Basic Data Manipulation

Nicholas Engel

March 17, 2016

```
library(dplyr)
```

Introduction

This is an RMarkdown-generated document that reports the various manipulations I made to the refine.xls data in order to accomplish the following five tasks:

1. normalize the brand information;
2. separate the product code and product number information into distinct categories;
3. Create a new column with the product category information;
4. concatenate the address information; and
5. Add dummy variables for the company and product category data.

Here is the raw data that we started with:

```
refine <- read.csv("refine.csv", header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE)

print(refine)
```

```
##       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
```

Problems with this data include the messiness of particular data points and the concatenation of product code and number information into the same row variable.

However, the data does have one advantage: it is already clearly divided into columns that indicate distinct variables, and rows that indicate distinct observations. Thus there is no need to use tidyr's gather or spread functions in order to get this data into normal form.

1. Normalizing Brand Information

The brand information indicated in the "company" column is misspelled and does not have consistent capitalization.

We can resolve the issue of inconsistent capitalization in two steps.

First, we can universally lowercase all the characters in the relevant data columns. The function to use for this is the "tolower" function, which converts all upper-case characters in a character vector to lower-case characters. Second, we need to correct all the spelling errors.

1.1 Universal lowercase

First, lets universally lowercase the relevant "company" column:

```
refine[, 1] <- tolower(refine[, 1])
print(refine[, 1])
```

```
##  [1] "phillips"   "phillips"   "philips"    "phllips"    "phillps"
##  [6] "phillips"   "akzo"       "akzo"       "akzo"       "akz0"
## [11] "ak zo"      "akzo"       "akzo"       "phillips"   "fillips"
## [16] "phlips"     "van houten" "van houten" "van houten" "van houten"
## [21] "van houten" "unilver"    "unilever"   "unilever"   "unilever"
```

1.2 Spelling correction

There is probably no automatic way to handle the many spelling errors in the "company" column of our dataset. However, the particular idiosyncracies of the errors admit of patterns that we can exploit. For example, every "phillips" entry correctly has the last two characters; every "akzo" entry correctly has the fist two characters; every "unilever" entry correctly has the first three characters; and every "van houten" entry is correct.

We can exploit this with a series of ifelse functions:

```
 for (i in 1:25 ) {
  ifelse(substring(refine[i, 1], 1, 2) == "ak", refine[i, 1] <- "akzo", NA)
  ifelse(substring(refine[i, 1], 1, 3) == "uni", refine[i, 1] <- "unilever", NA)
  j <- nchar(refine[i, 1])
  ifelse(substring(refine[i, 1], j-1, j) == "ps", refine[i, 1] <- "phillips", NA)
}
print(refine[1:25, 1])
```

```
##  [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"
```

Since "van houten" entries do not have any misspellings, we don't need to test for them. We leave all the entries in lowercase because that is what the assignment asked us to do.

2. Separating the Product Code and Product Number

Each letter in the "Product Code and Product Number" column represents a product type, and each number represents a product number. Since these are separate pieces of information, they should be indicated as such with distinct column variables.

We can do this easily by manipulating the strings and using dplyr to mutate the data frame. I also wrote a helper function to isolate the numerical content from the "Product Code and Product Number" column.

Here's the helper function:

```
productnumber <- function(x) {
  for (i in 1:25)
   j <- nchar(as.character(x))
   substr(x, 3, j)
}
```

And the mutate function:

```
refine <- dplyr::mutate(refine,
  product_code = substr(refine[, 2], 1,1),
  product_number = productnumber(refine[, 2])
)
print(refine[, 7:8])
```

```
##    product_code product_number
## 1             p              5
## 2             p             43
## 3             x              3
## 4             x             34
## 5             x             12
## 6             p             23
## 7             v             43
## 8             v             12
## 9             x              5
## 10            p             34
## 11            q              5
## 12            q              9
## 13            x              8
## 14            p             56
## 15            v             67
## 16            v             21
## 17            x             45
## 18            v             56
## 19            v             65
## 20            x             21
## 21            p             23
## 22            x              3
## 23            q              4
## 24            q              6
## 25            q              8
```

Adding Product Categories

The product category information we generated above represents the following product categories:

p = smartphone v = tv x = laptop q = tablet

Let's add this information into our data frame with the help of another helper function.

