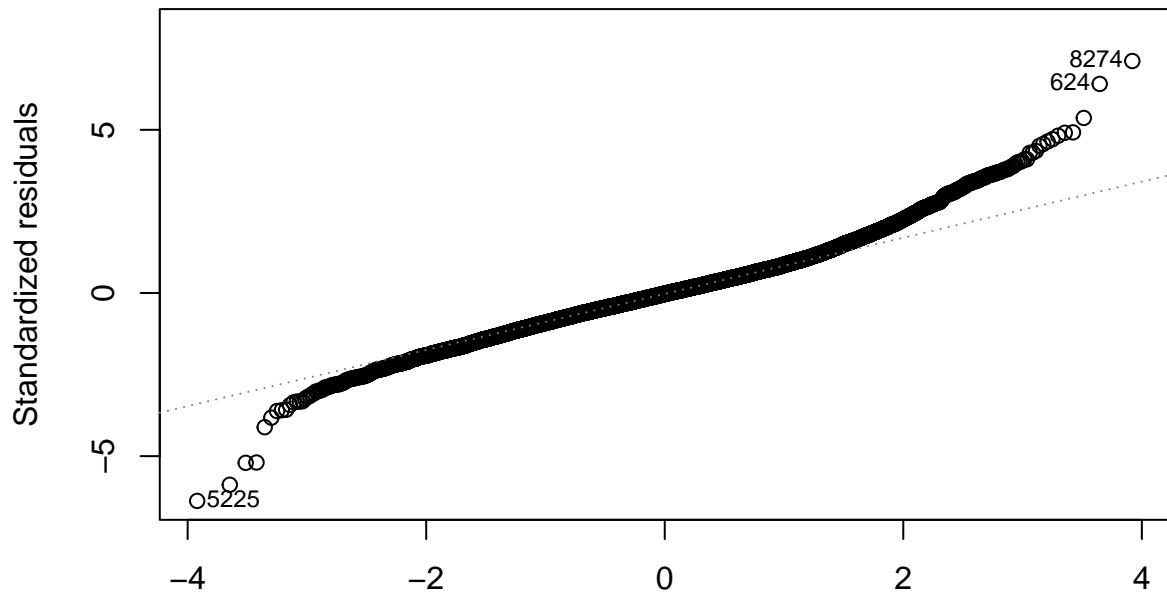
```

Residuals vs Fitted



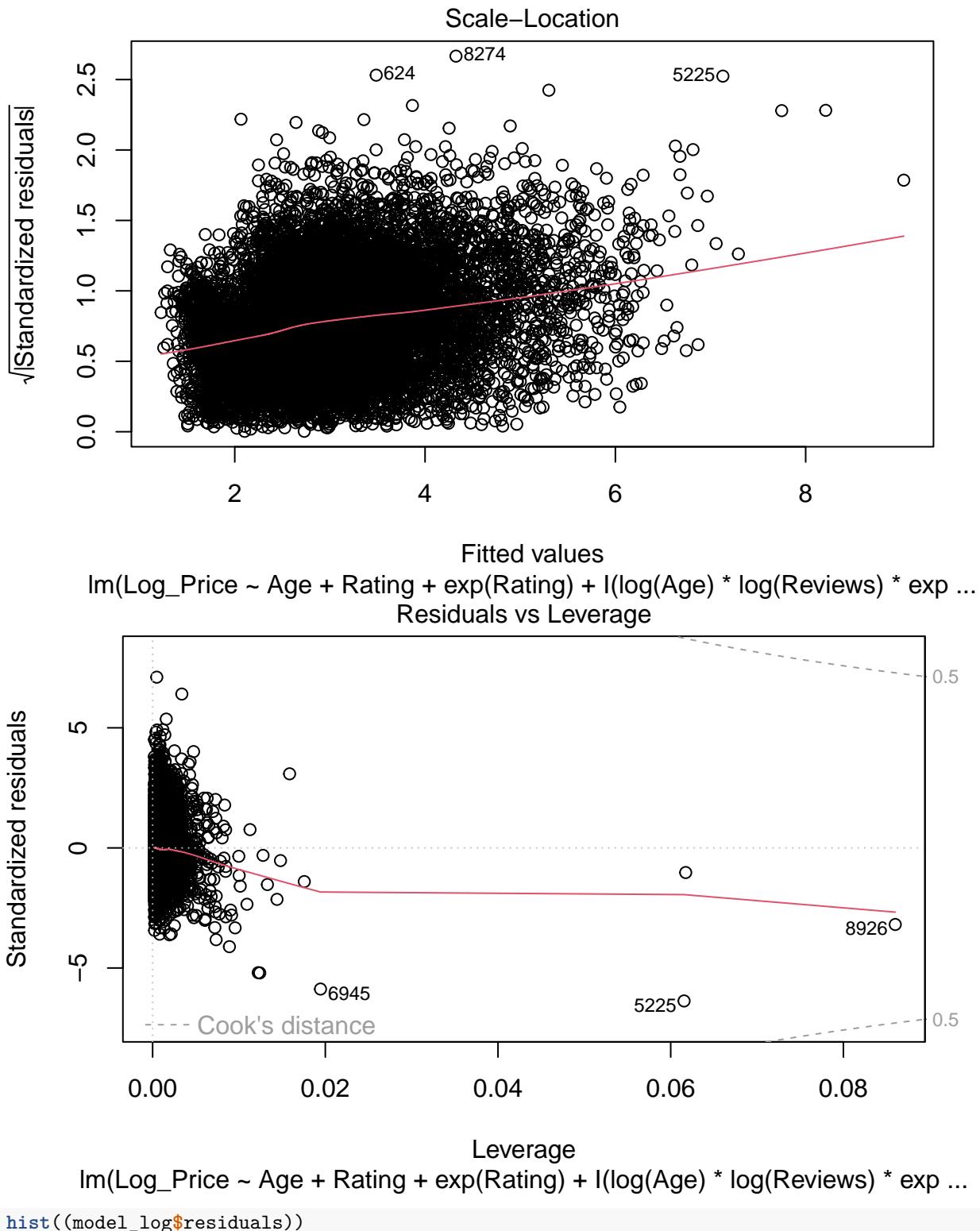
Fitted values

lm(Log\_Price ~ Age + Rating + exp(Rating) + I(log(Age)) \* log(Reviews) \* exp ...  
Q-Q Residuals

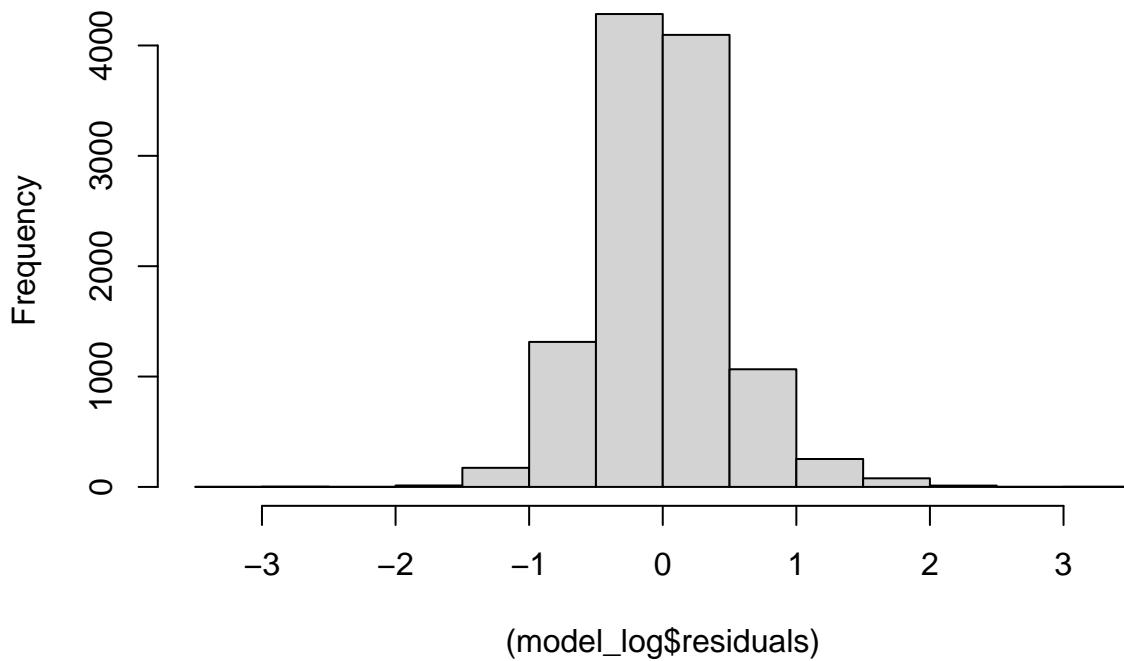


Theoretical Quantiles

lm(Log\_Price ~ Age + Rating + exp(Rating) + I(log(Age)) \* log(Reviews) \* exp ...

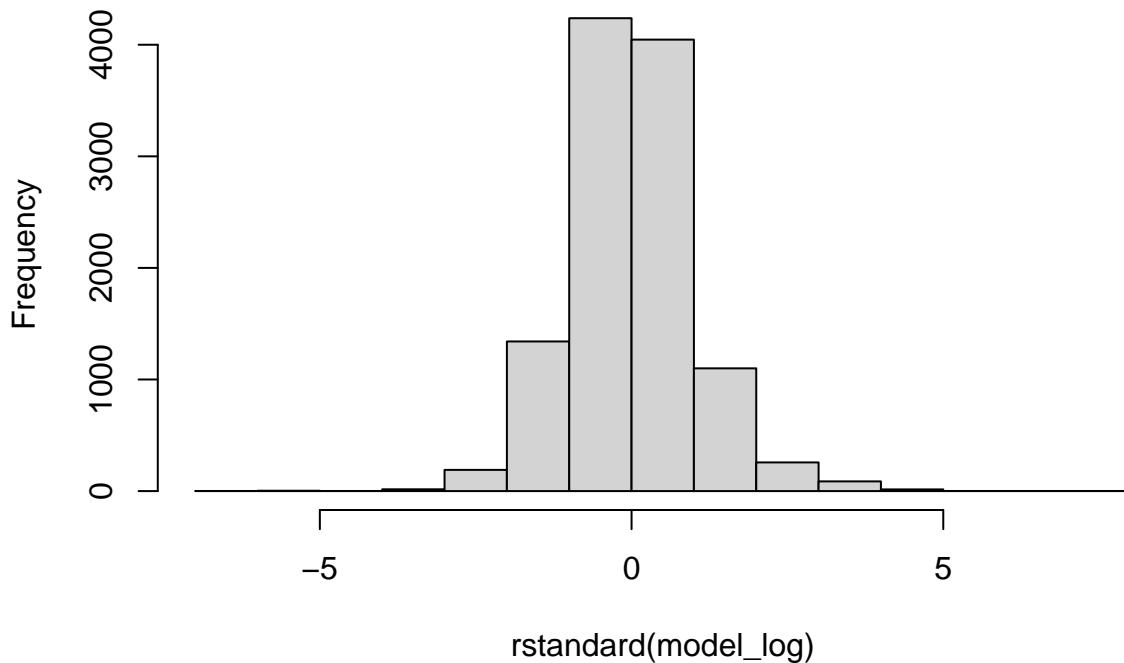


### Histogram of (model\_log\$residuals)



