Final_report_Fialko_Shvets

Testing coffee quality

Final report

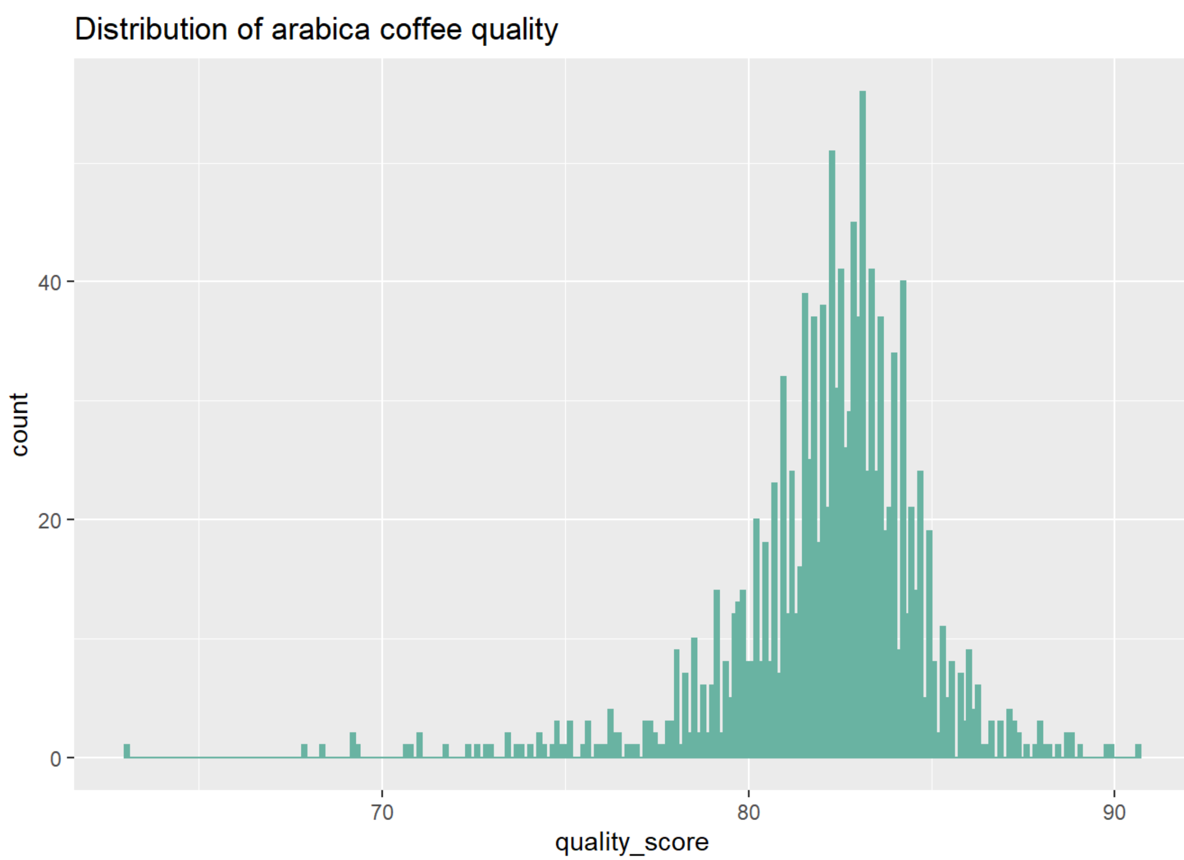
Team members: Yaryna Fialko, Shvets Anastasiia

Libraries

Let’s include libraries we need.

Reading data

```
coffee <- head(read.csv("arabica_ratings_raw.csv", header = TRUE), -3) # -1 because the last one sample is just 00 and it spoils data

ggplot(coffee, aes(x=quality_score)) +
  geom_histogram( binwidth=0.121, fill="#69b3a2", color="#69b3a2") +
  ggtitle("Distribution of arabica coffee quality") +
  theme(plot.title = element_text(size=))
```



Cleaning data

Let’s take only those data we will examine.

