Cocoa Percentage in Chocolate

Sheli Feigin, Daniel Yohan

2022-05-29

Background

Who doesn’t like chocolate?

We very much do, and believe (almost) everyone too. Therefore, we find that the questions should be asked about the preferences people have regarding chocolate.

In this markdown, we’ll analyze the chart from the following site: http://flavorsofcacao.com/chocolate_database.html

Goals

In our research we’ll focus on 2 main questions:

1. Can we assume there’s a linear regression between the cocoa percentage - as an independent variable, type of ingredients - as “dummies”, and the chocolate rating - as a dependent variable?

2. Are the percentage of cocoa, and the manufactures’ location (in which continent they’re based) are independent variables?

Libraries

```
library(tidyverse)
library(readxl)
library(ggplot2)
```

Importing the Dataset

```
cocoa_data <- read_excel("cocoa_data.xlsx")

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i
= sheet, :
## Expecting numeric in A2536 / R2536C1: got 'Asiaa'
```

Early Filtering and Tidying

The data set is extracted from an Excel file, imported from the linked site above. We’ve dropped columns that are irrelevant to our tests: ‘REF’ and ‘Most Memorable Characteristics’. Likewise, we’ve omitted rows with missing data, and the given cocoa percentages had to be converted to actual numbers.

```
drop <- c("REF", "Most Memorable Characteristics")
cocoa_data_clean = cocoa_data[,!(names(cocoa_data) %in% drop)]
cocoa_data_clean <- na.omit(cocoa_data_clean)
cocoa_data_clean[6] <- 100 * (cocoa_data_clean[6])
```

Some header names had to be changed:

```
names(cocoa_data_clean)[1] <- 'Company_(Manufacturer)'
names(cocoa_data_clean)[2] <- 'Company_Location'
names(cocoa_data_clean)[3] <- 'Review_Year'
names(cocoa_data_clean)[4] <- 'Bean_Origin'
names(cocoa_data_clean)[5] <- 'Bean_Origin_and_or_Bar_name'
names(cocoa_data_clean)[6] <- 'Cocoa_percent'
```

First test

In a significance level of 5%:

Our Null hypothesis is that all of the coefficients equal to 0, which means that we can't state there's a linear regression between any of the variables above.

Our Alternative hypothesis is that at least one coefficients isn't 0, which means that we can approve that there's a linear regression between some of the variables.

Transformation, Modeling and Visualization

Transformation

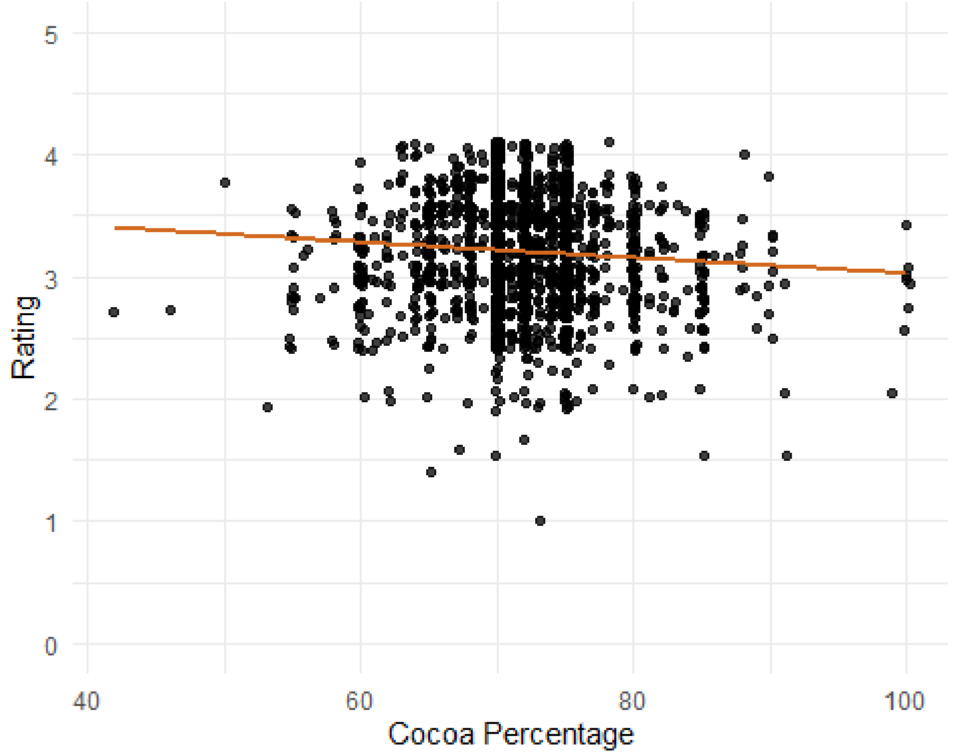
We've added 7 columns for each ingredient, indicating whether it's in the chocolate bar or not. Then we extracted the types of ingredients of each bar and filled the new columns accordingly.

```
cocoa_data_clean[7] <- gsub("^.*?- ", "", cocoa_data_clean$Ingredients)
cocoa_data_clean[c("Beans", "Sugar", "Sweetener", "Cocoa_Butter",
"Vanilla", "Lecithin", "Salt")] <- 0
cocoa_data_clean$Ingredients <- strsplit(cocoa_data_clean$Ingredients,
split = ",")
for (i in 1:length(cocoa_data_clean$Ingredients)) {
  curr_ingredients <- cocoa_data_clean$Ingredients[[i]]
  if ("B" %in% curr_ingredients) {
    cocoa_data_clean$Beans[i] <- 1
  }
   if ("S" %in% curr_ingredients) {
    cocoa_data_clean$Sugar[i] <- 1
   }
   if ("S*" %in% curr_ingredients) {
    cocoa_data_clean$Sweetener[i] <- 1
   }
   if ("C" %in% curr_ingredients) {
    cocoa_data_clean$Cocoa_Butter[i] <- 1
   }
   if ("V" %in% curr_ingredients) {
    cocoa_data_clean$Vanilla[i] <- 1
   }
   if ("L" %in% curr_ingredients) {
    cocoa_data_clean$Lecithin[i] <- 1
   }
   if ("Sa" %in% curr_ingredients) {
    cocoa_data_clean$Salt[i] <- 1
   }
}
```

