exploring-redhat-data

August 14, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

```
library(ggplot2)
library(data.table)
```

Functions for rendering HTML and PDF documents

```
render_pdf <-  function() {
    rmarkdown::render('exploring_redhat_data.Rmd',
                      output_file = 'markdown/exploring_redhat.pdf')
}

render_html <-  function() {
    rmarkdown::render('exploring_redhat_data.Rmd',
                      output_file = 'markdown/exploring_redhat.html')
}
```

Read data

```
activities_raw <- fread('../data/raw/act_train.csv')
```

```
##
Read 9.6% of 2197291 rows
Read 29.1% of 2197291 rows
Read 48.2% of 2197291 rows
Read 69.6% of 2197291 rows
Read 92.4% of 2197291 rows
Read 2197291 rows and 15 (of 15) columns from 0.131 GB file in 00:00:07
```

```
people_raw <- fread('../data/raw/people.csv')
```

```
##
Read 74.0% of 189118 rows
Read 189118 rows and 41 (of 41) columns from 0.046 GB file in 00:00:03
```

Data summary

```
summary(activities_raw)
```

```
##   people_id         activity_id            date
##  Length:2197291     Length:2197291     Length:2197291
##  Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  activity_category     char_1            char_2
##  Length:2197291     Length:2197291     Length:2197291
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     char_3            char_4            char_5
##  Length:2197291     Length:2197291     Length:2197291
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     char_6            char_7            char_8
##  Length:2197291     Length:2197291     Length:2197291
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     char_9           char_10            outcome
##  Length:2197291     Length:2197291     Min.   :0.000
##  Class :character   Class :character   1st Qu.:0.000
##  Mode  :character   Mode  :character   Median :0.000
##                                        Mean   :0.444
##                                        3rd Qu.:1.000
##                                        Max.   :1.000
```

```
summary(people_raw)
```