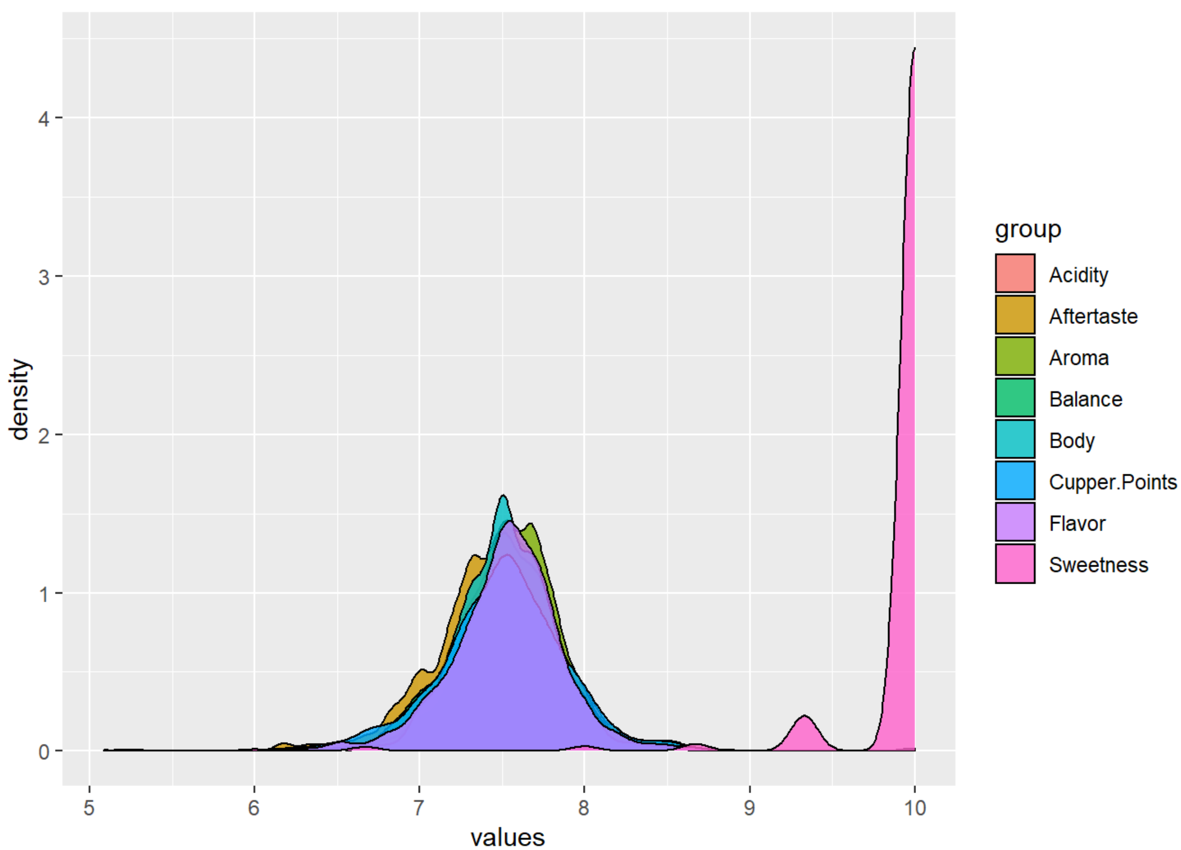
```
#select only necessary columns
coffee.to.analyze <- subset(coffee, select=c(quality_score, Aroma, Flavor, Aftertaste, Acidity, Sweetness,
Moisture, Body, Balance, Cupper.Points, Number.of.Bags))

coffee.to.analyze$Moisture = as.double(substr(coffee.to.analyze$Moisture, 1, nchar(coffee.to.analyze$Moistu
re)-2))
#colnames(coffee.to.analyze)[7] <- "Moisture_%"
```

```
summary(coffee.to.analyze)
```

```
##  quality_score       Aroma          Flavor        Aftertaste       Acidity
##  Min.   :63.08   Min.   :5.08   Min.   :6.080   Min.   :6.170   Min.   :5.250
##  1st Qu.:81.17   1st Qu.:7.42   1st Qu.:7.330   1st Qu.:7.250   1st Qu.:7.330
##  Median :82.50   Median :7.58   Median :7.580   Median :7.420   Median :7.500
##  Mean   :82.20   Mean   :7.57   Mean   :7.524   Mean   :7.404   Mean   :7.539
##  3rd Qu.:83.67   3rd Qu.:7.75   3rd Qu.:7.750   3rd Qu.:7.580   3rd Qu.:7.750
##  Max.   :90.58   Max.   :8.75   Max.   :8.830   Max.   :8.670   Max.   :8.750
##    Sweetness         Moisture           Body          Balance
##  Min.   : 6.000   Min.   : 0.000   Min.   :5.250   Min.   :6.080
##  1st Qu.:10.000   1st Qu.: 9.000   1st Qu.:7.330   1st Qu.:7.330
##  Median :10.000   Median :11.000   Median :7.500   Median :7.500
##  Mean   : 9.917   Mean   : 8.883   Mean   :7.524   Mean   :7.524
##  3rd Qu.:10.000   3rd Qu.:12.000   3rd Qu.:7.670   3rd Qu.:7.750
##  Max.   :10.000   Max.   :28.000   Max.   :8.580   Max.   :8.750
##  Cupper.Points    Number.of.Bags
##  Min.   : 5.170   Min.   :   0.0
##  1st Qu.: 7.250   1st Qu.:  14.0
##  Median : 7.500   Median : 170.0
##  Mean   : 7.504   Mean   : 153.7
##  3rd Qu.: 7.750   3rd Qu.: 275.0
##  Max.   :10.000   Max.   :1062.0
```

Plotting different variables

```
param <- c("Aroma", "Flavor", "Aftertaste", "Acidity", "Sweetness", "Body", "Balance", "Cupper.Points")

dat <- data.frame(matrix(nrow = 0, ncol = length(c("value", "group"))))

for (value in param)

{

  df1 <- subset(coffee, select=c(value))

  colnames(df1)[1] ="values"

  df2 <- data.frame(group = value)

  updated <- cbind(df1, df2)

  dat <- rbind(dat, updated)

}

ggplot(dat, aes(x = values, fill = group)) + geom_density(alpha = 0.8 )
```



We can see that all farctos except **Sweetness** are spread in the same interval with really close density.

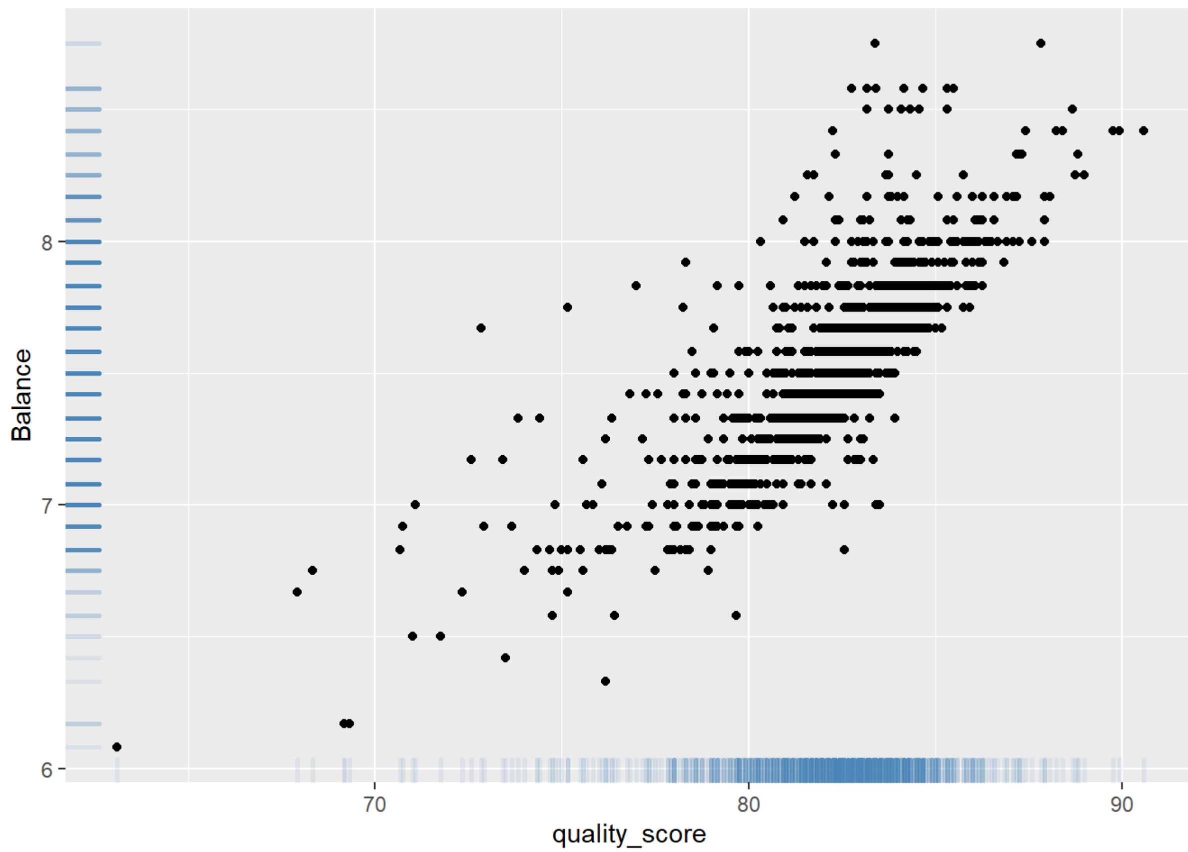
Let's look on correlation between **quality** of coffee and factors that affect it.

```
correlation <- cor(coffee.to.analyze)
print(correlation[,"quality_score"])
```

