

# 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)
```

```
## data.table 1.9.6 For help type ?data.table or https://github.com/Rdatatable/data.table/wiki
```

```
## The fastest way to learn (by data.table authors): https://www.datacamp.com/courses/data-analysis-the
```

```
library(dplyr)
```

```
## -----
```

```
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
```

```
## -----
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##   between, last
```

```
## The following objects are masked from 'package:stats':
##
##   filter, lag
```

```
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
source('../src/data/merge_data_to_disk.R')
```

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')
}
```

Function that removes all r objects from memory

```
clear <- function() {  
  rm(list=ls())  
}
```

## Read data

```
merge_and_write_data_to_disk()
```

```
## [1] "File ../data/processed/merged_data.csv already sourced."
```

```
merged_raw <- fread('../data/processed/merged_data.csv')
```

```
##  
Read 0.0% of 2197291 rows  
Read 7.3% of 2197291 rows  
Read 14.1% of 2197291 rows  
Read 20.9% of 2197291 rows  
Read 27.3% of 2197291 rows  
Read 33.7% of 2197291 rows  
Read 37.3% of 2197291 rows  
Read 43.7% of 2197291 rows  
Read 50.1% of 2197291 rows  
Read 56.4% of 2197291 rows  
Read 62.3% of 2197291 rows  
Read 68.3% of 2197291 rows  
Read 74.2% of 2197291 rows  
Read 80.1% of 2197291 rows  
Read 86.0% of 2197291 rows  
Read 91.9% of 2197291 rows  
Read 95.6% of 2197291 rows  
Read 2197291 rows and 56 (of 56) columns from 0.766 GB file in 00:00:28
```

## Data summary

```
head(merged_raw, 5)
```

```
##      V1 people_id people_char_1 people_group_1 people_char_2 people_date  
## 1:  1    ppl_100          type 2    group 17304          type 2 2021-06-29  
## 2:  2    ppl_100          type 2    group 17304          type 2 2021-06-29  
## 3:  3    ppl_100          type 2    group 17304          type 2 2021-06-29  
## 4:  4    ppl_100          type 2    group 17304          type 2 2021-06-29  
## 5:  5    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  
## 3:          type 5          type 5          type 5          type 3          type 11
```

```

## 4:      type 5      type 5      type 5      type 3      type 11
## 5:      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
## 3:      type 2      type 2      TRUE      FALSE
## 4:      type 2      type 2      TRUE      FALSE
## 5:      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
## 3:      FALSE      TRUE      TRUE      FALSE
## 4:      FALSE      TRUE      TRUE      FALSE
## 5:      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
## 3:      TRUE      FALSE      FALSE      FALSE
## 4:      TRUE      FALSE      FALSE      FALSE
## 5:      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
## 3:      FALSE      TRUE      FALSE      FALSE
## 4:      FALSE      TRUE      FALSE      FALSE
## 5:      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
## 3:      FALSE      FALSE      FALSE      TRUE
## 4:      FALSE      FALSE      FALSE      TRUE
## 5:      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
## 3:      TRUE      FALSE      TRUE      TRUE
## 4:      TRUE      FALSE      TRUE      TRUE
## 5:      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
## 3:      FALSE      FALSE      TRUE      TRUE
## 4:      FALSE      FALSE      TRUE      TRUE
## 5:      FALSE      FALSE      TRUE      TRUE
##  people_char_36 people_char_37 people_char_38 activity_id activity_date
## 1:      TRUE      FALSE      36 act2_1734928 2023-08-26
## 2:      TRUE      FALSE      36 act2_2434093 2022-09-27
## 3:      TRUE      FALSE      36 act2_3404049 2022-09-27
## 4:      TRUE      FALSE      36 act2_3651215 2023-08-04
## 5:      TRUE      FALSE      36 act2_4109017 2023-08-26
##  activity_category activity_char_1 activity_char_2 activity_char_3
## 1:      type 4
## 2:      type 2
## 3:      type 2

