CBR

Made by Maja
Updated 19/7/2021
Made in R/R Studio

Introduction

You can read the details of the story here - https://www.kaggle.com/rtatman/chocolate-bar-ratings

In this repository, I'll perform Exploratory Data Analysis (Descriptive statistics), and try to see if I have the prerequisites to do a regression analysis.

Exploratory Data Analysis (EDA) - numerical variables

Exploratory Data Analysis (EDA) or descriptive analysis consists of making summaries and explaining the results, in the matter of the shape and the variability of distribution. Also, I can do graphics like histogram, barplot and boxplots to see outliers and other things.

There is a lot more options for quantitative variables, and those here are Cocoa percentage and Rating, whereas all others are other types. We can check types by using the following formula:

```
str (dfrm) # code to see the structure of the data frame

## 'data.frame':   886 obs. of  9 variables:
##  $ company           : chr  "A. Morin" "A. Morin" "A. Morin" "A. Morin" ...
##  $ specific_bean_origin: chr  "Carenero" "Sur del Lago" "Puerto Cabello" "Madagascar" ...
##  $ REF               : int  1315 1315 1319 1011 1015 1470 705 705 705 705 ...
##  $ renew_date        : int  2014 2014 2014 2013 2013 2015 2011 2011 2011 2011 ...
##  $ cocoa_percentage  : num  0.7 0.7 0.7 0.7 0.7 0.7 0.6 0.8 0.88 0.72 ...
##  $ company_location  : chr  "France" "France" "France" "France" ...
##  $ rating            : num  2.75 3.5 3.75 3 4 3.75 2.75 3.25 3.5 3.5 ...
##  $ bean_type         : chr  "Criollo" "Criollo" "Criollo" "Criollo" ...
##  $ broad_bean_origin  : chr  "Venezuela" "Venezuela" "Venezuela" "Madagascar" ...
```

As I said, Rating and Cocoa percentage (not really a %, but a relative frequency) are numeric values. We have no use here of Renew data and REF variables, so we'll keep them at bay for now. There is a lot of different bean types, broad bean origin, specific bean origin, and company, so I'll think of a way that those are being shown as a geospatial analysis, or as a colour variable.

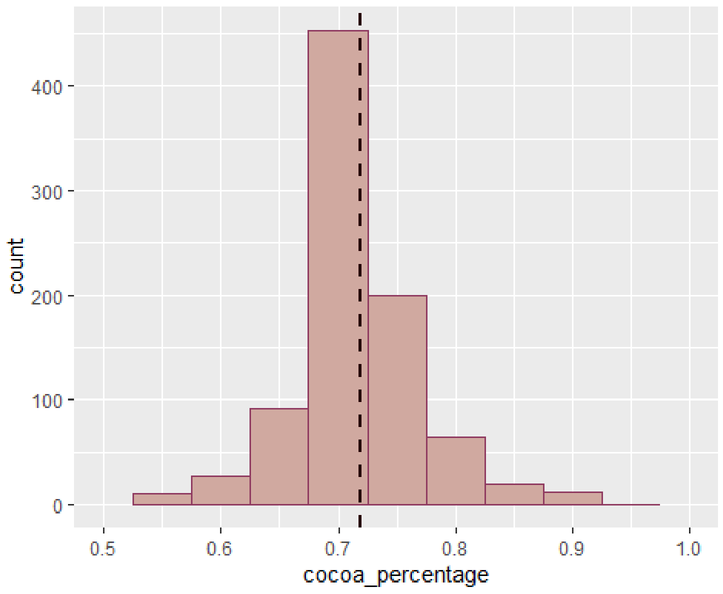
First, I'm going to start with the summaries of numerical variables of Rating and Cocoa percentage.

```
summary (cocoa_percentage) # code to get five most important numbers

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   0.700   0.700   0.719   0.750   1.000

sd (cocoa_percentage) # code for standard deviation (total variability)

## [1] 0.06116981

IQR (cocoa_percentage) # code for interquartile range (middle variability)

## [1] 0.05
```