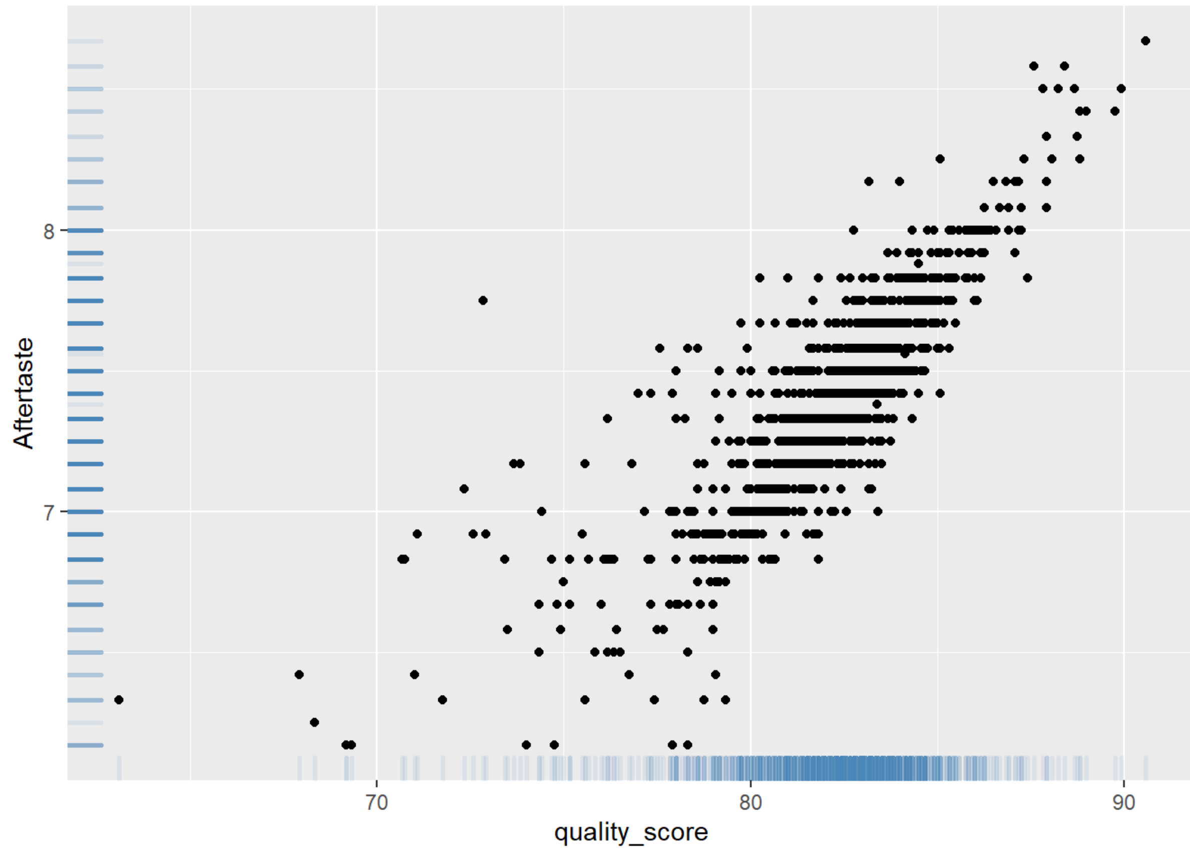
```
##   quality_score          Aroma          Flavor      Aftertaste         Acidity
##      1.00000000     0.71014842      0.84010990      0.82976857      0.72333697
##       Sweetness       Moisture            Body         Balance  Cupper.Points
##      0.37903044    -0.15130760      0.66070525      0.77856263      0.76632225
## Number.of.Bags
##      0.04352998
```

As we can see, **Balance, Aftertaste, Flavor** are the most correlated ones. Let's look on its graphics:

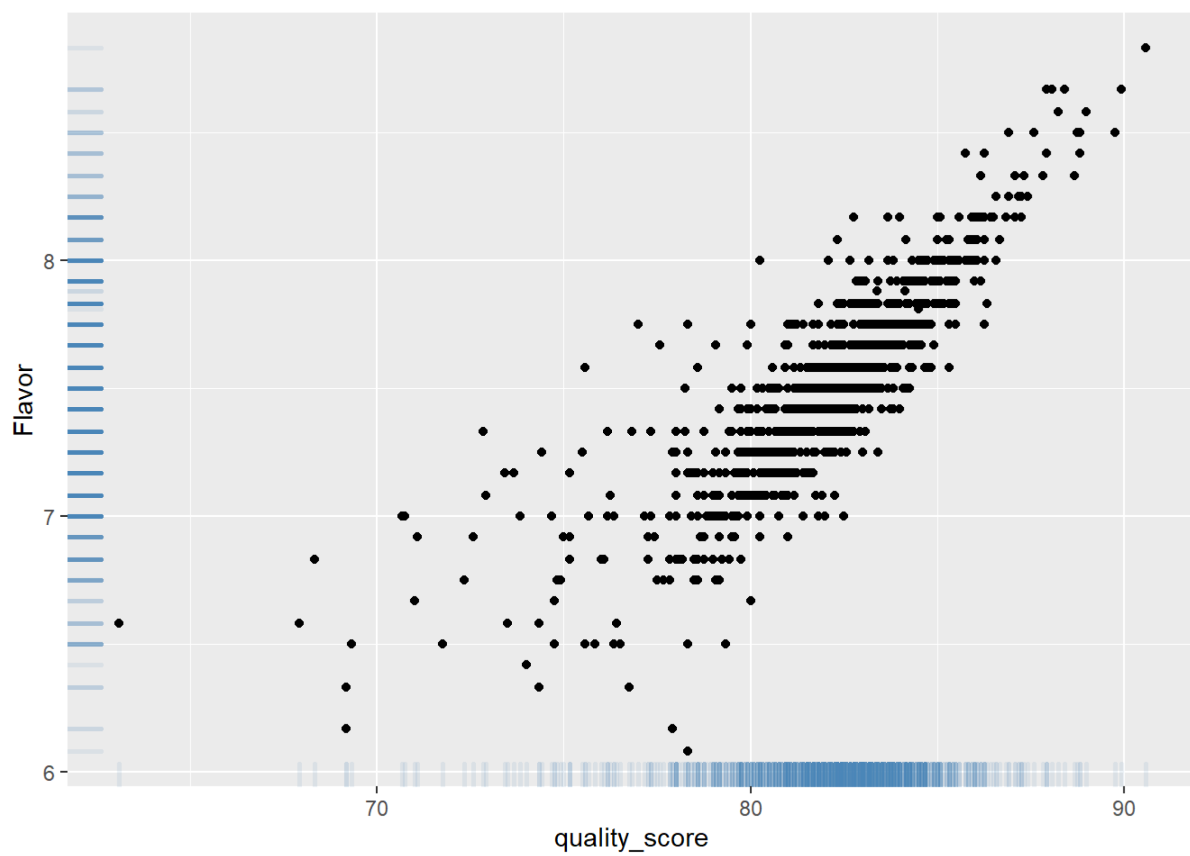
```
ggplot(data=coffee.to.analyze, aes(x=quality_score, Balance)) +
  geom_point() +
  geom_rug(col="steelblue",alpha=0.1, size=1)
```




```
ggplot(data=coffee.to.analyze, aes(x=quality_score, Aftertaste)) +
  geom_point() +
  geom_rug(col="steelblue",alpha=0.1, size=1)
```




```
ggplot(data=coffee.to.analyze, aes(x=quality_score, Flavor)) +
  geom_point() +
  geom_rug(col="steelblue",alpha=0.1, size=1)
```



Let’s look if we can say that they are **linearly dependent**.

```
y <- coffee.to.analyze$quality_score

x_aftertaste <- coffee.to.analyze$Aftertaste
reg_aftertaste <- lm(y~x_aftertaste)
summary(reg_aftertaste)
```

```
##
## Call:
## lm(formula = y ~ x_aftertaste)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4468  -0.4067   0.2108   0.7568   3.8032
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.2178     0.8564   42.29   <2e-16 ***
## x_aftertaste   6.2100     0.1155   53.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 1307 degrees of freedom
## Multiple R-squared:  0.6885, Adjusted R-squared:  0.6883
## F-statistic:  2889 on 1 and 1307 DF,  p-value: < 2.2e-16
```

```
x_balance <- coffee.to.analyze$Balance
reg_balance <- lm(y~x_balance)
summary(reg_balance)
```

```
##
## Call:
## lm(formula = y ~ x_balance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6838  -0.3132   0.3711   0.8711   4.4364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.2587     0.9807   39.01   <2e-16 ***
## x_balance     5.8397     0.1302   44.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.642 on 1307 degrees of freedom
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.6059
## F-statistic:  2012 on 1 and 1307 DF,  p-value: < 2.2e-16
```

```
x_flavor <- coffee.to.analyze$Flavor
reg_flavor <- lm(y~x_flavor)
summary(reg_flavor)
```