```

```
## 4:          type 2
## 5:          type 2
##   activity_char_4 activity_char_5 activity_char_6 activity_char_7
## 1:
## 2:
## 3:
## 4:
## 5:
##   activity_char_8 activity_char_9 activity_char_10 outcome
## 1:                                type 76          0
## 2:                                type 1           0
## 3:                                type 1           0
## 4:                                type 1           0
## 5:                                type 1           0
```

```
head(merged_raw[which(merged_raw$activity_char_1 != ''), ])
```

```
##      V1  people_id people_char_1 people_group_1 people_char_2 people_date
## 1:  53 ppl_100025      type 2      group 36096      type 3 2022-08-26
## 2: 106 ppl_100033      type 2      group 17304      type 2 2022-07-26
## 3: 107 ppl_100033      type 2      group 17304      type 2 2022-07-26
## 4: 108 ppl_100033      type 2      group 17304      type 2 2022-07-26
## 5: 109 ppl_100033      type 2      group 17304      type 2 2022-07-26
## 6: 125 ppl_100035      type 2      group 9439       type 3 2022-01-22
##   people_char_3 people_char_4 people_char_5 people_char_6 people_char_7
## 1:      type 14      type 6      type 8      type 3      type 9
## 2:      type 10      type 7      type 6      type 3      type 9
## 3:      type 10      type 7      type 6      type 3      type 9
## 4:      type 10      type 7      type 6      type 3      type 9
## 5:      type 10      type 7      type 6      type 3      type 9
## 6:      type 4      type 10      type 4      type 1      type 23
##   people_char_8 people_char_9 people_char_10 people_char_11
## 1:      type 6      type 6      FALSE      FALSE
## 2:      type 3      type 3      FALSE      FALSE
## 3:      type 3      type 3      FALSE      FALSE
## 4:      type 3      type 3      FALSE      FALSE
## 5:      type 3      type 3      FALSE      FALSE
## 6:      type 2      type 2      FALSE      TRUE
##   people_char_12 people_char_13 people_char_14 people_char_15
## 1:      FALSE      FALSE      FALSE      FALSE
## 2:      FALSE      FALSE      FALSE      FALSE
## 3:      FALSE      FALSE      FALSE      FALSE
## 4:      FALSE      FALSE      FALSE      FALSE
## 5:      FALSE      FALSE      FALSE      FALSE
## 6:      FALSE      FALSE      FALSE      FALSE
##   people_char_16 people_char_17 people_char_18 people_char_19
## 1:      FALSE      FALSE      FALSE      FALSE
## 2:      FALSE      FALSE      FALSE      FALSE
## 3:      FALSE      FALSE      FALSE      FALSE
## 4:      FALSE      FALSE      FALSE      FALSE
## 5:      FALSE      FALSE      FALSE      FALSE
## 6:      FALSE      FALSE      FALSE      TRUE
##   people_char_20 people_char_21 people_char_22 people_char_23
## 1:      FALSE      FALSE      FALSE      FALSE
```