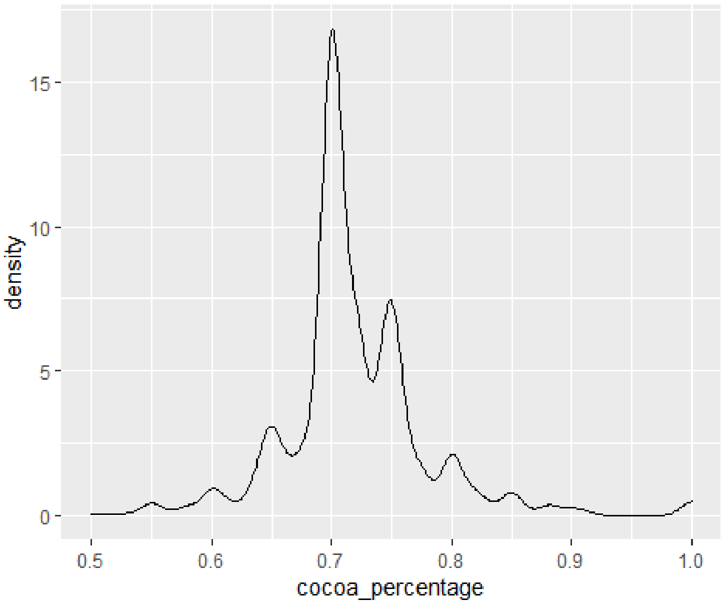
Let’s repeat real quick - cocoa percentage is the darkness of the chocolate bar being reviewed. When looking at the summary, we can see that the minimal value is 50%, whereas the maximum is 100%. The median and the mean are close together, around 70%, with the standard deviation of 6%. That means that the typical cocoa percentage is 71,9%, whereas 50% of chocolate has cocoa percentage lower than 70%, and 50% of chocolate has cocoa percentage higher than 70% (the median). Since the first and third quartile aren’t almost the same, that means we have some kind of skewness in this variable.

Let us see the graphics for distribution look and possible outliers. For this I’m going to use basic histogram and boxplot.

```
binsize = diff(range(dfrm$cocoa_percentage))/10 # code for bin number on histogram
ggplot(dfrm, aes(x = cocoa_percentage)) + geom_histogram(binwidth = binsize, fill = "#d0a9a0",colour = "#8f3a60")+
 geom_vline(aes(xintercept = mean(cocoa_percentage)), colour = "#240004",linetype = "dashed",size = 1)+
 xlim(0.5,1)

## Warning: Removed 2 rows containing missing values (geom_bar).
```



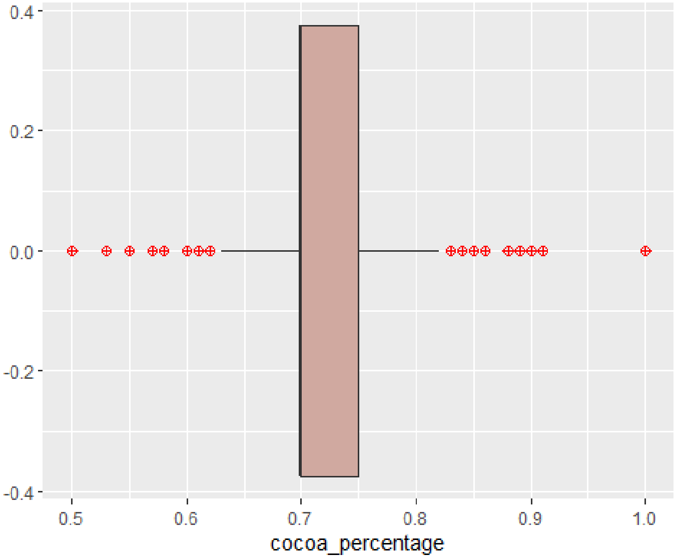
```r
ggplot(dfrm, aes(x = cocoa_percentage)) + geom_density()
```



Above, you can see what we data scientists call - histogram, or a graph representing a distribution of a viewed variable (in this case cocoa percentage). The black dotted line is representing mean value, whereas the median lies just where the tallest "tower" starts (so, a little bit left of the mean). Although the difference between the median and mean is small, we can see on the graph that difference isn't that small. Again, since the median is slightly lower than the mean, we can talk about a positive skewness, which isn't visible here. That is because both tails look the same.