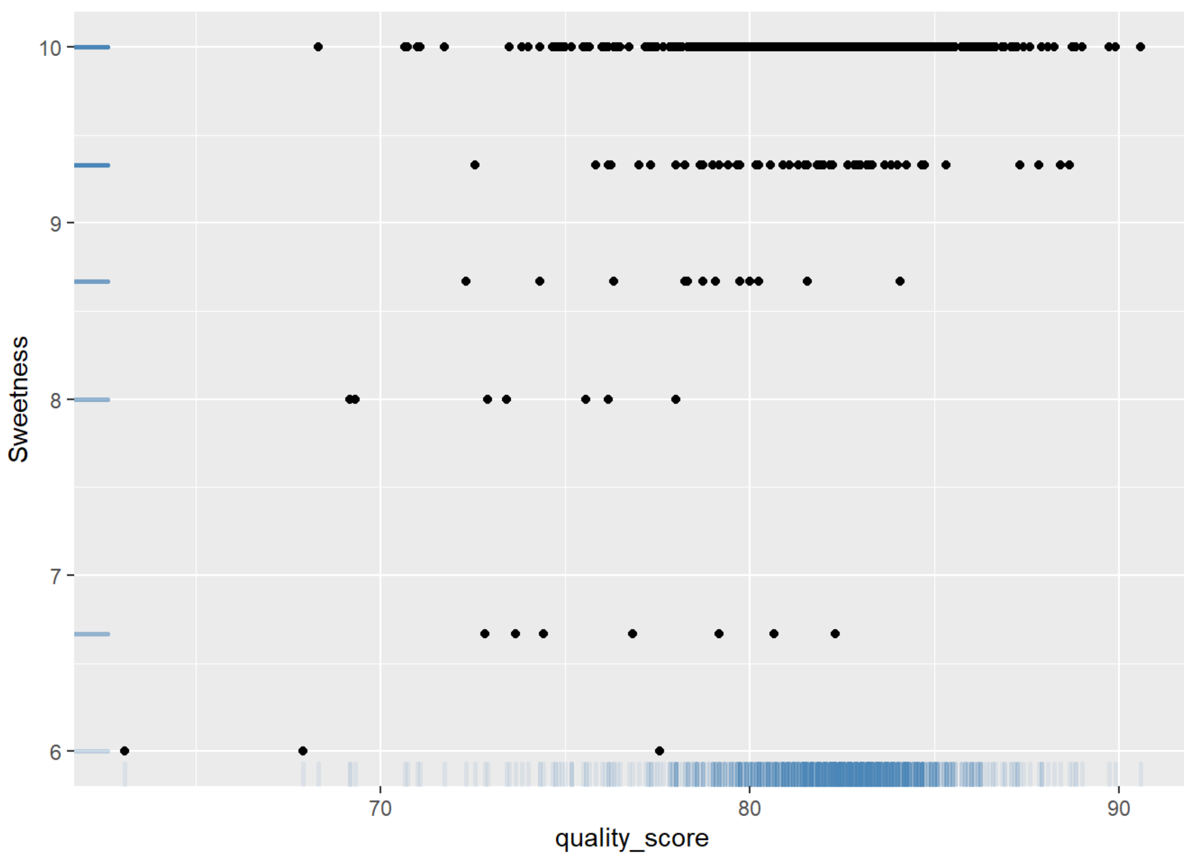
```
##
## Call:
## lm(formula = y ~ x_flavor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0348  -0.3682   0.1967   0.6973   5.4344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7496     0.8661   38.97   <2e-16 ***
## x_flavor      6.4385     0.1150   55.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.419 on 1307 degrees of freedom
## Multiple R-squared:  0.7058, Adjusted R-squared:  0.7056
## F-statistic:  3135 on 1 and 1307 DF,  p-value: < 2.2e-16
```

As we can see r^2 in **Balance** is about 0.6, so we can't really conclude that quality depends on this factors strongly linearly. But actually there is some sort of linear correlation.

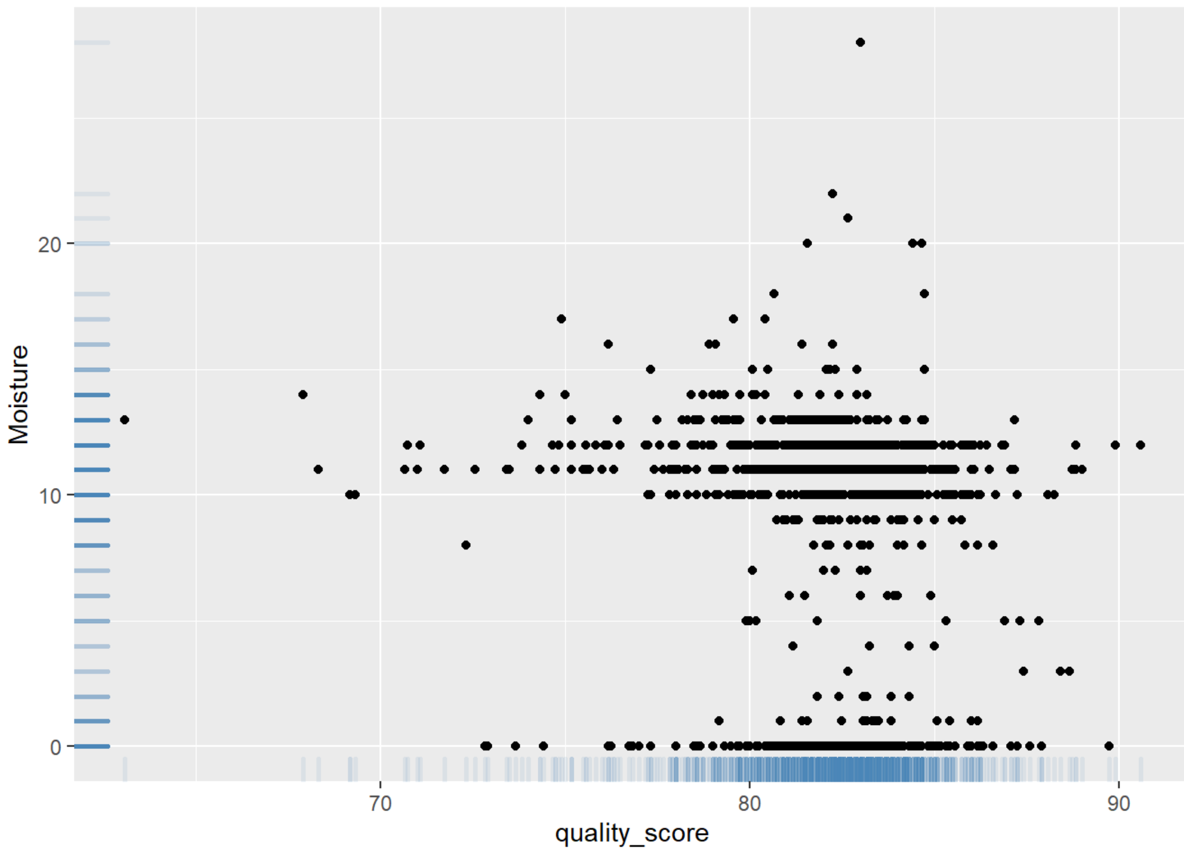
And **Aftertaste** and **Flavor** have r^2 about 0.7, so they correlate with quality more linearly.

Sweetness, Number.of.Bags and **Moisture** are the less correlated.