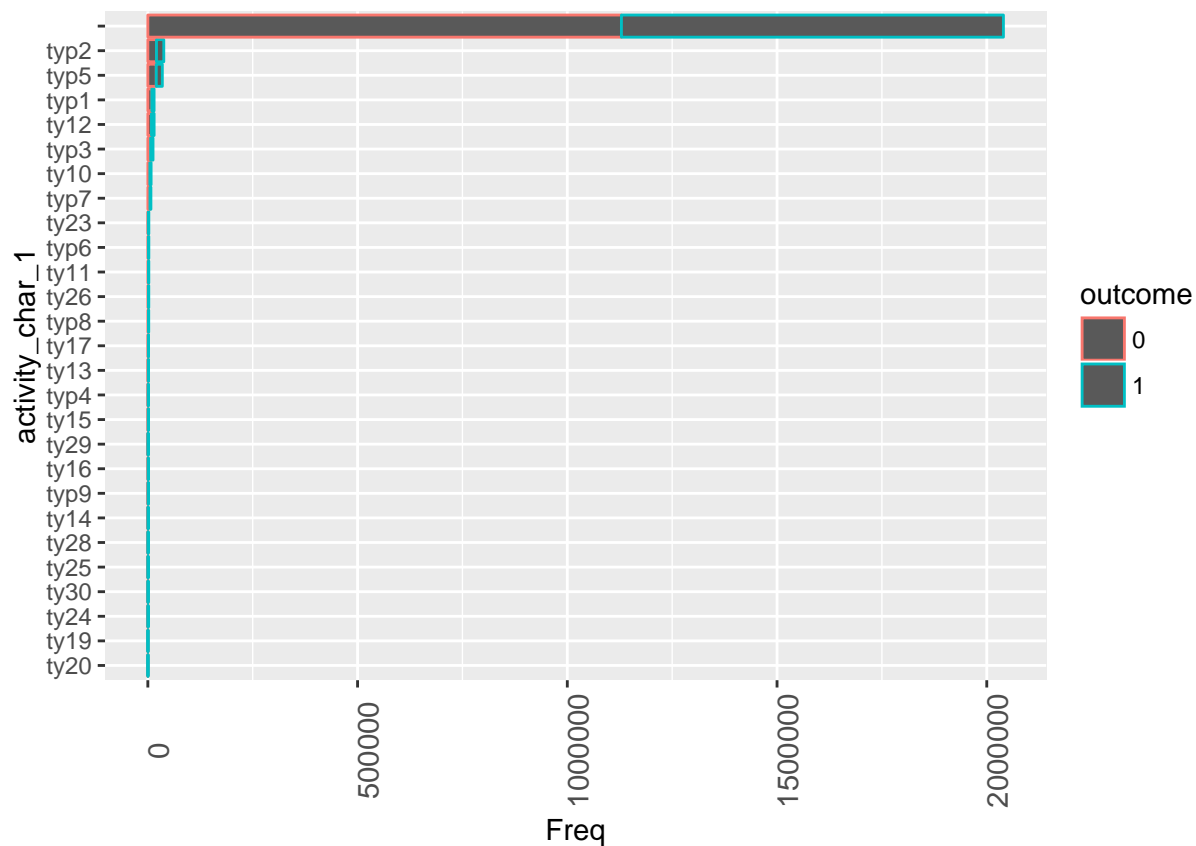
## 2:	FALSE	FALSE	FALSE	FALSE	
## 3:	FALSE	FALSE	FALSE	FALSE	
## 4:	FALSE	FALSE	FALSE	FALSE	
## 5:	FALSE	FALSE	FALSE	FALSE	
## 6:	TRUE	TRUE	TRUE	TRUE	
##	people_char_24	people_char_25	people_char_26	people_char_27	
## 1:	FALSE	FALSE	FALSE	FALSE	
## 2:	FALSE	FALSE	FALSE	FALSE	
## 3:	FALSE	FALSE	FALSE	FALSE	
## 4:	FALSE	FALSE	FALSE	FALSE	
## 5:	FALSE	FALSE	FALSE	FALSE	
## 6:	TRUE	TRUE	FALSE	FALSE	
##	people_char_28	people_char_29	people_char_30	people_char_31	
## 1:	FALSE	FALSE	FALSE	FALSE	
## 2:	FALSE	FALSE	FALSE	FALSE	
## 3:	FALSE	FALSE	FALSE	FALSE	
## 4:	FALSE	FALSE	FALSE	FALSE	
## 5:	FALSE	FALSE	FALSE	FALSE	
## 6:	FALSE	FALSE	FALSE	FALSE	
##	people_char_32	people_char_33	people_char_34	people_char_35	
## 1:	FALSE	FALSE	FALSE	FALSE	
## 2:	FALSE	FALSE	FALSE	FALSE	
## 3:	FALSE	FALSE	FALSE	FALSE	
## 4:	FALSE	FALSE	FALSE	FALSE	
## 5:	FALSE	FALSE	FALSE	FALSE	
## 6:	FALSE	FALSE	FALSE	FALSE	
##	people_char_36	people_char_37	people_char_38	activity_id	activity_date
## 1:	FALSE	FALSE	76	act1_9923	2022-11-25
## 2:	FALSE	FALSE	0	act1_198174	2022-07-26
## 3:	FALSE	FALSE	0	act1_214090	2023-06-15
## 4:	FALSE	FALSE	0	act1_230588	2023-02-28
## 5:	FALSE	FALSE	0	act1_271874	2022-07-26
## 6:	FALSE	TRUE	100	act1_104259	2023-07-28
##	activity_category	activity_char_1	activity_char_2	activity_char_3	
## 1:	type 1	type 3	type 5	type 1	
## 2:	type 1	type 36	type 11	type 5	
## 3:	type 1	type 24	type 6	type 6	
## 4:	type 1	type 2	type 2	type 3	
## 5:	type 1	type 2	type 5	type 3	
## 6:	type 1	type 5	type 2	type 7	
##	activity_char_4	activity_char_5	activity_char_6	activity_char_7	
## 1:	type 1	type 6	type 3	type 3	
## 2:	type 1	type 6	type 1	type 1	
## 3:	type 3	type 1	type 3	type 4	
## 4:	type 3	type 5	type 2	type 2	
## 5:	type 2	type 6	type 1	type 1	
## 6:	type 3	type 1	type 3	type 5	
##	activity_char_8	activity_char_9	activity_char_10	outcome	
## 1:	type 6	type 8		0	
## 2:	type 4	type 1		0	
## 3:	type 5	type 1		0	
## 4:	type 4	type 2		0	
## 5:	type 6	type 8		0	
## 6:	type 4	type 7		1	

