Data\_Cleaning.R

skhim

2023-04-06

##Chocolate Analysis  
library(readr) #to read the csv file

## Warning: package 'readr' was built under R version 4.2.3

chocolate <- read\_csv("data/raw\_chocolate.csv")

## New names:  
## Rows: 2224 Columns: 21  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (15): company, company\_location, country\_of\_bean\_origin, specific\_bean\_o... dbl  
## (6): ...1, ref, review\_date, cocoa\_percent, rating, counts\_of\_ingredients  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `` -> `...1`

View(chocolate)  
  
#View column names  
column\_names = colnames(chocolate)  
column\_names

## [1] "...1" "ref"   
## [3] "company" "company\_location"   
## [5] "review\_date" "country\_of\_bean\_origin"   
## [7] "specific\_bean\_origin\_or\_bar\_name" "cocoa\_percent"   
## [9] "rating" "counts\_of\_ingredients"   
## [11] "beans" "cocoa\_butter"   
## [13] "vanilla" "lecithin"   
## [15] "salt" "sugar"   
## [17] "sweetener\_without\_sugar" "first\_taste"   
## [19] "second\_taste" "third\_taste"   
## [21] "fourth\_taste"

#Removing columns not being used  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## 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

chocolate <- select(chocolate, -c(...1,fourth\_taste,specific\_bean\_origin\_or\_bar\_name))  
  
chocolate <- data.frame(chocolate)  
  
#Modifying the columns  
#Changing have/have not to 1/0  
  
chocolate = data.frame(chocolate)  
colnames(chocolate)[9] <- "bean"  
colnames(chocolate)[11] <- "vanila"  
  
#Columns to be changed: 9 to 15  
bool\_columns = colnames(chocolate[9:15])  
bool\_columns

## [1] "bean" "cocoa\_butter"   
## [3] "vanila" "lecithin"   
## [5] "salt" "sugar"   
## [7] "sweetener\_without\_sugar"

for (column in bool\_columns){  
 have\_not <- paste("have\_not\_", column, sep = "")  
 chocolate[column] = replace(chocolate[column], chocolate[column] == have\_not, 0);  
  
 have <- paste("have\_", column, sep = "")  
 chocolate[column] <- replace(chocolate[column], chocolate[column] == have, 1)  
}  
  
colnames(chocolate)[11] <- "vanilla"  
  
#Writing the cleaned data into a new file  
write.csv(chocolate, "data/chocolate.csv")