```
hist(rstandard(model_log))
```

### Histogram of rstandard(model\_log)



Kako bih lakše pridočili razliku u modelima ovdje računamo razliku prosječne pogreške između modela s CIjenum i log(cijenom)

```

predictions_original <- predict(model_original, newdata = dataset_vina.d)

predictions_log <- exp(predict(model_log, newdata = dataset_vina.d))

actual <- dataset_vina.d$Price

mae_original <- mean(abs(predictions_original - actual))
mae_log <- mean(abs(predictions_log - actual))

cat("MAE (Original Scale):", mae_original, "\n")

## MAE (Original Scale): 37.75807
cat("MAE (Log Scale):", mae_log, "\n")

```

## MAE (Log Scale): 21.65873

Kako bi poboljšali model ponekad je dobro izbaciti stršeće vrijednosti, jer one znatno utječu na model a često nisu reprezentativna za većinu uzorka. Iako smo ovdje to prikazali ponekad izbacivanje stršećih vrijednosti izbacuje bitne podatke o anomalijama u modelu što možda poboljšava model na podacima na kojima je treniran, ali gubi mogućnost generalizacije na neviđenim podacima. To je očenito preporučeno kada se smatra da su stršeće vrijednosti rezultat pogreške u podacima što ovdje nije nužno slučaj.

```

standardized_residuals <- rstandard(model_log)

outliers <- which(abs(standardized_residuals) > 3)

dataset_vina.cleaned <- dataset_vina.d[-outliers, ]

model_log <- lm(Log_Price ~ Age + Rating + exp(Rating) +
  I(log(Age) * log(Reviews) * exp(Rating)) + log(Reviews) +
  Country_Code_AUS + Country_Code_FRA + Country_Code_ESP +
  Country_Code_USA, data = dataset_vina.cleaned)
summary(model_log)

##
## Call:
## lm(formula = Log_Price ~ Age + Rating + exp(Rating) + I(log(Age) *
##   log(Reviews) * exp(Rating)) + log(Reviews) + Country_Code_AUS +
##   Country_Code_FRA + Country_Code_ESP + Country_Code_USA, data = dataset_vina.cleaned)
##
## Residuals:
##       Min     1Q     Median      3Q     Max 
## -1.58193 -0.28240 -0.00429  0.27397  1.50080 
##
## Coefficients:
## (Intercept)          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          3.860e+00  2.043e-01 18.893 <2e-16  
## Age                  1.550e-02  1.674e-03  9.260 <2e-16  
## Rating               -5.966e-01  7.018e-02 -8.502 <2e-16  
## exp(Rating)          3.043e-02  1.588e-03 19.171 <2e-16  
## I(log(Age) * log(Reviews) * exp(Rating)) 1.441e-03  6.112e-05 23.571 <2e-16  
## log(Reviews)          -2.189e-01  8.058e-03 -27.164 <2e-16  
## Country_Code_AUS      3.910e-01  2.393e-02 16.340 <2e-16  
## Country_Code_FRA      3.487e-01  1.248e-02 27.953 <2e-16

```

```

## Country_Code_ESP           -5.508e-01  1.313e-02 -41.953   <2e-16
## Country_Code_USA          6.250e-01  1.783e-02  35.058   <2e-16
##
## (Intercept)                 ***
## Age                         ***
## Rating                       ***
## exp(Rating)                  ***
## I(log(Age) * log(Reviews) * exp(Rating)) ***
## log(Reviews)                  ***
## Country_Code_AUS             ***
## Country_Code_FRA             ***
## Country_Code_ESP              ***
## Country_Code_USA              ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4539 on 11162 degrees of freedom
## Multiple R-squared:  0.7944, Adjusted R-squared:  0.7943
## F-statistic:  4793 on 9 and 11162 DF,  p-value: < 2.2e-16

```

Vidi se da se greška modela bez outliera dodatno smanjila, ali na punom skupu podataka s stržećim vrijednostima performa nešto gore.

```

predictions_original <- predict(model_original, newdata = dataset_vina.d)

predictions_log <- exp(predict(model_log, newdata = dataset_vina.cleaned))
predictions_log_full <- exp(predict(model_log, newdata = dataset_vina.d))

actual <- dataset_vina.d$Price
actual_cleaned <- dataset_vina.cleaned$Price
actual_full <- dataset_vina.d$Price

mae_original <- mean(abs(predictions_original - actual))
mae_log <- mean(abs(predictions_log - actual_cleaned))
mae_log_full <- mean(abs(predictions_log_full - actual_full))

cat("MAE (Original Scale):", mae_original, "\n")

## MAE (Original Scale): 37.75807
cat("MAE (Log Scale Cleaned):", mae_log, "\n")

## MAE (Log Scale Cleaned): 17.17211
cat("MAE (Log Scale Full):", mae_log_full, "\n")

## MAE (Log Scale Full): 22.60075

```

Pa, može li se na temelju dostupnih podataka predvidjeti cijena vina? Naš odgovor je, donekle, ali ne pouzdano (barem koristeći linearnu regresiju). Smatramo da neke druge varijable kao možda ime proizvođača i vina (marka) dosta utječu na samu cijenu pa nije moguće napraviti model linearne regresije koji je iznimno dobro predviđa. S tim da najbolji model ima prosječnu grešku od čak 20 dolara ne bih vjerovali toj predikciji da kupujemo vino na bez da vidimo pravu cijenu.

Posljednje pitanje na koje nas zanima odgovor jest: "Koja vinska regija nudi najbolji omjer ocjene i cijene vina? Koju vinariju bismo posjetili iz te regije?"

Započinjemo sa pregledom relevantnih podataka, u nastavku možemo vidjeti broj regija koje svaka država sadrži.

```
library(dplyr)

region_counts <- dataset_vina %>%
  group_by(Country) %>%
  summarize(Region_Count = n_distinct(Region))

print(region_counts)

## # A tibble: 10 x 2
##   Country      Region_Count
##   <chr>          <int>
## 1 Argentina        21
## 2 Australia         29
## 3 Chile             22
## 4 France            278
## 5 Germany            36
## 6 Italy              214
## 7 Portugal            38
## 8 South Africa       38
## 9 Spain                74
## 10 United States       60
```

Za početak, omjer ocijene i cijene izračunat ćemo tako da podatke grupiramo po regijama te za svaku regiju odredimo prosječnu ocijenu (Avg\_Rating), prosječnu cijenu vina u toj regiji (Avg\_Price) te na temelju njih izračunamo njihov omjer (Avg\_Rating\_Price\_Ratio). Regije u nastavku sortirane su silazno na temelju njihovog omjera.

Prvih 10 regija koje imaju najveći omjer ocjene i cijene:

```
result <- dataset_vina %>%
  group_by(Region) %>%
  summarize(
    Total_Vines = n(),
    Avg_Rating = mean(Rating, na.rm = TRUE),
    Avg_Price = mean(Price, na.rm = TRUE),
    Avg_Rating_Price_Ratio = Avg_Rating / Avg_Price
  ) %>%
  arrange(desc(Avg_Rating_Price_Ratio))

head(result, 10)

## # A tibble: 10 x 5
##   Region      Total_Vines  Avg_Rating  Avg_Price  Avg_Rating_Price_Ratio
##   <chr>          <int>      <dbl>      <dbl>                  <dbl>
## 1 Valle del Cinca      3       3.53      3.70                  0.954
## 2 Ribatejo            1       3.5       3.79                  0.923
## 3 Castelló            1       3.8       4.25                  0.894
## 4 Castelli Romani     1       3.2       4.13                  0.775
## 5 Valdepeñas          8       3.52      4.75                  0.742
## 6 La Mancha           19      3.55      5.37                  0.661
## 7 Cariñena            57      3.60      5.47                  0.658
## 8 Sierra de Salamanca 1       3.2       4.98                  0.643
## 9 Campo de Borja       29      3.64      5.81                  0.626
```

```

## 10 Almansa           13      3.75      5.99      0.626

Prikaz prosječnih ocijena i cijena vinarija u regiji sa najboljim omjerom ocijene i kvalitete:

library(dplyr)
library(ggplot2)
library(tidyr)

valle_del_cinca_data <- dataset_vina %>%
  filter(Region == "Valle del Cinca")

print(valle_del_cinca_data)

##   Winery Year Wine_ID                               Wine Rating
## 1 Nuviana 2019 1656567 Cabernet Sauvignon - Tempranillo Rosado 2019 3.7
## 2 Nuviana 2020 1656567 Cabernet Sauvignon - Tempranillo Rosado 2020 3.6
## 3 Nuviana 2020 5250039                           Tinto 2020 3.3
##   Reviews Price          Region Primary_Grape Natural Country Style
## 1     491  3.17 Valle del Cinca    Tempranillo  False Spain Rose
## 2      40  3.95 Valle del Cinca    Tempranillo  False Spain Rose
## 3      27  3.99 Valle del Cinca    Tempranillo  False Spain Red
##   Country_Code Age Log_Price Log_Age Log_Reviews
## 1           ESP  6  1.153732  1.791759  6.196444
## 2           ESP  5  1.373716  1.609438  3.688879
## 3           ESP  5  1.383791  1.609438  3.295837