That is why we can also make a boxplot. It is essentially a graphic notion of the FIVENUM calculated above, but it can show outliers. Outliers are extreme values in the distribution, which can distort the mean value.

```
ggplot(dfrm, aes(x = cocoa_percentage)) + geom_boxplot (fill = "#d0a9a0", outlier.colour = "red",
outlier.shape = 10, outlier.size = 2)
```



From the boxplot above, we can see outliers on both side of the boxplot (red dots). By theory, extreme outliers are being those who fall beyond the Q1-IQR * 1.5 and Q3 + IQR * 1,5 line. IQR is here 0.05, which means that everything that is beyond is being considered an extreme outlier. For this That means that everything outside [0.625, 0.825] range is an outlier, as seen on boxplot. With the following code, I'm going to delete those outliers that are going outside the range we just calculated.

```
boxplot(dfrm$cocoa_percentage, plot = FALSE)$out # this is the formula with which you can see the actual
outliers in the variable cocoa percentage. I plotted false because there is no need to see the boxplot again.

##  [1] 0.60 0.88 0.55 0.60 0.85 0.85 0.50 0.60 0.83 0.83 0.88 0.86 1.00 0.85 0.91
## [16] 0.60 0.55 0.55 0.55 0.55 0.61 1.00 0.88 0.85 0.62 0.83 0.85 0.60 0.58 0.60
## [31] 0.88 1.00 0.60 0.61 0.58 0.61 0.55 0.58 0.60 0.85 0.60 0.55 0.57 0.58 0.58
## [46] 1.00 0.90 1.00 0.60 0.90 0.55 0.62 0.85 0.60 0.60 1.00 0.85 0.85 1.00 0.85
## [61] 0.53 0.60 1.00 0.85 0.61 0.88 0.60 0.62 0.84 0.91 0.60 0.85 1.00 0.90 0.89
## [76] 0.88 0.85 0.85

outliers = boxplot(dfrm$cocoa_percentage, plot = FALSE)$out # we are going to attach the formula above to
a variable named outliers.
dfrm = dfrm[-which(dfrm$cocoa_percentage %in% outliers),] # with this formula, I removed the outliers from
the data
```

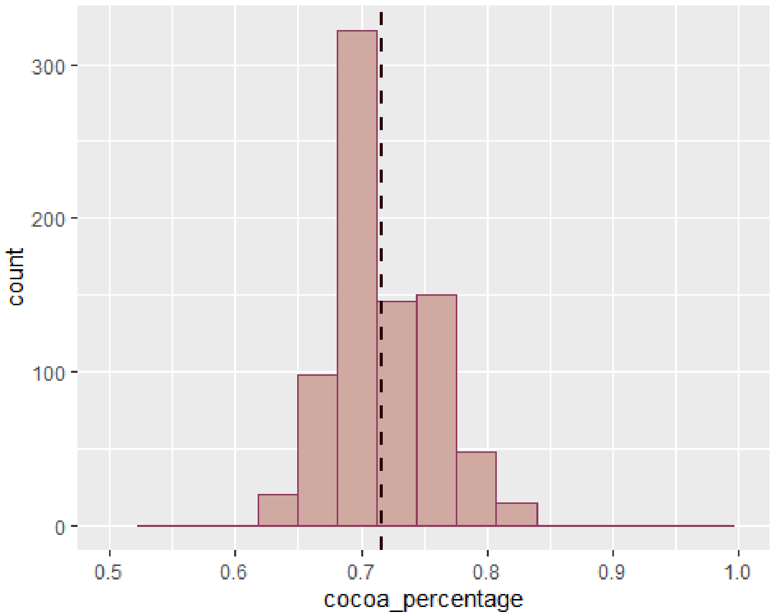
I'm assuming that the variable Rating also has outliers, so I'll copy the same formula for it.

```
boxplot(dfrm$rating, plot = FALSE)$out # this is the formula with which you can see the actual outliers in the
variable rating. I plotted false because there is no need to see the boxplot again.

## [1] 5.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 1.5
```

