Data Science Exercise 1

Kathryn Gadberry

April 5, 2016

Project Set-Up

```
electronics_clean <- read.csv("electronics_clean.csv")
refine_original <- read.csv("refine_original.csv")
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## 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
```

```
require(tidyr)
```

```
## Loading required package: tidyr
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
glimpse(refine_original)
```

```
## Observations: 25
## Variables: 6
## $ company               (fctr) Phillips, phillips, philips, phllips, p...
## $ Product.code...number (fctr) p-5, p-43, x-3, x-34, x-12, p-23, v-43,...
## $ address               (fctr) Groningensingel 147, Groningensingel 14...
## $ city                  (fctr) arnhem, arnhem, arnhem, arnhem, arnhem,...
## $ country               (fctr) the netherlands, the netherlands, the n...
## $ name                  (fctr) dhr p. jansen, dhr p. hansen, dhr j. Ga...
```

Importing Data

```
refine_original <- read.csv("C:/Users/KGadberry/Desktop/DATA ANALYSIS/GitHub/Data-Wrangli
ng/refine_original.csv")
View(refine_original)
electronics_clean <- data.frame(refine_original)
suppressMessages(library(dplyr))
suppressMessages(library(tidyr))
```

Cleaning Up Brand Names

```
length(grep("^[P|p|f]+", electronics_clean$company))
```

```
## [1] 9
```

```
comp <- electronics_clean$company
new_comp <-NULL
new_comp <- vector(mode = "character", length = length(comp))
i <- 1;
for(compn in comp)
 {
 cat("\n Currently Processing:", compn);
 ifelse(grepl("^[P|p|f]+", compn),
 new_comp[i] <- "phillips",
 ifelse(grepl("^[A|a]+", compn),
        new_comp[i] <- "akzo",
        ifelse(grepl("^[Va|va]+", compn),
               new_comp[i] <- "van houten",
               ifelse(grepl("^[u|U]+", compn),
                      new_comp[i] <- "unilever","NA")
        )
 )
 );
 cat(" Replaced By ", new_comp[i]);
 i <- i + 1;
 }
```

```
##
##  Currently Processing: Phillips Replaced By  phillips
##  Currently Processing: phillips Replaced By  phillips
##  Currently Processing: philips Replaced By  phillips
##  Currently Processing: phllips Replaced By  phillips
##  Currently Processing: phillps Replaced By  phillips
##  Currently Processing: phillipS Replaced By  phillips
##  Currently Processing: akzo Replaced By  akzo
##  Currently Processing: Akzo Replaced By  akzo
##  Currently Processing: AKZO Replaced By  akzo
##  Currently Processing: akz0 Replaced By  akzo
##  Currently Processing: ak zo Replaced By  akzo
##  Currently Processing: akzo Replaced By  akzo
##  Currently Processing: akzo Replaced By  akzo
##  Currently Processing: phillips Replaced By  phillips
##  Currently Processing: fillips Replaced By  phillips
##  Currently Processing: phlips Replaced By  phillips
##  Currently Processing: Van Houten Replaced By  van houten
##  Currently Processing: van Houten Replaced By  van houten
##  Currently Processing: van houten Replaced By  van houten
##  Currently Processing: van houten Replaced By  van houten
##  Currently Processing: Van Houten Replaced By  van houten
##  Currently Processing: unilver Replaced By  unilever
##  Currently Processing: unilever Replaced By  unilever
##  Currently Processing: Unilever Replaced By  unilever
##  Currently Processing: unilever Replaced By  unilever
```

```
# Assign new variable
electronics_clean$company <- new_comp
electronics_clean$company
```

```
##  [1] "phillips"   "phillips"   "phillips"   "phillips"   "phillips"
##  [6] "phillips"   "akzo"       "akzo"       "akzo"       "akzo"
## [11] "akzo"       "akzo"       "akzo"       "phillips"   "phillips"
## [16] "phillips"   "van houten" "van houten" "van houten" "van houten"
## [21] "van houten" "unilever"   "unilever"   "unilever"   "unilever"
```

Separating Product Code and Number

Creating A Product Code Column

```
product_code <- NULL
product_code <- vector(mode = "character", length = length(product_code))
product_code <- electronics_clean$Product.code...number
product_code <- substr(product_code, 1, 1)
product_code
```

```
##  [1] "p" "p" "x" "x" "x" "p" "v" "v" "x" "p" "q" "q" "x" "p" "v" "v" "x"
## [18] "v" "v" "x" "p" "x" "q" "q" "q"
```

```
electronics_clean$product_code <- product_code
electronics_clean$product_code
```

```
##  [1] "p" "p" "x" "x" "x" "p" "v" "v" "x" "p" "q" "q" "x" "p" "v" "v" "x"
## [18] "v" "v" "x" "p" "x" "q" "q" "q"
```