Inspect outcomes for variable activity\_char\_1

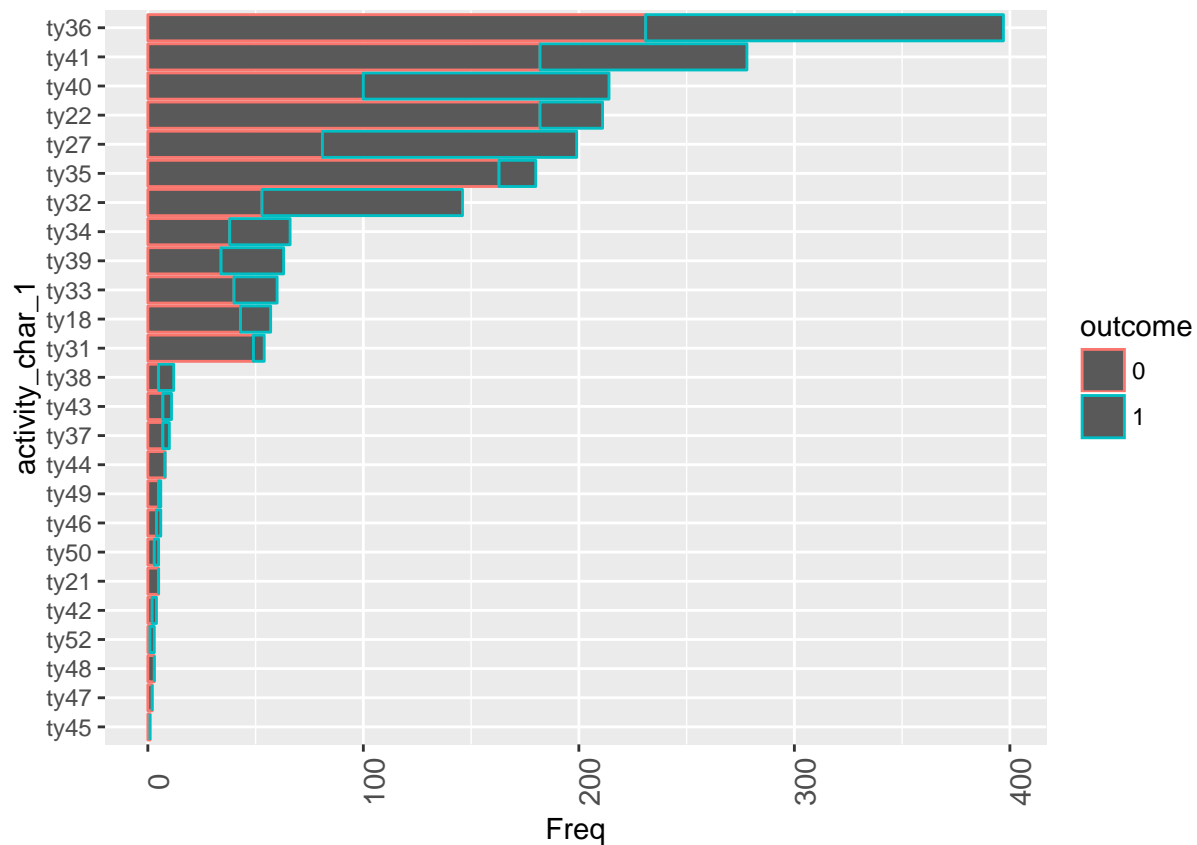
```
counts <- table(merged_raw[, c('activity_char_1', 'outcome'), with=F])

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$activity_char_1)[
  (ind_split+1): length(levels(activities_df$activity_char_1))]
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()
```

