Untitled

Objective

Red Hat Customer Survey

The objective of this project is to classify customer potential based on a survey conducted by Red Hat. The database for this project has been downloaded from Kaggle. The main goal of this project it to predict the potential business value of a person based on an activity performed by that persion over a given time frame. The business value outcome of each person is defined by Yes/No for each unique activity. This is used to indicate if a person has finished an activity within a time frame.

Library

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

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

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

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

## Loading required package: lattice

##
## Attaching package: 'kernlab'

## The following object is masked from 'package:ggplot2':
##
##     alpha

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

Load data

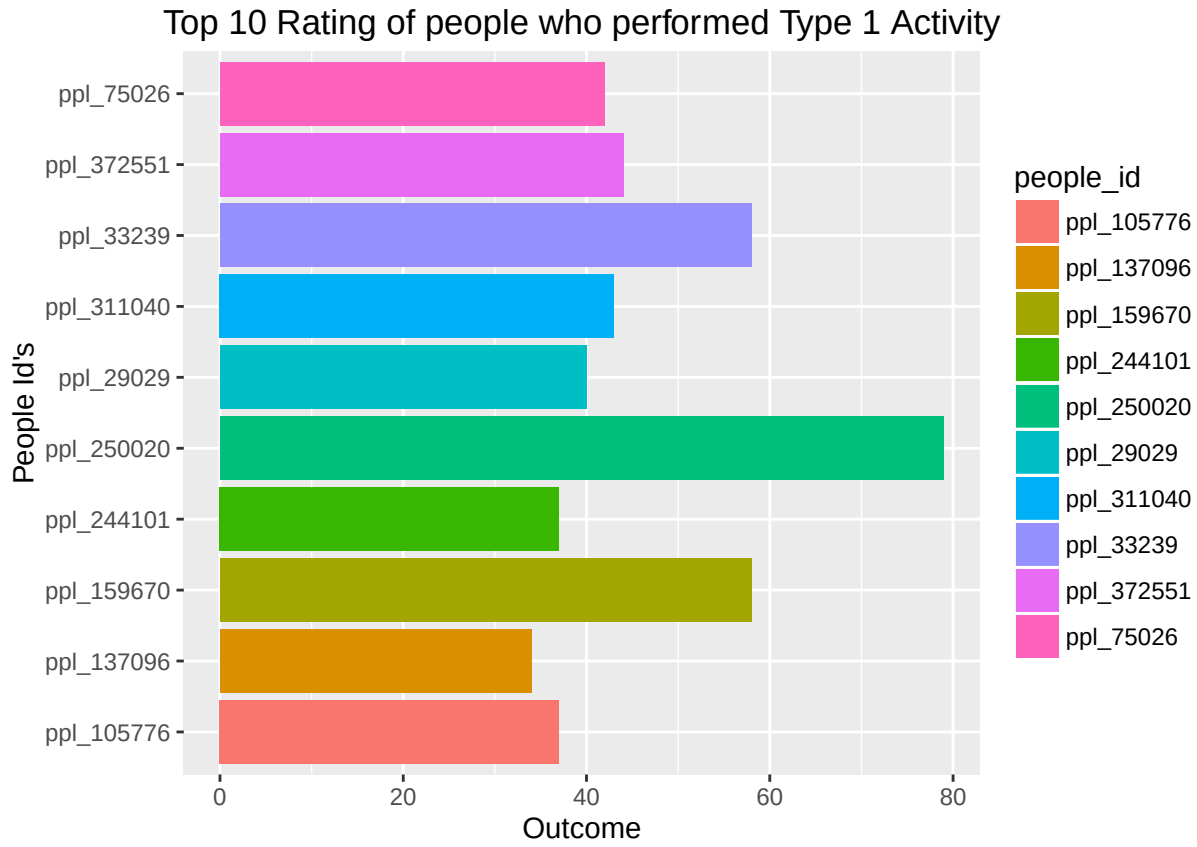
Cleaning and Prepare training data set

Analysis 1

Cleaning and Prepare testing data set

Split data in traing and testing sets for Type 1 activity

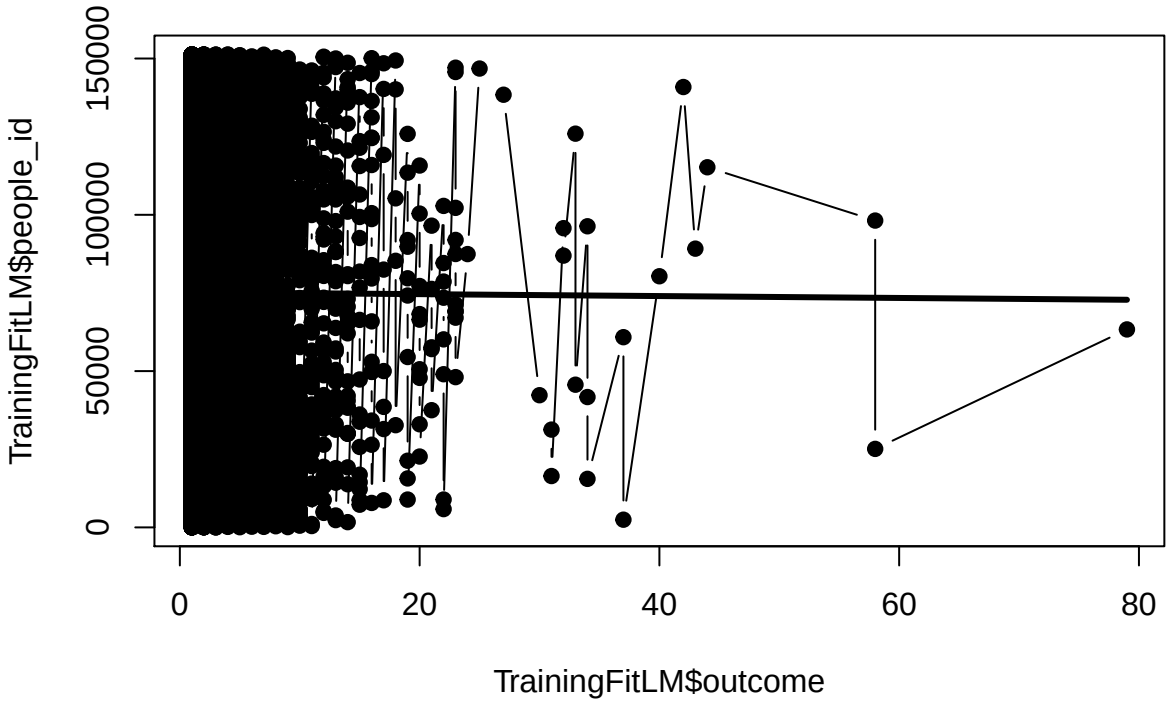
Prediction with Regression for Type 1 activity



Fit a linear model for Type 1 Activity

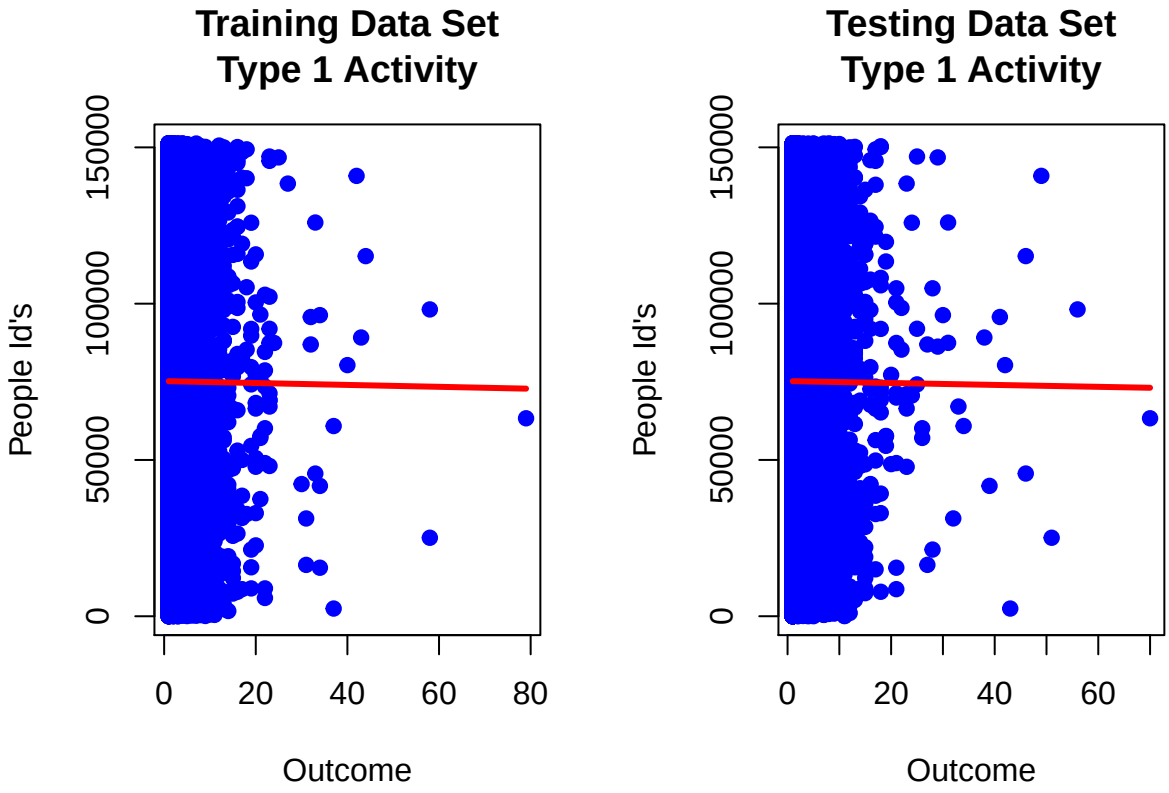
```
##
## Call:
## lm(formula = people_id ~ outcome, data = TrainingFitLM)
##
## Coefficients:
## (Intercept)      outcome
##     75226.6        -30.5
```

