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)
library(dplyr)
```

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')
}
```

TODO: Merge activities and people

Read data

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

```
##
Read 28.2% of 2197291 rows
Read 53.2% of 2197291 rows
Read 78.7% of 2197291 rows
Read 2197291 rows and 15 (of 15) columns from 0.131 GB file in 00:00:05
```

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

# Rename columns
cols <- paste0('char_', 1:10)
for (c in cols) {
    col_ind <- which(colnames(activities_raw) == c)
    colnames(activities_raw)[col_ind] <- paste0('activity_', c)
}

cols <- paste0('char_', 1:38)
for (c in cols) {
    col_ind <- which(colnames(people_raw) == c)
    colnames(people_raw)[col_ind] <- paste0('people_', c)
}
```

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  activity_char_1    activity_char_2   
##  Length:2197291     Length:2197291     Length:2197291    
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##  activity_char_3    activity_char_4    activity_char_5   
##  Length:2197291     Length:2197291     Length:2197291    
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##  activity_char_6    activity_char_7    activity_char_8   
##  Length:2197291     Length:2197291     Length:2197291    
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##  activity_char_9    activity_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          people_char_1        group_1         
##  Length:189118      Length:189118      Length:189118     
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##  people_char_2          date           people_char_3     
##  Length:189118      Length:189118      Length:189118     
##  Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  people_char_4      people_char_5      people_char_6
##  Length:189118      Length:189118      Length:189118
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  people_char_7      people_char_8      people_char_9      people_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
##
##
##  people_char_11  people_char_12  people_char_13  people_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
##
##
##  people_char_15  people_char_16  people_char_17  people_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
##
##
##  people_char_19  people_char_20  people_char_21  people_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
##
##
##  people_char_23  people_char_24  people_char_25  people_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
##
##
##  people_char_27  people_char_28  people_char_29  people_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
##
##
##  people_char_31  people_char_32  people_char_33  people_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
##
##
##  people_char_35  people_char_36  people_char_37  people_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
```

```
head(activities_raw,2)
```

```
##    people_id  activity_id       date activity_category activity_char_1
## 1:   ppl_100 act2_1734928 2023-08-26            type 4
## 2:   ppl_100 act2_2434093 2022-09-27            type 2
##    activity_char_2 activity_char_3 activity_char_4 activity_char_5
## 1:
## 2:
##    activity_char_6 activity_char_7 activity_char_8 activity_char_9
## 1:
## 2:
##    activity_char_10 outcome
## 1:          type 76       0
## 2:           type 1       0
```

```
head(people_raw,2)
```

```
##     people_id people_char_1     group_1 people_char_2       date
## 1:    ppl_100        type 2 group 17304        type 2 2021-06-29
## 2: ppl_100002        type 2  group 8688        type 3 2021-01-06
##    people_char_3 people_char_4 people_char_5 people_char_6 people_char_7
## 1:        type 5        type 5        type 5        type 3       type 11
## 2:       type 28        type 9        type 5        type 3       type 11
##    people_char_8 people_char_9 people_char_10 people_char_11
## 1:        type 2        type 2           TRUE          FALSE
## 2:        type 2        type 4          FALSE          FALSE
##    people_char_12 people_char_13 people_char_14 people_char_15
## 1:          FALSE           TRUE           TRUE          FALSE
## 2:           TRUE           TRUE          FALSE          FALSE
##    people_char_16 people_char_17 people_char_18 people_char_19
## 1:           TRUE          FALSE          FALSE          FALSE
## 2:          FALSE           TRUE          FALSE          FALSE
##    people_char_20 people_char_21 people_char_22 people_char_23
## 1:          FALSE           TRUE          FALSE          FALSE
## 2:          FALSE          FALSE          FALSE           TRUE
##    people_char_24 people_char_25 people_char_26 people_char_27
## 1:          FALSE          FALSE          FALSE           TRUE
## 2:          FALSE           TRUE           TRUE           TRUE
##    people_char_28 people_char_29 people_char_30 people_char_31
```

```
## 1:           TRUE          FALSE           TRUE           TRUE
## 2:          FALSE          FALSE           TRUE           TRUE
##    people_char_32 people_char_33 people_char_34 people_char_35
## 1:          FALSE          FALSE           TRUE           TRUE
## 2:           TRUE           TRUE           TRUE           TRUE
##    people_char_36 people_char_37 people_char_38
## 1:           TRUE          FALSE             36
## 2:           TRUE          FALSE             76
```

```
merged_raw <- merge(x=people_raw, y=activities_raw, by='people_id')
head(merged_raw,2)
```

```
##    people_id people_char_1     group_1 people_char_2     date.x
## 1:   ppl_100        type 2 group 17304        type 2 2021-06-29
## 2:   ppl_100        type 2 group 17304        type 2 2021-06-29
##    people_char_3 people_char_4 people_char_5 people_char_6 people_char_7
## 1:        type 5        type 5        type 5        type 3       type 11
## 2:        type 5        type 5        type 5        type 3       type 11
##    people_char_8 people_char_9 people_char_10 people_char_11
## 1:        type 2        type 2           TRUE          FALSE
## 2:        type 2        type 2           TRUE          FALSE
##    people_char_12 people_char_13 people_char_14 people_char_15
## 1:          FALSE           TRUE           TRUE          FALSE
## 2:          FALSE           TRUE           TRUE          FALSE
##    people_char_16 people_char_17 people_char_18 people_char_19
## 1:           TRUE          FALSE          FALSE          FALSE
## 2:           TRUE          FALSE          FALSE          FALSE
##    people_char_20 people_char_21 people_char_22 people_char_23
## 1:          FALSE           TRUE          FALSE          FALSE
## 2:          FALSE           TRUE          FALSE          FALSE
##    people_char_24 people_char_25 people_char_26 people_char_27
## 1:          FALSE          FALSE          FALSE           TRUE
## 2:          FALSE          FALSE          FALSE           TRUE
##    people_char_28 people_char_29 people_char_30 people_char_31
## 1:           TRUE          FALSE           TRUE           TRUE
## 2:           TRUE          FALSE           TRUE           TRUE
##    people_char_32 people_char_33 people_char_34 people_char_35
## 1:          FALSE          FALSE           TRUE           TRUE
## 2:          FALSE          FALSE           TRUE           TRUE
##    people_char_36 people_char_37 people_char_38  activity_id     date.y
## 1:           TRUE          FALSE             36 act2_1734928 2023-08-26
## 2:           TRUE          FALSE             36 act2_2434093 2022-09-27
##    activity_category activity_char_1 activity_char_2 activity_char_3
## 1:            type 4
## 2:            type 2
##    activity_char_4 activity_char_5 activity_char_6 activity_char_7
## 1:
## 2:
##    activity_char_8 activity_char_9 activity_char_10 outcome
## 1:                                         type 76       0
## 2:                                          type 1       0
```