NOTE: This is surely not the most efficient way to do this.

```
productcategory <- function(x) {

  ifelse(x == "p", j <- "smartphone",
  ifelse(x == "v", j <- "tv",
  ifelse(x == "x", j <- "laptop",
  ifelse(x == "q", j <- "tablet", NA))))
 j
}

refine <- dplyr::mutate(refine,
  product_category = refine[, 7])
```

```
for (i in 1:25) {
  refine[i, 9] <- productcategory(as.character(refine[i, 9]))
}
```

Here's the result:

```
print(refine[c(7, 9)])
```

```
##    product_code product_category
## 1             p       smartphone
## 2             p       smartphone
## 3             x           laptop
## 4             x           laptop
## 5             x           laptop
## 6             p       smartphone
## 7             v               tv
## 8             v               tv
## 9             x           laptop
## 10            p       smartphone
## 11            q           tablet
## 12            q           tablet
## 13            x           laptop
## 14            p       smartphone
## 15            v               tv
## 16            v               tv
## 17            x           laptop
## 18            v               tv
## 19            v               tv
## 20            x           laptop
## 21            p       smartphone
## 22            x           laptop
## 23            q           tablet
## 24            q           tablet
## 25            q           tablet
```

4. Concatenating the Address Information

We can concatenate the address information contained in the "address", "city", and "country" columns by using the "paste" function. As before, the output of our paste function can be added as an additional column in our data set with dplyr's "mutate" function.

```
refine <- dplyr::mutate(refine,
  full_address = paste(address, city, country, sep=", ")
)
```

And the result:

```
print(refine[c(3:5, 10)])
```

```
##                address   city         country
```

```
## 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
##                                    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
```

5. Adding Dummy Variables

For this last part of this assignment, we are asked to produce eight binary columns: one for each company, and one for each product category. I'll do this again with ifthen loops defined into helper functions:

```
phillipstest <- function(x) {
  ifelse(x == "phillips", 1, 0)
}

akzotest <- function(x) {
  ifelse(x == "akzo", 1, 0)
}

unilevertest <- function(x) {
  ifelse(x == "unilever", 1, 0)
}

vanhoutentest <- function(x) {
  ifelse(x == "van houten", 1, 0)
}

smartphonetest <- function(x) {
  ifelse(x == "smartphone", 1, 0)
}

tvtest <- function(x) {
  ifelse(x == "tv", 1, 0)
}

laptoptest <- function(x) {
  ifelse(x == "laptop", 1, 0)
}

tablettest <- function(x) {
  ifelse(x == "tablet", 1, 0)
}
```

Now we can easily add the eight columns into our dataframe with the mutate function:

```
refine <- dplyr::mutate(refine,
 company_phillips = phillipstest(as.character(company)),
 company_akzo = akzotest(as.character(company)),
 company_unilever = unilevertest(as.character(company)),
 company_van_houten = vanhoutentest(as.character(company)),
 product_smartphone = smartphonetest(as.character(refine[, 9])),
 product_tv = tvtest(as.character(refine[, 9])),
 product_laptop = laptoptest(as.character(refine[, 9])),
 product_tv = tvtest(as.character(refine[, 9])),
 product_tablet = tablettest(as.character(refine[, 9]))
)
```