Model Fit for Type 1 Activity



Predict a new value for Type 1 Activity

Plot predictions - Testing and Training for Type 1 Activity



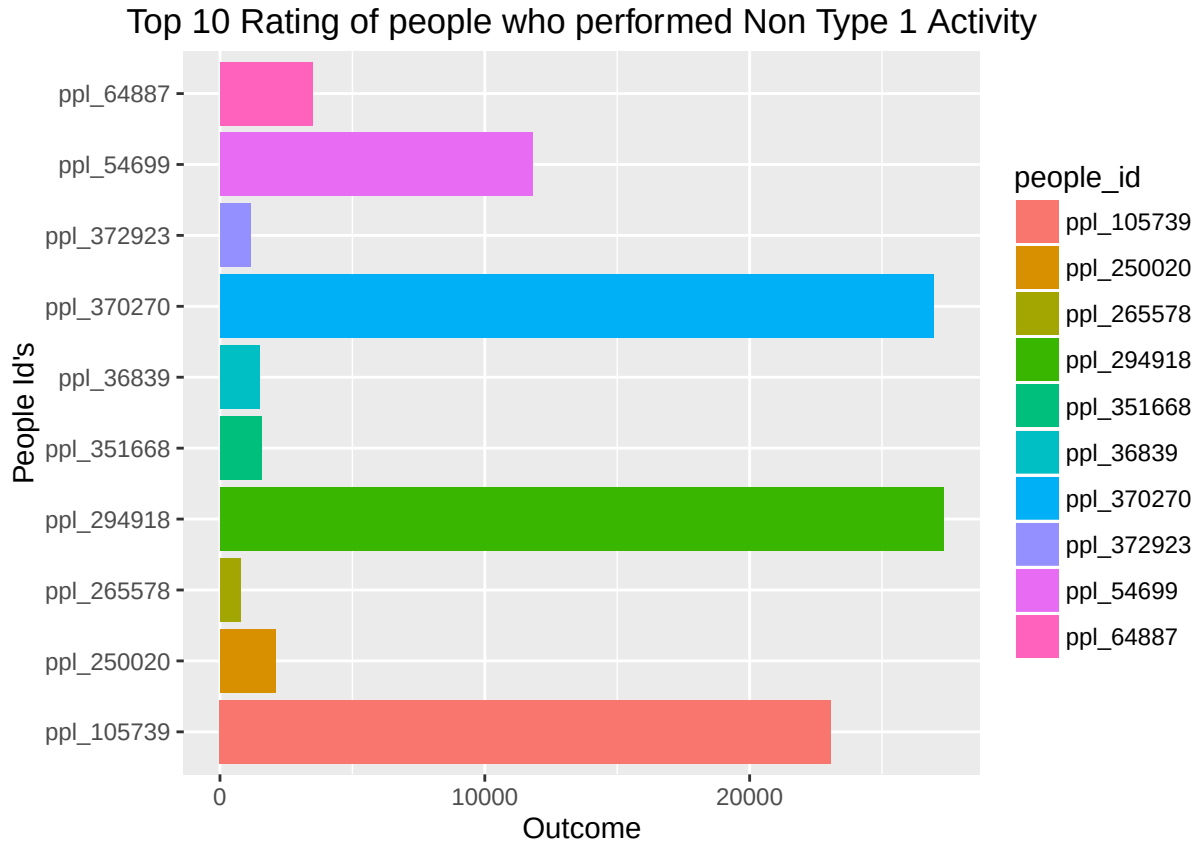
Calculate Traing and Testing data set errors for Type 1 Activity

```
## [1] 9503898
```

```
## [1] 9518819
```