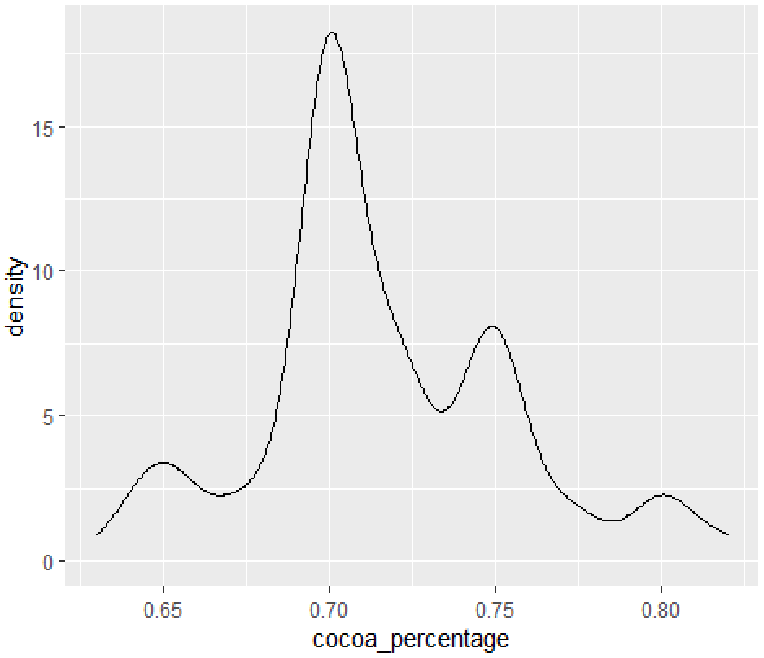
```
outliers = boxplot(dfrm$rating, plot = FALSE)$out # we are going to attach the formula above to a variable
named outliers.
dfrm = dfrm[-which(dfrm$rating %in% outliers),] # with this formula, I removed the outliers from the data.
```

Now, in total we have 799 observations, which means that 87 of them were outliers in those two variables. Other variables are character, so there are no outliers to delete. Now we should have a normal distribution to continue with, but let us check it again.

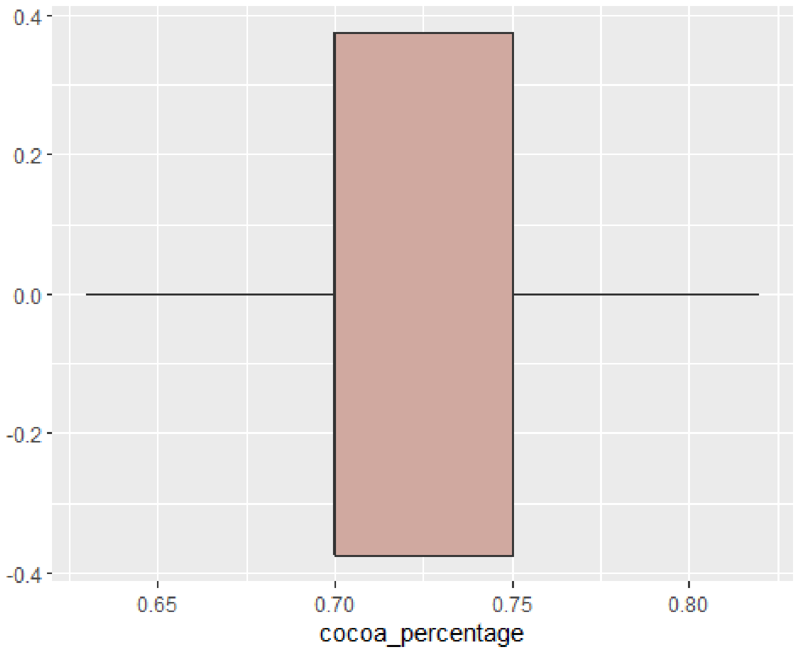
```
summary (cocoa_percentage)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   0.700   0.700   0.719   0.750   1.000

sd (cocoa_percentage)

## [1] 0.06116981

binsize = diff(range(dfrm$cocoa_percentage))/6
ggplot(dfrm, aes(x = cocoa_percentage)) + geom_histogram(binwidth = binsize, fill = "#d0a9a0", colour =
"#8f3a60")+
 geom_vline(aes(xintercept = mean(cocoa_percentage)), colour = "#240004", linetype = "dashed", size = 1)+
 xlim(0.5,1)

## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
ggplot (dfrm, aes(x = cocoa_percentage)) + geom_density()
```



```
ggplot (dfrm, aes(x = cocoa_percentage)) + geom_boxplot (fill = "#d0a9a0", outlier.colour = "red",
outlier.shape = 10, outlier.size = 2)
```



