DW-Exercise1

Diane K

September 7, 2016

Load libraries

0.Load dataset

```
my_dataset <- read.csv(file="refine_original.csv"
                  , head = TRUE, sep=",")
str(my_dataset)
```

```
## 'data.frame':    25 obs. of  6 variables:
##  $ company              : Factor w/ 19 levels "ak zo","akz0",..: 10 8 7 13 11 9 3 4 5 2 ...
##  $ Product.code...number: Factor w/ 23 levels "p-23","p-34",..: 4 3 19 20 17 1 13 11 22 2 ...
##  $ address              : Factor w/ 25 levels "Delfzijlstraat 54",..: 9 10 11 12 13 14 19 20 21 22 .
##  $ city                 : Factor w/ 1 level "arnhem": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country              : Factor w/ 1 level "the netherlands": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name                 : Factor w/ 20 levels "dhr j. Gansen",..: 7 6 1 9 4 5 2 10 3 8 ...
```

```
kable(head(my_dataset) , format = "markdown", caption = "Refine dataset")
```

company	Product.code...number	address	city	country	name
Phillips	p-5	Groningensingel 147	arnhem	the netherlands	dhr p. janse
phillips	p-43	Groningensingel 148	arnhem	the netherlands	dhr p. hans
philips	x-3	Groningensingel 149	arnhem	the netherlands	dhr j. Gans
phllips	x-34	Groningensingel 150	arnhem	the netherlands	dhr p. mans
phillps	x-12	Groningensingel 151	arnhem	the netherlands	dhr p. frans
phillipS	p-23	Groningensingel 152	arnhem	the netherlands	dhr p. frans

1. Clean up brand names

```
( my_dataset$company <-  tolower(my_dataset$company) )
```

```
##  [1] "phillips"   "phillips"   "philips"    "phllips"    "phillps"
##  [6] "phillips"   "akzo"       "akzo"       "akzo"       "akz0"
## [11] "ak zo"      "akzo"       "akzo"       "phillips"   "fillips"
## [16] "phlips"     "van houten" "van houten" "van houten" "van houten"
## [21] "van houten" "unilver"    "unilever"   "unilever"   "unilever"
```

```
(  my_dataset$company[grepl("ps$", my_dataset$company)] <- "philips" )
```

```
## [1] "philips"
```

```
(  my_dataset$company[grepl("^ak", my_dataset$company)] <- "akzo" )
```

```
## [1] "akzo"
```

```
(  my_dataset$company[grepl("^van", my_dataset$company)] <- "van houten" )
```

```
## [1] "van houten"
```

```
(  my_dataset$company[grepl("ver$", my_dataset$company)] <- "unilever" )
```

```
## [1] "unilever"
```

2. Separate product code and product number

```
my_dataset <- separate( data= my_dataset, col=Product.code...number
            , into = c("product_code", "product_number"), sep="-")
str(my_dataset)
```

```
## 'data.frame':    25 obs. of  7 variables:
##  $ company       : chr  "philips" "philips" "philips" "philips" ...
##  $ product_code  : chr  "p" "p" "x" "x" ...
##  $ product_number: chr  "5" "43" "3" "34" ...
##  $ address       : Factor w/ 25 levels "Delfzijlstraat 54",..: 9 10 11 12 13 14 19 20 21 22 ...
##  $ city          : Factor w/ 1 level "arnhem": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country       : Factor w/ 1 level "the netherlands": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name          : Factor w/ 20 levels "dhr j. Gansen",..: 7 6 1 9 4 5 2 10 3 8 ...
```

3.Add product category

```
get_product_category <- function(productCode){
    productCode <- tolower(productCode)
    if( productCode == "p"){
      return("SmartPhone")
    }else if(productCode == "v"){
      return("TV")
    }
    else if(productCode == "x"){
      return("Laptop")
    }else if(productCode == "q"){
      return("Tablet")
    }
}

my_dataset <- my_dataset %>%
  mutate( product_category = sapply(product_code, get_product_category)  )
str(my_dataset)
```

```
## 'data.frame':    25 obs. of  8 variables:
##  $ company         : chr  "philips" "philips" "philips" "philips" ...
##  $ product_code    : chr  "p" "p" "x" "x" ...
##  $ product_number  : chr  "5" "43" "3" "34" ...
##  $ address         : Factor w/ 25 levels "Delfzijlstraat 54",..: 9 10 11 12 13 14 19 20 21 22 ...
##  $ city            : Factor w/ 1 level "arnhem": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country         : Factor w/ 1 level "the netherlands": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name            : Factor w/ 20 levels "dhr j. Gansen",..: 7 6 1 9 4 5 2 10 3 8 ...
##  $ product_category: Named chr  "SmartPhone" "SmartPhone" "Laptop" "Laptop" ...
##   ..- attr(*, "names")= chr  "p" "p" "x" "x" ...
```