Split data in traing and testing sets for Non Type 1 activity

Predict with regression for non Type 1 Activity

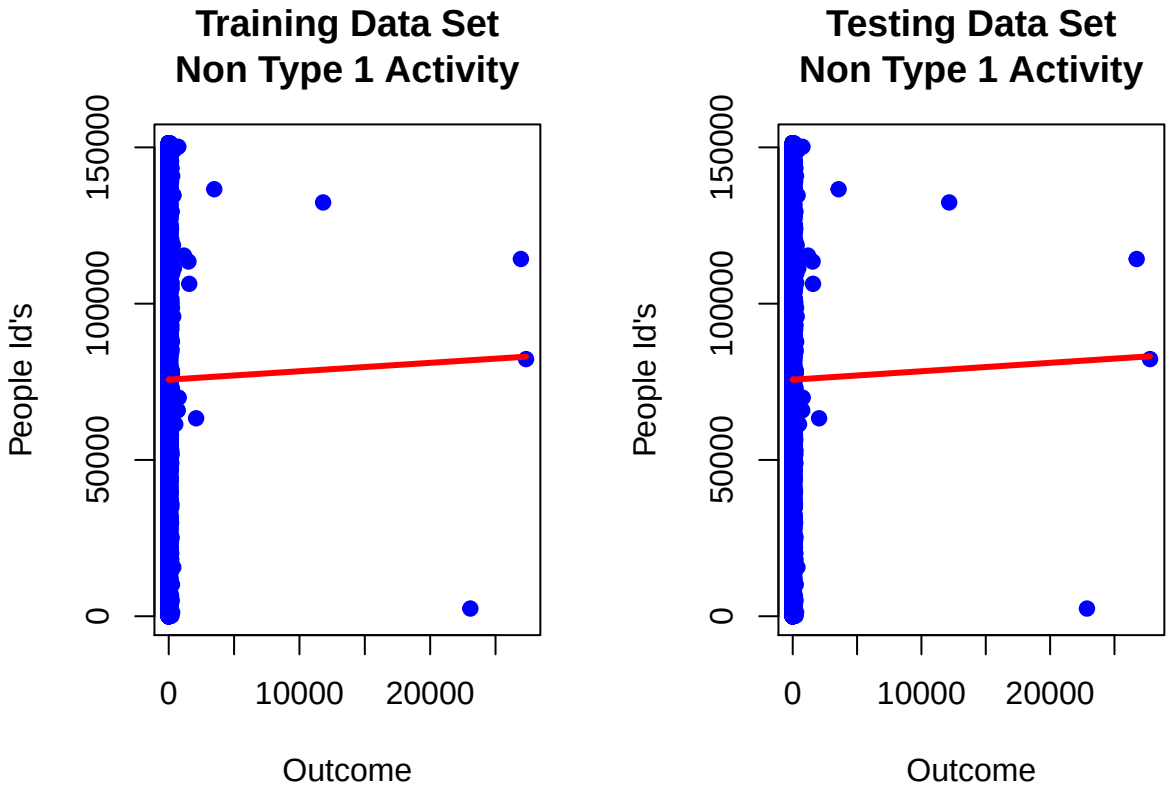


Fit a linear model for non Type 1 Activity

```
##
## Call:
## lm(formula = people_id ~ outcome, data = TrainingFitNonType1LM)
##
## Coefficients:
## (Intercept)      outcome
##   7.568e+04    2.695e-01
```