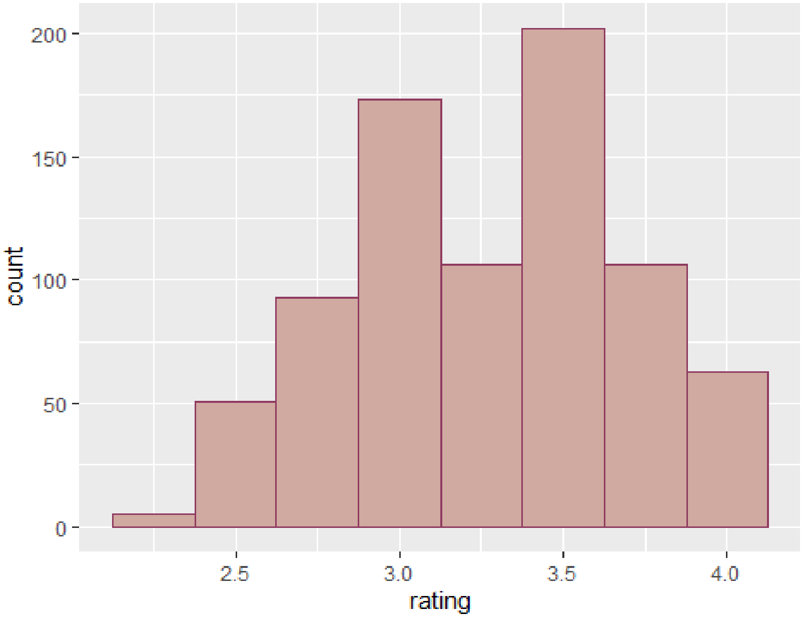
The distribution is still not normal, but there are no outliers that are extreme, according to boxplot, and total variability (standard deviation) went from 6.1% to 5%. I'll leave this distribution as it is, and go onto Rating variable.

The next numeric value worth taking a look at is rating.

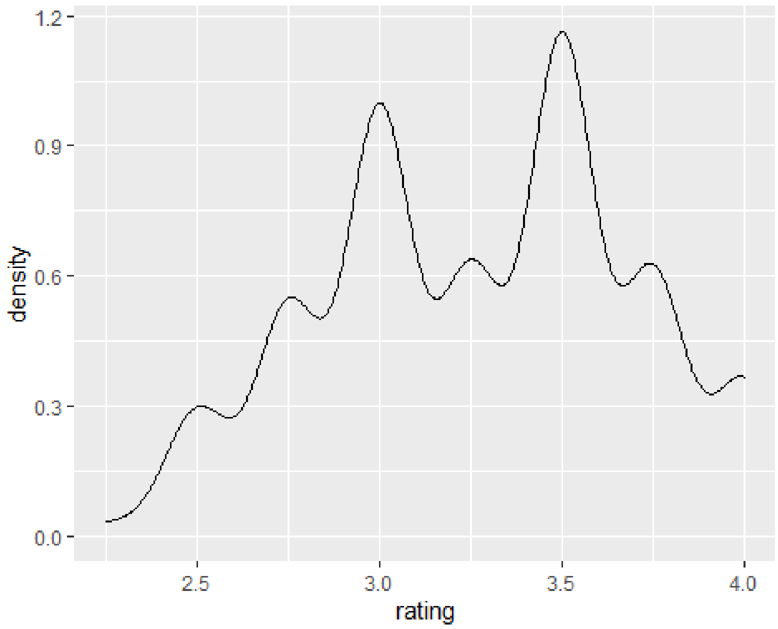
```
summary (rating)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.250   3.228   3.500   5.000

sd (rating)

## [1] 0.4662293

IQR (rating)

## [1] 0.5
```

Rating goes from 1 to 5, with typical value of 3,228. The mean and the median are very close together, but since the quartiles aren't the same, we can assume skewness. Let's check via histogram and boxplot what kind of distribution this is. Since we got rid of outliers with the formula above, we shouldn't have any extreme outliers now.

```
binsize = diff(range(dfrm$rating))/7
ggplot (dfrm, aes(x = rating)) + geom_histogram(binwidth = binsize, fill = "#d0a9a0", colour = "#8f3a60")
```