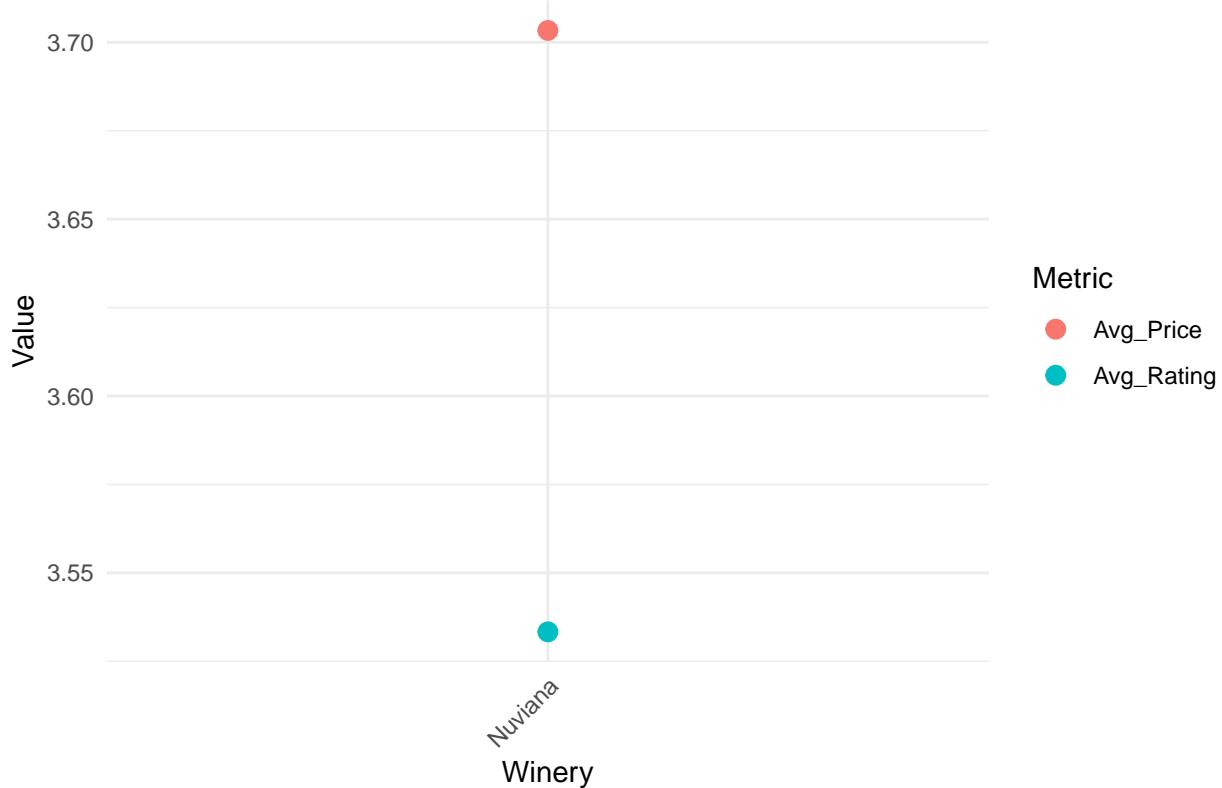
summary_data <- valle_del_cinca_data %>%
  group_by(Winery) %>%
  summarize(
    Avg_Rating = mean(Rating, na.rm = TRUE),
    Avg_Price = mean(Price, na.rm = TRUE)
  )

long_data <- summary_data %>%
  pivot_longer(
    cols = c(Avg_Rating, Avg_Price),
    names_to = "Metric",
    values_to = "Value"
  )

ggplot(long_data, aes(x = Winery, y = Value, color = Metric, group = Metric)) +
  geom_point(size = 3) +
  labs(
    title = "Average Review and Price by Winery",
    x = "Winery",
    y = "Value",
    color = "Metric"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

## Average Review and Price by Winery



Možemo vidjeti da kad smo izračunali samo omjer ocijene i cijene, dobili samo da najbolji omjer ima regija koja sadrži samo dvije različite vrste vina, pri čemu imamo prikaz jedne drste vina kroz dvije godine. Ocijene su također relativno niske, ali cijena vina bliska je ocijenama, stoga je omjer ocjene i cijene najveći, no to ne mora značiti da je uistinu i najbolji. Budući da regija sadrži mali broj zapisa, nema velikih fluktuacija u ocijenama što također doprinosi visokoj vrijednosti omjera.

Kako bi prosječne vrijednosti adekvatnije prikazivale stanje u regiji, promotrimo sada prvih 10 regija koje imaju najveći omjer ocijene i cijene, a da imaju više od 10 ocijena vina unutar regije:

```
min_vines <- 10
result <- dataset_vina %>%
  group_by(Region) %>%
  summarize(
    Total_Vines = n(),
    Avg_Rating = mean(Rating, na.rm = TRUE),
    Avg_Price = mean(Price, na.rm = TRUE),
    Avg_Rating_Price_Ratio = Avg_Rating / Avg_Price
  ) %>%
  filter(Total_Vines > min_vines) %>%
  arrange(desc(Avg_Rating_Price_Ratio))

head(result, 10)

## # A tibble: 10 x 5
##   Region      Total_Vines  Avg_Rating  Avg_Price Avg_Rating_Price_Ratio
##   <chr>          <int>       <dbl>       <dbl>             <dbl>
## 1 La Mancha      19        3.55       5.37            0.661
## 2 Cariñena       57        3.60       5.47            0.658
```

```

## 3 Campo de Borja      29     3.64     5.81     0.626
## 4 Almansa             13     3.75     5.99     0.626
## 5 Castilla            49     3.61     5.87     0.616
## 6 Navarra             81     3.56     5.91     0.603
## 7 Yecla               14     3.64     6.28     0.580
## 8 Utiel-Requena       25     3.68     6.41     0.574
## 9 Valencia            50     3.61     6.31     0.571
## 10 Somontano           65     3.64     6.48     0.562

best <- head(result, 1)

```

Vinarije u regiji koja ima najveći omjer ocijene i cijene, a sadrže više od 10 zapisa o vinima unutar regije:

```

library(dplyr)
library(ggplot2)
library(tidyr)

data_best <- dataset_vina %>%
  filter(Region %in% best$Region)

print(data_best)

##                               Winery Year Wine_ID          Wine
## 1 Finca Venta de Don Quijote 2020 1939609 Sauvignon Blanc - Macabeo 2020
## 2                         Ayuso 2015 2381489 Castillo de Benízar Tempranillo 2015
## 3 Finca La Estacada 2014 5376743 Tempranillo Crianza 2014
## 4 Vinoletto 2015 4585737 Garnacha Seco 2015
## 5 Vinoletto 2018 4585737 Garnacha Seco 2018
## 6 Ayuso 2017 2198505 Estola Crianza 2017
## 7 Oristan 2015 78569 Bronze Crianza 2015
## 8 Finca Los Trenzones 2020 77084 Verdejo 2020
## 9 Ayuso 2016 1138136 Estola Reserva 2016
## 10 Campos de Dulcinea 2019 5453192 Selección de Familia 2019
## 11 Finca Antigua 2016 1513516 Garnacha 2016
## 12 Finca Antigua 2017 13940 Petit Verdot 2017
## 13 Volver 2019 1542510 Paso a Paso Organic Red 2019
## 14 Pago de la Jaraba 2016 2490463 Azagador Reserva Tinto 2016
## 15 Hacienda Albae 2017 2801731 Cabernet Sauvignon 2017
## 16 Hacienda Albae 2018 1459107 Syrah 2018
## 17 Finca Antigua 2014 13943 Crianza 2014
## 18 Bodegas Quinta de Aves 2019 6247755 Syrah 2019
## 19 Manuel Manzaneque Suárez 2017 1860918 ;Ea! Tinto 2017

##   Rating Reviews Price Region Primary_Grape Natural Country Style
## 1    4.0     317 2.720 La Mancha Tempranillo  False Spain  White
## 2    3.3     151 3.130 La Mancha Tempranillo  False Spain   Red
## 3    3.6     363 3.625 La Mancha Tempranillo  False Spain   Red
## 4    3.1      33 3.750 La Mancha Tempranillo  False Spain   Red
## 5    3.4      28 3.750 La Mancha Tempranillo  False Spain   Red
## 6    3.4     240 3.850 La Mancha Tempranillo  False Spain   Red
## 7    3.4     308 4.000 La Mancha Tempranillo  False Spain   Red
## 8    3.6      39 4.050 La Mancha Tempranillo  False Spain  White
## 9    3.6     530 4.660 La Mancha Tempranillo  False Spain   Red
## 10   3.3      35 5.375 La Mancha Tempranillo  False Spain   Red
## 11   3.6      70 5.600 La Mancha Tempranillo  False Spain   Red
## 12   3.6      47 5.950 La Mancha Tempranillo  False Spain   Red

```

```

## 13    3.5    42 6.500 La Mancha    Tempranillo  False  Spain  Red
## 14    3.6    30 6.500 La Mancha    Tempranillo  False  Spain  Red
## 15    3.7    29 6.900 La Mancha    Tempranillo  False  Spain  Red
## 16    3.6    28 6.900 La Mancha    Tempranillo  False  Spain  Red
## 17    3.8    97 7.200 La Mancha    Tempranillo  False  Spain  Red
## 18    3.6    49 8.650 La Mancha    Tempranillo  False  Spain  Red
## 19    3.8    59 8.950 La Mancha    Tempranillo  False  Spain  Red
##   Country_Code Age Log_Price Log_Age Log_Reviews
## 1          ESP  5  1.000632 1.609438  5.758902
## 2          ESP 10  1.141033 2.302585  5.017280
## 3          ESP 11  1.287854 2.397895  5.894403
## 4          ESP 10  1.321756 2.302585  3.496508
## 5          ESP  7  1.321756 1.945910  3.332205
## 6          ESP  8  1.348073 2.079442  5.480639
## 7          ESP 10  1.386294 2.302585  5.730100
## 8          ESP  5  1.398717 1.609438  3.663562
## 9          ESP  9  1.539015 2.197225  6.272877
## 10         ESP  6  1.681759 1.791759  3.555348
## 11         ESP  9  1.722767 2.197225  4.248495
## 12         ESP  8  1.783391 2.079442  3.850148
## 13         ESP  6  1.871802 1.791759  3.737670
## 14         ESP  9  1.871802 2.197225  3.401197
## 15         ESP  8  1.931521 2.079442  3.367296
## 16         ESP  7  1.931521 1.945910  3.332205
## 17         ESP 11  1.974081 2.397895  4.574711
## 18         ESP  6  2.157559 1.791759  3.891820
## 19         ESP  8  2.191654 2.079442  4.077537

summary_data <- data_best %>%
  group_by(Winery) %>%
  summarize(
    Avg_Rating = mean(Rating, na.rm = TRUE),
    Avg_Price = mean(Price, na.rm = TRUE)
  )

global_avg_price <- mean(data_best$Price, na.rm = TRUE)
global_avg_rating <- mean(data_best$Rating, na.rm = TRUE)

long_data <- summary_data %>%
  pivot_longer(
    cols = c(Avg_Rating, Avg_Price),
    names_to = "Metric",
    values_to = "Value"
  )

ggplot(long_data, aes(x = Winery, y = Value, color = Metric, group = Metric)) +
  geom_point(size = 3) +
  geom_hline(yintercept = global_avg_price, color = "red", linetype = "dashed",
             size = 1, show.legend = TRUE) +
  geom_hline(yintercept = global_avg_rating, color = "blue", linetype = "dashed",
             size = 1, show.legend = TRUE) +
  labs(

```

```

title = "Average Review and Price by Winery",
x = "Winery",
y = "Value",
color = "Metric",
caption = paste("Dashed blue line: Average Rating | Dashed red line:
                 Average Price | Region: ",best$Region)
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

```

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```



Kada promotrimo regije koje imaju više od 10 zapisa o vinima, prvo mjesto zauzima Španjolska regija Cariñena. Kada bismo birali vinariju u ovoj regiji odabrali bismo vinariju El Ciroo jer ima prosječnu ocijenu blisku prosječnoj ocijeni regije, dok je prosječna cijena niža naspram drugih vinarija.

