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)

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')
}
```

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 6.8% of 2197291 rows
Read 13.7% of 2197291 rows
Read 17.7% of 2197291 rows
Read 24.1% of 2197291 rows
Read 30.5% of 2197291 rows
Read 33.7% of 2197291 rows
Read 40.0% of 2197291 rows
Read 44.1% of 2197291 rows
Read 50.5% of 2197291 rows
Read 56.4% of 2197291 rows
Read 62.8% of 2197291 rows
Read 68.7% of 2197291 rows
Read 71.5% of 2197291 rows
Read 77.8% of 2197291 rows
Read 84.2% of 2197291 rows
Read 89.7% of 2197291 rows
Read 95.6% of 2197291 rows
Read 2197291 rows and 56 (of 56) columns from 0.766 GB file in 00:00:34
```

Data summary

```r
head(merged_raw,2)
```

```
##    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
##    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 activity_date
## 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
```

Inspect outomces for variable activity_char_1

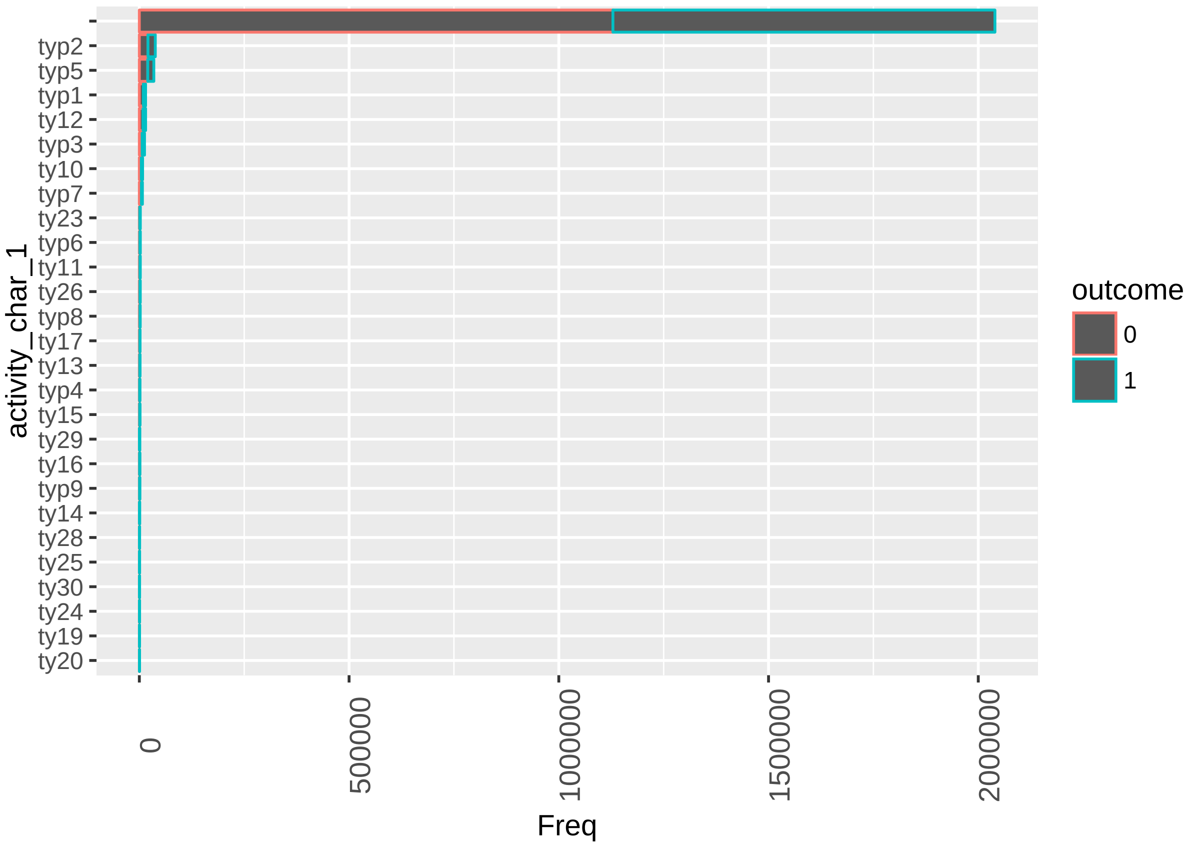
```r
gc()
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 592991 31.7     940480 50.3   750400 40.1
## Vcells 804624  6.2    1650153 12.6  1085240  8.3
```