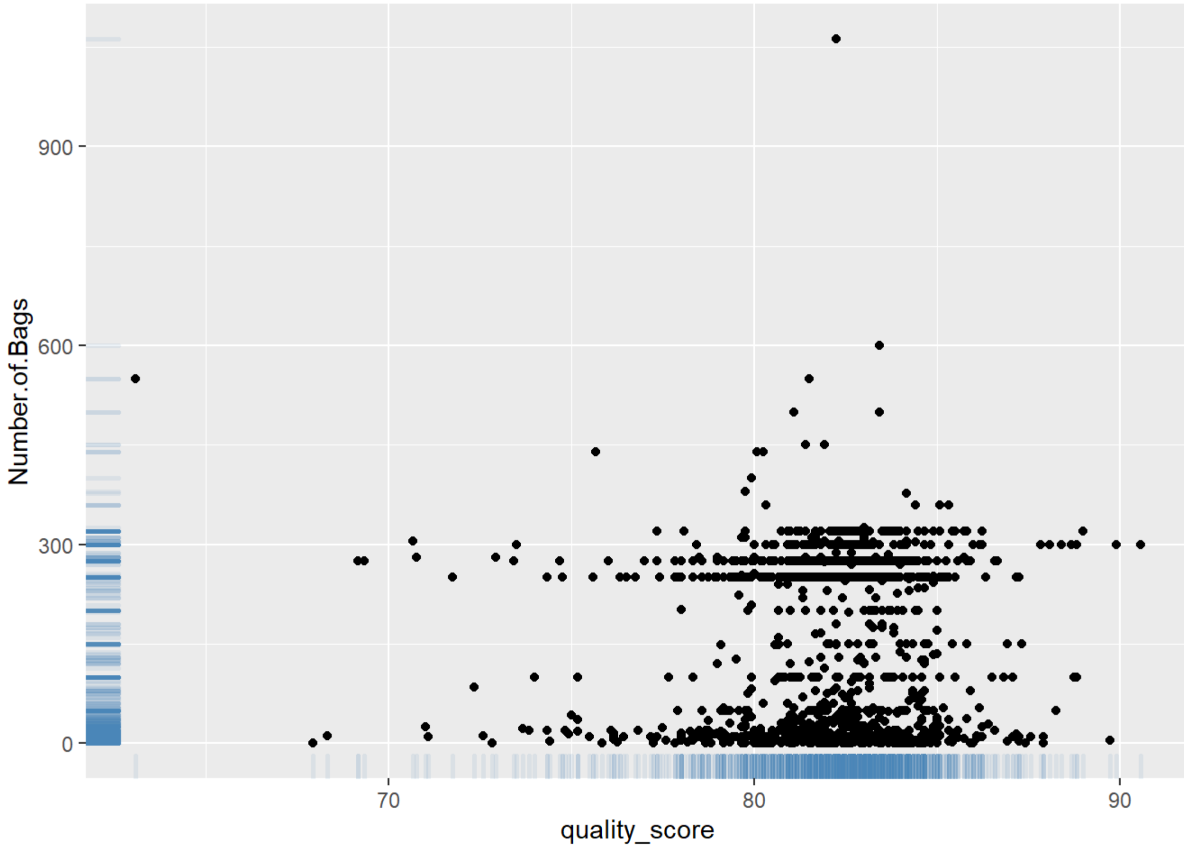
```
ggplot(data=coffee.to.analyze, aes(x=quality_score, Sweetness)) +
  geom_point() +
  geom_rug(col="steelblue",alpha=0.1, size=1)
```



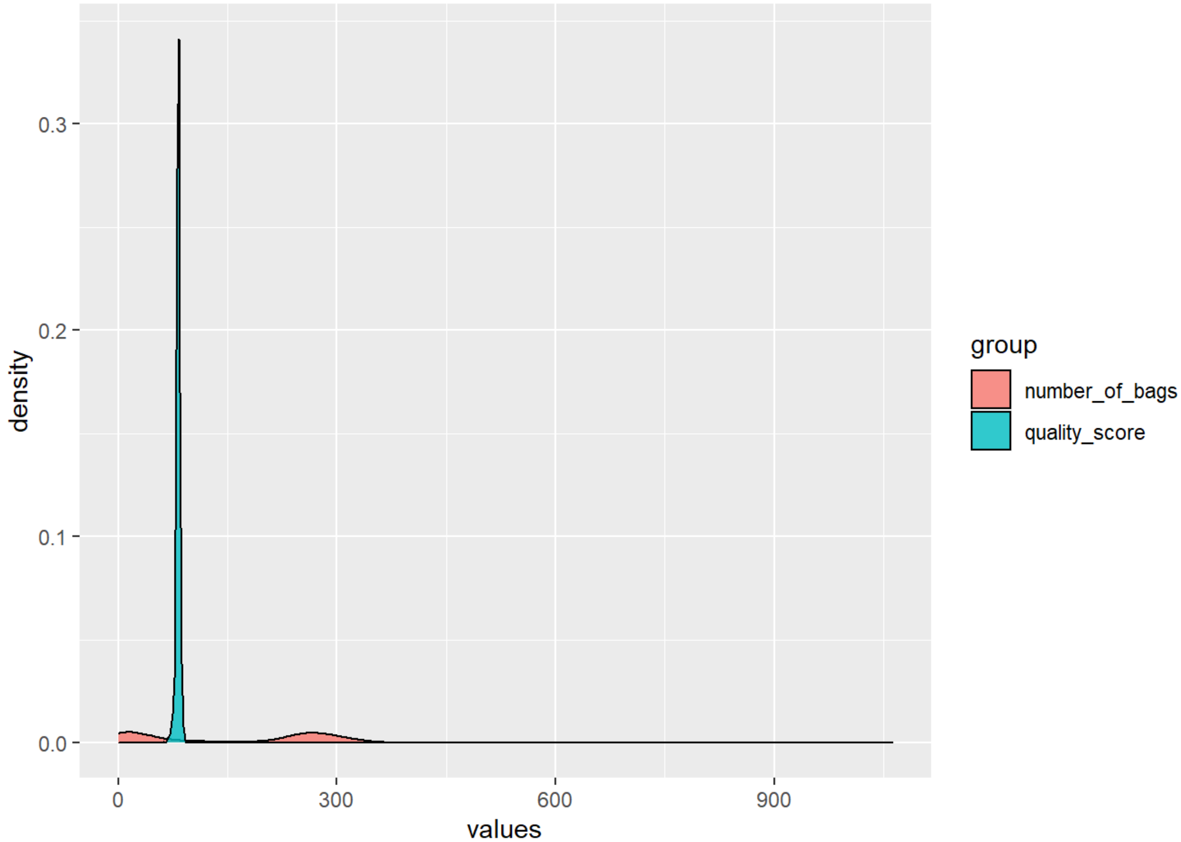
```
ggplot(data=coffee.to.analyze, aes(x=quality_score, Moisture)) +
  geom_point() +
  geom_rug(col="steelblue",alpha=0.1, size=1)
```




```
ggplot(data=coffee.to.analyze, aes(x=quality_score, Number.of.Bags)) +
  geom_point() +
  geom_rug(col="steelblue",alpha=0.1, size=1)
```




```r
#Sample data


dataframe1 <- subset(coffee, select=c(quality_score))
colnames(dataframe1)[1] ="values"
dataframe2 <- data.frame(group = "quality_score")


dataframe3 <- subset(coffee, select=c(Number.of.Bags))
colnames(dataframe3)[1] ="values"
dataframe4 <- data.frame(group = "number_of_bags")


updated1 <- cbind(dataframe1, dataframe2)
updated2 <- cbind(dataframe3, dataframe4)


dat <- rbind(updated1, updated2)

ggplot(dat, aes(x = values, fill = group)) + geom_density(alpha = 0.8 )
```