Ipak, možemo uvidjeti da mali brojnik (ocijene su iz intervala od 0-5) znatno utječe na omjer, jer su bolja vina često i skuplja, skuplja vina znatno smanjuju omjer ocijene i cijene jer ocijena i cijena ne rastu proporcionalno. Stoga ćemo proučiti promjenu omjera ovisno o rastu prosječne ocijene regije. U filtraciji zatražit ćemo da ocijena regije iterativno raste sa 3.5 na 4.5 kako bismo pronašli kvalitetnije vino po pristojnoj cijeni.

Radimo iteracije u kojima mijenjamo vrijednost filtera filter(Avg\_Rating > x), pri čemu x poprima vrijednosti od 3.5 do 4.5 uz korak 0.1:

```

library(dplyr)

x <- 3.5
max_x <- 4.5
increment <- 0.1
best_list <- list()
iteration <- 1

while (x < max_x) {
  cat("\nIteration with Avg_Rating >", x, "\n")

  result <- dataset_vina %>%
    group_by(Region) %>%
    summarize(
      Total_Vines = n(),
      Avg_Rating = mean(Rating, na.rm = TRUE),
      Avg_Price = mean(Price, na.rm = TRUE),
      Avg_Rating_Price_Ratio = Avg_Rating / Avg_Price
    ) %>%
    filter(Total_Vines > 10) %>%
    filter(Avg_Rating > x) %>%
    arrange(desc(Avg_Rating_Price_Ratio))

  print(head(result, 10))

  best_list[[iteration]] <- head(result, 1)

  x <- x + increment
  iteration <- iteration + 1
}

## 
## Iteration with Avg_Rating > 3.5
## # A tibble: 10 x 5
##   Region     Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>        <int>     <dbl>     <dbl>             <dbl>
## 1 La Mancha       19      3.55      5.37            0.661
## 2 Cariñena        57      3.60      5.47            0.658
## 3 Campo de Borja    29      3.64      5.81            0.626
## 4 Almansa         13      3.75      5.99            0.626
## 5 Castilla          49      3.61      5.87            0.616
## 6 Navarra           81      3.56      5.91            0.603
## 7 Yecla              14      3.64      6.28            0.580
## 8 Utiel-Requena      25      3.68      6.41            0.574
## 9 Valencia            50      3.61      6.31            0.571
## 10 Somontano          65      3.64      6.48            0.562
## 
## Iteration with Avg_Rating > 3.6
## # A tibble: 10 x 5
##   Region     Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>        <int>     <dbl>     <dbl>             <dbl>
## 1 Campo de Borja      29      3.64      5.81            0.626
## 2 Almansa            13      3.75      5.99            0.626
## 3 Castilla            49      3.61      5.87            0.616

```

```

## 4 Yecla           14    3.64    6.28      0.580
## 5 Utiel-Requena 25    3.68    6.41      0.574
## 6 Valencia        50    3.61    6.31      0.571
## 7 Somontano       65    3.64    6.48      0.562
## 8 Rueda           169   3.74    6.78      0.552
## 9 Alicante         26    3.68    6.68      0.552
## 10 Jumilla        59    3.71    6.74      0.550
##
## Iteration with Avg_Rating > 3.7
## # A tibble: 10 x 5
##   Region          Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>            <int>     <dbl>      <dbl>                <dbl>
## 1 Almansa          13      3.75      5.99      0.626
## 2 Rueda            169     3.74      6.78      0.552
## 3 Jumilla          59      3.71      6.74      0.550
## 4 Ribeiro          13      3.72      7.15      0.520
## 5 Aragón           12      3.72      7.29      0.511
## 6 Rías Baixas     23      3.75      8.04      0.467
## 7 Vinho verde      67      3.79      10.9      0.348
## 8 Puglia           28      3.84      12        0.320
## 9 Beiras           13      3.72      11.8      0.315
## 10 Primitivo di Manduria 11      4.09      13.4      0.306
##
## Iteration with Avg_Rating > 3.8
## # A tibble: 10 x 5
##   Region          Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>            <int>     <dbl>      <dbl>                <dbl>
## 1 Puglia           28      3.84      12        0.320
## 2 Primitivo di Manduria 11      4.09      13.4      0.306
## 3 Península de Setúbal 38      3.84      13.0      0.294
## 4 Lisboa           59      3.85      13.1      0.293
## 5 Campania         12      3.84      13.3      0.289
## 6 Languedoc        14      3.87      13.4      0.289
## 7 Languedoc-Rosellón 14      3.89      14.2      0.274
## 8 Durbanville      31      3.92      14.8      0.265
## 9 Sicilia          39      3.81      14.6      0.262
## 10 Côtes du Roussillon 16      3.89      15.9      0.244
##
## Iteration with Avg_Rating > 3.9
## # A tibble: 10 x 5
##   Region          Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>            <int>     <dbl>      <dbl>                <dbl>
## 1 Primitivo di Manduria 11      4.09      13.4      0.306
## 2 Durbanville       31      3.92      14.8      0.265
## 3 Overberg          20      3.96      17.3      0.228
## 4 Beira Interior    19      3.94      17.7      0.223
## 5 Haut-Médoc        12      3.93      18.3      0.214
## 6 Monção e Melgaço  37      4.07      20.1      0.202
## 7 Gavi              14      3.93      20.1      0.196
## 8 Collio            11      3.92      21.3      0.184
## 9 Côtes de Provence 30      4.01      22.0      0.183
## 10 Duriense         13      4.18      23.4      0.179
##
## Iteration with Avg_Rating > 4

```

```

## # A tibble: 10 x 5
##   Region           Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>              <int>      <dbl>       <dbl>                  <dbl>
## 1 Primitivo di Manduria     11       4.09      13.4                 0.306
## 2 Monção e Melgaço         37       4.07      20.1                 0.202
## 3 Côtes de Provence        30       4.01      22.0                 0.183
## 4 Duriense                  13       4.18      23.4                 0.179
## 5 Tulbagh                   18       4.14      26.5                 0.157
## 6 Douro                      467      4.00      29.3                 0.136
## 7 Sancerre                   30       4.04      30.1                 0.134
## 8 Brauneberg                  16       4.08      30.9                 0.132
## 9 Mosel                      217      4.02      33.0                 0.122
## 10 Nahe                     83        4.02      34.0                 0.118
##
## Iteration with Avg_Rating > 4.1
## # A tibble: 10 x 5
##   Region           Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>              <int>      <dbl>       <dbl>                  <dbl>
## 1 Duriense                  13       4.18      23.4                 0.179
## 2 Tulbagh                   18       4.14      26.5                 0.157
## 3 Graach                     14       4.16      36.6                 0.113
## 4 Pouilly-Fuissé                20       4.20      37.7                 0.111
## 5 Wehlen                      11       4.23      38.2                 0.111
## 6 Valle de Cachapoal          11       4.17      49.0                 0.0852
## 7 Amarone della Valpol~        14       4.24      53.6                 0.0792
## 8 Pouilly-Fumé                  15       4.11      55.5                 0.0741
## 9 Valle de Aconcagua            30       4.12      57.0                 0.0724
## 10 Dundee Hills                  15       4.12      57.4                 0.0718
##
## Iteration with Avg_Rating > 4.2
## # A tibble: 10 x 5
##   Region           Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>              <int>      <dbl>       <dbl>                  <dbl>
## 1 Wehlen                     11       4.23      38.2                 0.111
## 2 Amarone della Valpol~        14       4.24      53.6                 0.0792
## 3 Gualtallary                  11       4.2       59.8                 0.0703
## 4 Paso Robles                  54       4.27      72.2                 0.0592
## 5 Barolo                      122      4.21      93.5                 0.0450
## 6 Sauternes                   35       4.21      95.5                 0.0440
## 7 Russian River Valley          46       4.27      99.5                 0.0429
## 8 Brunello di Montalcini~       56       4.27      104.                  0.0411
## 9 Châteauneuf-du-Pape            68       4.26      107.                  0.0398
## 10 Los Carneros                  26       4.22      110.                  0.0384
##
## Iteration with Avg_Rating > 4.3
## # A tibble: 10 x 5
##   Region           Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>              <int>      <dbl>       <dbl>                  <dbl>
## 1 Amarone della Valpol~         15       4.31      114.                 0.0378
## 2 Champán                      90       4.32      145.                 0.0298
## 3 Rutherford                   21       4.40      152.                 0.0290
## 4 Sonoma Coast                   52       4.33      161.                 0.0269
## 5 Hermitage                      17       4.32      173.                 0.0250
## 6 Valle de Napa                  182      4.35      211.                 0.0207