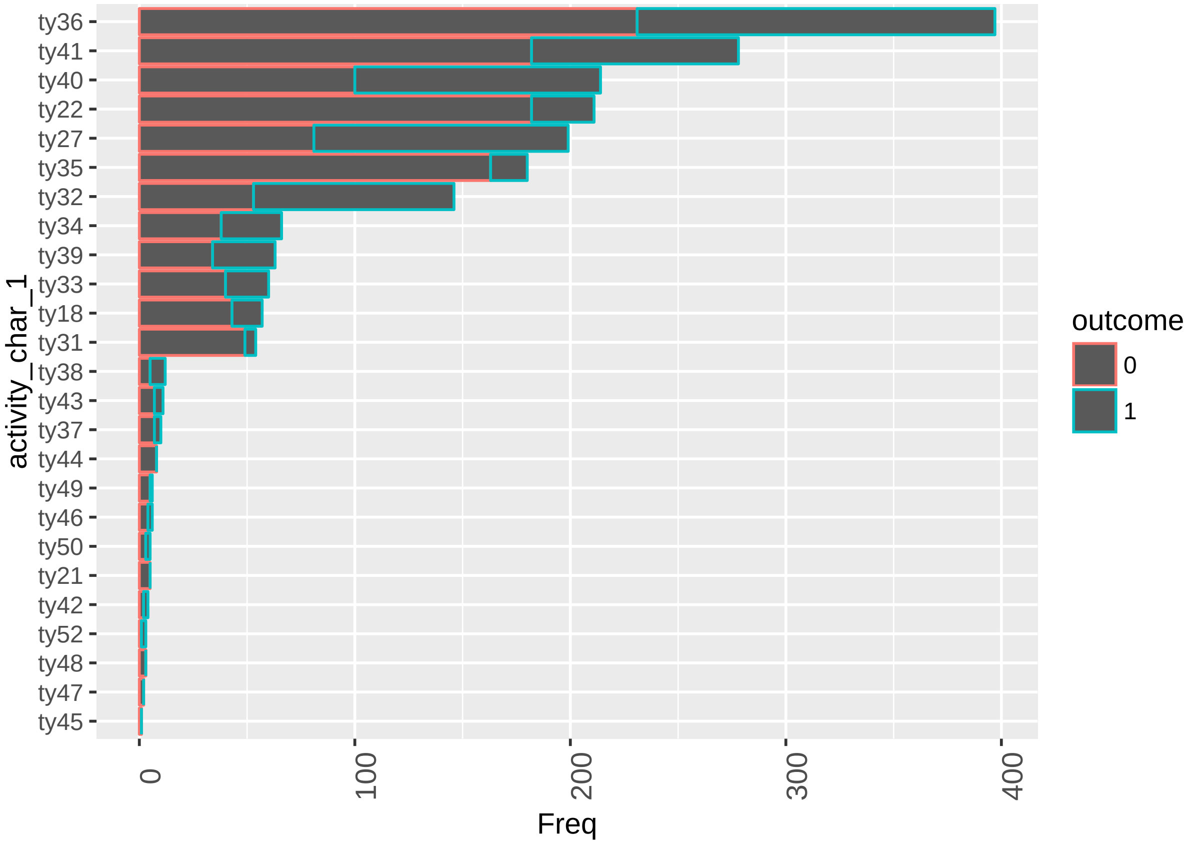
```r
counts <- table(merged_raw[, c('activity_char_1', 'outcome'), with=F])

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()
```