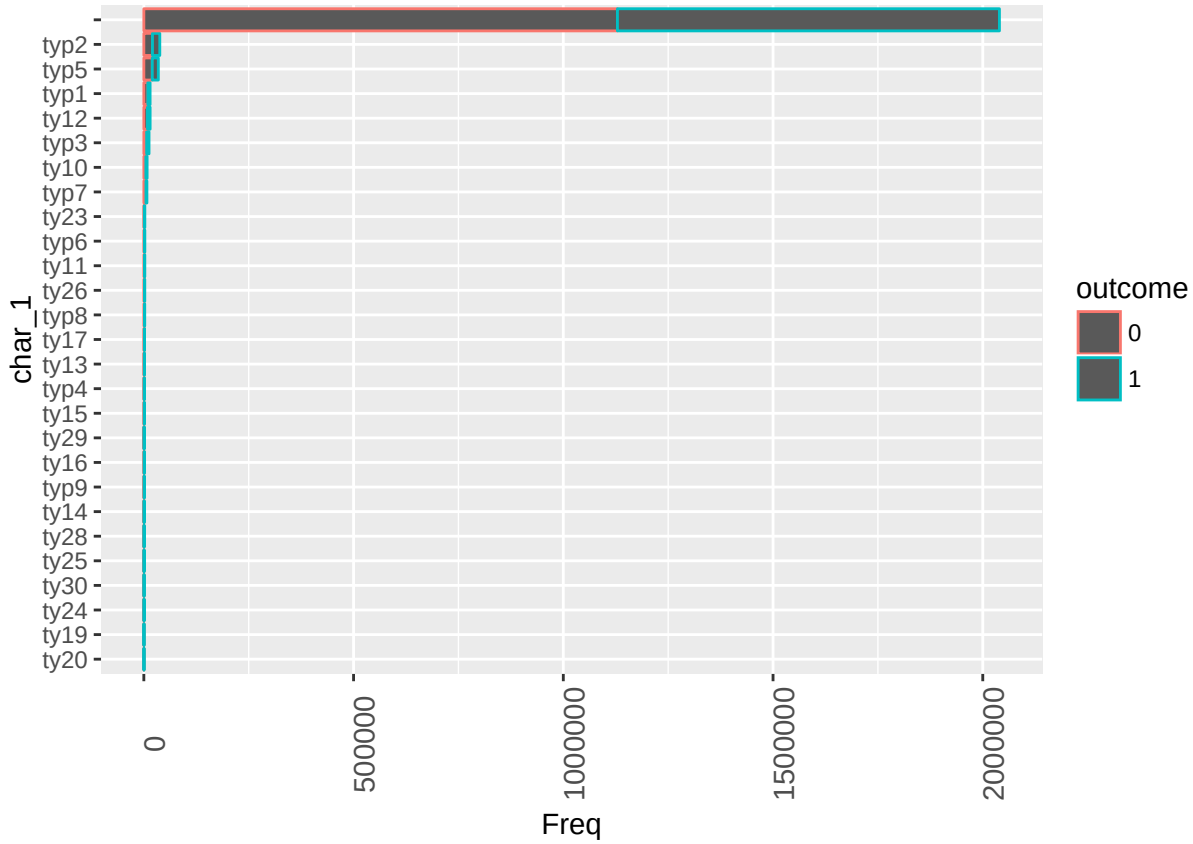
```
##   people_id            char_1           group_1
##  Length:189118      Length:189118      Length:189118
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     char_2              date            char_3
##  Length:189118      Length:189118      Length:189118
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     char_4            char_5            char_6
##  Length:189118      Length:189118      Length:189118
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##     char_7             char_8             char_9           char_10
##  Length:189118      Length:189118      Length:189118      Mode :logical
##  Class :character   Class :character   Class :character   FALSE:141660
##  Mode  :character   Mode  :character   Mode  :character   TRUE :47458
##                                                           NA's :0
##
##
##   char_11          char_12          char_13          char_14
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:148363    FALSE:143664    FALSE:120076    FALSE:139985
##  TRUE :40755     TRUE :45454     TRUE :69042     TRUE :49133
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##   char_15          char_16          char_17          char_18
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:138148    FALSE:135772    FALSE:133903    FALSE:153635
##  TRUE :50970     TRUE :53346     TRUE :55215     TRUE :35483
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##   char_19          char_20          char_21          char_22
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:135284    FALSE:145788    FALSE:135213    FALSE:134074
##  TRUE :53834     TRUE :43330     TRUE :53905     TRUE :55044
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##   char_23          char_24          char_25          char_26
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:132668    FALSE:153101    FALSE:127128    FALSE:157530
##  TRUE :56450     TRUE :36017     TRUE :61990     TRUE :31588
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##   char_27          char_28          char_29          char_30
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:144098    FALSE:134484    FALSE:157281    FALSE:149983
##  TRUE :45020     TRUE :54634     TRUE :31837     TRUE :39135
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##   char_31          char_32          char_33          char_34
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:136433    FALSE:135237    FALSE:147920    FALSE:121701
##  TRUE :52685     TRUE :53881     TRUE :41198     TRUE :67417
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##   char_35          char_36          char_37             char_38
##  Mode :logical   Mode :logical   Mode :logical   Min.   :  0.00
##  FALSE:149351    FALSE:124118    FALSE:135134    1st Qu.: 10.00
```

```
##  TRUE :39767     TRUE :65000     TRUE :53984     Median : 58.00
##  NA's :0         NA's :0         NA's :0         Mean   : 50.33
##                                                  3rd Qu.: 83.00
##                                                  Max.   :100.00
```