Write merged data to disk

```
# colnames(merged_raw)[which(colnames(merged_raw) == 'date.x')] <- 'people_date'
# colnames(merged_raw)[which(colnames(merged_raw) == 'date.y')] <- 'activity_date'
# write.csv(x=merged_raw, file='../data/processed/merged_data.csv')
merged_raw <- fread('../data/processed/merged_data.csv')
```

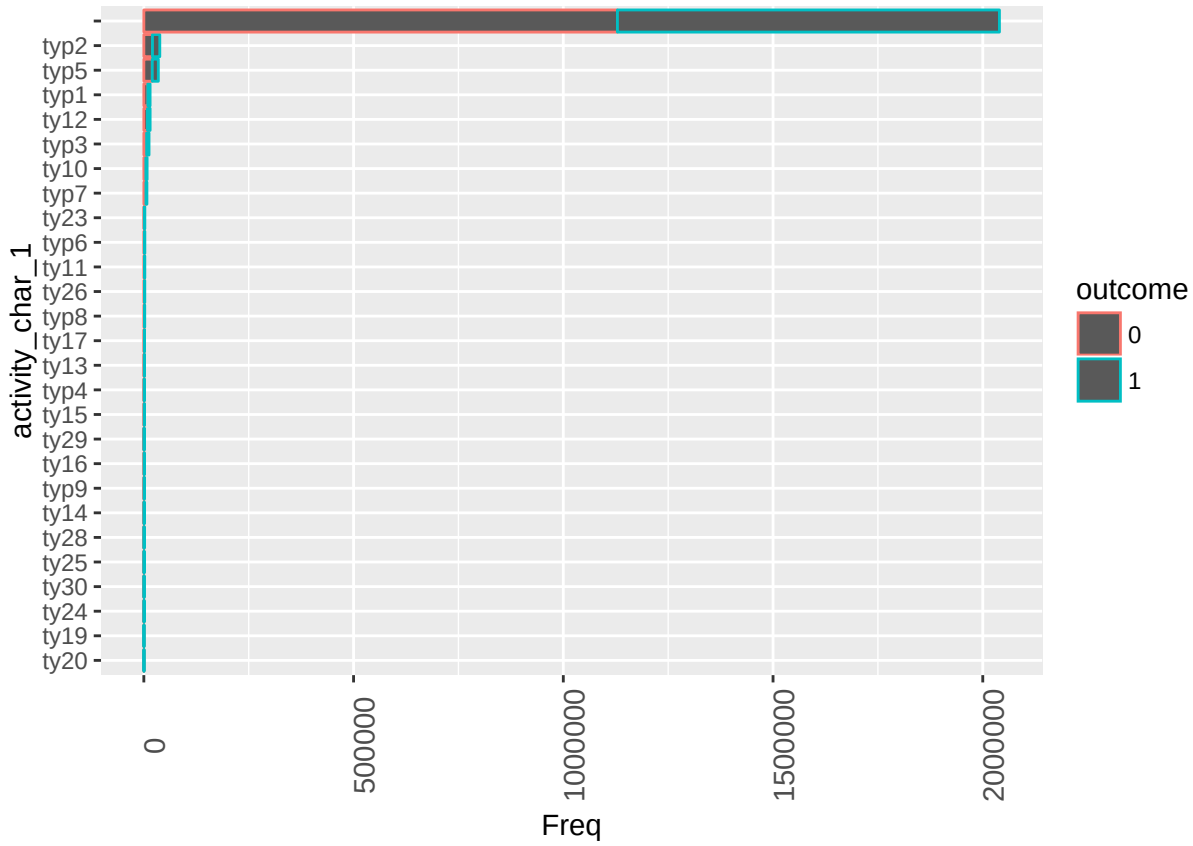
```
##
Read 0.0% of 2197291 rows
Read 7.7% of 2197291 rows
Read 15.9% of 2197291 rows
Read 16.4% of 2197291 rows
Read 24.1% of 2197291 rows
Read 27.8% of 2197291 rows
Read 36.4% of 2197291 rows
Read 38.7% of 2197291 rows
Read 46.9% of 2197291 rows
Read 53.7% of 2197291 rows
Read 61.4% of 2197291 rows
Read 69.2% of 2197291 rows
Read 71.5% of 2197291 rows
Read 78.7% of 2197291 rows
Read 86.0% of 2197291 rows
Read 92.8% of 2197291 rows
Read 2197291 rows and 55 (of 55) columns from 0.729 GB file in 00:00:27
```