Visualization

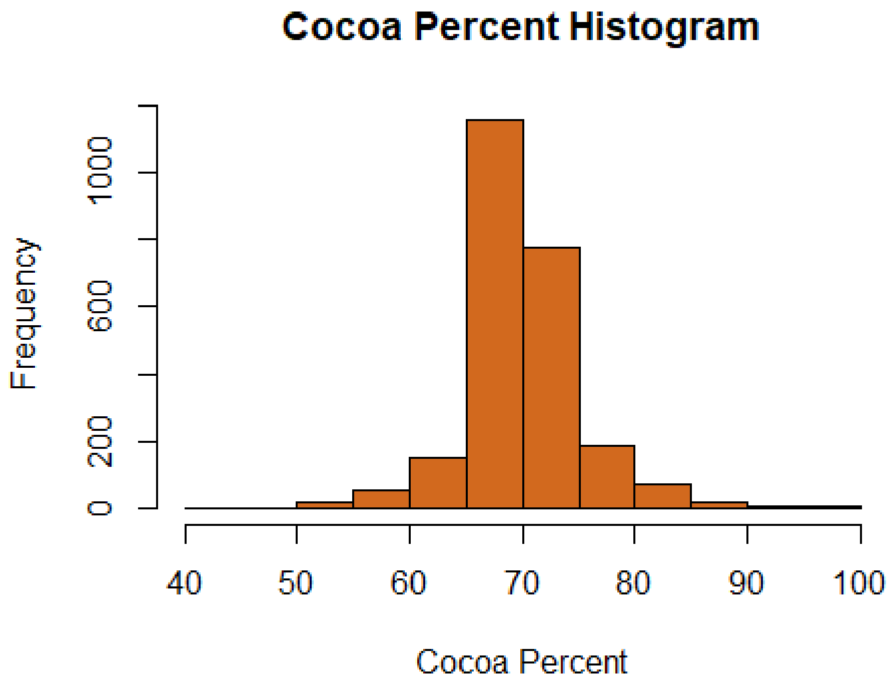
Let’s get a first impression regarding the cocoa percentage and ratings in out data set:

```
cocoa_data_clean %>%
ggplot(aes(x = Cocoa_percent, y = Rating)) +
  geom_jitter(alpha = .75) +
  coord_cartesian(ylim = c(0,5)) +
  labs(x = 'Cocoa Percentage', y = 'Rating') +
  theme_minimal() +
  geom_smooth(method = 'lm', se = FALSE, col = 'chocolate')

## `geom_smooth()` using formula 'y ~ x'
```



We can’t notice a strong relationship between the cocoa percentage and the chocolates’ rating.