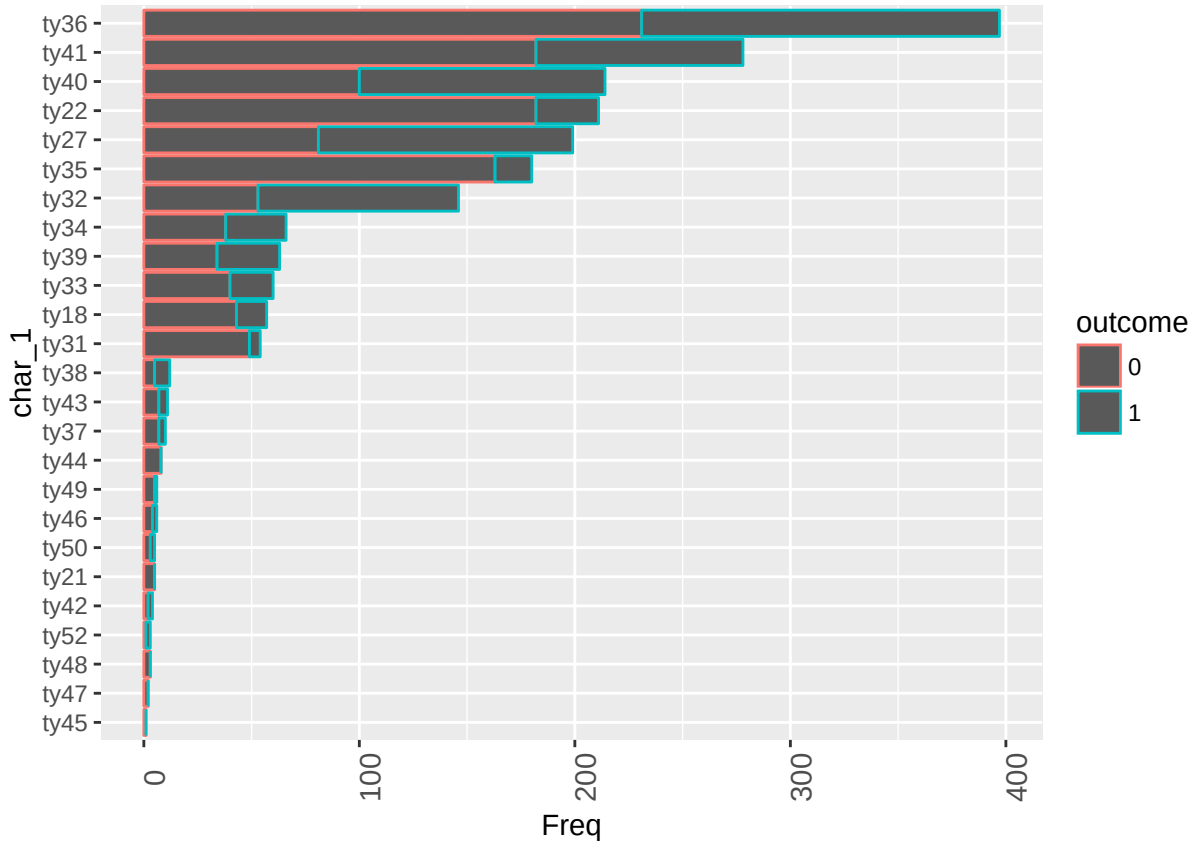
Inspect outomces for variables char_1

```
counts <- table(activities_raw[, c('char_1', 'outcome'), with=F])

activities_df <- as.data.frame(counts)
activities_df$char_1 <- reorder(activities_df$char_1, activities_df$Freq)
ind_split <- as.integer((length(levels(activities_df$char_1))-1) / 2)
most_frequent_levels <- levels(activities_df$char_1)[(ind_split+1): length(levels(activities_df$char_1))]
second_frequent_levels <- levels(activities_df$char_1)[1:ind_split]

ggplot(data=activities_df[activities_df$char_1 %in% most_frequent_levels, ],
       aes(x=char_1, y=Freq,  color=outcome)) +
    geom_bar(stat='identity') +
    theme(axis.text.x=element_text(angle=90, size=11)) +
    scale_x_discrete(labels=abbreviate) +
    coord_flip()
```