Here is the result for the company variables:

```
print(refine[c(1, 11:14)])
```

```
##       company company_phillips company_akzo company_unilever
## 1    phillips                1            0                 0
## 2    phillips                1            0                 0
## 3    phillips                1            0                 0
## 4    phillips                1            0                 0
## 5    phillips                1            0                 0
## 6    phillips                1            0                 0
## 7        akzo                0            1                 0
## 8        akzo                0            1                 0
## 9        akzo                0            1                 0
## 10       akzo                0            1                 0
## 11       akzo                0            1                 0
## 12       akzo                0            1                 0
## 13       akzo                0            1                 0
## 14   phillips                1            0                 0
## 15   phillips                1            0                 0
## 16   phillips                1            0                 0
## 17 van houten                0            0                 0
## 18 van houten                0            0                 0
## 19 van houten                0            0                 0
## 20 van houten                0            0                 0
## 21 van houten                0            0                 0
## 22   unilever                0            0                 1
## 23   unilever                0            0                 1
## 24   unilever                0            0                 1
## 25   unilever                0            0                 1
##    company_van_houten
## 1                   0
## 2                   0
## 3                   0
## 4                   0
## 5                   0
## 6                   0
## 7                   0
## 8                   0
## 9                   0
## 10                  0
## 11                  0
## 12                  0
## 13                  0
## 14                  0
## 15                  0
## 16                  0
## 17                  1
## 18                  1
## 19                  1
## 20                  1
## 21                  1
## 22                  0
## 23                  0
## 24                  0
```

```
## 25                  0
```

And for the product variables:

```
print(refine[c(9, 15:18)])
```

```
##    product_category product_smartphone product_tv product_laptop
## 1        smartphone                  1          0              0
## 2        smartphone                  1          0              0
## 3            laptop                  0          0              1
## 4            laptop                  0          0              1
## 5            laptop                  0          0              1
## 6        smartphone                  1          0              0
## 7                tv                  0          1              0
## 8                tv                  0          1              0
## 9            laptop                  0          0              1
## 10       smartphone                  1          0              0
## 11           tablet                  0          0              0
## 12           tablet                  0          0              0
## 13           laptop                  0          0              1
## 14       smartphone                  1          0              0
## 15               tv                  0          1              0
## 16               tv                  0          1              0
## 17           laptop                  0          0              1
## 18               tv                  0          1              0
## 19               tv                  0          1              0
## 20           laptop                  0          0              1
## 21       smartphone                  1          0              0
## 22           laptop                  0          0              1
## 23           tablet                  0          0              0
## 24           tablet                  0          0              0
## 25           tablet                  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
```