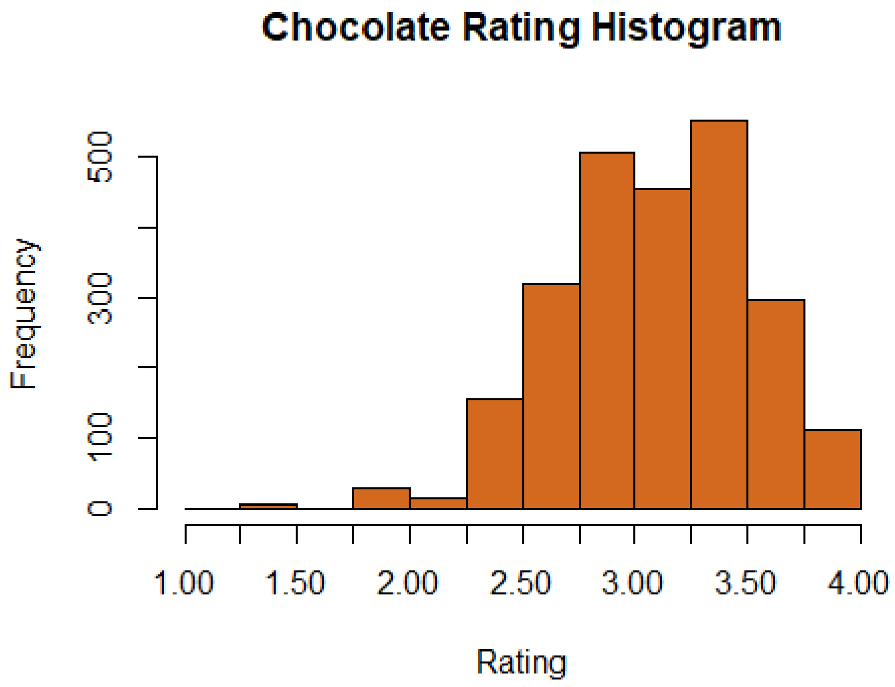
```
hist(cocoa_data_clean$Cocoa_percent,
     col = "chocolate",
     main = "Cocoa Percent Histogram",
     xlab = "Cocoa Percent")
```



Here we can roughly see that the frequency of the cocoa percentage distribute normally. Moreover, most of the documented chocolate bars consist between 65% - 75% cocoa.

Note: in order to convert the normal distribution plot from density to frequency, we had to multiply its y values by: the histogram bars width (5); the number of observations (nrow).

```
hist(cocoa_data_clean$Rating,
     col = "chocolate",
     main = "Chocolate Rating Histogram",
     xlab = "Rating",
     breaks = seq(1.0, 4.0, 0.25),
     xaxp = c(1.0, 4.0, 12)
     )
```



From the second Histogram we understand that the vast majority of the chocolate bars have a rating greater than 2.25.

Modeling

We've created some vectors from the relevant columns as variables for the modeling:

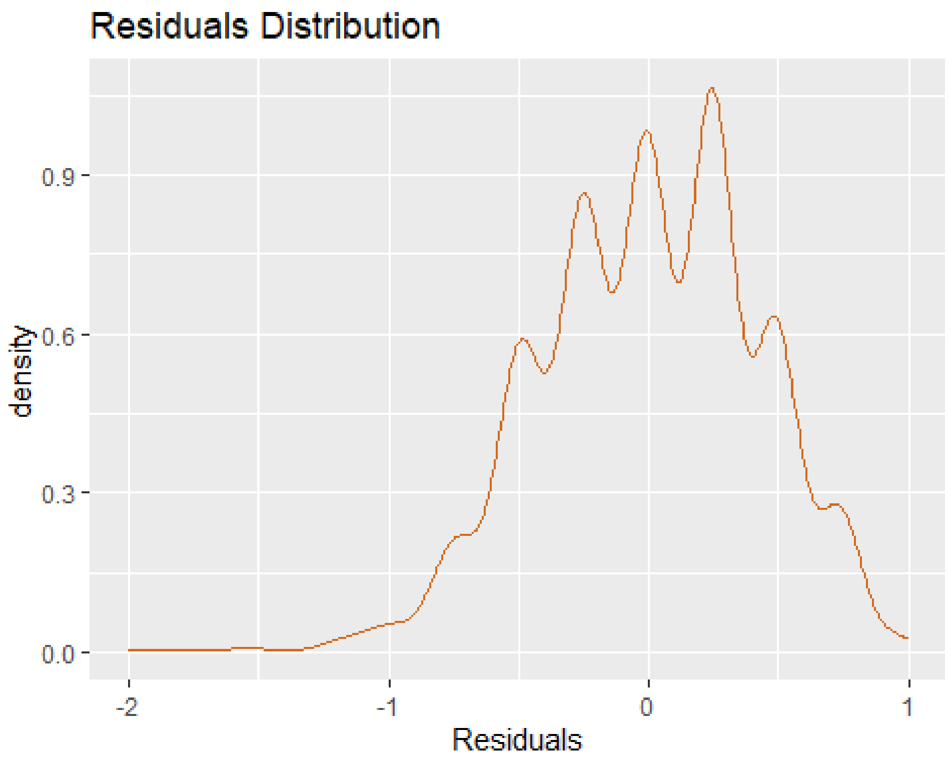
```
rating <- cocoa_data_clean$Rating
cocoa_percent <- cocoa_data_clean$Cocoa_percent
sugar <- cocoa_data_clean$Sugar
sweetener <- cocoa_data_clean$Sweetener
cocoa_Butter <- cocoa_data_clean$Cocoa_Butter
vanilla <- cocoa_data_clean$Vanilla
lecithin <- cocoa_data_clean$Lecithin
salt <- cocoa_data_clean$Salt

model <- lm( formula = rating ~ cocoa_percent + sugar + sweetener +
cocoa_Butter + vanilla + lecithin + salt, data = cocoa_data_clean)
model %>% summary()

##
## Call:
## lm(formula = rating ~ cocoa_percent + sugar + sweetener +
cocoa_Butter +
##     vanilla + lecithin + salt, data = cocoa_data_clean)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9982 -0.2747  0.0085  0.2607  0.9936
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.779734   0.235461  16.052  < 2e-16 ***
## cocoa_percent -0.008219   0.001744  -4.713 2.58e-06 ***
## sugar          0.034878   0.165946   0.210   0.8335
## sweetener     -0.165496   0.173361  -0.955   0.3399
## cocoa_Butter   0.045146   0.019518   2.313   0.0208 *
## vanilla       -0.218608   0.026234  -8.333  < 2e-16 ***
## lecithin      -0.042929   0.023464  -1.830   0.0674 .
## salt          -0.074684   0.075548  -0.989   0.3230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4182 on 2435 degrees of freedom
## Multiple R-squared:  0.04964,    Adjusted R-squared:  0.04691
## F-statistic: 18.17 on 7 and 2435 DF,  p-value: < 2.2e-16
```

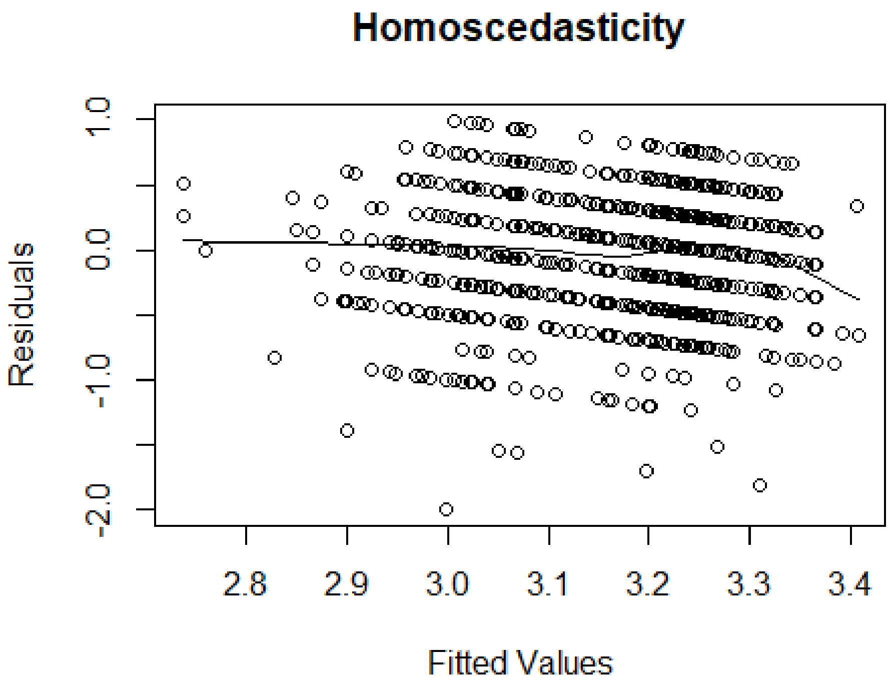
From the regression model above, we can observe that as the cocoa percentage gets higher the chocolates’ rating slightly decreases. Moreover, the presence of sweeteners, vanilla, lecithin and salt also decrease the rating - whereas cocoa butter and lecithin increase the rating.