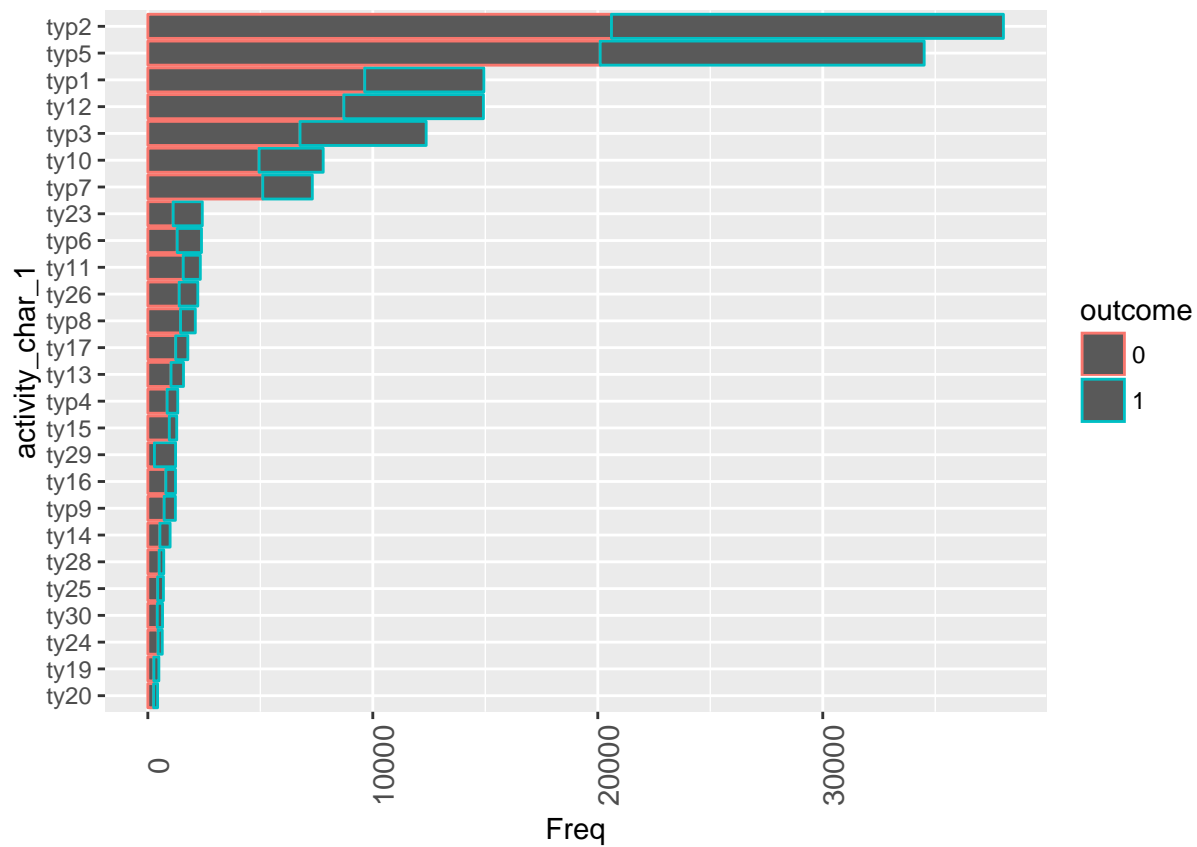


```
ggplot(data=activities_df[activities_df$activity_char_1 %in% second_frequent_levels, ],
  aes(x=activity_char_1, y=Freq, color=outcome)) +
  geom_bar(stat='identity') +
  theme(axis.text.x=element_text(angle=90, size=11)) +
  scale_x_discrete(labels=abbreviate) +
  coord_flip()
```