```

```

## 7 Pomerol          32    4.31    215.      0.0201
## 8 Margaux          19    4.33    217.      0.0199
## 9 Sonoma Mountain   13    4.42    262.      0.0169
## 10 Pauillac         29    4.33    357.      0.0121
##
## Iteration with Avg_Rating > 4.4
## # A tibble: 3 x 5
##   Region      Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>        <int>     <dbl>      <dbl>                <dbl>
## 1 Rutherford      21      4.40      152.      0.0290
## 2 Sonoma Mountain  13      4.42      262.      0.0169
## 3 Oakville        15      4.55      772.      0.00589
##
## Iteration with Avg_Rating > 4.5
## # A tibble: 1 x 5
##   Region      Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>        <int>     <dbl>      <dbl>                <dbl>
## 1 Oakville       15      4.55      772.      0.00589
best_combined <- bind_rows(best_list)

print(best_combined)

## # A tibble: 11 x 5
##   Region      Total_Vines Avg_Rating Avg_Price Avg_Rating_Price_Ratio
##   <chr>        <int>     <dbl>      <dbl>                <dbl>
## 1 La Mancha      19      3.55      5.37      0.661
## 2 Campo de Borja 29      3.64      5.81      0.626
## 3 Almansa        13      3.75      5.99      0.626
## 4 Puglia          28      3.84      12        0.320
## 5 Primitivo di Manduria 11      4.09      13.4      0.306
## 6 Primitivo di Manduria 11      4.09      13.4      0.306
## 7 Duriense        13      4.18      23.4      0.179
## 8 Wehlen          11      4.23      38.2      0.111
## 9 Amarone della Valpol~ 15      4.31      114.      0.0378
## 10 Rutherford     21      4.40      152.      0.0290
## 11 Oakville        15      4.55      772.      0.00589

```

Grafički prikaz vinarija koje imaju prosječnu ocjenu regije između 3.5-4.5:

```

library(dplyr)
library(ggplot2)
library(tidyr)

x_br <- 3.5
max_x <- 4.5
increment <- 0.1
i <- 1

while (x_br < max_x) {
  best_combined_row <- best_combined[i, , drop = FALSE]

  data_best <- dataset_vina %>%
    filter(Region %in% best_combined_row$Region)

  summary_data <- data_best %>%

```

```

group_by(Winery) %>%
summarize(
  Avg_Rating = mean(Rating, na.rm = TRUE),
  Avg_Price = mean(Price, na.rm = TRUE)
)

global_avg_price <- mean(data_best$Price, na.rm = TRUE)
global_avg_rating <- mean(data_best$Rating, na.rm = TRUE)

long_data <- summary_data %>%
pivot_longer(
  cols = c(Avg_Rating, Avg_Price),
  names_to = "Metric",
  values_to = "Value"
)

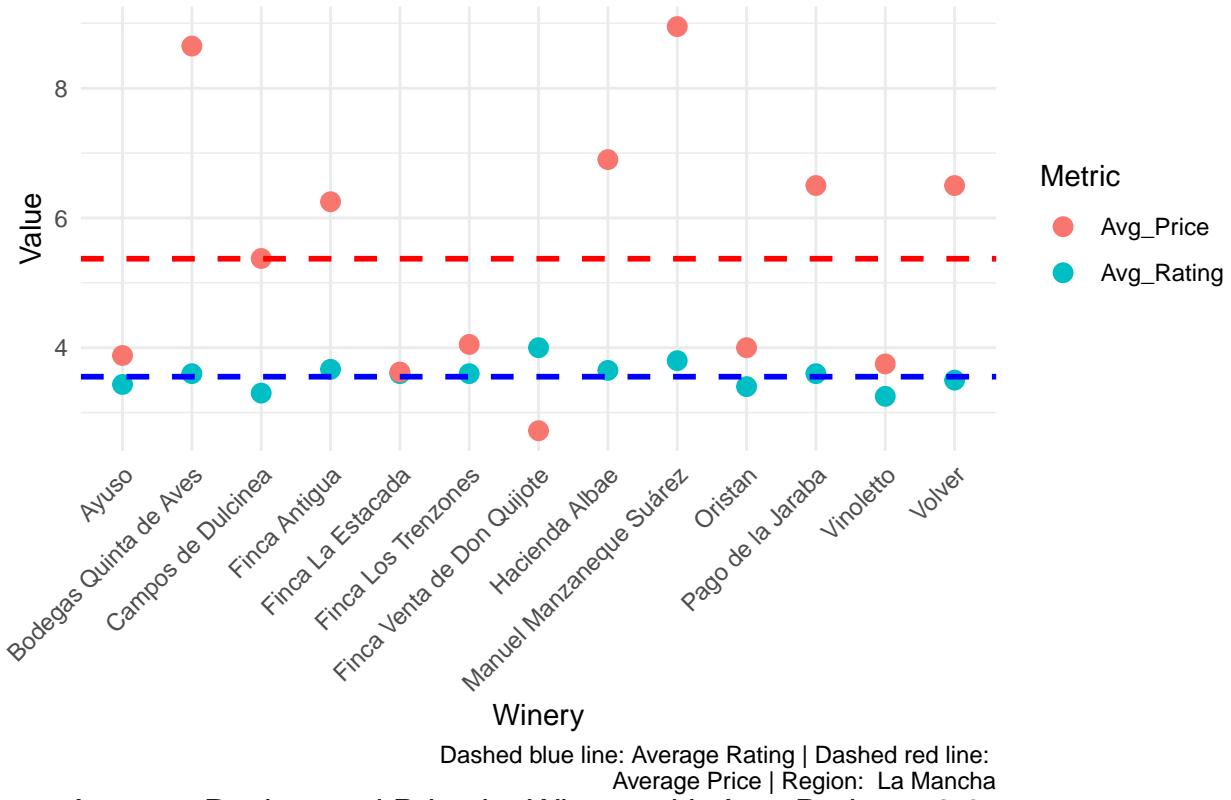
plot <- ggplot(long_data, aes(x = Winery, y = Value, color = Metric, group = Metric)) +
  geom_point(size = 3) +
  geom_hline(yintercept = global_avg_price, color = "red", linetype = "dashed",
             size = 1) +
  geom_hline(yintercept = global_avg_rating, color = "blue", linetype = "dashed",
             size = 1) +
  labs(
    title = paste("Average Review and Price by Winery with Avg_Rating >", x_br),
    x = "Winery",
    y = "Value",
    color = "Metric",
    caption = paste("Dashed blue line: Average Rating | Dashed red line:
                    Average Price | Region: ",best_combined_row)
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(plot)

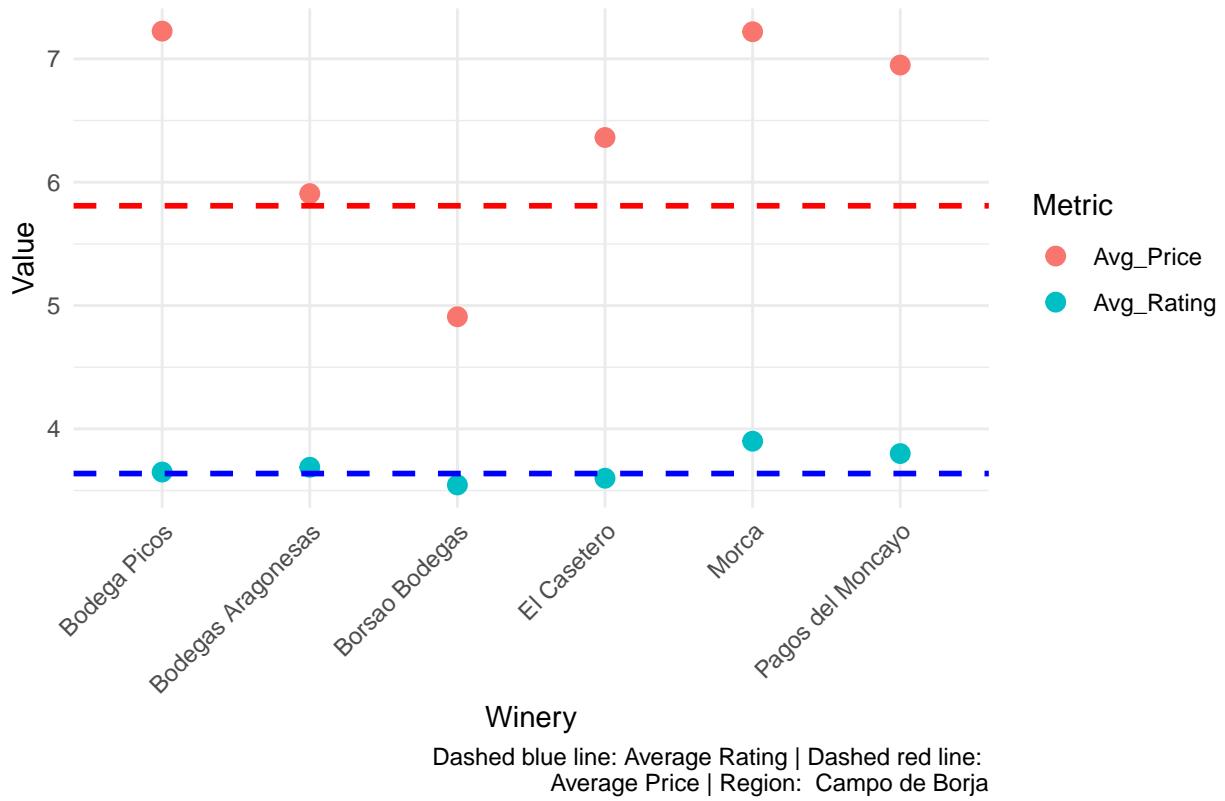
x_br <- x_br + increment
i <- i + 1
}