In the following chunk we’ll examine if homoscedasticity exists in our model:

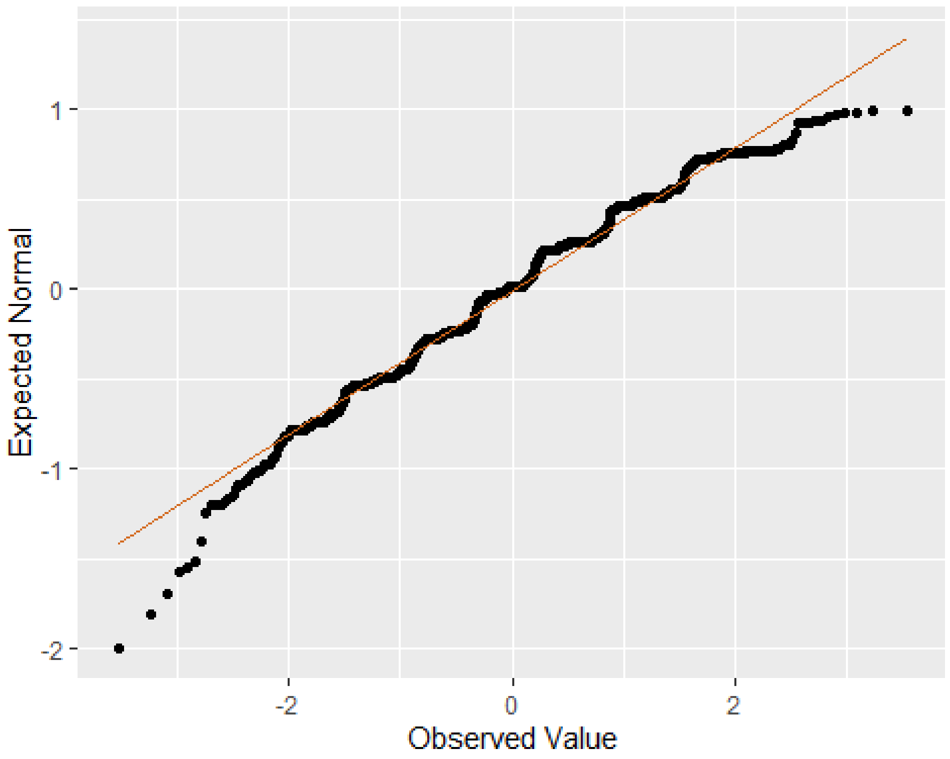
```
res <- model$residuals
fit <- model$fitted.values
res_table= tibble(res,fit)

ggplot(res_table,aes(x=res)) + geom_density(color="chocolate") +
xlab("Residuals") + ggtitle("Residuals Distribution")
```



```
scatter.smooth(x=fit, y=res, main="Homoscedasticity", xlab = "Fitted
Values", ylab = "Residuals")
```



```
ggplot(res_table, aes(sample=res)) + geom_qq() +
geom_qq_line(color="chocolate") + xlab("Observed Value") +
ylab("Expected Normal")
```



We notice from the second plot that the variance is not quite consistent along the fitted values. Moreover, it's seen from the third that the residuals can't be found on the line - which means they don't distribute normally. Consequently, we infer there's no homoscedasticity in our model. In any case we'll conclude the test as we could assume homoscedasticity.

Conclusion

Our PV is smaller then 5%, therefore we deny the null hypothesis, and accept the alternative. We approve there's a linear connections between some of the variables.

Worth mentioning that (almost) every dummy variable has a negative coefficient, excluding the variable of chocolate bars with 6 ingredients.

Second Test

In a significance level of 5%:

Our Null hypothesis is that the percentage of cocoa and the continents of the chocolate manufacturers are independent variables.