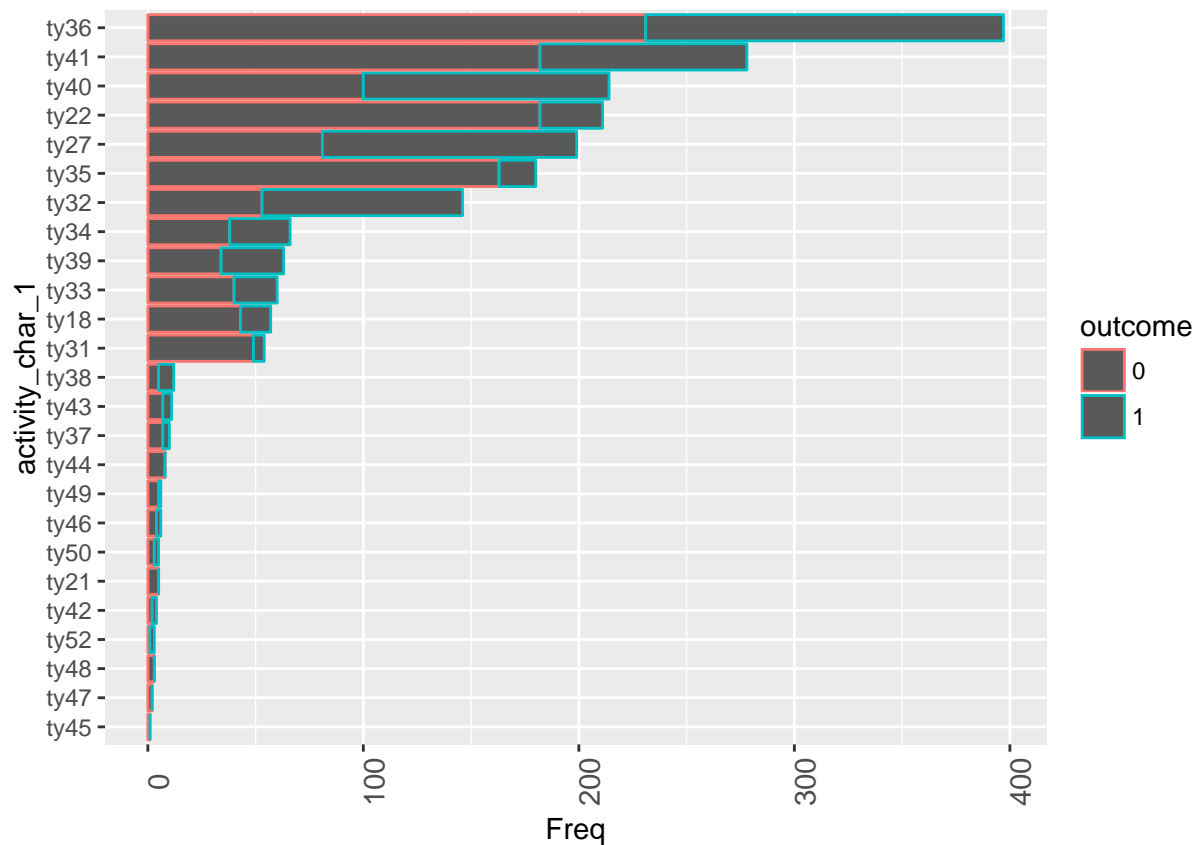
```
df_without_blanks <- activities_df[activities_df$activity_char_1 != '',]
df_without_blanks$activity_char_1 <- as.factor(df_without_blanks$activity_char_1)

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(merged_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
```

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      people_id  people_char_1  people_group_1
##      0      0      0      0
##  people_char_2  people_date  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  activity_id  activity_date
##          0          0          0          0
## activity_category  activity_char_1  activity_char_2  activity_char_3
##          0          2039676          2039676          2039676
##  activity_char_4  activity_char_5  activity_char_6  activity_char_7
##          2039676          2039676          2039676          2039676
##  activity_char_8  activity_char_9  activity_char_10  outcome
##          2039676          2039676          157615          0
```

Notice that the number of blanks for variables `activity_char_1` up to 9 is constant. This indicates that each record contains data associated to one specific activity.

Number of unqie values for each variable and when grouping over outcome

```
merged_raw[, lapply(.SD, function(x) length(unique(x))))]
```

```
##          V1  people_id  people_char_1  people_group_1  people_char_2
## 1: 2197291  151295          2          29899          3
##  people_date  people_char_3  people_char_4  people_char_5  people_char_6
## 1:    1196          43          25          9          7
##  people_char_7  people_char_8  people_char_9  people_char_10  people_char_11
## 1:    25          8          9          2          2
##  people_char_12  people_char_13  people_char_14  people_char_15
## 1:    2          2          2          2
##  people_char_16  people_char_17  people_char_18  people_char_19
## 1:    2          2          2          2
##  people_char_20  people_char_21  people_char_22  people_char_23
## 1:    2          2          2          2
##  people_char_24  people_char_25  people_char_26  people_char_27
## 1:    2          2          2          2
##  people_char_28  people_char_29  people_char_30  people_char_31
## 1:    2          2          2          2
##  people_char_32  people_char_33  people_char_34  people_char_35
## 1:    2          2          2          2
##  people_char_36  people_char_37  people_char_38  activity_id  activity_date
## 1:    2          2          101    2197291          411
##  activity_category  activity_char_1  activity_char_2  activity_char_3
```

