Data Wrangling Exercise-1

Pratik Gandhi

March 6, 2016

```
# 0: Load the data:
library(rio)
library(tidyr)
library(dplyr)

## Warning: package 'dplyr' was built under R version 3.2.4

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

setwd("C:/Users/Pratik Gandhi/Downloads")
convert("refine.xlsx","refine_original.csv")

# Loading the data
refine_data <- read.csv(file = "refine_original.csv",header = TRUE)
head(refine_data,n=3)

##    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
##           country          name
## 1 the netherlands dhr p. jansen
## 2 the netherlands dhr p. hansen
## 3 the netherlands dhr j. Gansen

# 1: Clean up brand names:

refine_data[1:6,1] <- "philips"
refine_data[7:13,1] <- "akzo"
refine_data[14:16,1] <- "philips"
refine_data[17:21,1] <- "van houten"
refine_data[22:25,1] <- "unilever"
head(refine_data,n=3)
```

```
##   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
##           country          name
## 1 the netherlands dhr p. jansen
## 2 the netherlands dhr p. hansen
## 3 the netherlands dhr j. Gansen
```

```
# 2: Separate product code and number

refine_data <-
separate(refine_data,Product.code...number.,c("product_code","product_number"
),sep="-")
head(refine_data,n=3)
```

```
##   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
##           country          name
## 1 the netherlands dhr p. jansen
## 2 the netherlands dhr p. hansen
## 3 the netherlands dhr j. Gansen
```

```
# 3: Add product categories ---- p=Smartphone,v=TV,x=Laptop,q=Tabley

refine_data$product_category <- 0

y1=c("p","v","x","q")
y2=c("Smartphone","TV","Laptop","Tablet")
y=data.frame(cbind(y1,y2))
names(y)=c("product_code","product_category")

refine_data <-
refine_data[,c("company","product_code","product_category","product_number","
address","city","country","name")]

refine_data$product_category <-
y[match(refine_data$product_code,y$product_code),2]

# 4: Add full address for geocoding

refine_data <- unite(refine_data,"full_address",address,city,country,sep =
',')
head(refine_data,n=3)
```

```
##   company product_code product_category product_number
## 1 philips            p       Smartphone              5
## 2 philips            p       Smartphone             43
```

```
## 3 philips            x           Laptop              3
##                               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
```

```
# 5: Create dummy variables for company and product category

names_com <- unique(refine_data$company)
for (i in 1:length(names_com)){
  company_name <- names_com[i]
  com_colname <- paste("company", company_name, sep = "_")
  #refine_data$com_colnames <- 0
  refine_data[[paste0(com_colname)]]<- as.numeric(refine_data$company ==
company_name)
}

names_product <- unique(refine_data$product_category)
for (i in 1:length(names_product)){
  product_name <- names_product[i]
  pro_colname <- paste("product", product_name, sep = "_")
  #refine_data$com_colnames <- 0
  refine_data[[paste0(pro_colname)]]<-
as.numeric(refine_data$product_category == product_name)
}

head(refine_data,n=3)
```

```
##   company product_code product_category product_number
## 1 philips            p       Smartphone              5
## 2 philips            p       Smartphone             43
## 3 philips            x           Laptop              3
##                               full_address          name company_philips
## 1 Groningensingel 147,arnhem,the netherlands dhr p. jansen               1
## 2 Groningensingel 148,arnhem,the netherlands dhr p. hansen               1
## 3 Groningensingel 149,arnhem,the netherlands dhr j. Gansen               1
##   company_akzo company_van houten company_unilever product_Smartphone
## 1            0                 0                0                   1
## 2            0                 0                0                   1
## 3            0                 0                0                   0
##   product_Laptop product_TV product_Tablet
## 1              0          0              0
## 2              0          0              0
## 3              1          0              0
```

```
write.csv(refine_data,"C:/Users/Pratik
Gandhi/Downloads/refine_clean.csv",row.names = FALSE)
```