Hypothesis testing

H_0 : - **Number of bags** and **quality** are assigned independently.

H_1 : - There is a dependence between the **number of bags** and coffee **quality**.

```
quality_production <- subset(coffee.to.analyze, select=c(quality_score, Number.of.Bags))
chisq.test(quality_production)
```

```
##
##  Pearson's Chi-squared test
##
## data:  quality_production
## X-squared = 74981, df = 1308, p-value < 2.2e-16
```

Conclusion:

P-value is close to zero, so we can reject H_0. There is a dependence between the **number of bags** and coffee **quality**.

Hypothesis testing

H_0 : - **Sweetness** and **quality** are assigned independently.

H_1 : - There is a dependence between the **sweetness** and coffee **quality**.

```
quality_sweetness <- subset(coffee.to.analyze, select=c(quality_score, Sweetness))
chisq.test(quality_sweetness)
```

```
##
##  Pearson's Chi-squared test
##
## data:  quality_sweetness
## X-squared = 19.596, df = 1308, p-value = 1
```

Conclusion:

P-value is 1, so we cannot reject H_0. **Sweetness** and **quality** are assigned independently.

Hypothesis testing

H_0 : - **Moisture** and **quality** are assigned independently.

H_1 : - There is a dependence between the **moisture** and **coffee quality**.

```
quality_moisture <- subset(coffee.to.analyze, select=c(quality_score, Moisture))
chisq.test(quality_moisture)
```

```
##
##  Pearson's Chi-squared test
##
## data:  quality_moisture
## X-squared = 3306.7, df = 1308, p-value < 2.2e-16
```

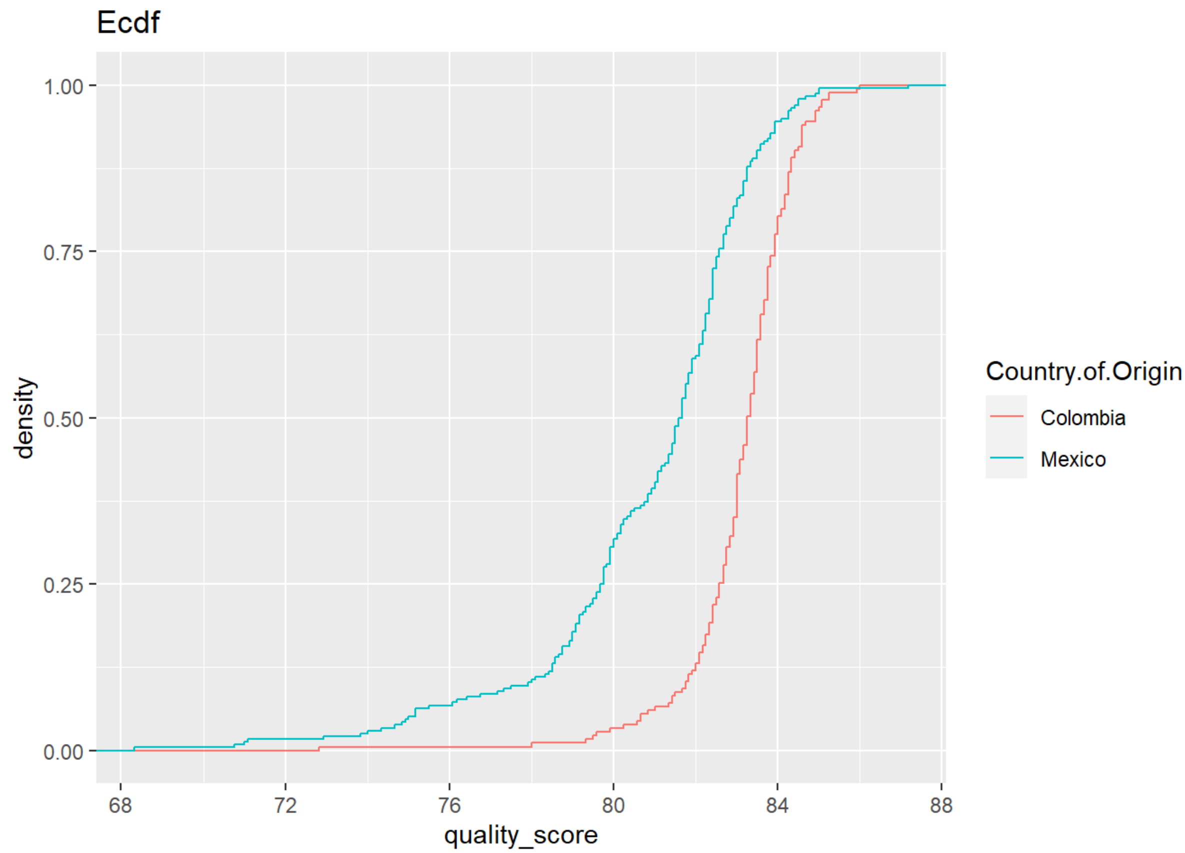
Conclusion:

P-value is almost zero, so we reject H_0. The **quality** of coffee and **moisture** are dependent.