Inspect outomces for variable activity_char_1

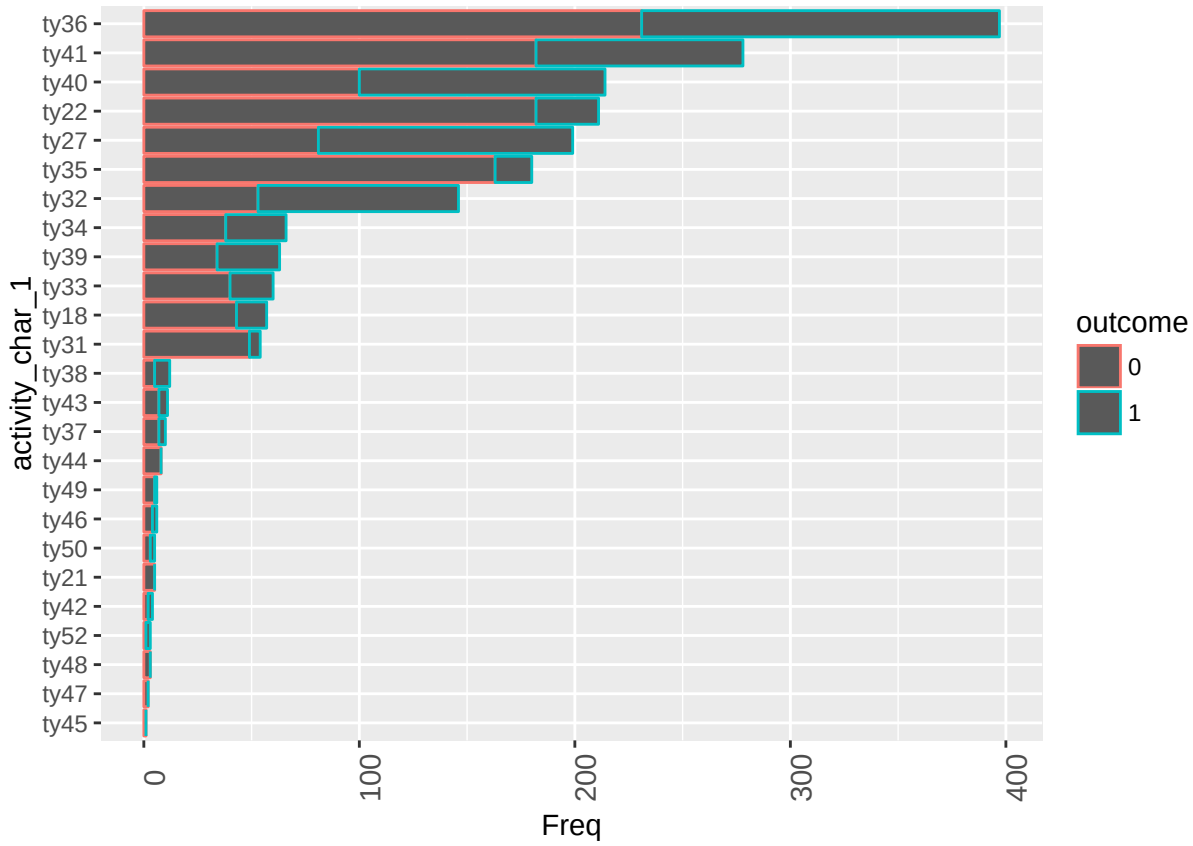
```
counts <- table(activities_raw[, c('activity_char_1', 'outcome'), with=F])

activities_df <- as.data.frame(counts)
activities_df$activity_char_1 <- reorder(activities_df$activity_char_1, activities_df$Freq)
ind_split <- as.integer((length(levels(activities_df$activity_char_1))-1) / 2)
most_frequent_levels <- levels(activities_df$activity_char_1)[(ind_split+1): length(levels(activities_d
second_frequent_levels <- levels(activities_df$activity_char_1)[1:ind_split]

ggplot(data=activities_df[activities_df$activity_char_1 %in% most_frequent_levels, ],
       aes(x=activity_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()
```