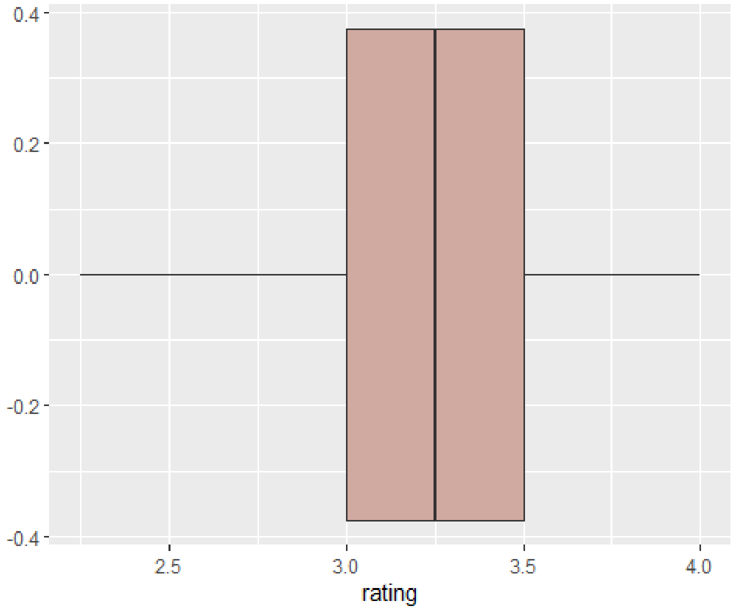


```
ggplot(dfrm, aes(x = rating)) + geom_density()
```



As seen in the table and histogram above, this distribution has a negative skew or a left tail, but also that is a bimodal. I won't touch bimodality here, and leave it as it is. Let's draw a boxplot to see if there are any significant outliers.

```
ggplot(dfrm, aes(x = rating)) + geom_boxplot (fill = "#d0a9a0", outlier.colour = "red", outlier.shape = 10,
outlier.size = 2)
```



There are no outliers in the boxplot.

EDA - other variables

```
str (dfrm) # code to see new structure

## 'data.frame':    799 obs. of  9 variables:
##  $ company           : chr  "A. Morin" "A. Morin" "A. Morin" "A. Morin" ...
##  $ specific_bean_origin: chr  "Carenero" "Sur del Lago" "Puerto Cabello" "Madagascar" ...
##  $ REF               : int  1315 1315 1319 1011 1015 1470 705 705 370 316 ...
##  $ renew_date        : int  2014 2014 2014 2013 2013 2015 2011 2011 2009 2009 ...
##  $ cocoa_percentage   : num  0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.72 0.7 0.7 ...
##  $ company_location   : chr  "France" "France" "France" "France" ...
##  $ rating            : num  2.75 3.5 3.75 3 4 3.75 3.25 3.5 3 3 ...
##  $ bean_type         : chr  "Criollo" "Criollo" "Criollo" "Criollo" ...
##  $ broad_bean_origin   : chr  "Venezuela" "Venezuela" "Venezuela" "Madagascar" ...
```

From the above mentioned variables, interesting ones to see would be company location, bean type and specific/broad bean origin. All except bean type are locations, so we can use geospatial analysis, to show the frequency. For this, I'm going to use Excel maps, and you can find the result in the repository as well.

After we are done with geospatial analysis/maps in Excel, I'm going to make a correlation and regression analysis.

