ANA 515 Assignment 2, Loading, Saving, and Describing Data

# Section 1: Description of the data

I’m working with the Amazon dataset. This dataset contains over 10,000 different Amazon products and their information, user reviews, and so on for the year of 2020. The data is downloaded from <https://www.kaggle.com/datasets/nguyenngocphung/10000-amazon-products-dataset> in csv format.

# Section 2: Reading the data into R

# Using the fread() function from 'data.table' package  
# Using it because the function is very fast at reading large volume of data   
# the amazon data set has so many rows and its size is more than 20 MB  
# reading first 14 columns since they contain information  
amazon <- data.table::fread("Amazon\_Products.csv")[,1:14]

# Section 3: Clean the data

amazon <- amazon[,c(1:9)] # taking the first 9 columns for convenience  
# Renaming variables  
colnames(amazon) <- c("ID","product","manufacturer","price","num\_available","num\_reviews","num\_ans\_ques","avg\_rating","cat\_and\_sub\_cat")  
# Changing data types of some variables  
amazon$price <- as.numeric(substring(amazon$price, 3))  
amazon$avg\_rating <- as.numeric(substring(amazon$avg\_rating, 1, 4))  
amazon$num\_ans\_ques <- as.numeric(amazon$num\_ans\_ques)  
amazon$num\_reviews <- as.numeric(amazon$num\_reviews)

# Section 4: Characteristics of the data

This dataframe has 10004 rows and 9 columns. The names of the columns and a brief description of each are in the table below:

|  |  |
| --- | --- |
| Column Names | Brief Descriptions |
| ID | Unique ID of the product |
| product | Product name |
| manufacturer | Manufacturer name |
| price | Price of the product in £ |
| num\_available | Number of products available in the stock |
| num\_reviews | Number of reviews |
| num\_ans\_ques | Number of answered questions |
| avg\_rating | Average rating of reviews |
| cat\_and\_sub\_cat | Amazon category and sub category |

# Section 5: Summary statistics

# Selecting the three columns  
amazon\_sub <- amazon[,c("price","num\_reviews","avg\_rating")]  
# Creating a function that calculates the required summary statistics  
summ <- function(df){  
 su <- data.frame(Minimum = apply(df, 2, min, na.rm = TRUE),  
 Maximum = apply(df, 2, max, na.rm = TRUE),  
 Mean = apply(df, 2, mean, na.rm = TRUE),  
 No\_Missing = apply(df, 2, function(x) sum(is.na(x))))  
 return(su)   
}  
# Assigning the result of summary statistics of the dataset  
amazon\_summary <- summ(amazon\_sub)  
knitr::kable(amazon\_summary)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Minimum | Maximum | Mean | No\_Missing |
| price | 0.01 | 995.11 | 19.968272 | 1458 |
| num\_reviews | 1.00 | 802.00 | 8.897395 | 24 |
| avg\_rating | 2.30 | 5.00 | 4.707283 | 22 |