Creating A Product Number Column

```
product_number <- NULL
product_number <- vector(mode = "numeric", length = length(product_number))
product_number <- electronics_clean$Product.code...number
product_number <-substr(product_number, 3, 4)
product_number
```

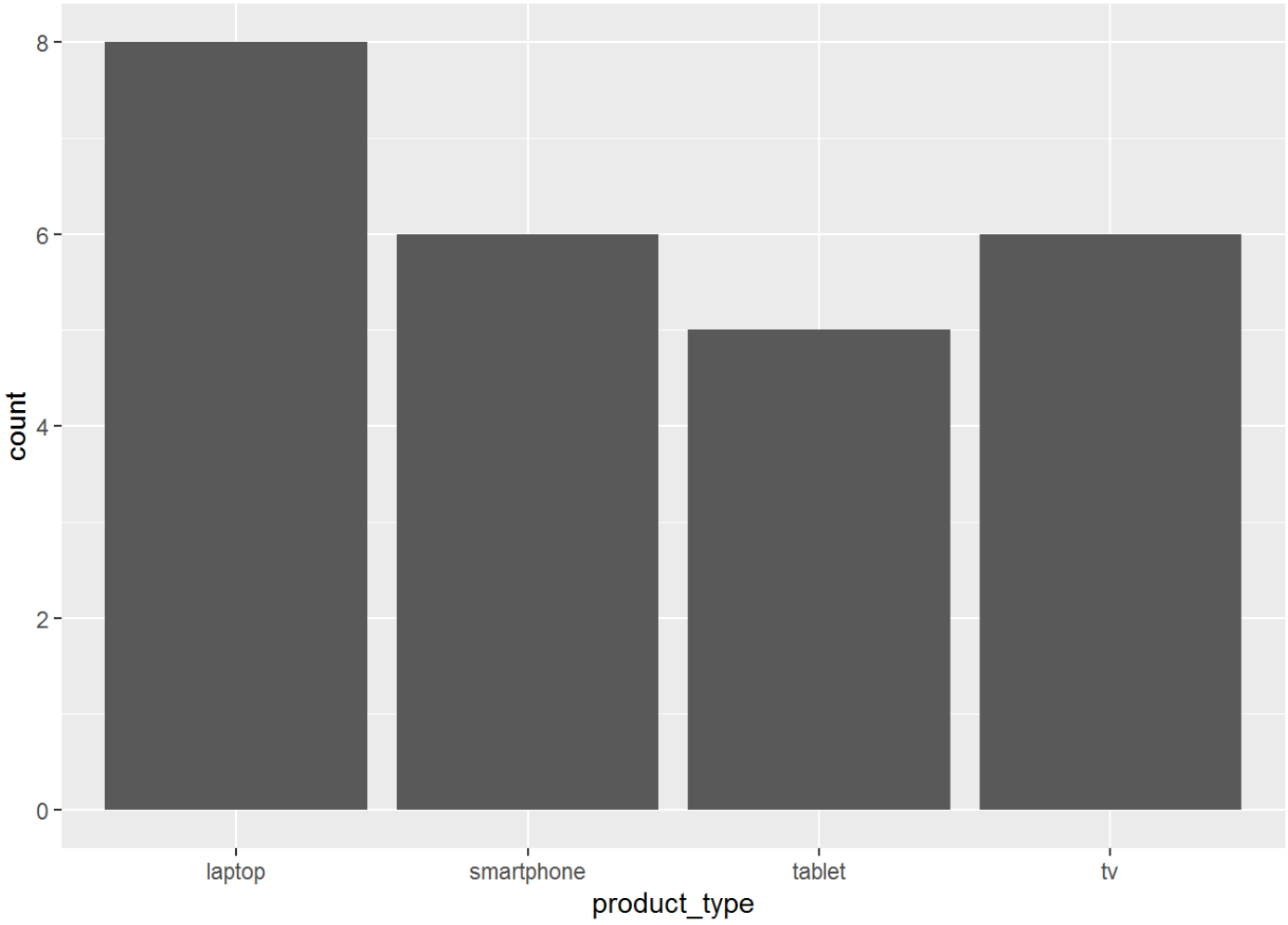
```
##  [1] "5"  "43" "3"  "34" "12" "23" "43" "12" "5"  "34" "5"  "9"  "8"  "56"
## [15] "67" "21" "45" "56" "65" "21" "23" "3"  "4"  "6"  "8"
```