Predict a new value for non Type 1 Activity

Plot prediction - testing and traing for non Type 1 activity

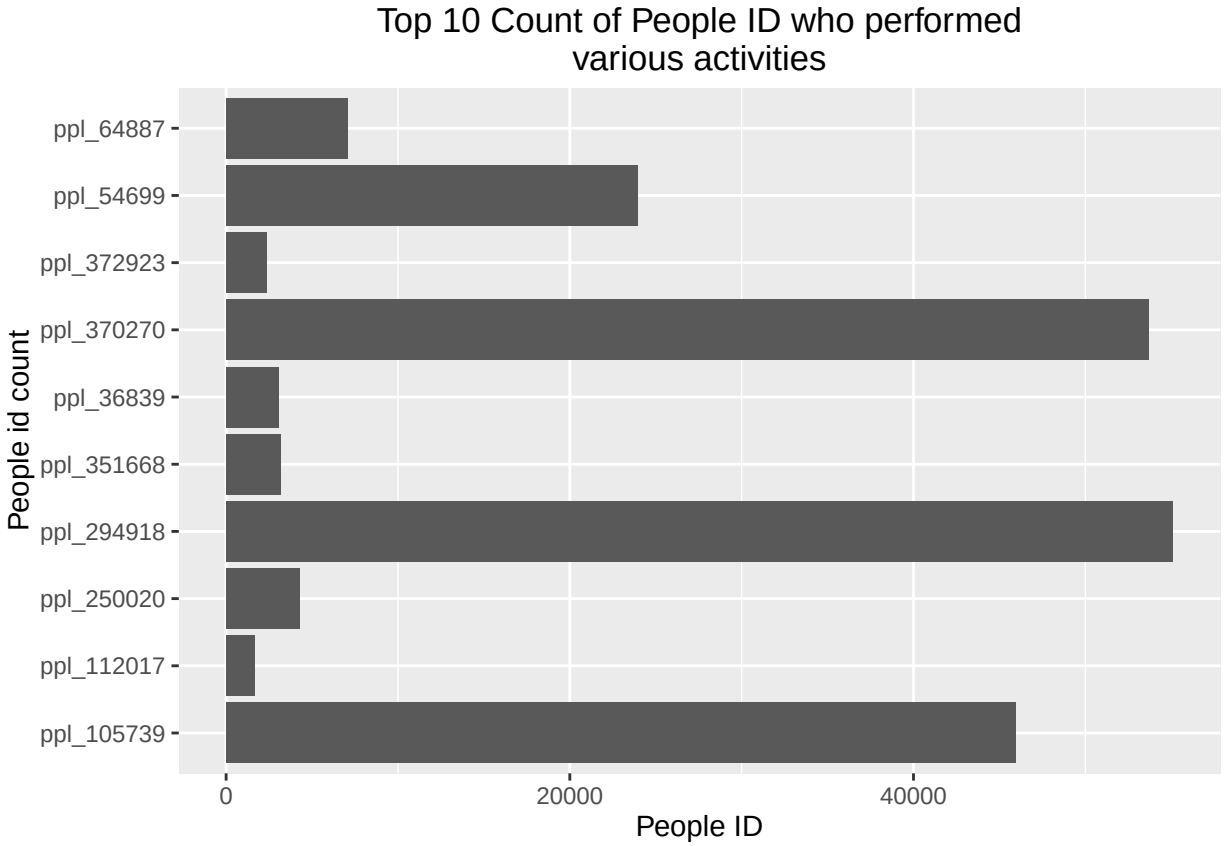