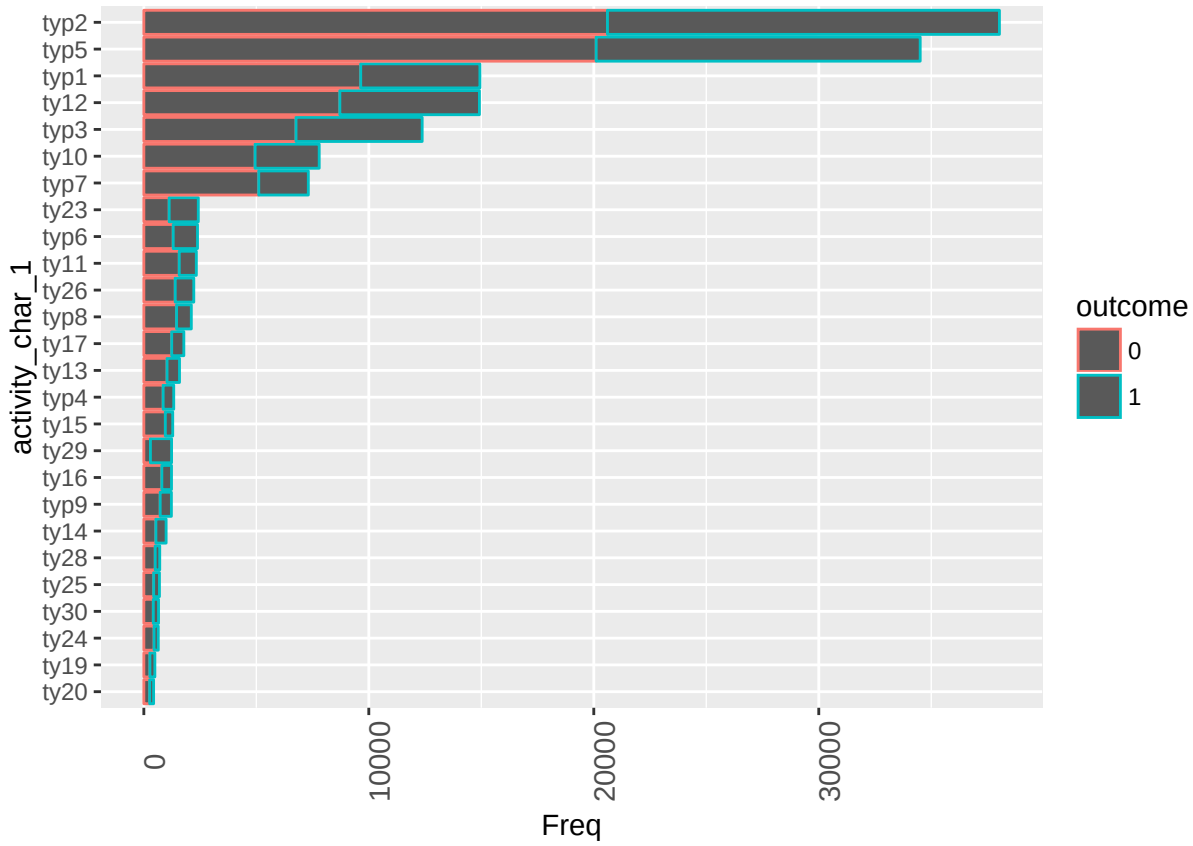



```
ggplot(data=activities_df[activities_df$activity_char_1 %in% second_frequent_levels,],
       aes(x=activity_char_1, y=Freq, color=outcome)) +
    geom_bar(stat='identity') +
    theme(axis.text.x=element_text(angle=90, size=11)) +
    scale_x_discrete(labels=abbreviate) +
    coord_flip()
```

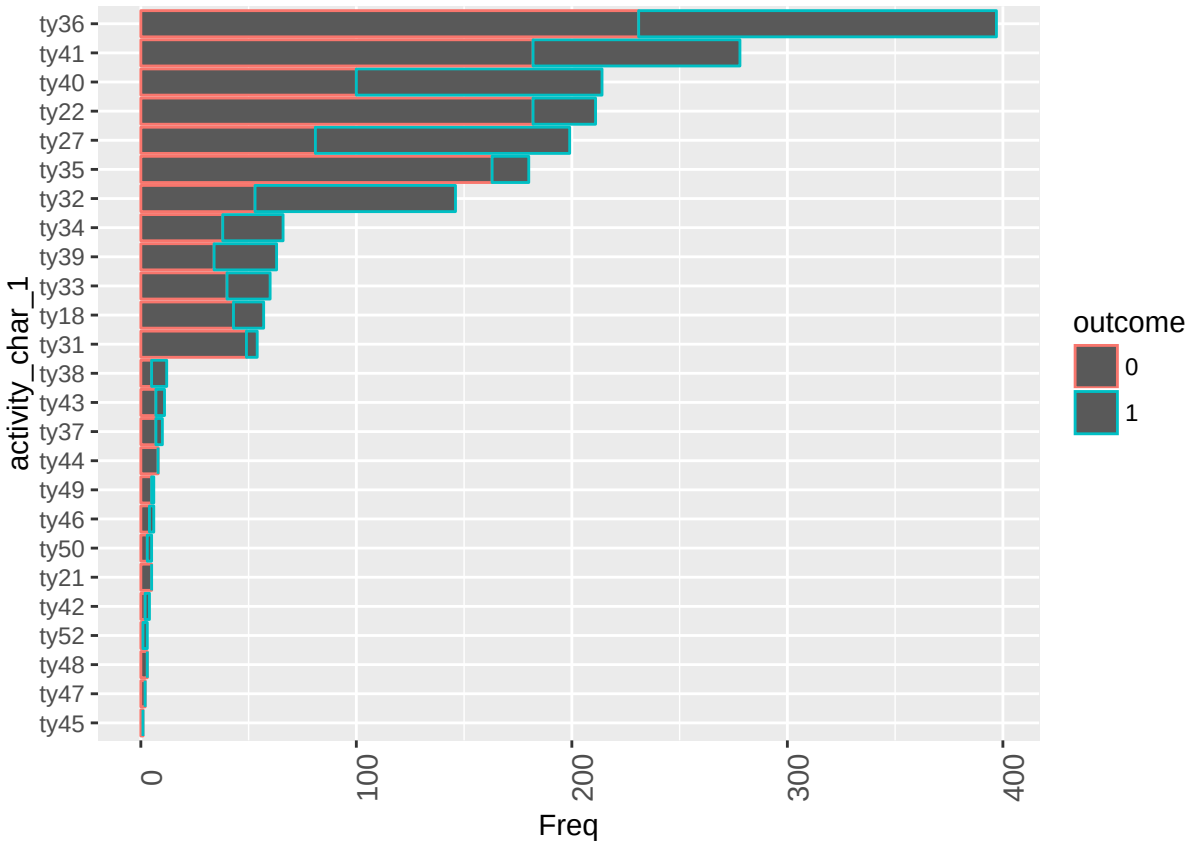



```
df_without_blanks <- activities_df[activities_df$activity_char_1 != '',]
df_without_blanks$activity_char_1 <- as.factor(df_without_blanks$activity_char_1)

ggplot(data=df_without_blanks[df_without_blanks$activity_char_1 %in% most_frequent_levels, ],
       aes(x=activity_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$activity_char_1 %in% second_frequent_levels, ],
       aes(x=activity_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$activity_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
```

```
##           used (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells  590704 31.6    3127137  167.1   3086178  164.9
## Vcells 1787650 13.7  352338892 2688.2 600177631 4579.0
```