4. Geocode addresses

```
my_dataset <- my_dataset %>%
  mutate( full_address = paste(address ,city,country, sep= ", " )  )
str(my_dataset)
```

```
## 'data.frame':    25 obs. of  9 variables:
##  $ company         : chr  "philips" "philips" "philips" "philips" ...
##  $ product_code    : chr  "p" "p" "x" "x" ...
##  $ product_number  : chr  "5" "43" "3" "34" ...
##  $ address         : Factor w/ 25 levels "Delfzijlstraat 54",..: 9 10 11 12 13 14 19 20 21 22 ...
##  $ city            : Factor w/ 1 level "arnhem": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country         : Factor w/ 1 level "the netherlands": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name            : Factor w/ 20 levels "dhr j. Gansen",..: 7 6 1 9 4 5 2 10 3 8 ...
##  $ product_category: chr  "SmartPhone" "SmartPhone" "Laptop" "Laptop" ...
##  $ full_address    : chr  "Groningensingel 147, arnhem, the netherlands" "Groningensingel 148, arnhe
```

5. Create dummy variables for company and category columns

```
#Keeping copies of original columns for verification
my_dataset <- data.frame(append(my_dataset
                            ,list(product = my_dataset$product_category )
                            ,after=match("product_category", names(my_dataset))))
my_dataset <- data.frame(append(my_dataset
                  ,list(brand = my_dataset$company )
                  ,after=0))

my_dataset <- dummy.data.frame( names=c("company","product")
                                , data = my_dataset, sep="_", drop=TRUE
                                , fun=as.integer, verbose=FALSE)
str( my_dataset)
```

```
## 'data.frame':    25 obs. of  17 variables:
##  $ brand             : Factor w/ 4 levels "akzo","philips",..: 2 2 2 2 2 2 1 1 1 1 ...
##  $ company_akzo      : int  0 0 0 0 0 0 1 1 1 1 ...
##  $ company_philips   : int  1 1 1 1 1 1 0 0 0 0 ...
##  $ company_unilever  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ company_van houten: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ product_code      : Factor w/ 4 levels "p","q","v","x": 1 1 4 4 4 1 3 3 4 1 ...
##  $ product_number    : Factor w/ 15 levels "12","21","23",..: 9 7 4 5 1 3 7 1 9 5 ...
##  $ address           : Factor w/ 25 levels "Delfzijlstraat 54",..: 9 10 11 12 13 14 19 20 21 22 ...
##  $ city              : Factor w/ 1 level "arnhem": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country           : Factor w/ 1 level "the netherlands": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name              : Factor w/ 20 levels "dhr j. Gansen",..: 7 6 1 9 4 5 2 10 3 8 ...
##  $ product_category  : Factor w/ 4 levels "Laptop","SmartPhone",..: 2 2 1 1 1 2 4 4 1 2 ...
##  $ product_Laptop    : int  0 0 1 1 1 0 0 0 1 0 ...
##  $ product_SmartPhone: int  1 1 0 0 0 1 0 0 0 1 ...
##  $ product_Tablet    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ product_TV        : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ full_address      : Factor w/ 25 levels "Delfzijlstraat 54, arnhem, the netherlands",..: 9 10 11
##  - attr(*, "dummies")=List of 2
##   ..$ company: int  2 3 4 5
##   ..$ product: int  13 14 15 16
```

Output

```r
kable(head(my_dataset[,1:6]) , format = "markdown", caption = "Refine dataset cleaned")
```

brand	company_akzo	company_philips	company_unilever	company_van houten	product_code
philips	0	1	0	0	p
philips	0	1	0	0	p
philips	0	1	0	0	x
philips	0	1	0	0	x
philips	0	1	0	0	x
philips	0	1	0	0	p

```r
kable(head(my_dataset[,7:12]) , format = "markdown")
```

product_number	address	city	country	name	product_ca
5	Groningensingel 147	arnhem	the netherlands	dhr p. jansen	SmartPhon
43	Groningensingel 148	arnhem	the netherlands	dhr p. hansen	SmartPhon
3	Groningensingel 149	arnhem	the netherlands	dhr j. Gansen	Laptop
34	Groningensingel 150	arnhem	the netherlands	dhr p. mansen	Laptop
12	Groningensingel 151	arnhem	the netherlands	dhr p. fransen	Laptop
23	Groningensingel 152	arnhem	the netherlands	dhr p. franssen	SmartPhon

```r
kable(head(my_dataset[,13:17]) , format = "markdown")
```

product_Laptop	product_SmartPhone	product_Tablet	product_TV	full_address
0	1	0	0	Groningensingel 147, arnhem, t
0	1	0	0	Groningensingel 148, arnhem, t
1	0	0	0	Groningensingel 149, arnhem, t
1	0	0	0	Groningensingel 150, arnhem, t
1	0	0	0	Groningensingel 151, arnhem, t

product_Laptop	product_SmartPhone	product_Tablet	product_TV	full_address
0	1	0	0	Groningensingel 152, arnhem, t

```r
write.csv(my_dataset, file = "refine_clean.csv"
        , row.names = FALSE, append = FALSE)
```

```
## Warning in write.csv(my_dataset, file = "refine_clean.csv", row.names =
## FALSE, : attempt to set 'append' ignored
```