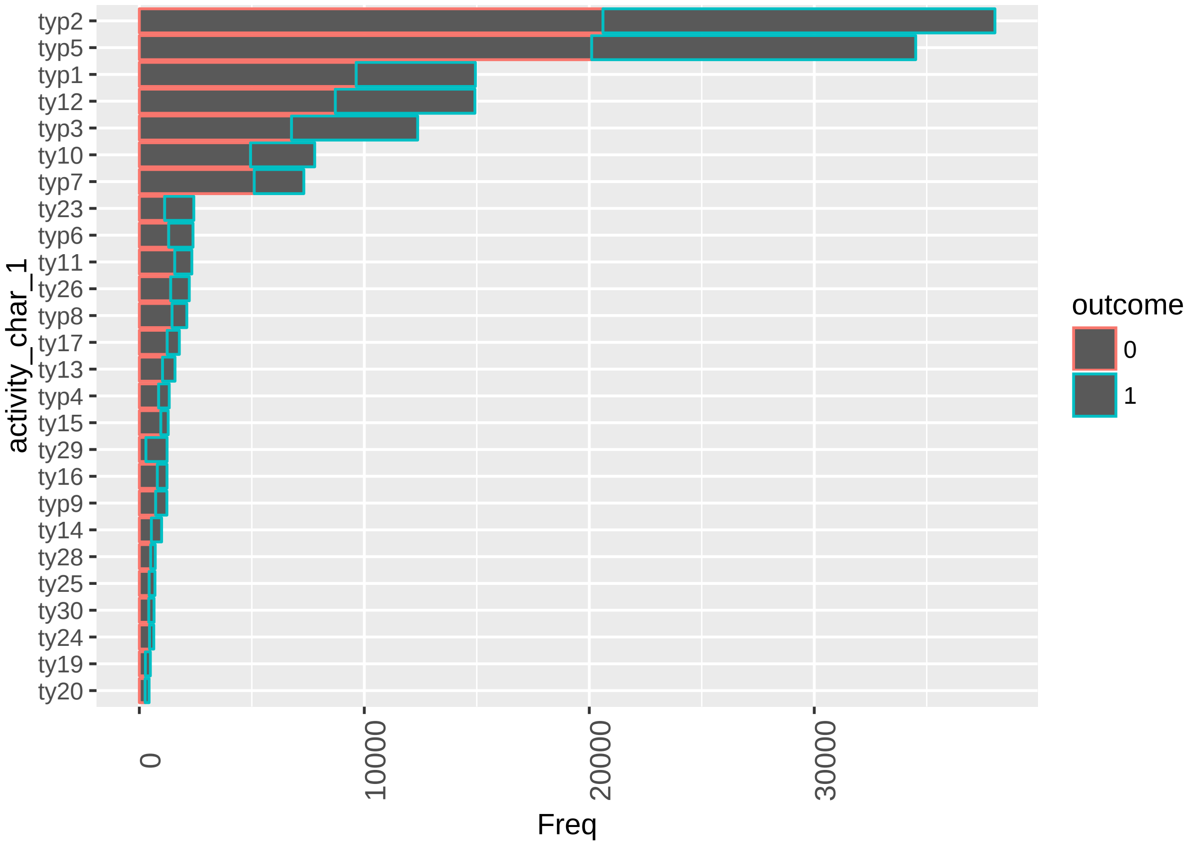


```r
ggplot(data=activities_df[activities_df$activity_char_1 %in% second_frequent_levels,],
       aes(x=activity_char_1, y=Freq, color=outcome)) +
    geom_bar(stat='identity') +
    theme(axis.text.x=element_text(angle=90, size=11)) +
    scale_x_discrete(labels=abbreviate) +
    coord_flip()
```

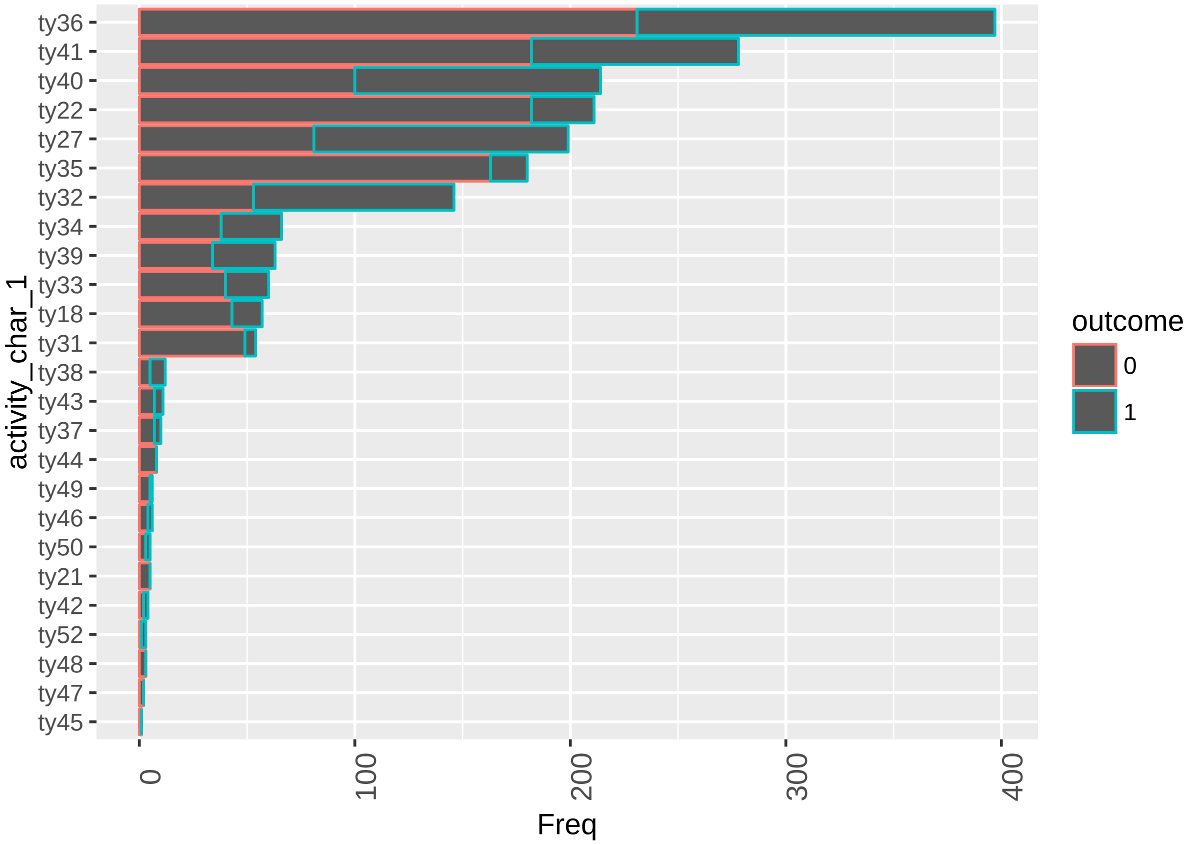


```
df_without_blanks <- activities_df[activities_df$activity_char_1 != '',]
df_without_blanks$activity_char_1 <- as.factor(df_without_blanks$activity_char_1)

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                  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
```