Most outcomes for variable activity_char_1 are blanks. Counting the number of blanks for each variable is easily done by the colSums function

```
colSums(merged_raw == '')
```

```
##                V1                id     people_char_1      people_group
##                 0                 0                 0                 0
##     people_char_2       date_people     people_char_3     people_char_4
##                 0                 0                 0                 0
##     people_char_5     people_char_6     people_char_7     people_char_8
##                 0                 0                 0                 0
##     people_char_9    people_char_10    people_char_11    people_char_12
##                 0                 0                 0                 0
##    people_char_13    people_char_14    people_char_15    people_char_16
##                 0                 0                 0                 0
##    people_char_17    people_char_18    people_char_19    people_char_20
##                 0                 0                 0                 0
##    people_char_21    people_char_22    people_char_23    people_char_24
##                 0                 0                 0                 0
##    people_char_25    people_char_26    people_char_27    people_char_28
##                 0                 0                 0                 0
##    people_char_29    people_char_30    people_char_31    people_char_32
##                 0                 0                 0                 0
##    people_char_33    people_char_34    people_char_35    people_char_36
##                 0                 0                 0                 0
##    people_char_37    people_char_38     date_activity activity_category
##                 0                 0                 0                 0
##   activity_char_1   activity_char_2   activity_char_3   activity_char_4
##           2039676           2039676           2039676           2039676
##   activity_char_5   activity_char_6   activity_char_7   activity_char_8
##           2039676           2039676           2039676           2039676
##   activity_char_9  activity_char_10           outcome
##           2039676            157615                 0
```

```
##           used (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells  590658 31.6    2501709  133.7   3086178  164.9
## Vcells 1787860 13.7  281871113 2150.6 600177631 4579.0
```

