Data Wrangling Exercise 1

Shafeeq Ur Rahaman

July 06 2017

knitr::opts\_chunk$set(echo = TRUE)

library(dplyr)  
library(tidyr)  
library(readr)  
library(plyr)  
library(magrittr)

#### Requirement over here is to clean the dataset provided in order to make it easier to visualize and analyze.

# Step 0: Load the data in RStudio

#### Loading the CSV file that needs to cleaned

refine\_original <- read.csv("refine\_original.csv", header = TRUE)

## company Product.code...number address city  
## 1 Phillips p-5 Groningensingel 147 arnhem  
## 2 phillips p-43 Groningensingel 148 arnhem  
## 3 philips x-3 Groningensingel 149 arnhem  
## 4 phllips x-34 Groningensingel 150 arnhem  
## 5 phillps x-12 Groningensingel 151 arnhem  
## 6 phillipS p-23 Groningensingel 152 arnhem  
## 7 akzo v-43 Leeuwardenweg 178 arnhem  
## 8 Akzo v-12 Leeuwardenweg 179 arnhem  
## 9 AKZO x-5 Leeuwardenweg 180 arnhem  
## 10 akz0 p-34 Leeuwardenweg 181 arnhem  
## 11 ak zo q-5 Leeuwardenweg 182 arnhem  
## 12 akzo q-9 Leeuwardenweg 183 arnhem  
## 13 akzo x-8 Leeuwardenweg 184 arnhem  
## 14 phillips p-56 Delfzijlstraat 54 arnhem  
## 15 fillips v-67 Delfzijlstraat 55 arnhem  
## 16 phlips v-21 Delfzijlstraat 56 arnhem  
## 17 Van Houten x-45 Delfzijlstraat 57 arnhem  
## 18 van Houten v-56 Delfzijlstraat 58 arnhem  
## 19 van houten v-65 Delfzijlstraat 59 arnhem  
## 20 van houten x-21 Delfzijlstraat 60 arnhem  
## 21 Van Houten p-23 Delfzijlstraat 61 arnhem  
## 22 unilver x-3 Jourestraat 23 arnhem  
## 23 unilever q-4 Jourestraat 24 arnhem  
## 24 Unilever q-6 Jourestraat 25 arnhem  
## 25 unilever q-8 Jourestraat 26 arnhem  
## country name  
## 1 the netherlands dhr p. jansen  
## 2 the netherlands dhr p. hansen  
## 3 the netherlands dhr j. Gansen  
## 4 the netherlands dhr p. mansen  
## 5 the netherlands dhr p. fransen  
## 6 the netherlands dhr p. franssen  
## 7 the netherlands dhr p. bansen  
## 8 the netherlands dhr p. vansen  
## 9 the netherlands dhr p. bransen  
## 10 the netherlands dhr p. janssen  
## 11 the netherlands mevr l. rokken  
## 12 the netherlands mevr l. lokken  
## 13 the netherlands mevr l. mokken  
## 14 the netherlands mevr l. mokken  
## 15 the netherlands mevr l. mokken  
## 16 the netherlands mevr l. mokken  
## 17 the netherlands mevr l. sokken  
## 18 the netherlands mevr l. wokken  
## 19 the netherlands mevr l. kokken  
## 20 the netherlands mevr l. Bokken  
## 21 the netherlands mevr l. dokken  
## 22 the netherlands mevr l. gokken  
## 23 the netherlands mevr l. stokken  
## 24 the netherlands mevr l. rokken  
## 25 the netherlands mevr l. rokken

# Step 1: Clean up brand names

#### Cleaning the company column for misspellings

refine\_original$company %<>% tolower()  
refine\_original %<>% mutate(company = replace(company, company %in% c('phillips','phllips','philips','phillps','fillips','phlips'), 'philips'))  
refine\_original %<>% mutate(company = replace(company, company %in% c('akzo','akz0','ak zo'), 'akzo'))  
refine\_original %<>% mutate(company = replace(company, company %in% c('unilver','unilever'), 'unilever'))

