Data Wrangling Project 1

Vamsi K Mamidi

September 24, 2017

### Loading Packages

library (tidyr)  
library (readr)   
library (dplyr)

## Load the data in RStudio

refine\_original <- read\_csv("refine\_original.csv",col\_names = TRUE)

## Parsed with column specification:  
## cols(  
## company = col\_character(),  
## `Product code / number` = col\_character(),  
## address = col\_character(),  
## city = col\_character(),  
## country = col\_character(),  
## name = col\_character()  
## )

tbl\_df(refine\_original)

## # A tibble: 25 x 6  
## company `Product code / number` address city  
## <chr> <chr> <chr> <chr>  
## 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  
## # ... with 15 more rows, and 2 more variables: country <chr>, name <chr>

## Clean up brand names

refine\_original$company<-replace(refine\_original$company,agrep("philips",refine\_original$company,max=2,fixed=FALSE,ignore.case=TRUE,value=FALSE),"philips")  
  
refine\_original$company<-replace(refine\_original$company,agrep("akzo",refine\_original$company,max=2,fixed=FALSE,ignore.case=TRUE,value=FALSE),"akzo")  
  
refine\_original$company<-replace(refine\_original$company,agrep("van houten",refine\_original$company,max=2,fixed=FALSE,ignore.case=TRUE,value=FALSE),"van houten")  
  
refine\_original$company<-replace(refine\_original$company,agrep("unilever",refine\_original$company,max=2,fixed=FALSE,ignore.case=TRUE,value=FALSE),"unilever")

### Output:

tbl\_df(refine\_original)

## # A tibble: 25 x 6  
## company `Product code / number` address city  
## <chr> <chr> <chr> <chr>  
## 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  
## # ... with 15 more rows, and 2 more variables: country <chr>, name <chr>

## Separate product code and number

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

### Output:

tbl\_df(refine\_original\_sep)

## # A tibble: 25 x 7  
## company product\_code product\_number address city  
## \* <chr> <chr> <chr> <chr> <chr>  
## 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  
## # ... with 15 more rows, and 2 more variables: country <chr>, name <chr>

## Add product categories

levels<-c("p","x","v","q")  
labels<-c("Smartphone","Laptop","TV","Tablet")  
refine\_original\_sep$product\_category<-labels[match(refine\_original\_sep$product\_code,levels)]

### Output:

refine\_original\_sep%>%select(product\_category,product\_code)

## # A tibble: 25 x 2  
## product\_category product\_code  
## \* <chr> <chr>  
## 1 Smartphone p  
## 2 Smartphone p  
## 3 Laptop x  
## 4 Laptop x  
## 5 Laptop x  
## 6 Smartphone p  
## 7 TV v  
## 8 TV v  
## 9 Laptop x  
## 10 Smartphone p  
## # ... with 15 more rows

## Add full address for geocoding

refine\_original\_add<- refine\_original\_sep %>% unite("Full\_address",address,city,country,sep=",")

### Output:

refine\_original\_add%>%select(Full\_address)

## # A tibble: 25 x 1  
## Full\_address  
## \* <chr>  
## 1 Groningensingel 147,arnhem,the netherlands  
## 2 Groningensingel 148,arnhem,the netherlands  
## 3 Groningensingel 149,arnhem,the netherlands  
## 4 Groningensingel 150,arnhem,the netherlands  
## 5 Groningensingel 151,arnhem,the netherlands  
## 6 Groningensingel 152,arnhem,the netherlands  
## 7 Leeuwardenweg 178,arnhem,the netherlands  
## 8 Leeuwardenweg 179,arnhem,the netherlands  
## 9 Leeuwardenweg 180,arnhem,the netherlands  
## 10 Leeuwardenweg 181,arnhem,the netherlands  
## # ... with 15 more rows

## Create dummy variable for company

refine\_original\_add$company\_philips<-as.numeric(agrepl("philips",refine\_original\_add$company,max=2,fixed=FALSE,ignore.case=TRUE))  
  
refine\_original\_add$company\_akzo<-as.numeric(agrepl("akzo",refine\_original\_add$company,max=2,fixed=FALSE,ignore.case=TRUE))  
  
refine\_original\_add$company\_van\_houten<-as.numeric(agrepl("van houten",refine\_original\_add$company,max=2,fixed=FALSE,ignore.case=TRUE))  
  
refine\_original\_add$company\_unilever<-as.numeric(agrepl("unilever",refine\_original\_add$company,max=2,fixed=FALSE,ignore.case=TRUE))

### Output:

refine\_original\_add%>%select(company\_philips,company\_unilever,company\_van\_houten,company\_akzo,company)

## # A tibble: 25 x 5  
## company\_philips company\_unilever company\_van\_houten company\_akzo  
## \* <dbl> <dbl> <dbl> <dbl>  
## 1 1 0 0 0  
## 2 1 0 0 0  
## 3 1 0 0 0  
## 4 1 0 0 0  
## 5 1 0 0 0  
## 6 1 0 0 0  
## 7 0 0 0 1  
## 8 0 0 0 1  
## 9 0 0 0 1  
## 10 0 0 0 1  
## # ... with 15 more rows, and 1 more variables: company <chr>

## Create dummy variable for product category

refine\_original\_add$product\_smartphone<-as.numeric(agrepl("Smartphone",refine\_original\_add$product\_category,max=2,fixed=FALSE,ignore.case=TRUE))  
  
refine\_original\_add$product\_tv<-as.numeric(agrepl("TV",refine\_original\_add$product\_category,max=0.1,fixed=TRUE,ignore.case=TRUE,costs = 1))  
  
refine\_original\_add$product\_Laptop<-as.numeric(agrepl("Laptop",refine\_original\_add$product\_category,max=2,fixed=FALSE,ignore.case=TRUE))  
  
refine\_original\_add$product\_Tablet<-as.numeric(agrepl("Tablet",refine\_original\_add$product\_category,max=2,fixed=FALSE,ignore.case=TRUE))

### Output:

refine\_original\_add%>%select(product\_Laptop,product\_category)

## # A tibble: 25 x 2  
## product\_Laptop product\_category  
## \* <dbl> <chr>  
## 1 0 Smartphone  
## 2 0 Smartphone  
## 3 1 Laptop  
## 4 1 Laptop  
## 5 1 Laptop  
## 6 0 Smartphone  
## 7 0 TV  
## 8 0 TV  
## 9 1 Laptop  
## 10 0 Smartphone  
## # ... with 15 more rows

## Write a CSV File

write\_csv(refine\_original\_add,'refine\_clean.csv')