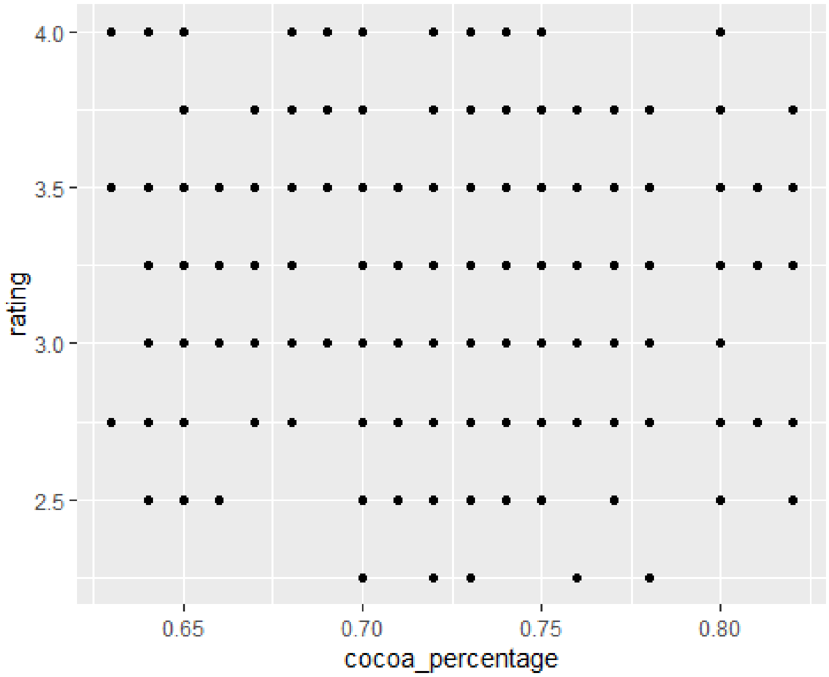
Correlation analysis

Correlation is a measure that represents the relationship that two variables have, but not showing cause and effect of one another. For example, the correlation between height and weight is pretty positively high, meaning that the relationship between those two variables is positive (when a child grows, his weight also grows - both variables go in the same direction, but that doesn't mean that the height causes the weight to grow too, the weight itself is being connected to how much a person eats, exercises etc.).

Real correlation can be calculated for numeric values/variables, and we have two of them here (rating and cocoa percentage). When I say "real", i mean you can back it up with number (percentage of correlation). You can still check correlation between non-numeric or qualitative variables, but then you can only look up the p-value and compare it to your posed p-value, and then get an answer whether two variables have a relationship or not. It is just not backed up by a specific number. Sadly, the qualitative variables need to be in a "factor" class, and our qualitative variables aren't, and can't be turned into factor.

Next, we can check the real correlation between the numeric values, and immediately create a scatter plot which puts together both numeric values on the graph and shows their movement.

```
cor(dfrm$rating, dfrm$cocoa_percentage) # code for correlation

## [1] -0.1453325

cor.test(dfrm$rating, dfrm$cocoa_percentage, method="pearson") # code for correlation summary

##
##  Pearson's product-moment correlation
##
## data:  dfrm$rating and dfrm$cocoa_percentage
## t = -4.1469, df = 797, p-value = 3.731e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.21254758 -0.07674857
## sample estimates:
##        cor
## -0.1453325

ggplot(dfrm, aes(x = cocoa_percentage, y = rating)) + geom_point() # code for scatter plot
```



Sadly, from the correlation test we conducted, we can see that the correlation is only -0.14, which is too low to be considered significant (it should be around minimum 0.4). Nevertheless, the p-value is lower than 0.05, so by that measure, the correlation can be seen as significant.

We can see on a scatter plot that the relationship isn't showing any way of movement, so hence the low correlation, and the scatter plot is backing that up.

Still, it would be neat to make a regression analysis as well, as we can show the relationship between numerical variables in a form of a linear equation.

Regression analysis

We have two numeric variables in this dataset - cocoa percentage and rating, so we can use regression analysis to see a form of a linear equation between them.

```
reg1 = lm(dfrm$rating~dfrm$cocoa_percentage) # code for regression analysis
print (summary(reg1)) # code for regression summary

##
## Call:
## lm(formula = dfrm$rating ~ dfrm$cocoa_percentage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04842 -0.29842  0.03084  0.28084  0.86010
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.4081     0.2743  16.070  < 2e-16 ***
## dfrm$cocoa_percentage  -1.5852     0.3823  -4.147 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4175 on 797 degrees of freedom
## Multiple R-squared:  0.02112,    Adjusted R-squared:  0.01989
## F-statistic:  17.2 on 1 and 797 DF,  p-value: 3.731e-05
```

R adjusted (coefficient of determination) means a fraction of the variance (variability/dispersion) of y-variable that is explained by the regression model (variables that are in the regression model itself), and the remaining variance/variability isn't explained by the variables in the model and that must be due to some other factors.

The R squared here is very low, only 0.019, which means that only 3% of the variance of the rating is explained by the cocoa percentage. That means that some other factors influence and are taken into account when the rating is given, and those variables aren't in this model. The p-value is lower than 0.05, which makes the model significant, and the FIVENUM of the residuals (difference between real data in the dataset and the data predicted by this model) is distributed well.

If we wish to use this regression model, the equation would be => rating = 4.4081 – 1.5852*cocoa_percentage

But, since the R squared and the correlation are very low, the model wouldn't give precise output.