**Exercise**

Using R, clean this data set to make it easier to visualize and analyze. Specifically, these are the tasks you need to do:

**0: Load the data in RStudio**

Save the data set as a CSV file called refine\_original.csv and load it in RStudio into a data frame.

I downloaded the excel file and then renamed it as refine\_original and saved it as a csv file on my Desktop. I then opened Rstudio, on the right top found “Import Dataset”>”From Local File”>”Desktop” and found the file refine\_original\_csv. This command automatically came into the console window:

>refine\_original < -read.csv("C:/Users/srivats/Desktop/refine\_original.csv")

>View(refine\_original)

I then typed refine\_original and it opened up the table as a text format in the console window itself when I was testing to see if it had opened.

**1: Clean up brand names**

Clean up the 'company' column, so all of the misspellings of the brand names are standardized. For example, you can transform the values in the column to be: philips, akzo, van houten and unilever (all lowercase).

I first started by typing install.packages("dplyr") and install.packages("tidyr") in Rconsole.

**2: Separate product code and number – Getting an error!!**

Separate the product code and product number into separate columns i.e. add two new columns called *product\_code* and *product\_number*, containing the product code and number respectively.

> library("tidyr")

> separate (refine\_original, 'Product code/number', into=c("Productcode", "Product number"), sep="-", remove=TRUE, convert=FALSE, extra="warn", fill="warn")

I’m getting an error:

Error in if (!after) c(values, x) else if (after >= lengx) c(x, values) else c(x[1L:after], :

argument is of length zero

**3: Add product categorie**s

You learn that the product codes actually represent the following product categories:

* p = Smartphone
* v = TV
* x = Laptop
* q = Tablet

In order to make the data more readable, add a column with the product category for each record.

**4: Add full address for geocoding**

You'd like to view the customer information on a map. In order to do that, the addresses need to be in a form that can be easily geocoded. Create a new column *full\_address* that concatenates the three address fields (*address, city, country*), separated by commas.

**5: Create dummy variables for company and product category**

Both the company name and product category are categorical variables i.e. they take only a fixed set of values. In order to use them in further analysis you need to create dummy variables. Create dummy binary variables for each of them with the prefix *company\_* and *product\_* i.e.

1. Add four binary (1 or 0) columns for company: *company\_philips, company\_akzo, company\_van\_houten* and *company\_unilever*
2. Add four binary (1 or 0) columns for product category:*product\_smartphone, product\_tv, product\_laptop* and *product\_tablet*