```

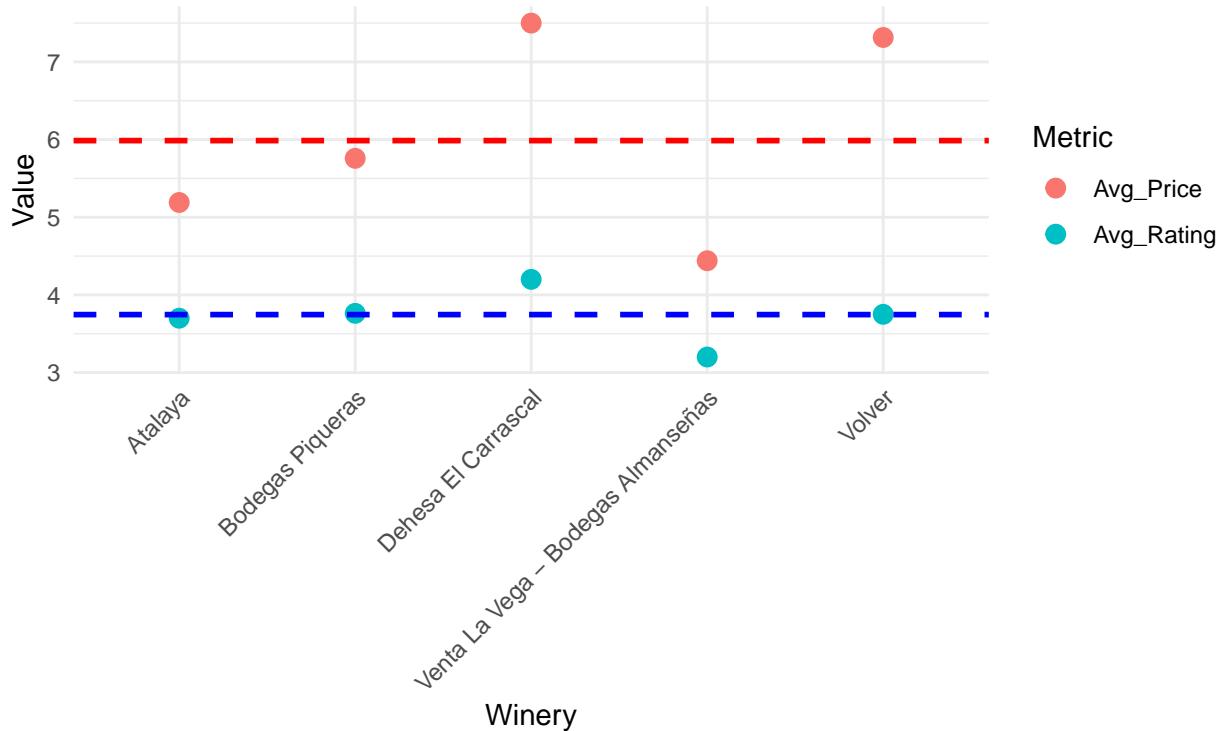
### Average Review and Price by Winery with Avg\_Rating > 3.5



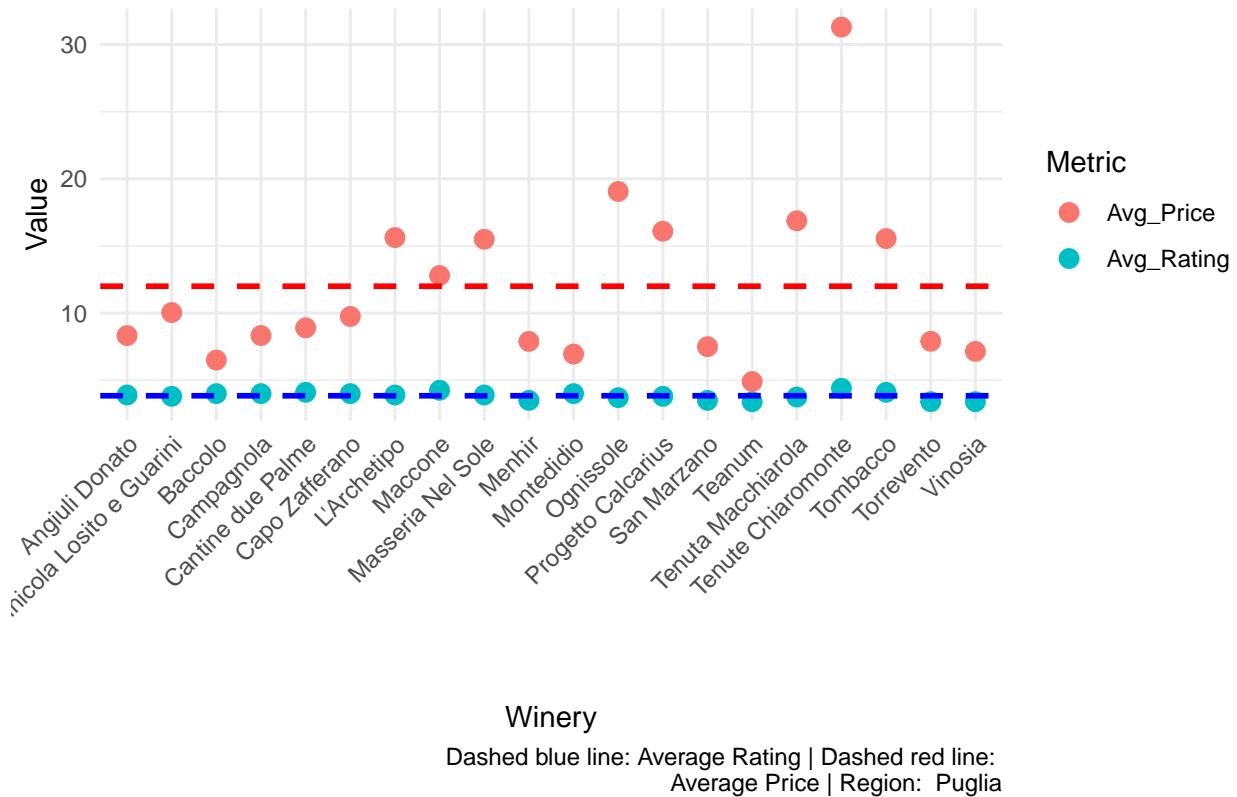
### Average Review and Price by Winery with Avg\_Rating > 3.6



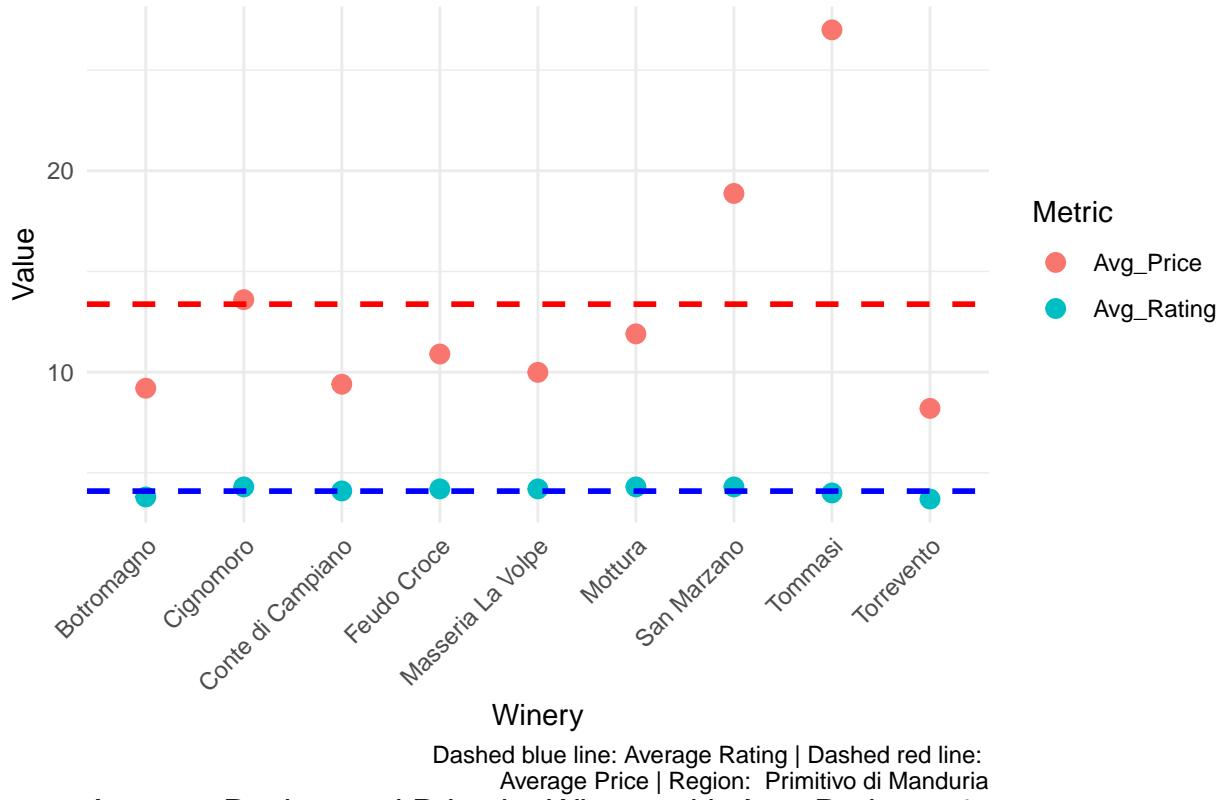
### Average Review and Price by Winery with Avg\_Rating > 3.7



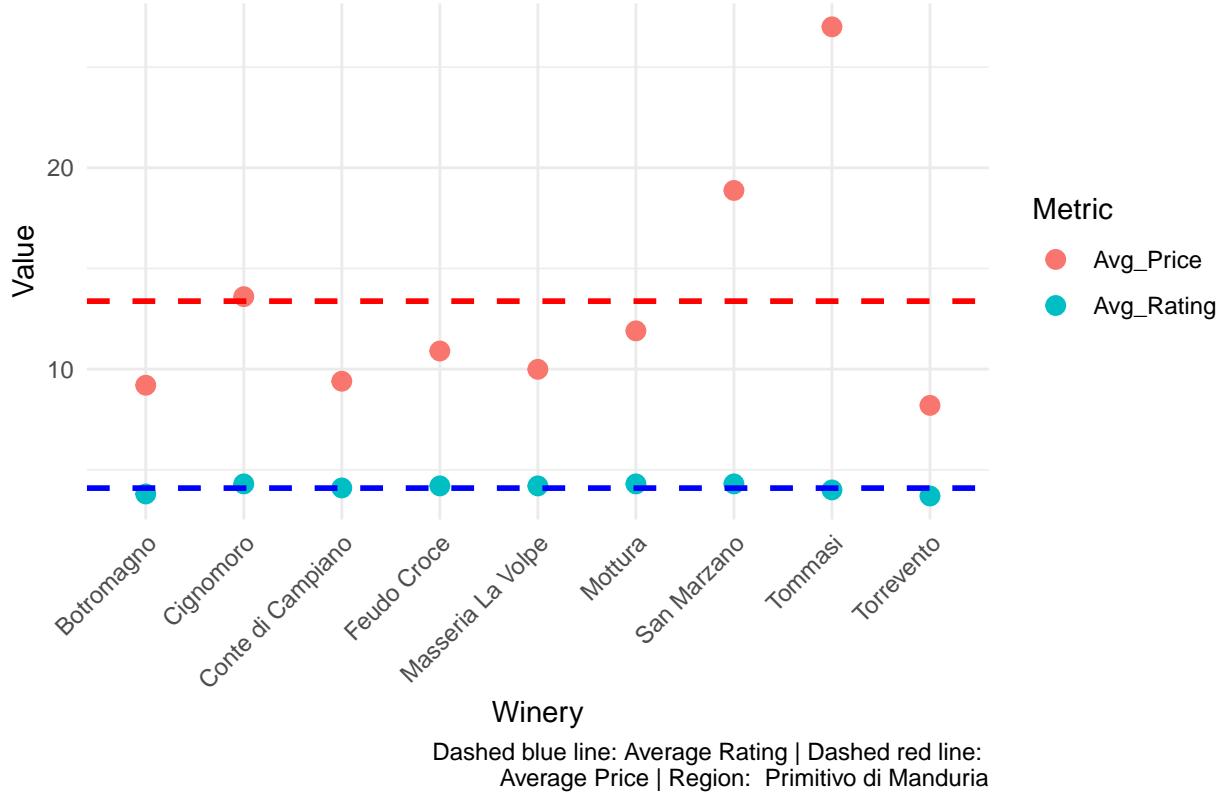
### Average Review and Price by Winery with Avg\_Rating > 3.8



