Exploring scanner and web scraped data sets

Exercises

# Packages to use

library(PriceIndices)  
library(dplyr)

# Exercises

Let’s practice on a scanner data set contained in the R PriceIndices package. Use the expertise that you have learned in the example of the exploration of the Dominick’s data set to answer the questions respective the **milk** data set.

To do that, load the PriceIndices package and note that the data set can be accessed after loading the package. You will need to explore the data set using some codes in R to find the answer.

library(PriceIndices)  
head(milk)

## time prices quantities prodID retID description  
## 1 2018-12-01 8.78 9.0 14215 2210 powdered milk  
## 2 2019-01-01 8.78 13.5 14215 2210 powdered milk  
## 3 2019-02-01 8.78 0.5 14215 1311 powdered milk  
## 4 2019-02-01 8.78 8.0 14215 2210 powdered milk  
## 5 2019-03-01 8.78 0.5 14215 1311 powdered milk  
## 6 2019-03-01 8.78 1.5 14215 2210 powdered milk

1. What is the number of rows in the whole data set (hint: use the glimpse function)?

() 6 () 4386 () 1236

1. What is the number of variables in the data set?

() 6 () 4 () 7

1. How is data being aggregated in time?

() daily () weekly () hourly () monthly

1. The variables retID (it is interesting to note here that the data is being given at the retailer level, and not at the stores level) and prodID identify a given retailer and product, respectively. Find out what is the respective number of distinct retailers and products in the data set.

() 5, 61 () 6, 66 () 7, 69 () 5, 68

1. Use the ‘mutate()’ verb of dplyr to create a new column called turnover which is the result of the product of the prices and quantities columns and answer what is the retailer with the largest amount of sales accumulated in the whole period.

() 1311 () 2210 () 6610 () 7611 () 8910

1. For the retailer with the largest turnover, what is the product with the largest expenditure in the period?

() 15404 () 34540 () 401350 () 406245 () 404770

1. For the product found above, what is the period where the sales were lower and higher for the retailer with largest turnover found above, respectively?

() 2018-12-01 () 2019-12-01 () 2019-08-01 () 2020-08-01

1. Have all the retailers stayed for the same number of months in the panel?

() yes () no

1. How many months the products with shorter lifetime stayed in the panel?

() 1 month () 2 months () 3 months () 4 months () 5 months