Adding Product Categories

```
type <- electronics_clean$product_code
product_type <- NULL
product_type <- vector(mode = "character", length = length(product_type))
i <- 1;
for(typen in type)
{
    cat("\n Currently Processing : ", typen);
    ifelse(grepl("^[p]+", typen) ,
           product_type[i] <- "smartphone",
           ifelse(grepl("^[v]+", typen),
                  product_type[i] <- "tv",
                  ifelse(grepl("^[x]+", typen),
                         product_type[i] <- "laptop",
                         ifelse(grepl("^[q]+", typen),
                                product_type[i] <- "tablet","NA")
                  )
           )
    );
    cat(" Replaced By ", product_type[i]);
    i <- i + 1;
}
```

```
##
##  Currently Processing :  p Replaced By  smartphone
##  Currently Processing :  p Replaced By  smartphone
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  p Replaced By  smartphone
##  Currently Processing :  v Replaced By  tv
##  Currently Processing :  v Replaced By  tv
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  p Replaced By  smartphone
##  Currently Processing :  q Replaced By  tablet
##  Currently Processing :  q Replaced By  tablet
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  p Replaced By  smartphone
##  Currently Processing :  v Replaced By  tv
##  Currently Processing :  v Replaced By  tv
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  v Replaced By  tv
##  Currently Processing :  v Replaced By  tv
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  p Replaced By  smartphone
##  Currently Processing :  x Replaced By  laptop
##  Currently Processing :  q Replaced By  tablet
##  Currently Processing :  q Replaced By  tablet
##  Currently Processing :  q Replaced By  tablet
```

```
qplot(x = product_type, data = electronics_clean)
```



Concatenating Variables: Full Adress for Geocoding

```
full_address <- NULL
full_address <- vector(mode = "character", length = length(electronics_clean$address))
electronics_clean$full_address <- paste(electronics_clean$address, electronics_clean$cit
y, electronics_clean$country, sep=', ')
electronics_clean$full_address
```

```
##  [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"
## [11] "Leeuwardenweg 182, arnhem, the netherlands"
## [12] "Leeuwardenweg 183, arnhem, the netherlands"
## [13] "Leeuwardenweg 184, arnhem, the netherlands"
## [14] "Delfzijlstraat 54, arnhem, the netherlands"
## [15] "Delfzijlstraat 55, arnhem, the netherlands"
## [16] "Delfzijlstraat 56, arnhem, the netherlands"
## [17] "Delfzijlstraat 57, arnhem, the netherlands"
## [18] "Delfzijlstraat 58, arnhem, the netherlands"
## [19] "Delfzijlstraat 59, arnhem, the netherlands"
## [20] "Delfzijlstraat 60, arnhem, the netherlands"
## [21] "Delfzijlstraat 61, arnhem, the netherlands"
## [22] "Jourestraat 23, arnhem, the netherlands"
## [23] "Jourestraat 24, arnhem, the netherlands"
## [24] "Jourestraat 25, arnhem, the netherlands"
## [25] "Jourestraat 26, arnhem, the netherlands"
```

Creating Dummy Variables

Add four binary columns for company

```
comp_dummy <- electronics_clean$company
comp_dummy <- as.numeric(comp_dummy == "phillips")
electronics_clean$company_phillips <- comp_dummy
electronics_clean$company_phillips
```

```
##  [1] 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0
```

```
comp_dummy <- electronics_clean$company
comp_dummy <- as.numeric(comp_dummy == "akzo")
electronics_clean$company_akzo <- comp_dummy
electronics_clean$company_akzo
```

```
##  [1] 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
```

```
comp_dummy <- electronics_clean$company
comp_dummy <- as.numeric(comp_dummy == "van houten")
electronics_clean$company_van_houten <- comp_dummy
electronics_clean$company_van_houten
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0
```

```
comp_dummy <- electronics_clean$company
comp_dummy <- as.numeric(comp_dummy == "unilever")
electronics_clean$company_unilever <- comp_dummy
electronics_clean$company_unilever
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
```