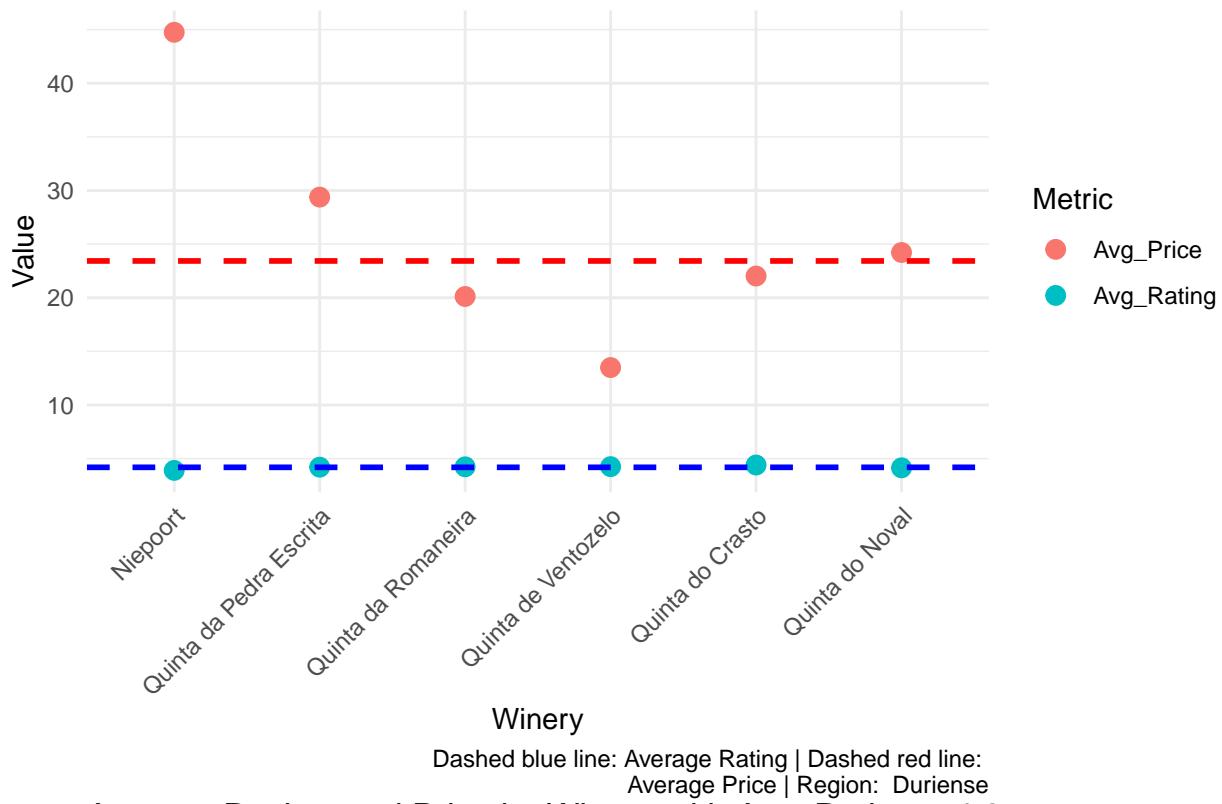
## Average Review and Price by Winery with Avg\_Rating > 3.9



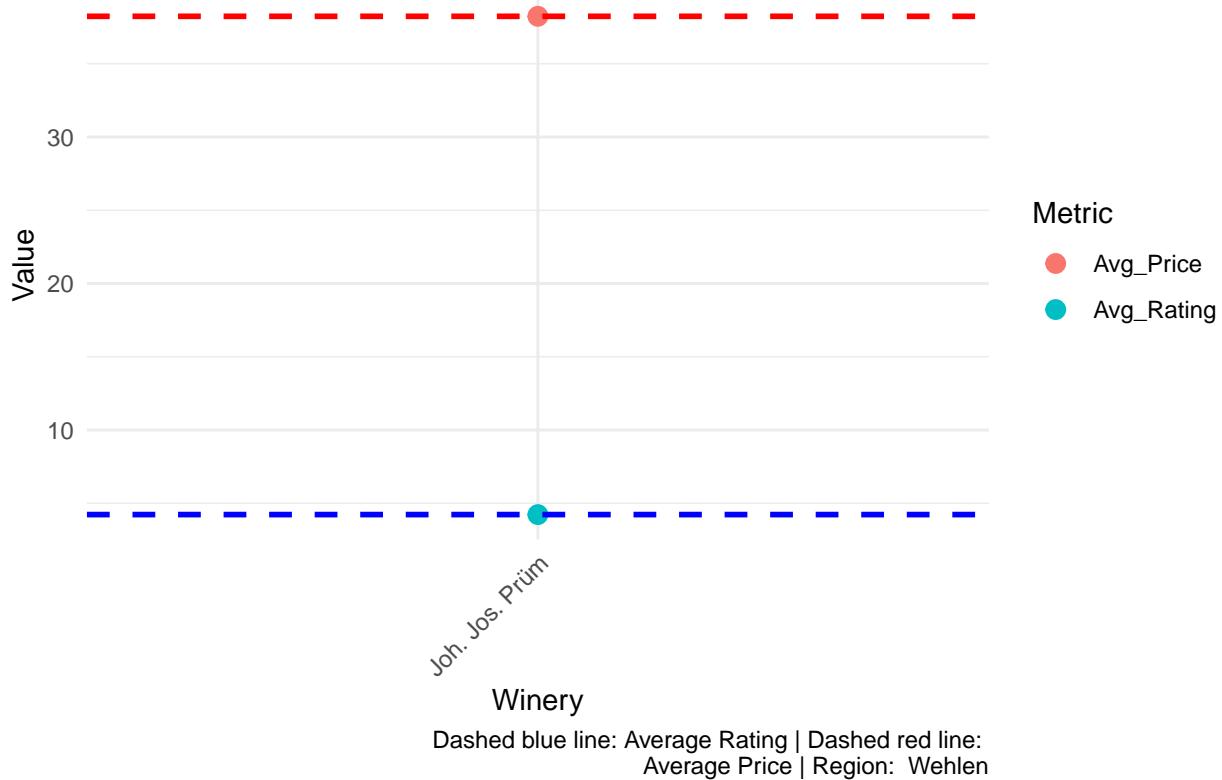
## Average Review and Price by Winery with Avg\_Rating > 4



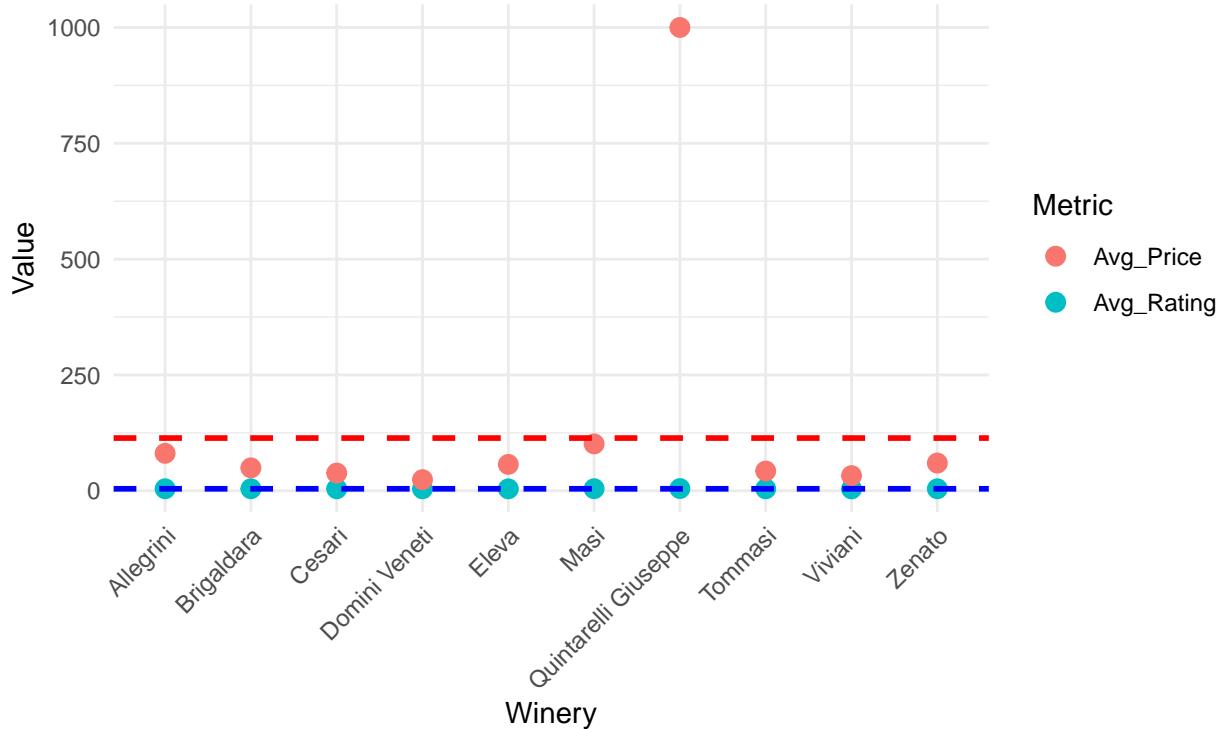
### Average Review and Price by Winery with Avg\_Rating > 4.1