And the entire result:

```
print(refine)
```

```
##       company Product.code...number             address   city
## 1    phillips                   p-5 Groningensingel 147 arnhem
## 2    phillips                  p-43 Groningensingel 148 arnhem
## 3    phillips                   x-3 Groningensingel 149 arnhem
## 4    phillips                  x-34 Groningensingel 150 arnhem
## 5    phillips                  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       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   phillips                  p-56   Delfzijlstraat 54 arnhem
## 15   phillips                  v-67   Delfzijlstraat 55 arnhem
## 16   phillips                  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_code product_number
## 1  the netherlands    dhr p. jansen            p              5
## 2  the netherlands    dhr p. hansen            p             43
## 3  the netherlands    dhr j. Gansen            x              3
## 4  the netherlands    dhr p. mansen            x             34
## 5  the netherlands   dhr p. fransen            x             12
## 6  the netherlands  dhr p. franssen            p             23
## 7  the netherlands    dhr p. bansen            v             43
## 8  the netherlands    dhr p. vansen            v             12
## 9  the netherlands   dhr p. bransen            x              5
## 10 the netherlands   dhr p. janssen            p             34
## 11 the netherlands  mevr l.  rokken            q              5
## 12 the netherlands  mevr l.  lokken            q              9
## 13 the netherlands  mevr l.  mokken            x              8
## 14 the netherlands  mevr l.  mokken            p             56
## 15 the netherlands  mevr l.  mokken            v             67
## 16 the netherlands  mevr l.  mokken            v             21
```

```
## 17 the netherlands  mevr l.  sokken            x             45
## 18 the netherlands  mevr l.  wokken            v             56
## 19 the netherlands  mevr l.  kokken            v             65
## 20 the netherlands  mevr l.  Bokken            x             21
## 21 the netherlands  mevr l.  dokken            p             23
## 22 the netherlands  mevr l.  gokken            x              3
## 23 the netherlands mevr l.  stokken            q              4
## 24 the netherlands  mevr l.  rokken            q              6
## 25 the netherlands  mevr l.  rokken            q              8
##    product_category                                  full_address
## 1        smartphone Groningensingel 147, arnhem, the netherlands
## 2        smartphone Groningensingel 148, arnhem, the netherlands
## 3            laptop Groningensingel 149, arnhem, the netherlands
## 4            laptop Groningensingel 150, arnhem, the netherlands
## 5            laptop Groningensingel 151, arnhem, the netherlands
## 6        smartphone Groningensingel 152, arnhem, the netherlands
## 7                tv   Leeuwardenweg 178, arnhem, the netherlands
## 8                tv   Leeuwardenweg 179, arnhem, the netherlands
## 9            laptop   Leeuwardenweg 180, arnhem, the netherlands
## 10       smartphone   Leeuwardenweg 181, arnhem, the netherlands
## 11           tablet   Leeuwardenweg 182, arnhem, the netherlands
## 12           tablet   Leeuwardenweg 183, arnhem, the netherlands
## 13           laptop   Leeuwardenweg 184, arnhem, the netherlands
## 14       smartphone   Delfzijlstraat 54, arnhem, the netherlands
## 15               tv   Delfzijlstraat 55, arnhem, the netherlands
## 16               tv   Delfzijlstraat 56, arnhem, the netherlands
## 17           laptop   Delfzijlstraat 57, arnhem, the netherlands
## 18               tv   Delfzijlstraat 58, arnhem, the netherlands
## 19               tv   Delfzijlstraat 59, arnhem, the netherlands
## 20           laptop   Delfzijlstraat 60, arnhem, the netherlands
## 21       smartphone   Delfzijlstraat 61, arnhem, the netherlands
## 22           laptop      Jourestraat 23, arnhem, the netherlands
## 23           tablet      Jourestraat 24, arnhem, the netherlands
## 24           tablet      Jourestraat 25, arnhem, the netherlands
## 25           tablet      Jourestraat 26, arnhem, the netherlands
##    company_phillips company_akzo company_unilever company_van_houten
## 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            1                0                  0
## 8                 0            1                0                  0
## 9                 0            1                0                  0
## 10                0            1                0                  0
## 11                0            1                0                  0
## 12                0            1                0                  0
## 13                0            1                0                  0
## 14                1            0                0                  0
## 15                1            0                0                  0
## 16                1            0                0                  0
## 17                0            0                0                  1
## 18                0            0                0                  1
```

```
## 19                0            0                0                1
## 20                0            0                0                1
## 21                0            0                0                1
## 22                0            0                1                0
## 23                0            0                1                0
## 24                0            0                1                0
## 25                0            0                1                0
##    product_smartphone product_tv product_laptop product_tablet
## 1                   1          0              0              0
## 2                   1          0              0              0
## 3                   0          0              1              0
## 4                   0          0              1              0
## 5                   0          0              1              0
## 6                   1          0              0              0
## 7                   0          1              0              0
## 8                   0          1              0              0
## 9                   0          0              1              0
## 10                  1          0              0              0
## 11                  0          0              0              1
## 12                  0          0              0              1
## 13                  0          0              1              0
## 14                  1          0              0              0
## 15                  0          1              0              0
## 16                  0          1              0              0
## 17                  0          0              1              0
## 18                  0          1              0              0
## 19                  0          1              0              0
## 20                  0          0              1              0
## 21                  1          0              0              0
## 22                  0          0              1              0
## 23                  0          0              0              1
## 24                  0          0              0              1
## 25                  0          0              0              1
```