Add four binary columns for product

```
product_dummy <- electronics_clean$product_code
product_dummy <- as.numeric(product_dummy == "p")
electronics_clean$product_smartphone <- product_dummy
electronics_clean$product_smartphone
```

```
##  [1] 1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0
```

```
product_dummy <- electronics_clean$product_code
product_dummy <- as.numeric(product_dummy == "v")
electronics_clean$product_tv <- product_dummy
electronics_clean$product_tv
```

```
##  [1] 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 0
```

```
product_dummy <- electronics_clean$product_code
product_dummy <- as.numeric(product_dummy == "x")
electronics_clean$product_laptop <- product_dummy
electronics_clean$product_laptop
```

```
##  [1] 0 0 1 1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0
```

```
product_dummy <- electronics_clean$product_code
product_dummy <- as.numeric(product_dummy == "q")
electronics_clean$product_tablet <- product_dummy
electronics_clean$product_tablet
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1
```

Project Summary

```
glimpse(electronics_clean)
```

```
## Observations: 25
## Variables: 16
## $ company               (chr) "phillips", "phillips", "phillips", "phi...
## $ Product.code...number (fctr) p-5, p-43, x-3, x-34, x-12, p-23, v-43,...
## $ address               (fctr) Groningensingel 147, Groningensingel 14...
## $ city                  (fctr) arnhem, arnhem, arnhem, arnhem, arnhem,...
## $ country               (fctr) the netherlands, the netherlands, the n...
## $ name                  (fctr) dhr p. jansen, dhr p. hansen, dhr j. Ga...
## $ product_code          (chr) "p", "p", "x", "x", "x", "p", "v", "v", ...
## $ full_address          (chr) "Groningensingel 147, arnhem, the nether...
## $ company_phillips      (dbl) 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1...
## $ company_akzo          (dbl) 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0...
## $ company_van_houten    (dbl) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ company_unilever      (dbl) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ product_smartphone    (dbl) 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1...
## $ product_tv            (dbl) 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0...
## $ product_laptop        (dbl) 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0...
## $ product_tablet        (dbl) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0...
```

```
summary(electronics_clean)
```

```
##    company          Product.code...number             address
##  Length:25          p-23   : 2           Delfzijlstraat 54: 1
##  Class :character   x-3    : 2           Delfzijlstraat 55: 1
##  Mode  :character   p-34   : 1           Delfzijlstraat 56: 1
##                     p-43   : 1           Delfzijlstraat 57: 1
##                     p-5    : 1           Delfzijlstraat 58: 1
##                     p-56   : 1           Delfzijlstraat 59: 1
##                     (Other):17           (Other)         :19
##      city               country               name    product_code
##  arnhem:25   the netherlands:25   mevr l.  mokken: 4   Length:25
##                                   mevr l.  rokken: 3   Class :character
##                                   dhr j. Gansen  : 1   Mode  :character
##                                   dhr p. bansen  : 1
##                                   dhr p. bransen : 1
##                                   dhr p. fransen : 1
##                                   (Other)        :14
##  full_address       company_phillips  company_akzo  company_van_houten
##  Length:25          Min.   :0.00     Min.   :0.00   Min.   :0.0
##  Class :character   1st Qu.:0.00     1st Qu.:0.00   1st Qu.:0.0
##  Mode  :character   Median :0.00     Median :0.00   Median :0.0
##                     Mean   :0.36     Mean   :0.28   Mean   :0.2
##                     3rd Qu.:1.00     3rd Qu.:1.00   3rd Qu.:0.0
##                     Max.   :1.00     Max.   :1.00   Max.   :1.0
##
##  company_unilever product_smartphone   product_tv   product_laptop
##  Min.   :0.00     Min.   :0.00       Min.   :0.00   Min.   :0.00
##  1st Qu.:0.00     1st Qu.:0.00       1st Qu.:0.00   1st Qu.:0.00
##  Median :0.00     Median :0.00       Median :0.00   Median :0.00
##  Mean   :0.16     Mean   :0.24       Mean   :0.24   Mean   :0.32
##  3rd Qu.:0.00     3rd Qu.:0.00       3rd Qu.:0.00   3rd Qu.:1.00
##  Max.   :1.00     Max.   :1.00       Max.   :1.00   Max.   :1.00
##
##  product_tablet
##  Min.   :0.0
##  1st Qu.:0.0
##  Median :0.0
##  Mean   :0.2
##  3rd Qu.:0.0
##  Max.   :1.0
##
```