Our Alternative hypothesis is that those variables are dependent.

Transformation, Modeling and Visualization

Transformation

The transformation for this test was more complex.

Firstly, we’ve created vectors of continents for each chocolate companies’ location included in the data set.

```
in_africa = c("Sao Tome",  "Ghana",    "Madagascar",   "South Africa")
in_asia = c("Japan",    "India",    "Israel",   "Malaysia",
"Philippines",  "Russia",   "Singapore",    "South Korea",  "Taiwan",
"Thailand", "U.A.E.",   "Vietnam")
in_australia = c("Australia",   "Fiji", "New Zealand",  "Vanuatu")
in_europe = c("U.K.",   "France",   "Germany",  "Italy",
"Amsterdam",    "Austria",  "Belgium",  "Spain",    "Czech Republic",
"Denmark",  "Finland",  "Hungary",  "Iceland",  "Ireland",
"Lithuania",    "Netherlands",  "Norway",   "Poland",   "Portugal",
"Scotland", "Sweden",   "Switzerland",  "Wales", "8")
in_north_america = c("Canada",  "U.S.A.",   "Costa Rica",   "Dominican
Republic",   "El Salvador",  "Grenada",  "Guatemala",    "Honduras",
"Martinique",   "Mexico",   "Nicaragua",    "Puerto Rico",  "Sao Tome
0%_to_60%& Principe",  "St. Lucia",   "St.Vincent-Grenadines")
in_south_america = c("Ecuador", "Argentina",    "Bolivia",  "Brazil",
"Chile",    "Colombia", "Peru", "Suriname", "Venezuela")
```

Then we made a new table with observations as an input for the Chi-square test.

```
observed_table <- matrix(0, ncol = 4, nrow = 6)
continent_names <- c("Africa", "Asia", "Australia", "Europe", "North
America", "South America")
rownames(observed_table) <- continent_names
colnames(observed_table) <- c("40%-66%", "67%-74%", "75%-80%", "81%-
100%")
```

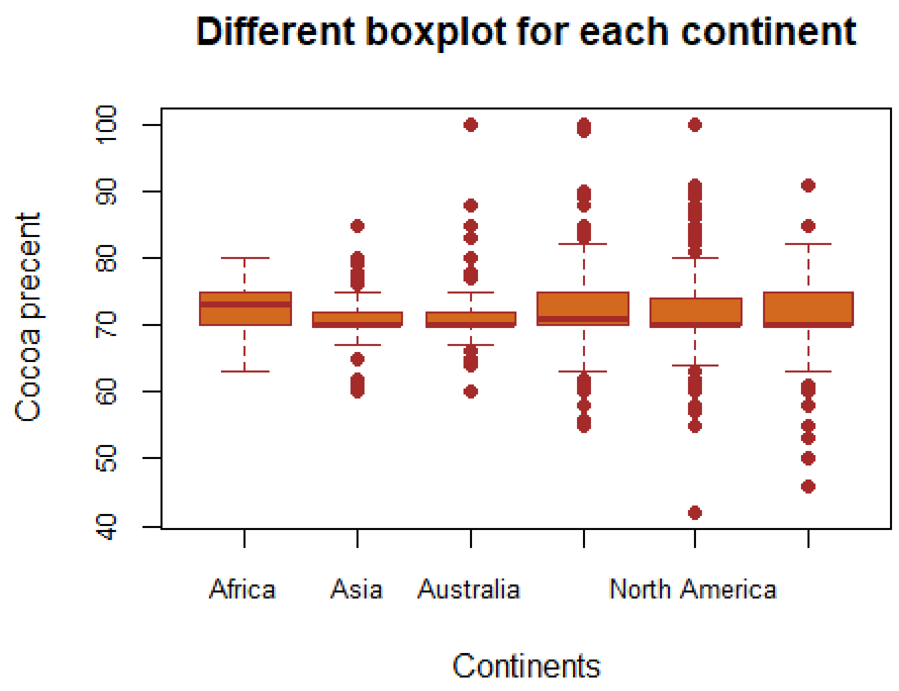
In the following chunk we’ve added the Continent column. We also created a for-loop that filled each row in the main data set with its’ appropriate continent, and simultaneously filled observed_table for the square chi test.

```
cocoa_data_clean$Continent <- NA
for (i in 1:length(cocoa_data_clean$Company_Location)) {
  if (cocoa_data_clean$Company_Location[i] %in% in_africa) {
    cocoa_data_clean$Continent[i] <- "Africa"
    fill_row = 1
  }
  else if (cocoa_data_clean$Company_Location[i] %in% in_asia) {
    cocoa_data_clean$Continent[i] <- "Asia"
    fill_row = 2
  }
  else if (cocoa_data_clean$Company_Location[i] %in% in_australia) {
    cocoa_data_clean$Continent[i] <- "Australia"
    fill_row = 3
  }
  else if (cocoa_data_clean$Company_Location[i] %in% in_europe) {
    cocoa_data_clean$Continent[i] <- "Europe"
    fill_row = 4
  }
  else if (cocoa_data_clean$Company_Location[i] %in% in_north_america)
{
    cocoa_data_clean$Continent[i] <- "North America"
    fill_row = 5
  }
  else {
    cocoa_data_clean$Continent[i] <- "South America"
    fill_row = 6
  }
  if (cocoa_data_clean$Cocoa_percent[i] %in% seq(40,68)) {
  observed_table[fill_row, 1] = observed_table[fill_row, 1] + 1
  }
  else if (cocoa_data_clean$Cocoa_percent[i] %in% seq(69,74)) {
  observed_table[fill_row, 2] = observed_table[fill_row, 2] + 1
  }
  else if (cocoa_data_clean$Cocoa_percent[i] %in% seq(75,80)) {
  observed_table[fill_row, 3] = observed_table[fill_row, 3] + 1
  }
  else {
  observed_table[fill_row, 4] = observed_table[fill_row, 4] + 1
  }
}
```