## company Product.code...number address city  
## 1 philips p-5 Groningensingel 147 arnhem  
## 2 philips p-43 Groningensingel 148 arnhem  
## 3 philips x-3 Groningensingel 149 arnhem  
## 4 philips x-34 Groningensingel 150 arnhem  
## 5 philips x-12 Groningensingel 151 arnhem  
## 6 philips p-23 Groningensingel 152 arnhem  
## 7 akzo v-43 Leeuwardenweg 178 arnhem  
## 8 akzo v-12 Leeuwardenweg 179 arnhem  
## 9 akzo x-5 Leeuwardenweg 180 arnhem  
## 10 akzo p-34 Leeuwardenweg 181 arnhem  
## 11 akzo q-5 Leeuwardenweg 182 arnhem  
## 12 akzo q-9 Leeuwardenweg 183 arnhem  
## 13 akzo x-8 Leeuwardenweg 184 arnhem  
## 14 philips p-56 Delfzijlstraat 54 arnhem  
## 15 philips v-67 Delfzijlstraat 55 arnhem  
## 16 philips v-21 Delfzijlstraat 56 arnhem  
## 17 van houten x-45 Delfzijlstraat 57 arnhem  
## 18 van houten v-56 Delfzijlstraat 58 arnhem  
## 19 van houten v-65 Delfzijlstraat 59 arnhem  
## 20 van houten x-21 Delfzijlstraat 60 arnhem  
## 21 van houten p-23 Delfzijlstraat 61 arnhem  
## 22 unilever x-3 Jourestraat 23 arnhem  
## 23 unilever q-4 Jourestraat 24 arnhem  
## 24 unilever q-6 Jourestraat 25 arnhem  
## 25 unilever q-8 Jourestraat 26 arnhem  
## country name  
## 1 the netherlands dhr p. jansen  
## 2 the netherlands dhr p. hansen  
## 3 the netherlands dhr j. Gansen  
## 4 the netherlands dhr p. mansen  
## 5 the netherlands dhr p. fransen  
## 6 the netherlands dhr p. franssen  
## 7 the netherlands dhr p. bansen  
## 8 the netherlands dhr p. vansen  
## 9 the netherlands dhr p. bransen  
## 10 the netherlands dhr p. janssen  
## 11 the netherlands mevr l. rokken  
## 12 the netherlands mevr l. lokken  
## 13 the netherlands mevr l. mokken  
## 14 the netherlands mevr l. mokken  
## 15 the netherlands mevr l. mokken  
## 16 the netherlands mevr l. mokken  
## 17 the netherlands mevr l. sokken  
## 18 the netherlands mevr l. wokken  
## 19 the netherlands mevr l. kokken  
## 20 the netherlands mevr l. Bokken  
## 21 the netherlands mevr l. dokken  
## 22 the netherlands mevr l. gokken  
## 23 the netherlands mevr l. stokken  
## 24 the netherlands mevr l. rokken  
## 25 the netherlands mevr l. rokken