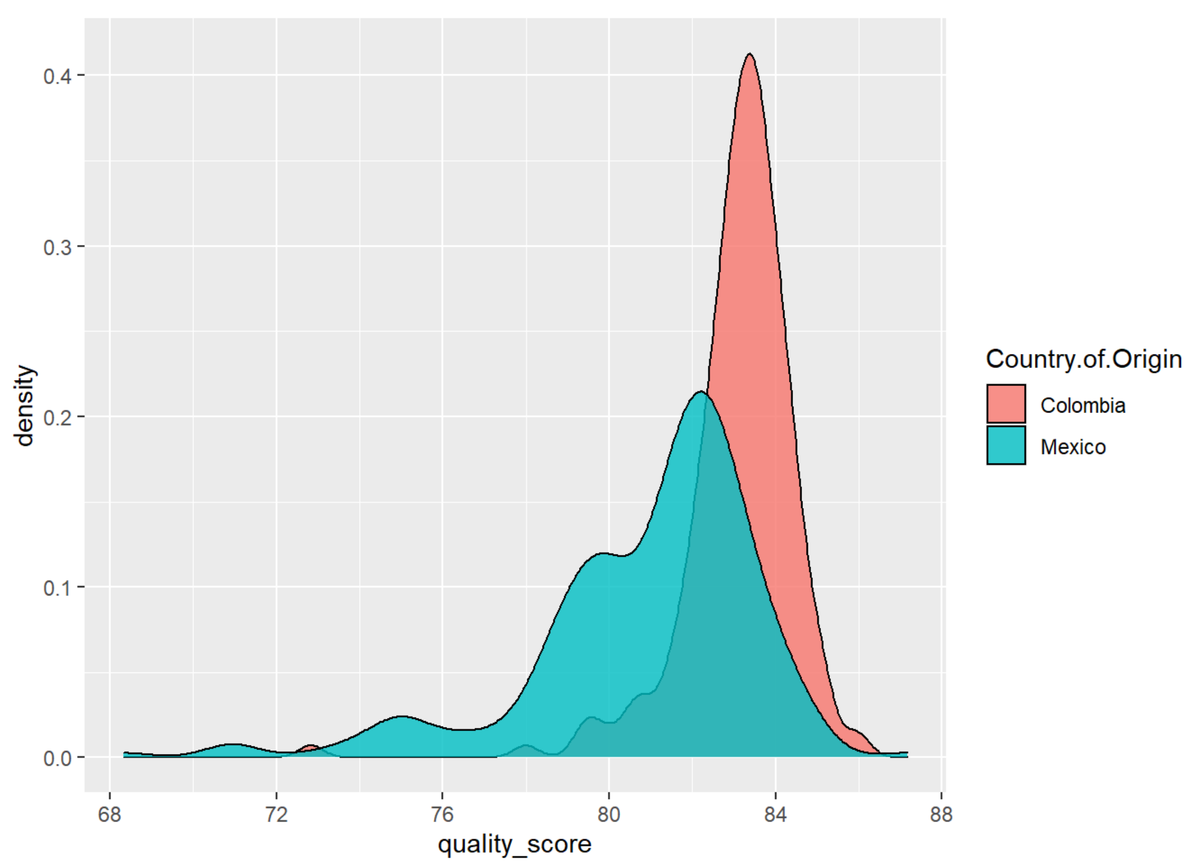
Quality of coffee and country of origin correlation

From the given data the biggest producers of coffee are **Mexico** and **Colombia**.

```
quality_country <- subset(coffee, select=c(quality_score, Country.of.Origin))

filtered <- subset(quality_country, Country.of.Origin %in% c('Colombia','Mexico'))

ggplot(filtered, aes(x=quality_score, col=Country.of.Origin)) +
  stat_ecdf() +
  ylab("density") +
  ggtitle("Ecdf")
```




```
ggplot(filtered, aes(x = quality_score, fill = Country.of.Origin)) + geom_density(alpha = 0.8 )
```



As we can see from density functions and ecdf, the quality of coffee produced in **Colombia** is almost in range $80 - 86$, and the most common value is about 84. At the same time, quality of coffee from **Mexico** is between 78 and 84, and the mode is 82.

Hypothesis testing

H_0 : - The quality of coffee produced in **Colombia** and **Mexico** is the same.

H_1 : - The quality of coffee produced in **Colombia** and **Mexico** differs.

Test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

```
Colombia <-  subset(quality_country, Country.of.Origin == 'Colombia')[1]
Mexico <-  subset(quality_country, Country.of.Origin == 'Mexico')[1]

t.test(Colombia, Mexico, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Colombia and Mexico
## t = 10.716, df = 367.83, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.809737 2.623208
## sample estimates:
## mean of x mean of y
##  83.10656  80.89008
```

P-value is almost zero, so we reject H_0. The quality of coffee produced in **Colombia** and **Mexico** differs.

Conclusion:

Overall, **Colombia** produce coffee with better quality comparing to **Mexico**.

General conclusion

Taking into account all above, we can say that **quality of coffee** mostly depends on its balance, aftertaste and flavor, and its correlation with aftertaste and flavor is really close to linear. Its moisture and sweetness affects quality of coffee the less. Though, we can’t reject that the moisture influence the quality, the sweetness and the quality of coffee are independent.

Also, among Colombia and Mexico - the biggest producers of coffee - Colombia produce coffee wuth higher quality.