Number of unqie values for each variable

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

```
##               V1        people_id    people_char_1   people_group_1
##          2197291           151295                2            29899
##    people_char_2      people_date    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      activity_id    activity_date
##                2              101          2197291              411
## activity_category  activity_char_1  activity_char_2  activity_char_3
##                 7                52                33                12
##   activity_char_4   activity_char_5   activity_char_6   activity_char_7
##                 8                 8                 6                 9
##   activity_char_8   activity_char_9  activity_char_10           outcome
##                19                20              6516                 2
```

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) length(unique(x)))
```

```
##                V1         people_id     people_char_1    people_group_1
##           2039676            141558                 2             28431
##     people_char_2       people_date     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       activity_id     activity_date
##                 2               101           2039676               386
## activity_category   activity_char_1   activity_char_2   activity_char_3
##                 6                 1                 1                 1
##   activity_char_4   activity_char_5   activity_char_6   activity_char_7
##                 1                 1                 1                 1
##   activity_char_8   activity_char_9  activity_char_10           outcome
##                 1                 1              6515                 2
```

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

```
##                V1         people_id     people_char_1    people_group_1
##           2039676            141558                 2             28431
##     people_char_2       people_date     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       activity_id     activity_date
##                 2               101           2039676               386
## activity_category   activity_char_1   activity_char_2   activity_char_3
##                 6                 1                 1                 1
##   activity_char_4   activity_char_5   activity_char_6   activity_char_7
##                 1                 1                 1                 1
##   activity_char_8   activity_char_9  activity_char_10           outcome
##                 1                 1              6515                 2
```

Number of unique values while fixing activity 1

```
apply(merged_raw[merged_raw$activity_char_2==merged_raw$activity_char_1[1], ],
      MARGIN=2, function(x) length(unique(x)))
```

```
##                V1         people_id     people_char_1    people_group_1
##           2039676            141558                 2             28431
##     people_char_2       people_date     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       activity_id     activity_date
##                 2               101           2039676               386
## activity_category   activity_char_1   activity_char_2   activity_char_3
##                 6                 1                 1                 1
##   activity_char_4   activity_char_5   activity_char_6   activity_char_7
##                 1                 1                 1                 1
##   activity_char_8   activity_char_9  activity_char_10           outcome
##                 1                 1              6515                 2
```

```
apply(merged_raw[merged_raw$activity_char_2==merged_raw$activity_char_1[250], ],
      MARGIN=2, function(x) length(unique(x)))
```

```
##                V1         people_id     people_char_1    people_group_1
##           2039676            141558                 2             28431
##     people_char_2       people_date     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       activity_id     activity_date
##                 2               101           2039676               386
## activity_category   activity_char_1   activity_char_2   activity_char_3
##                 6                 1                 1                 1
##   activity_char_4   activity_char_5   activity_char_6   activity_char_7
##                 1                 1                 1                 1
##   activity_char_8   activity_char_9  activity_char_10           outcome
##                 1                 1              6515                 2
```