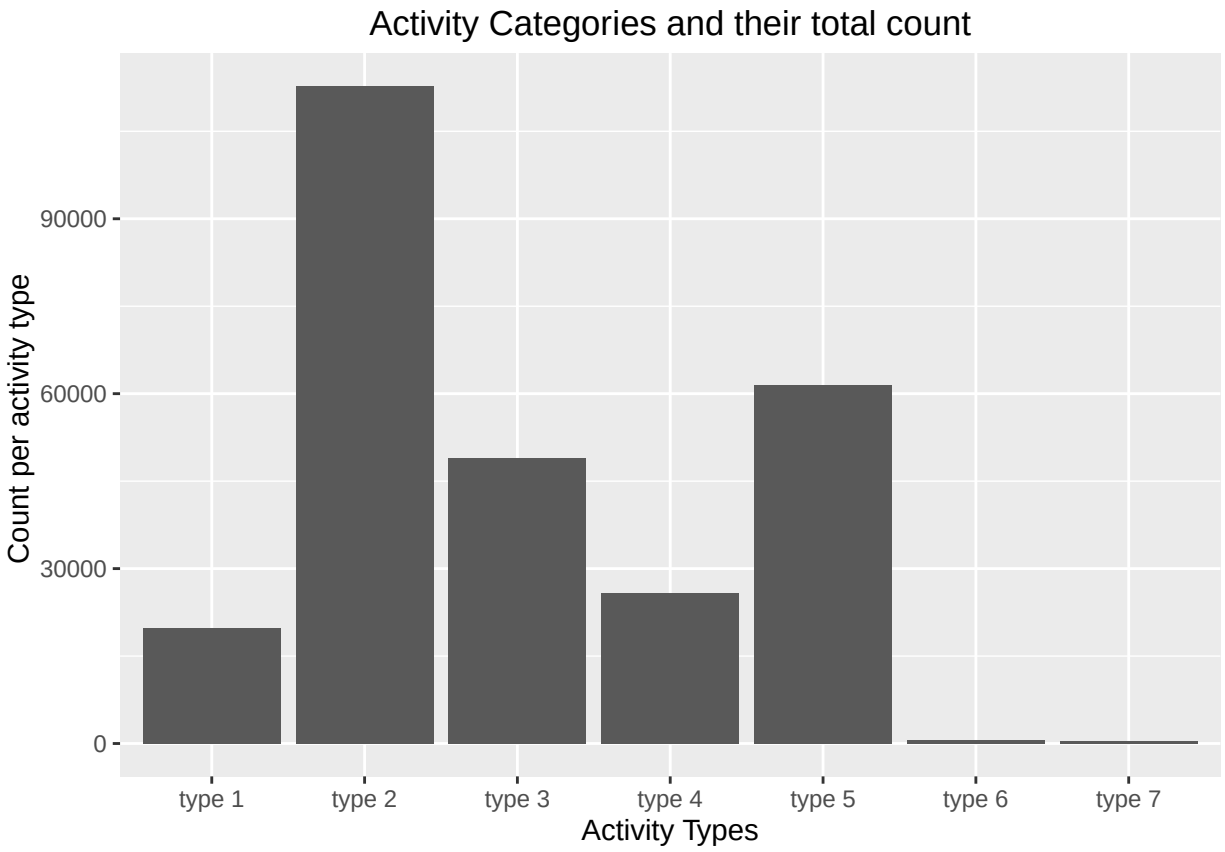
Calculate training and testing data set errors for non Type 1 activity