# Step 2: Separate Product Code and Number

#### Divide the 'Product.Code...Number' variable into two new variables

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

## company product\_code product\_number address city  
## 1 philips p 5 Groningensingel 147 arnhem  
## 2 philips p 43 Groningensingel 148 arnhem  
## 3 philips x 3 Groningensingel 149 arnhem  
## 4 philips x 34 Groningensingel 150 arnhem  
## 5 philips x 12 Groningensingel 151 arnhem  
## 6 philips p 23 Groningensingel 152 arnhem  
## 7 akzo v 43 Leeuwardenweg 178 arnhem  
## 8 akzo v 12 Leeuwardenweg 179 arnhem  
## 9 akzo x 5 Leeuwardenweg 180 arnhem  
## 10 akzo p 34 Leeuwardenweg 181 arnhem  
## 11 akzo q 5 Leeuwardenweg 182 arnhem  
## 12 akzo q 9 Leeuwardenweg 183 arnhem  
## 13 akzo x 8 Leeuwardenweg 184 arnhem  
## 14 philips p 56 Delfzijlstraat 54 arnhem  
## 15 philips v 67 Delfzijlstraat 55 arnhem  
## 16 philips v 21 Delfzijlstraat 56 arnhem  
## 17 van houten x 45 Delfzijlstraat 57 arnhem  
## 18 van houten v 56 Delfzijlstraat 58 arnhem  
## 19 van houten v 65 Delfzijlstraat 59 arnhem  
## 20 van houten x 21 Delfzijlstraat 60 arnhem  
## 21 van houten p 23 Delfzijlstraat 61 arnhem  
## 22 unilever x 3 Jourestraat 23 arnhem  
## 23 unilever q 4 Jourestraat 24 arnhem  
## 24 unilever q 6 Jourestraat 25 arnhem  
## 25 unilever q 8 Jourestraat 26 arnhem  
## country name  
## 1 the netherlands dhr p. jansen  
## 2 the netherlands dhr p. hansen  
## 3 the netherlands dhr j. Gansen  
## 4 the netherlands dhr p. mansen  
## 5 the netherlands dhr p. fransen  
## 6 the netherlands dhr p. franssen  
## 7 the netherlands dhr p. bansen  
## 8 the netherlands dhr p. vansen  
## 9 the netherlands dhr p. bransen  
## 10 the netherlands dhr p. janssen  
## 11 the netherlands mevr l. rokken  
## 12 the netherlands mevr l. lokken  
## 13 the netherlands mevr l. mokken  
## 14 the netherlands mevr l. mokken  
## 15 the netherlands mevr l. mokken  
## 16 the netherlands mevr l. mokken  
## 17 the netherlands mevr l. sokken  
## 18 the netherlands mevr l. wokken  
## 19 the netherlands mevr l. kokken  
## 20 the netherlands mevr l. Bokken  
## 21 the netherlands mevr l. dokken  
## 22 the netherlands mevr l. gokken  
## 23 the netherlands mevr l. stokken  
## 24 the netherlands mevr l. rokken  
## 25 the netherlands mevr l. rokken

# Step 3: Adding new variable to the dataset

#### Add Product Categories variable to the dataset that abbreviate the codes in product code variable.

refine\_original %<>% mutate(product\_category = case\_when(refine\_original$product\_code == 'p' ~ 'Smartphone', refine\_original$product\_code == 'v' ~ 'TV', refine\_original$product\_code =='x' ~ 'Laptop', refine\_original$product\_code == 'q' ~ 'Tablet'))