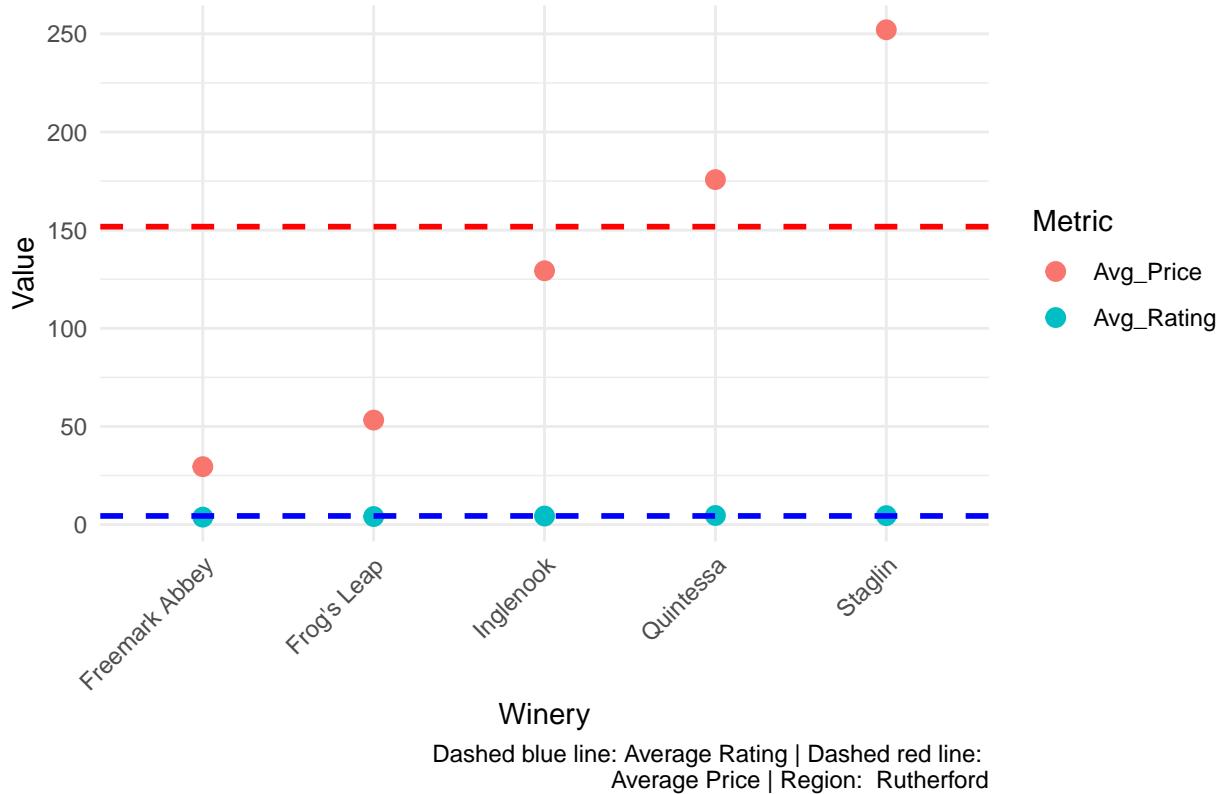
### Average Review and Price by Winery with Avg\_Rating > 4.2



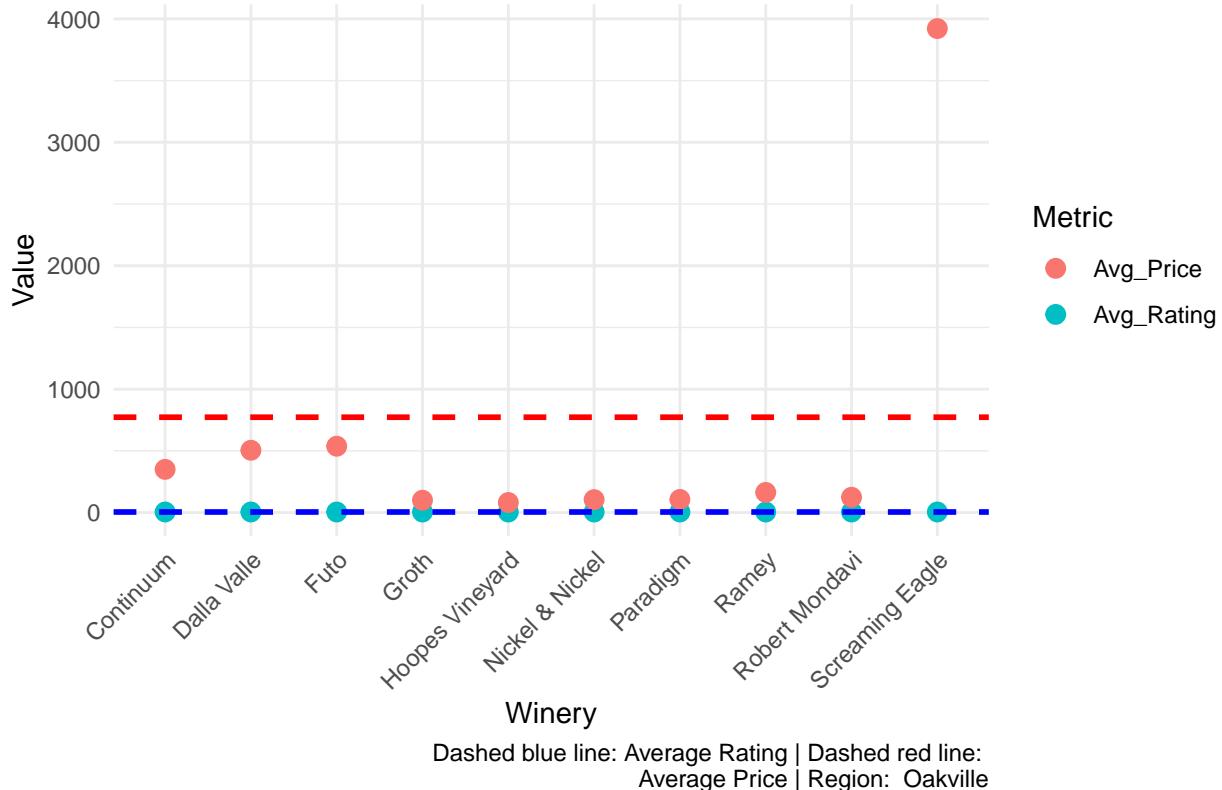
### Average Review and Price by Winery with Avg\_Rating > 4.3



### Average Review and Price by Winery with Avg\_Rating > 4.4



## Average Review and Price by Winery with Avg\_Rating > 4.5



U sljedećem prikazu podataka, uviđamo kako je najbolji omjer ocijene i cijene ponudila regija Primitivo di Manduria, sa prosječnom ocijenom 4.090909, te prosječnom cijenom 13.376364. Iako je sam omjer relativno mali (0.305831181), to je za očekivati budući da je nazivnik (prosječna ocijena) vrlo mali broj, a porast prosječne ocijene ne prati linearno porast cijene.

```
Primitivo_di_Manduria_data <- dataset_vina %>%
  filter(Region == "Primitivo di Manduria")

print(Primitivo_di_Manduria_data)

## #> #>   Winery Year Wine_ID
## #> 1 Conte di Campiano 2020 1228622
## #> 2 Masseria La Volpe 2020 3390935
## #> 3 Feudo Croce 2019 1425998
## #> 4 Mottura 2019 4572535
## #> 5 San Marzano 2020 2272631
## #> 6 San Marzano 2017 11890
## #> 7 Cignomoro 2017 1438695
## #> 8 Torrevento 2009 1148207
## #> 9 Botromagno 2014 1155543
## #> 10 Botromagno 2019 1155543
## #> 11 Tommasi 2013 3484137
## #>
## #>   Wine Rating Reviews
## #> 1 Primitivo di Manduria 2020 4.1 25
## #> 2 Uno Primitivo di Manduria 2020 4.2 259
## #> 3 Byzantium Primitivo di Manduria 2019 4.2 386
## #> 4 Stilio Primitivo di Manduria 2019 4.3 1913
## #> 5 Talò Primitivo di Manduria 2020 4.1 126
```

```

## 6      60 Sessantanni Old Vines Primitivo di Manduria 2017    4.5    7442
## 7                      Primitivo di Manduria 2017    4.3     223
## 8      Ghenos Primitivo di Manduria 2009    3.7     389
## 9                      Primitivo 2014    3.8     119
## 10                     Primitivo 2019    3.8      47
## 11 Masseria Surani Dionysos Primitivo di Manduria Riserva 2013    4.0     457
##   Price          Region Primary_Grape Natural Country Style Country_Code
## 1  9.40 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 2  9.99 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 3 10.90 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 4 11.90 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 5 12.40 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 6 25.35 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 7 13.60 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 8  8.20 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 9  9.00 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 10 9.40 Primitivo di Manduria Sangiovese False Italy Red    ITA
## 11 27.00 Primitivo di Manduria Sangiovese False Italy Red    ITA
##   Age Log_Price Log_Age Log_Reviews
## 1   5  2.240710 1.609438    3.218876
## 2   5  2.301585 1.609438    5.556828
## 3   6  2.388763 1.791759    5.955837
## 4   6  2.476538 1.791759    7.556428
## 5   5  2.517696 1.609438    4.836282
## 6   8  3.232779 2.079442    8.914895
## 7   8  2.610070 2.079442    5.407172
## 8  16  2.104134 2.772589    5.963579
## 9  11  2.197225 2.397895    4.779123
## 10  6  2.240710 1.791759    3.850148
## 11 12  3.295837 2.484907    6.124683

```

Kada bismo morali birati vinariju iz ove regije odabrali bismo San Marzano jer njihovo vino "60 Sessantanni Old Vines Primitivo di Manduria 2017" ima 7442 osvrta te visoku ocjenu od 4.5. Stoga možemo zaključiti ukoliko vinarija ima toliki publicitet i visoke ocijene, da su im i proizvodi kvalitetni i vrijedni degustacije.

Za osobe koje su spremne izdvojiti više, regije Duriense i Wehlen također mogu pružiti dobar omjer prosječne ocijene i cijene, dok svim regijama sa prosječnom ocijenom većom od 4.3, cijena izrazito naraste.