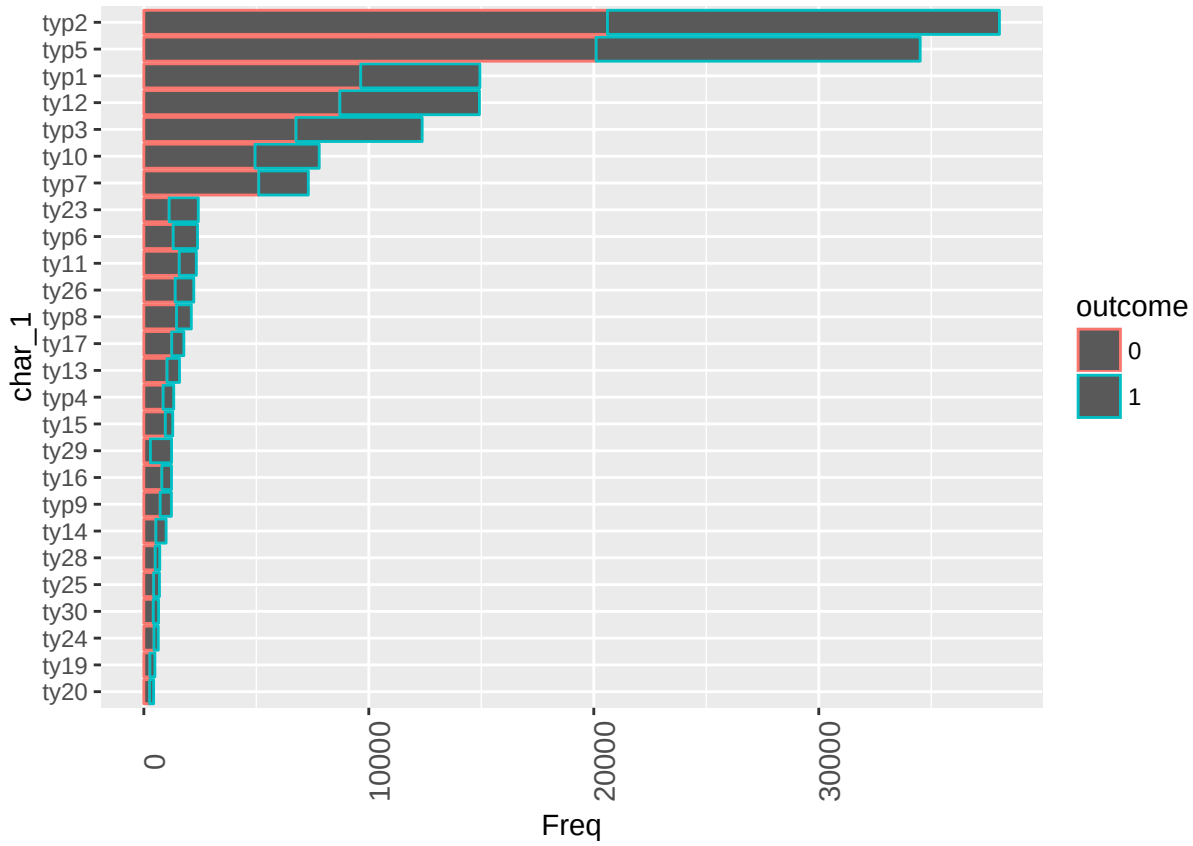



```
ggplot(data=activities_df[activities_df$char_1 %in% second_frequent_levels,],
       aes(x=char_1, y=Freq, color=outcome)) +
    geom_bar(stat='identity') +
    theme(axis.text.x=element_text(angle=90, size=11)) +
```