## company product\_code product\_number address city  
## 1 philips p 5 Groningensingel 147 arnhem  
## 2 philips p 43 Groningensingel 148 arnhem  
## 3 philips x 3 Groningensingel 149 arnhem  
## 4 philips x 34 Groningensingel 150 arnhem  
## 5 philips x 12 Groningensingel 151 arnhem  
## 6 philips p 23 Groningensingel 152 arnhem  
## 7 akzo v 43 Leeuwardenweg 178 arnhem  
## 8 akzo v 12 Leeuwardenweg 179 arnhem  
## 9 akzo x 5 Leeuwardenweg 180 arnhem  
## 10 akzo p 34 Leeuwardenweg 181 arnhem  
## 11 akzo q 5 Leeuwardenweg 182 arnhem  
## 12 akzo q 9 Leeuwardenweg 183 arnhem  
## 13 akzo x 8 Leeuwardenweg 184 arnhem  
## 14 philips p 56 Delfzijlstraat 54 arnhem  
## 15 philips v 67 Delfzijlstraat 55 arnhem  
## 16 philips v 21 Delfzijlstraat 56 arnhem  
## 17 van houten x 45 Delfzijlstraat 57 arnhem  
## 18 van houten v 56 Delfzijlstraat 58 arnhem  
## 19 van houten v 65 Delfzijlstraat 59 arnhem  
## 20 van houten x 21 Delfzijlstraat 60 arnhem  
## 21 van houten p 23 Delfzijlstraat 61 arnhem  
## 22 unilever x 3 Jourestraat 23 arnhem  
## 23 unilever q 4 Jourestraat 24 arnhem  
## 24 unilever q 6 Jourestraat 25 arnhem  
## 25 unilever q 8 Jourestraat 26 arnhem  
## country name product\_category  
## 1 the netherlands dhr p. jansen Smartphone  
## 2 the netherlands dhr p. hansen Smartphone  
## 3 the netherlands dhr j. Gansen Laptop  
## 4 the netherlands dhr p. mansen Laptop  
## 5 the netherlands dhr p. fransen Laptop  
## 6 the netherlands dhr p. franssen Smartphone  
## 7 the netherlands dhr p. bansen TV  
## 8 the netherlands dhr p. vansen TV  
## 9 the netherlands dhr p. bransen Laptop  
## 10 the netherlands dhr p. janssen Smartphone  
## 11 the netherlands mevr l. rokken Tablet  
## 12 the netherlands mevr l. lokken Tablet  
## 13 the netherlands mevr l. mokken Laptop  
## 14 the netherlands mevr l. mokken Smartphone  
## 15 the netherlands mevr l. mokken TV  
## 16 the netherlands mevr l. mokken TV  
## 17 the netherlands mevr l. sokken Laptop  
## 18 the netherlands mevr l. wokken TV  
## 19 the netherlands mevr l. kokken TV  
## 20 the netherlands mevr l. Bokken Laptop  
## 21 the netherlands mevr l. dokken Smartphone  
## 22 the netherlands mevr l. gokken Laptop  
## 23 the netherlands mevr l. stokken Tablet  
## 24 the netherlands mevr l. rokken Tablet  
## 25 the netherlands mevr l. rokken Tablet