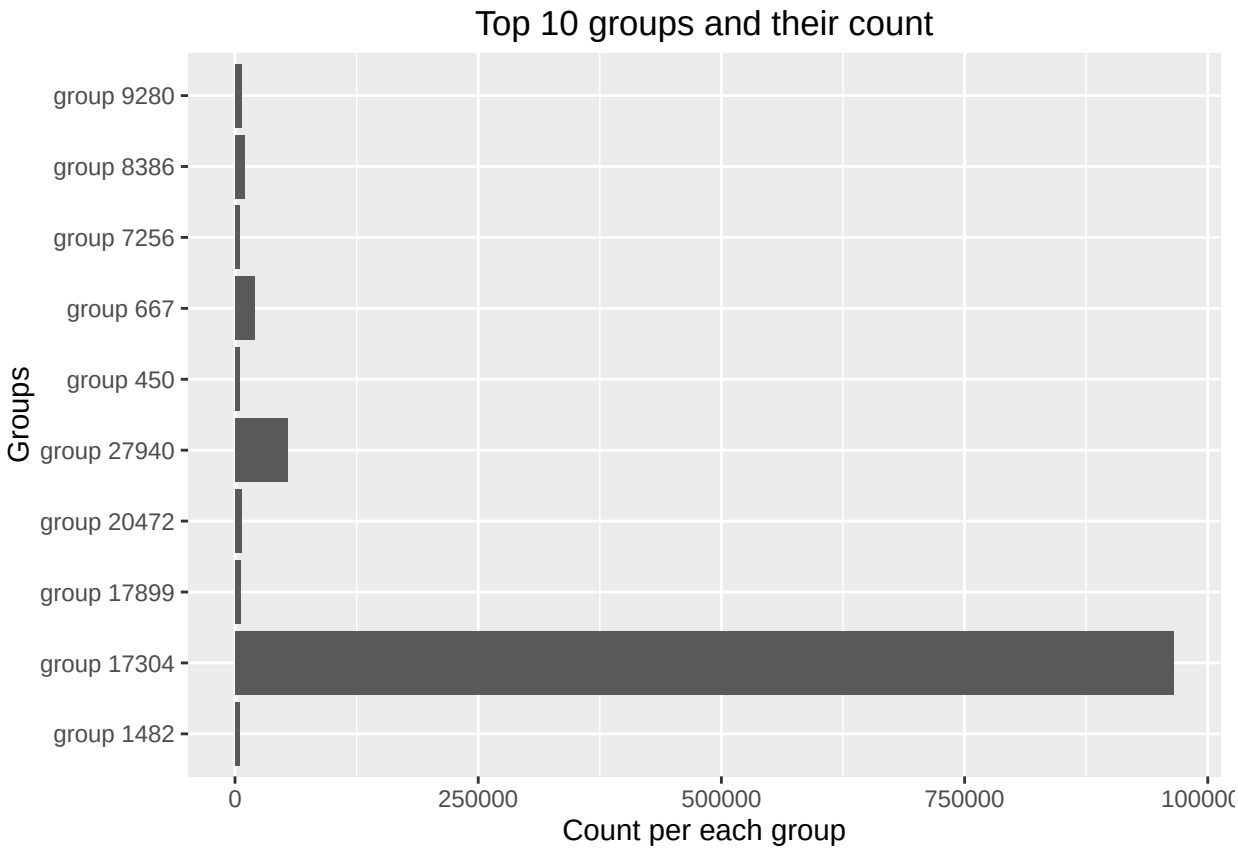
```
## [1] 15178340
```

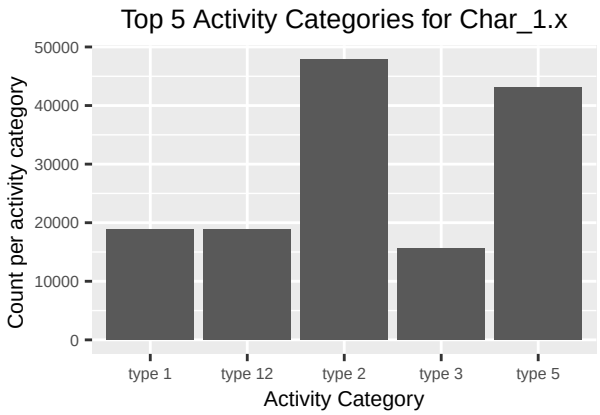
```
## [1] 15173857
```

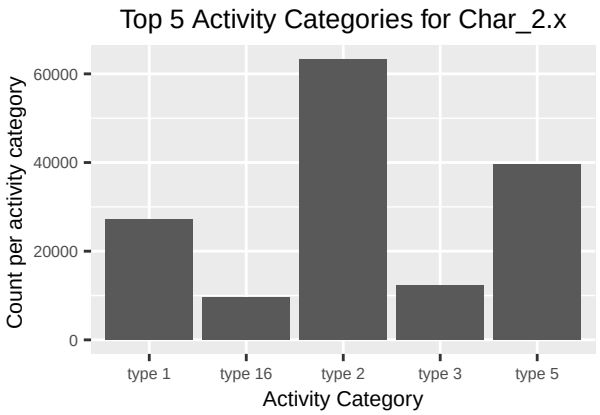
Analysis 2

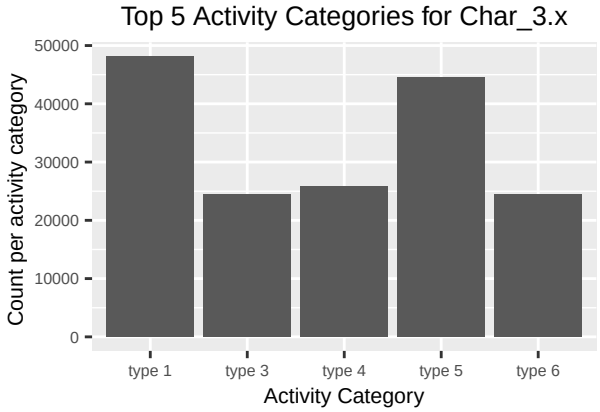

Top 10 Count of People ID who performed various activities
People id count
ppl_64887
ppl_54699
ppl_372923
ppl_370270
ppl_36839
ppl_351668
ppl_294918
ppl_250020
ppl_112017
ppl_105739
0
20000
40000
People ID

