Nirmal Patel Data Cleaning R

#will be used to filter data

library (dplyr)

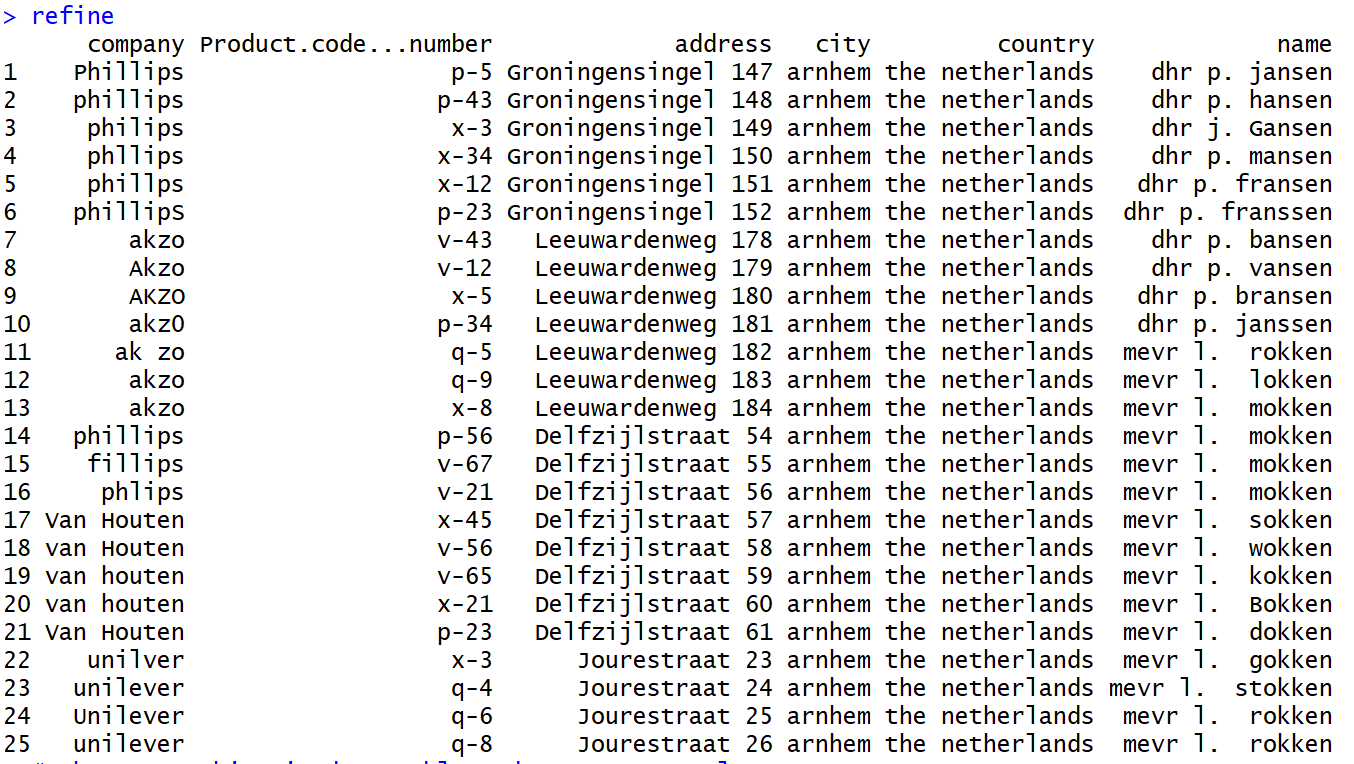
library (tidyr)

#easy way for me to find where my data is stored

refine<-read.csv(file.choose(),header=T)

#i can see the raw data

refine



#makes everything in data table under company to lowercase

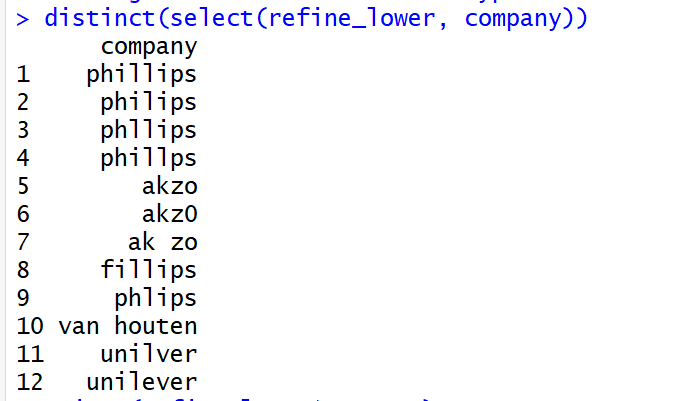
refine\_lower<-mutate (refine, company=tolower(company))

refine\_lower

#this gives the the different types of names i will have to make into 4

distinct(select(refine\_lower, company))

unique(refine\_lower$company)



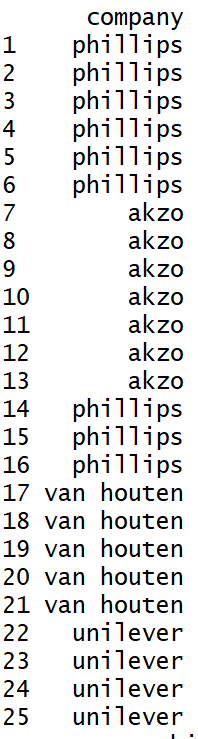
#refines the names into similar names for the same company (<http://astrostatistics.psu.edu/su07/R/html/base/html/grep.html>)

refine\_lower$company <- gsub(glob2rx("\*ps\*"),"phillips",refine\_lower$company)

refine\_lower$company <- gsub(glob2rx("\*ak\*$"),"akzo",refine\_lower$company)

refine\_lower$company <- gsub(glob2rx("\*van\*$"),"van houten",refine\_lower$company)

refine\_lower$company <- gsub(glob2rx("\*uni\*$"),"unilever",refine\_lower$company)