```
## 1:          7          52          33          12
##   activity_char_4 activity_char_5 activity_char_6 activity_char_7
## 1:          8          8          6          9
##   activity_char_8 activity_char_9 activity_char_10 outcome
## 1:         19         20         6516          2
```

```
merged_raw[, lapply(.SD, function(x) length(unique(x))), by=outcome]
```

```
##   outcome      V1 people_id people_char_1 people_group_1 people_char_2
## 1:      0 1221794    89180          2      16850          3
## 2:      1  975497    68771          2      17302          2
##   people_date people_char_3 people_char_4 people_char_5 people_char_6
## 1:      1195         43         25         9         7
## 2:      1173         41         25         9         6
##   people_char_7 people_char_8 people_char_9 people_char_10 people_char_11
## 1:         25          8          9          2          2
## 2:         25          8          9          2          2
##   people_char_12 people_char_13 people_char_14 people_char_15
## 1:          2          2          2          2
## 2:          2          2          2          2
##   people_char_16 people_char_17 people_char_18 people_char_19
## 1:          2          2          2          2
## 2:          2          2          2          2
##   people_char_20 people_char_21 people_char_22 people_char_23
## 1:          2          2          2          2
## 2:          2          2          2          2
##   people_char_24 people_char_25 people_char_26 people_char_27
## 1:          2          2          2          2
## 2:          2          2          2          2
##   people_char_28 people_char_29 people_char_30 people_char_31
## 1:          2          2          2          2
## 2:          2          2          2          2
##   people_char_32 people_char_33 people_char_34 people_char_35
## 1:          2          2          2          2
## 2:          2          2          2          2
##   people_char_36 people_char_37 people_char_38 activity_id activity_date
## 1:          2          2          101    1221794          411
## 2:          2          2          64    975497          410
##   activity_category activity_char_1 activity_char_2 activity_char_3
## 1:          7          52          33          12
## 2:          7          47          32          12
##   activity_char_4 activity_char_5 activity_char_6 activity_char_7
## 1:          8          8          6          9
## 2:          8          7          6          9
##   activity_char_8 activity_char_9 activity_char_10
## 1:         19         20         5315
## 2:         19         20         4733
```

Check if non blank activity values are recorded groupwise

```
for (ind in 2:10) {
  colname <- paste0("activity_char_", ind)
  if (sum((merged_raw$activity_char_1 != '') != (merged_raw[, colname, with=F] != '')) != 0) {
```

```

    print(paste("Non blank indices for activitiy_char_1 and activity_char", ind, "differ"))
  }
  else {
    print(paste("Non blank indices for activitiy_char_1 and activity_char", ind, "are equal"))
  }
}

```

```

## [1] "Non blank indices for activitiy_char_1 and activity_char 2 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 3 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 4 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 5 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 6 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 7 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 8 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 9 are equal"
## [1] "Non blank indices for activitiy_char_1 and activity_char 10 differ"

```

```
gc()
```

```

##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 572207 30.6      940480 50.3    750400 40.1
## Vcells 774551  6.0     1650153 12.6   1040048  8.0

```

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)

Count value distribution for the activity categories

```
table(merged_raw$activity_category)
```

```

##
## type 1 type 2 type 3 type 4 type 5 type 6 type 7
## 157615 904683 429408 207465 490710  4253  3157

```

Number of unique values grouped by activity category

```

cols <- c(paste0('activity_char_', 1:9), 'activity_category', 'outcome')

activities_dt <- merged_raw[, cols, with=F]
dt <- activities_dt[,
  , lapply(.SD, function(x) length(unique(x))), by=list(activity_category, outcome)]
colnames(dt) <- gsub('activity_', '', colnames(dt))

merge_cols <- colnames(dt)[which(!colnames(dt) %in% c('category', 'outcome'))]
long <- reshape(data=dt, varying=merge_cols,
  v.names='num_unique_values',
  timevar='variable', times=merge_cols, direction='long')

ggplot(data=long, aes(x=variable, y=num_unique_values, colour=factor(outcome))) +
  facet_grid(category ~ .) +
  theme(axis.text.x=element_text(angle=90, size=9)) +
  geom_bar(stat='identity') ## coord_flip()

```