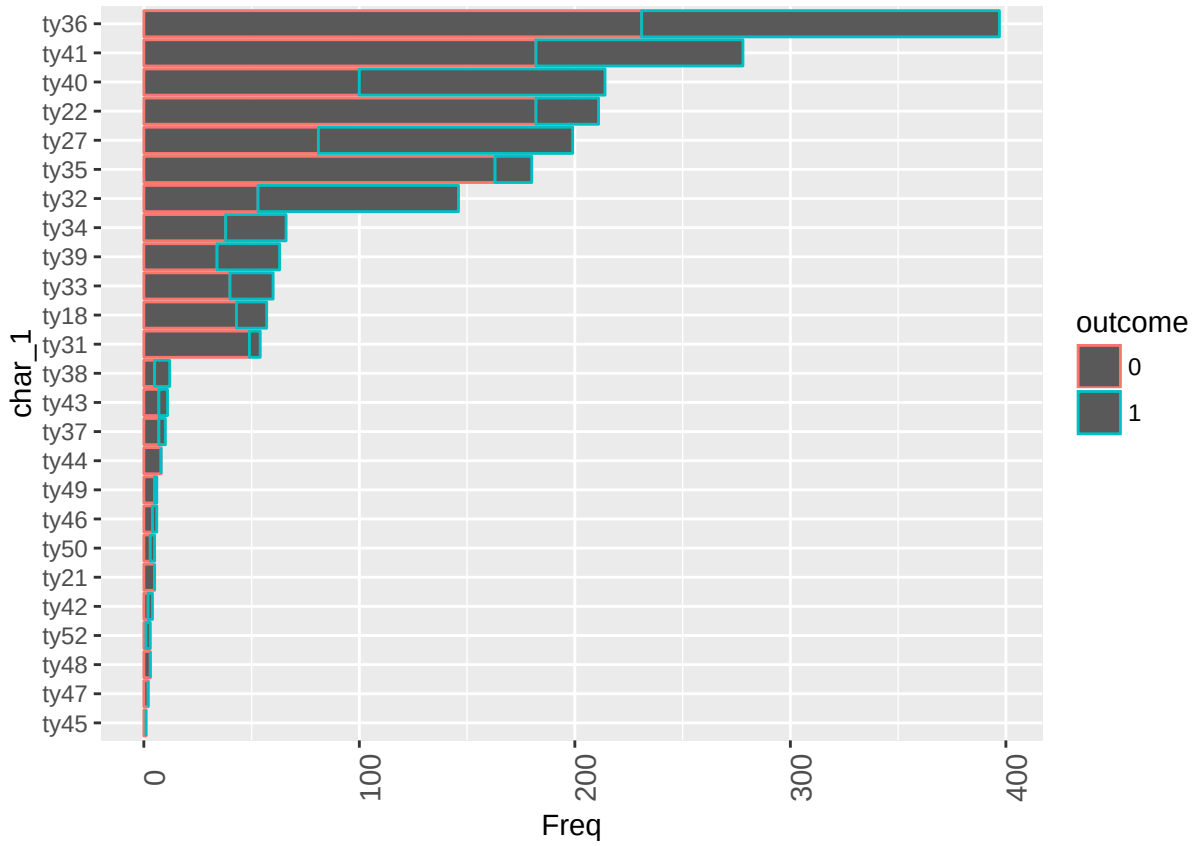
```
    scale_x_discrete(labels=abbreviate) +
    coord_flip()
```




```
df_without_blanks <- activities_df[activities_df$char_1 != '',]
df_without_blanks$char_1 <- as.factor(df_without_blanks$char_1)

ggplot(data=df_without_blanks[df_without_blanks$char_1 %in% most_frequent_levels, ],
       aes(x=char_1, y=Freq, color=outcome)) +
    geom_bar(stat='identity') +
    theme(axis.text.x=element_text(angle=90, size=11)) +
    scale_x_discrete(labels=abbreviate) +
    coord_flip()
```




```
ggplot(data=df_without_blanks[df_without_blanks$char_1 %in% second_frequent_levels, ],
       aes(x=char_1, y=Freq, color=outcome)) +
    geom_bar(stat='identity') +
    theme(axis.text.x=element_text(angle=90, size=11)) +
    scale_x_discrete(labels=abbreviate) +
    coord_flip()
```




```
counts <- table(activities_raw$char_1)
counts[order(counts, decreasing=T)]
```

```
##
##          type 2  type 5  type 1 type 12  type 3 type 10  type 7 type 23
## 2039676   38030   34509   14938   14917   12372    7795    7312    2420
##  type 6 type 11 type 26  type 8 type 17 type 13  type 4 type 15 type 29
##    2385    2333    2220    2110    1778    1586    1329    1284    1233
## type 16  type 9 type 14 type 28 type 25 type 30 type 24 type 19 type 20
##    1229    1225     990     706     694     653     641     491     434
## type 36 type 41 type 40 type 22 type 27 type 35 type 32 type 34 type 39
##     397     278     214     211     199     180     146      66      63
## type 33 type 18 type 31 type 38 type 43 type 37 type 44 type 46 type 49
##      60      57      54      12      11      10       8       6       6
## type 21 type 50 type 42 type 48 type 52 type 47 type 45
##       5       5       4       3       3       2       1
```

Most outcomes for variable char_1 are blanks. Counting number of blanks for each variable is easily done by the colSums function

```
colSums(activities_raw == '')
```

```
##         people_id       activity_id              date activity_category
##                 0                 0                 0                 0
##            char_1            char_2            char_3            char_4
```

```
##          2039676          2039676          2039676          2039676
##           char_5           char_6           char_7           char_8
##          2039676          2039676          2039676          2039676
##           char_9          char_10          outcome
##          2039676           157615                0
```