# Step 4: Add full address for geocoding

#### Combine all three variables, address, city, country to form a new variable Full Address

refine\_original %<>% unite(full\_address, c(address, city, country),sep = ',',remove=TRUE)

## company product\_code product\_number  
## 1 philips p 5  
## 2 philips p 43  
## 3 philips x 3  
## 4 philips x 34  
## 5 philips x 12  
## 6 philips p 23  
## 7 akzo v 43  
## 8 akzo v 12  
## 9 akzo x 5  
## 10 akzo p 34  
## 11 akzo q 5  
## 12 akzo q 9  
## 13 akzo x 8  
## 14 philips p 56  
## 15 philips v 67  
## 16 philips v 21  
## 17 van houten x 45  
## 18 van houten v 56  
## 19 van houten v 65  
## 20 van houten x 21  
## 21 van houten p 23  
## 22 unilever x 3  
## 23 unilever q 4  
## 24 unilever q 6  
## 25 unilever q 8  
## full\_address name  
## 1 Groningensingel 147,arnhem,the netherlands dhr p. jansen  
## 2 Groningensingel 148,arnhem,the netherlands dhr p. hansen  
## 3 Groningensingel 149,arnhem,the netherlands dhr j. Gansen  
## 4 Groningensingel 150,arnhem,the netherlands dhr p. mansen  
## 5 Groningensingel 151,arnhem,the netherlands dhr p. fransen  
## 6 Groningensingel 152,arnhem,the netherlands dhr p. franssen  
## 7 Leeuwardenweg 178,arnhem,the netherlands dhr p. bansen  
## 8 Leeuwardenweg 179,arnhem,the netherlands dhr p. vansen  
## 9 Leeuwardenweg 180,arnhem,the netherlands dhr p. bransen  
## 10 Leeuwardenweg 181,arnhem,the netherlands dhr p. janssen  
## 11 Leeuwardenweg 182,arnhem,the netherlands mevr l. rokken  
## 12 Leeuwardenweg 183,arnhem,the netherlands mevr l. lokken  
## 13 Leeuwardenweg 184,arnhem,the netherlands mevr l. mokken  
## 14 Delfzijlstraat 54,arnhem,the netherlands mevr l. mokken  
## 15 Delfzijlstraat 55,arnhem,the netherlands mevr l. mokken  
## 16 Delfzijlstraat 56,arnhem,the netherlands mevr l. mokken  
## 17 Delfzijlstraat 57,arnhem,the netherlands mevr l. sokken  
## 18 Delfzijlstraat 58,arnhem,the netherlands mevr l. wokken  
## 19 Delfzijlstraat 59,arnhem,the netherlands mevr l. kokken  
## 20 Delfzijlstraat 60,arnhem,the netherlands mevr l. Bokken  
## 21 Delfzijlstraat 61,arnhem,the netherlands mevr l. dokken  
## 22 Jourestraat 23,arnhem,the netherlands mevr l. gokken  
## 23 Jourestraat 24,arnhem,the netherlands mevr l. stokken  
## 24 Jourestraat 25,arnhem,the netherlands mevr l. rokken  
## 25 Jourestraat 26,arnhem,the netherlands mevr l. rokken  
## product\_category  
## 1 Smartphone  
## 2 Smartphone  
## 3 Laptop  
## 4 Laptop  
## 5 Laptop  
## 6 Smartphone  
## 7 TV  
## 8 TV  
## 9 Laptop  
## 10 Smartphone  
## 11 Tablet  
## 12 Tablet  
## 13 Laptop  
## 14 Smartphone  
## 15 TV  
## 16 TV  
## 17 Laptop  
## 18 TV  
## 19 TV  
## 20 Laptop  
## 21 Smartphone  
## 22 Laptop  
## 23 Tablet  
## 24 Tablet  
## 25 Tablet