Visualization

We wanted to take a look at the connection of cocoa percentage and the manufacturers' continent, so we've created box-plots graph to see how the cocoa percentages are scattered in each continent.

```
boxplot(cocoa_data_clean$Cocoa_percent ~ cocoa_data_clean$Continent,
data=cocoa_data_clean,
main="Different boxplot for each continent",
xlab="Continents",
ylab="Cocoa precent",
col="chocolate",
border="brown",
pch = 19,
cex.axis = 0.8
)
```



```
rowSums(observed_table)

##         Africa          Asia     Australia        Europe North 
America 
##             21           105            85           702 
1363 
## South America 
##           167
```

We can see that the continents have a similar median values of cocoa percentage (around 70%), and the two middle quarters of each continent are between 70% and 75%. A noticeable difference between the boxes is within the extreme values. The continents with relatively more occurrences in the data set (Europe and North America) have a greater number of extremes.

Modeling

Let's examine the expected values for our independence test.

```
chi_test <- chisq.test(observed_table)

## Warning in chisq.test(observed_table): Chi-squared approximation may 
be
## incorrect

chi_test$expected
```

```
##                    40%-66%   67%-74%   75%-80%  81%-100%
## Africa            2.810888  12.80802   4.22063  1.160458
## Asia             14.054441  64.04011  21.10315  5.802292
## Australia        11.377405  51.84200  17.08350  4.697094
## Europe           93.963979 428.15391 141.08964 38.792468
## North America   182.440033 831.30168 273.93901 75.319280
## South America    22.353254 101.85428  33.56406  9.228408
```

There are cells for Africa and Australia with expected data smaller then 5. Therefore, we've merged the data of the two continent into one.

```
for (i in 1:4) {
  observed_table[1, i] <- observed_table[1, i] + observed_table[3, i]
}
fixed_observed_table = observed_table[-3,]
rownames(fixed_observed_table)[1] <- "Africa & Australia"
```

Now each cell has a value greater then 5 and we can properly conduct our second test.

```
chi_test <- chisq.test(fixed_observed_table)
chi_test$expected
```

```
##                        40%-66%   67%-74%   75%-80%  81%-100%
## Africa & Australia  14.18829  64.65002  21.30413  5.857552
## Asia                14.05444  64.04011  21.10315  5.802292
## Europe              93.96398 428.15391 141.08964 38.792468
## North America      182.44003 831.30168 273.93901 75.319280
## South America       22.35325 101.85428  33.56406  9.228408
```

```
chi_test
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  fixed_observed_table
## X-squared = 62.293, df = 12, p-value = 8.594e-09
```

Conclusion

Our PV is smaller then 5%. In that significance level, we deny the null hypothesis and accept the alternative. Therefore, we can state that the percentage of cocoa, and the continents of the chocolate manufacturers are indeed dependent.