Data Cleaning

Anisha Dubhashi

11/11/2017

load all data from csv in working directory

```
df <- read_csv("catalog sales data.csv")
```

```
## Parsed with column specification:
## cols(
##   targdol = col_double(),
##   datead6 = col_character(),
##   datelp6 = col_character(),
##   lpuryear = col_integer(),
##   slstyr = col_integer(),
##   slslyr = col_integer(),
##   sls2ago = col_integer(),
##   sls3ago = col_integer(),
##   slshist = col_integer(),
##   ordtyr = col_integer(),
##   ordlyr = col_integer(),
##   ord2ago = col_integer(),
##   ord3ago = col_integer(),
##   ordhist = col_integer(),
##   falord = col_integer(),
##   sprord = col_integer(),
##   train = col_integer()
## )
```

convert dates to date type

```
df$datead6 <- as.Date(df$datead6,format = "%m/%d/%Y")
df$datelp6 <- as.Date(df$datelp6,format = "%m/%d/%Y")
```

compare datelp6 to lpuryear - think lpuryear isn't necessary

```
df2 <- df[df$targdol > 0, ]

table(df$lpuryear, format(df$datelp6, "%Y"), useNA = "ifany")
```

```
##
##          1980  2002  2003  2004  2005  2006  2007  2008  2009  2010  2011
##   0         0     0     0     0     0     0     0     0    62 14848     8
##   1         0     0     0     0     0     0     0     0     0  1988 17689
##   2         0     0     0     0     0     0     0     0     0     0  2169
##   3         0     0  1969    46     0     0     1     1     0     9     9
##   4         0     0     0  3108     4     0     0     0     0    15    16
##   5         0     0     0     0  5446     6     0     0     1    27    25
##   6         0     0     0     0     0  6734    11     0     0    41    35
##   7         0     0     0     0     0     0  9375     9     1    87    45
##   8         0     0     0     0     0     0     0 12027    18   147    91
##   9         0     0     0     0     0     0     0     0 15193   214   132
##   <NA>     18   687    12     6     0     0     0     0     0     2     0
##       
##         2012
##   0       13
##   1       56
##   2     6643
##   3     2312
##   4        9
##   5       14
##   6       12
##   7       28
##   8       44
##   9       66
##   <NA>     3
```

~200 orders have no sales value for each year

```r
#2012
table(df$slstyr > 0, df$ordtyr > 0, dnn = c("sales", "orders"))
```

```
##        orders
## sales   FALSE  TRUE
##   FALSE 78861   205
##   TRUE      0 22466
```

```r
#2011
table(df$slslyr > 0, df$ordlyr > 0, dnn = c("sales", "orders"))
```

```
##        orders
## sales   FALSE  TRUE
##   FALSE 77003   121
##   TRUE      0 24408
```

```r
#2010
table(df$sls2ago > 0, df$ord2ago > 0, dnn = c("sales", "orders"))
```

```
##        orders
## sales   FALSE  TRUE
##   FALSE 76392   124
##   TRUE      0 25016
```

```r
#2009
table(df$sls3ago > 0, df$ord3ago > 0, dnn = c("sales", "orders"))
```

```
##        orders
## sales   FALSE  TRUE
##   FALSE 78005   237
##   TRUE      0 23290
```

```r
#history
tab <- table(df$slshist > 0, df$ordhist > 0, dnn = c("sales", "orders"))
tab
```

```
##        orders
## sales    FALSE   TRUE
##   FALSE     18    499
##   TRUE       0 101015
```

```
#percent of discrepancy
sum(tab[1,2])/sum(tab)
```

```
## [1] 0.004914707
```

lots of discrepancies between orders and date of last purchase

```
#2012
table(format(df$datelp6, "%Y"), df$ordtyr > 0, useNA = "ifany", dnn = c("year last order", "2012 order"
```

```
##                2012 order
## year last order FALSE  TRUE
##            1980    18     0
##            2002   687     0
##            2003  1976     5
##            2004  3158     2
##            2005  5445     5
##            2006  6737     3
##            2007  9380     7
##            2008 12025    12
##            2009 15242    33
##            2010 17370     8
##            2011  5398 14821
##            2012  1425  7775
```

```
#2011
table(format(df$datelp6, "%Y"), df$ordlyr > 0, useNA = "ifany", dnn = c("year last order", "2011 order"
```

```
##                2011 order
## year last order FALSE  TRUE
##            1980    17     1
##            2002   687     0
##            2003  1974     7
##            2004  3159     1
##            2005  5447     3
##            2006  6738     2
##            2007  9384     3
##            2008 12023    14
##            2009 15227    48
##            2010  4543 12835
##            2011 11229  8990
##            2012  6575  2625
```

```
#2010
table(format(df$datelp6, "%Y"), df$ord2ago > 0, useNA = "ifany", dnn = c("year last order", "2010 order
```

```
##                2010 order
## year last order FALSE  TRUE
##            1980    18     0
##            2002   687     0
##            2003  1980     1
##            2004  3160     0
##            2005  5450     0
##            2006  6739     1
##            2007  9382     5
##            2008 11809   228
##            2009  4238 11037
##            2010 10018  7360
##            2011 16028  4191
##            2012  6883  2317
```

```
#2009
table(format(df$datelp6, "%Y"), df$ord3ago > 0, useNA = "ifany", dnn = c("year last order", "2009 order
```

```
##                2009 order
## year last order FALSE  TRUE
##            1980    18     0
##            2002   687     0
##            2003  1977     4
##            2004  3160     0
##            2005  5450     0
##            2006  6738     2
##            2007  9183   204
##            2008  3630  8407
##            2009  8885  6390
##            2010 14282  3096
##            2011 16817  3402
##            2012  7178  2022
```

fall & spring order history

```
ordhist_desc <- (df$falord + df$sprord) - df$ordhist
table(ordhist_desc, useNA = "ifany", dnn = c("falord + sprord - ordhist"))
```

```
## falord + sprord - ordhist
##    -9    -7    -6    -5    -4    -3    -2    -1     0     1     2     3
##     1     3     4    17    54   186   890  4952 92740  2132   441    82
##     4     5     6     7    98
##    16     8     2     1     3
```

```
#falord + sprord < ordhist
sum(ordhist_desc < 0)
```

```
## [1] 6107
```

```
#falord + sprord > ordhist
sum(ordhist_desc > 0)
```

```
## [1] 2685
```

```
#falord + sprord = ordhist
sum(ordhist_desc == 0)
```

```
## [1] 92740
```