# Step 5: Create dummy variables for company and product category.

#### Create dummy variables for company and product category as they are categorical variables.

refine\_original %<>% mutate(company\_philips = (if\_else(grepl(pattern='philips', x=refine\_original$company), 1, 0)))  
refine\_original %<>% mutate(company\_akzo = (if\_else(grepl(pattern='akzo', x=refine\_original$company), 1, 0)))  
refine\_original %<>% mutate(company\_vanhouten = (if\_else(grepl(pattern='van houten', x=refine\_original$company), 1, 0)))  
refine\_original %<>% mutate(company\_unilever = (if\_else(grepl(pattern='unilever', x=refine\_original$company), 1, 0)))  
refine\_original %<>% mutate(product\_smartphone = (if\_else(grepl(pattern='p', x=refine\_original$product\_code), 1, 0)))  
refine\_original %<>% mutate(product\_laptop = (if\_else(grepl(pattern='x', x=refine\_original$product\_code), 1, 0)))  
refine\_original %<>% mutate(product\_TV = (if\_else(grepl(pattern='v', x=refine\_original$product\_code), 1, 0)))  
refine\_original %<>% mutate(product\_Tablet = (if\_else(grepl(pattern='q', x=refine\_original$product\_code), 1, 0)))

## company product\_code product\_number  
## 1 philips p 5  
## 2 philips p 43  
## 3 philips x 3  
## 4 philips x 34  
## 5 philips x 12  
## 6 philips p 23  
## 7 akzo v 43  
## 8 akzo v 12  
## 9 akzo x 5  
## 10 akzo p 34  
## 11 akzo q 5  
## 12 akzo q 9  
## 13 akzo x 8  
## 14 philips p 56  
## 15 philips v 67  
## 16 philips v 21  
## 17 van houten x 45  
## 18 van houten v 56  
## 19 van houten v 65  
## 20 van houten x 21  
## 21 van houten p 23  
## 22 unilever x 3  
## 23 unilever q 4  
## 24 unilever q 6  
## 25 unilever q 8  
## full\_address name  
## 1 Groningensingel 147,arnhem,the netherlands dhr p. jansen  
## 2 Groningensingel 148,arnhem,the netherlands dhr p. hansen  
## 3 Groningensingel 149,arnhem,the netherlands dhr j. Gansen  
## 4 Groningensingel 150,arnhem,the netherlands dhr p. mansen  
## 5 Groningensingel 151,arnhem,the netherlands dhr p. fransen  
## 6 Groningensingel 152,arnhem,the netherlands dhr p. franssen  
## 7 Leeuwardenweg 178,arnhem,the netherlands dhr p. bansen  
## 8 Leeuwardenweg 179,arnhem,the netherlands dhr p. vansen  
## 9 Leeuwardenweg 180,arnhem,the netherlands dhr p. bransen  
## 10 Leeuwardenweg 181,arnhem,the netherlands dhr p. janssen  
## 11 Leeuwardenweg 182,arnhem,the netherlands mevr l. rokken  
## 12 Leeuwardenweg 183,arnhem,the netherlands mevr l. lokken  
## 13 Leeuwardenweg 184,arnhem,the netherlands mevr l. mokken  
## 14 Delfzijlstraat 54,arnhem,the netherlands mevr l. mokken  
## 15 Delfzijlstraat 55,arnhem,the netherlands mevr l. mokken  
## 16 Delfzijlstraat 56,arnhem,the netherlands mevr l. mokken  
## 17 Delfzijlstraat 57,arnhem,the netherlands mevr l. sokken  
## 18 Delfzijlstraat 58,arnhem,the netherlands mevr l. wokken  
## 19 Delfzijlstraat 59,arnhem,the netherlands mevr l. kokken  
## 20 Delfzijlstraat 60,arnhem,the netherlands mevr l. Bokken  
## 21 Delfzijlstraat 61,arnhem,the netherlands mevr l. dokken  
## 22 Jourestraat 23,arnhem,the netherlands mevr l. gokken  
## 23 Jourestraat 24,arnhem,the netherlands mevr l. stokken  
## 24 Jourestraat 25,arnhem,the netherlands mevr l. rokken  
## 25 Jourestraat 26,arnhem,the netherlands mevr l. rokken  
## product\_category company\_philips company\_akzo company\_vanhouten  
## 1 Smartphone 1 0 0  
## 2 Smartphone 1 0 0  
## 3 Laptop 1 0 0  
## 4 Laptop 1 0 0  
## 5 Laptop 1 0 0  
## 6 Smartphone 1 0 0  
## 7 TV 0 1 0  
## 8 TV 0 1 0  
## 9 Laptop 0 1 0  
## 10 Smartphone 0 1 0  
## 11 Tablet 0 1 0  
## 12 Tablet 0 1 0  
## 13 Laptop 0 1 0  
## 14 Smartphone 1 0 0  
## 15 TV 1 0 0  
## 16 TV 1 0 0  
## 17 Laptop 0 0 1  
## 18 TV 0 0 1  
## 19 TV 0 0 1  
## 20 Laptop 0 0 1  
## 21 Smartphone 0 0 1  
## 22 Laptop 0 0 0  
## 23 Tablet 0 0 0  
## 24 Tablet 0 0 0  
## 25 Tablet 0 0 0  
## company\_unilever product\_smartphone product\_laptop product\_TV  
## 1 0 1 0 0  
## 2 0 1 0 0  
## 3 0 0 1 0  
## 4 0 0 1 0  
## 5 0 0 1 0  
## 6 0 1 0 0  
## 7 0 0 0 1  
## 8 0 0 0 1  
## 9 0 0 1 0  
## 10 0 1 0 0  
## 11 0 0 0 0  
## 12 0 0 0 0  
## 13 0 0 1 0  
## 14 0 1 0 0  
## 15 0 0 0 1  
## 16 0 0 0 1  
## 17 0 0 1 0  
## 18 0 0 0 1  
## 19 0 0 0 1  
## 20 0 0 1 0  
## 21 0 1 0 0  
## 22 1 0 1 0  
## 23 1 0 0 0  
## 24 1 0 0 0  
## 25 1 0 0 0  
## product\_Tablet  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## 7 0  
## 8 0  
## 9 0  
## 10 0  
## 11 1  
## 12 1  
## 13 0  
## 14 0  
## 15 0  
## 16 0  
## 17 0  
## 18 0  
## 19 0  
## 20 0  
## 21 0  
## 22 0  
## 23 1  
## 24 1  
## 25 1

# Step 6: Save as a new dataframe and submit on Github.

#### Creating the cleaned file and submitting it into Github

write.csv(refine\_original, "refine\_clean.csv")