Number of unqie values for each variable

```
apply(merged_raw, MARGIN=2, function(x) length(unique(x)))
```

```
##                V1                id     people_char_1      people_group
##           2197291           2197291                 2             29899
##     people_char_2       date_people     people_char_3     people_char_4
##                 3              1196                43                25
##     people_char_5     people_char_6     people_char_7     people_char_8
##                 9                 7                25                 8
##     people_char_9    people_char_10    people_char_11    people_char_12
##                 9                 2                 2                 2
##    people_char_13    people_char_14    people_char_15    people_char_16
##                 2                 2                 2                 2
##    people_char_17    people_char_18    people_char_19    people_char_20
##                 2                 2                 2                 2
##    people_char_21    people_char_22    people_char_23    people_char_24
```

```
##                 2                 2                 2                 2
##    people_char_25    people_char_26    people_char_27    people_char_28
##                 2                 2                 2                 2
##    people_char_29    people_char_30    people_char_31    people_char_32
##                 2                 2                 2                 2
##    people_char_33    people_char_34    people_char_35    people_char_36
##                 2                 2                 2                 2
##    people_char_37    people_char_38     date_activity activity_category
##                 2               101               411                 7
##   activity_char_1   activity_char_2   activity_char_3   activity_char_4
##                52                33                12                 8
##   activity_char_5   activity_char_6   activity_char_7   activity_char_8
##                 8                 6                 9                19
##   activity_char_9  activity_char_10           outcome
##                20              6516                 2

##           used (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells  590660 31.6    2001367  106.9   3086178  164.9
## Vcells 1788155 13.7  225496890 1720.5 600177631 4579.0
```

By the data specification it is said that type 1 activities are different from type 2-7 activities in the sense that there are more known characteristics associated with type 1 activities (nine in total) than type 2-7 activities (which have only one associated characteristic)

Get number of unique values while fixing activity 2

```
apply(merged_raw[merged_raw$activity_char_2==merged_raw$activity_char_2[1], ], MARGIN=2, function(x) le
```

```
##                V1                id     people_char_1      people_group
##           2039676           2039676                 2             28431
##     people_char_2       date_people     people_char_3     people_char_4
##                 3              1195                43                25
##     people_char_5     people_char_6     people_char_7     people_char_8
##                 9                 7                25                 8
##     people_char_9    people_char_10    people_char_11    people_char_12
##                 9                 2                 2                 2
##    people_char_13    people_char_14    people_char_15    people_char_16
##                 2                 2                 2                 2
##    people_char_17    people_char_18    people_char_19    people_char_20
##                 2                 2                 2                 2
##    people_char_21    people_char_22    people_char_23    people_char_24
##                 2                 2                 2                 2
##    people_char_25    people_char_26    people_char_27    people_char_28
##                 2                 2                 2                 2
##    people_char_29    people_char_30    people_char_31    people_char_32
##                 2                 2                 2                 2
##    people_char_33    people_char_34    people_char_35    people_char_36
##                 2                 2                 2                 2
##    people_char_37    people_char_38     date_activity activity_category
##                 2               101               386                 6
##   activity_char_1   activity_char_2   activity_char_3   activity_char_4
##                 1                 1                 1                 1
##   activity_char_5   activity_char_6   activity_char_7   activity_char_8
##                 1                 1                 1                 1
```

```
##   activity_char_9  activity_char_10           outcome
##                 1             6515                 2
```

```
##           used (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells  590662 31.6    2001367  106.9   3086178  164.9
## Vcells 1788457 13.7  180397512 1376.4 600177631 4579.0
```

```
apply(merged_raw[merged_raw$activity_char_2==merged_raw$activity_char_2[250], ], MARGIN=2, function(x) l
```

```
##               V1               id     people_char_1      people_group
##          2039676          2039676                2             28431
##     people_char_2      date_people     people_char_3     people_char_4
##                 3             1195                43                25
##     people_char_5     people_char_6     people_char_7     people_char_8
##                 9                 7                25                 8
##     people_char_9    people_char_10    people_char_11    people_char_12
##                 9                 2                 2                 2
##    people_char_13    people_char_14    people_char_15    people_char_16
##                 2                 2                 2                 2
##    people_char_17    people_char_18    people_char_19    people_char_20
##                 2                 2                 2                 2
##    people_char_21    people_char_22    people_char_23    people_char_24
##                 2                 2                 2                 2
##    people_char_25    people_char_26    people_char_27    people_char_28
##                 2                 2                 2                 2
##    people_char_29    people_char_30    people_char_31    people_char_32
##                 2                 2                 2                 2
##    people_char_33    people_char_34    people_char_35    people_char_36
##                 2                 2                 2                 2
##    people_char_37    people_char_38     date_activity activity_category
##                 2               101               386                 6
##   activity_char_1   activity_char_2   activity_char_3   activity_char_4
##                 1                 1                 1                 1
##   activity_char_5   activity_char_6   activity_char_7   activity_char_8
##                 1                 1                 1                 1
##   activity_char_9  activity_char_10           outcome
##                 1             6515                 2
```