#YES We have to made it into 4

distinct(select(refine\_lower, company))

# Seperate product\_code and product\_number

refine\_lower <- refine\_lower %>% separate(Product.code...number, c("product\_code", "product\_number"), sep = "-")

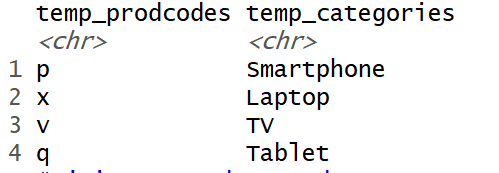
# We are makeing names of the product categories

temp\_prodcodes <- c("p","x","v","q")

temp\_categories <- c("Smartphone", "Laptop", "TV", "Tablet")

temp\_df <- tbl\_df(data\_frame(temp\_prodcodes, temp\_categories))

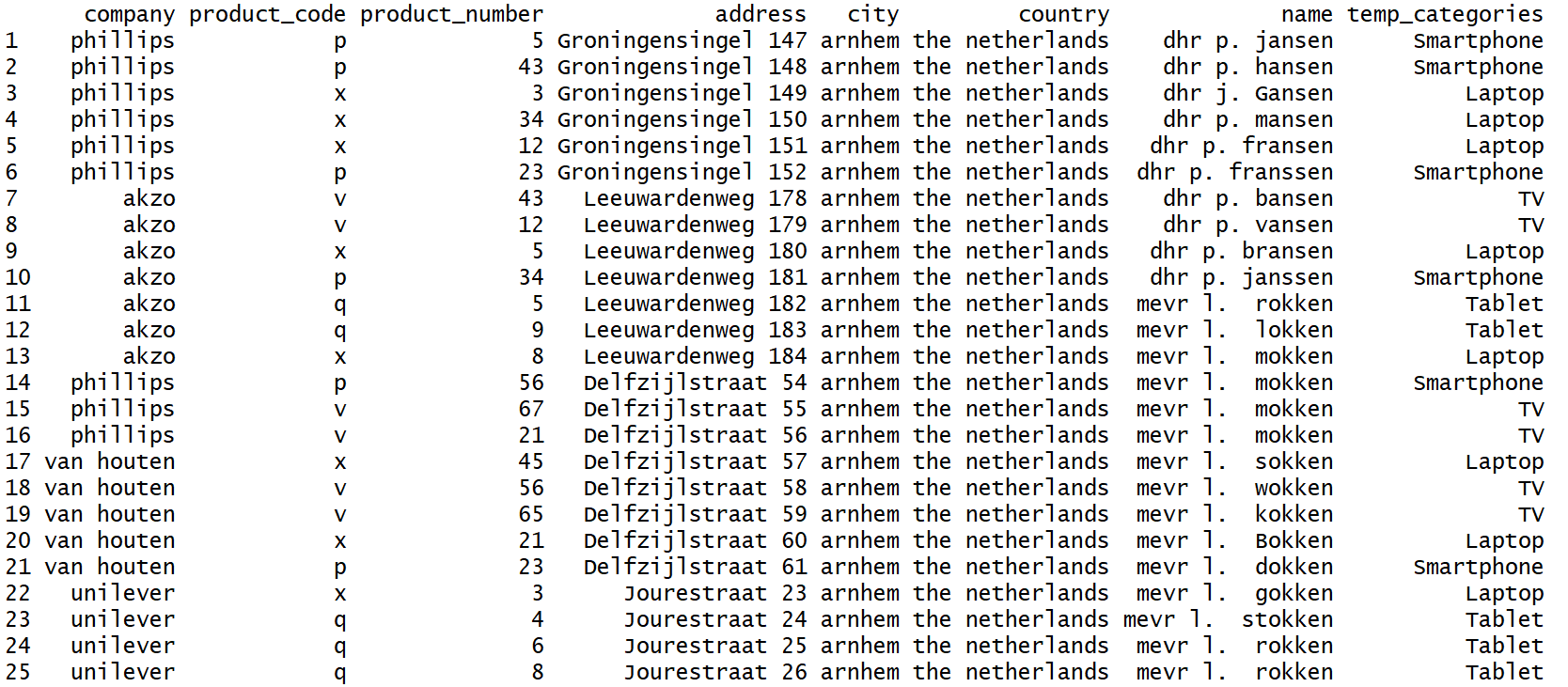
print(temp\_df)



# join on product code

refine\_lower <- inner\_join(refine\_lower, temp\_df, by = c("product\_code" = "temp\_prodcodes"))

refine\_lower



# rename temp\_categories column

refine\_lower <- rename(refine\_lower, product\_category = temp\_categories)

#4: Add full address for geocoding by address, city, country

refine\_lower$full\_address<- paste(refine\_lower$address,refine\_lower$city,refine\_lower$country, sep=",")

refine\_lower



#5: Create dummy variables for company and product category

refine\_lower <-refine\_lower %>% mutate(company\_phillips = ifelse(company == "phillips", 1, 0),

company\_akzo = ifelse(company == "akzo", 1, 0),

company\_van\_houten = ifelse(company == "van houten", 1, 0),

company\_unilever = ifelse(company == "unilever", 1, 0),

product\_smartphone = ifelse(product\_category == "Smartphone", 1, 0),

product\_tv = ifelse(product\_category == "TV", 1, 0),

product\_laptop = ifelse(product\_category == "Laptop", 1, 0),

product\_tablet = ifelse(product\_category == "Tablet", 1, 0))

refine\_lower