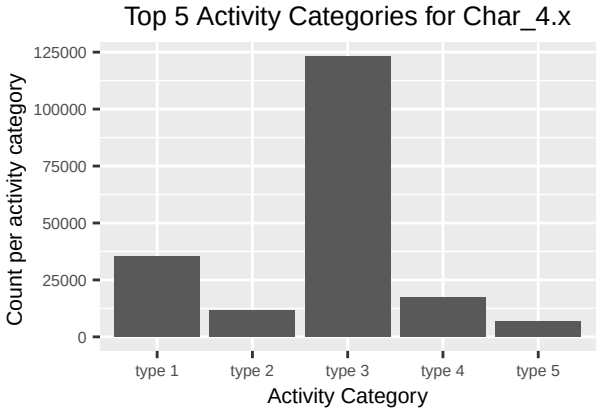

Activity Categories and their total count
Count per activity type
90000
60000
30000
0
type 1
type 2
type 3
type 4
type 5
type 6
type 7
Activity Types

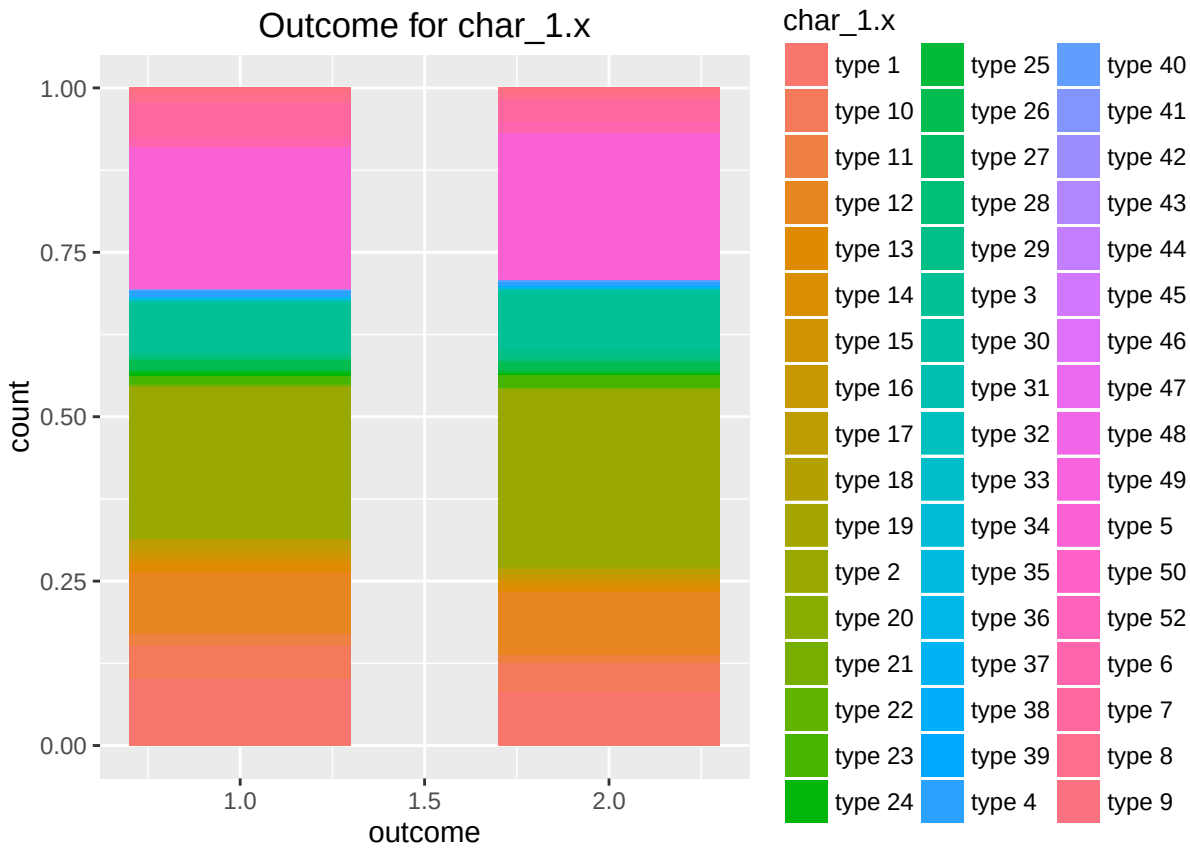

Top 10 groups and their count
group 9280
group 8386
group 7256
group 667
group 450
group 27940
group 20472
group 17899
group 17304
group 1482
Groups
0
250000
500000
750000
100000
Count per each group

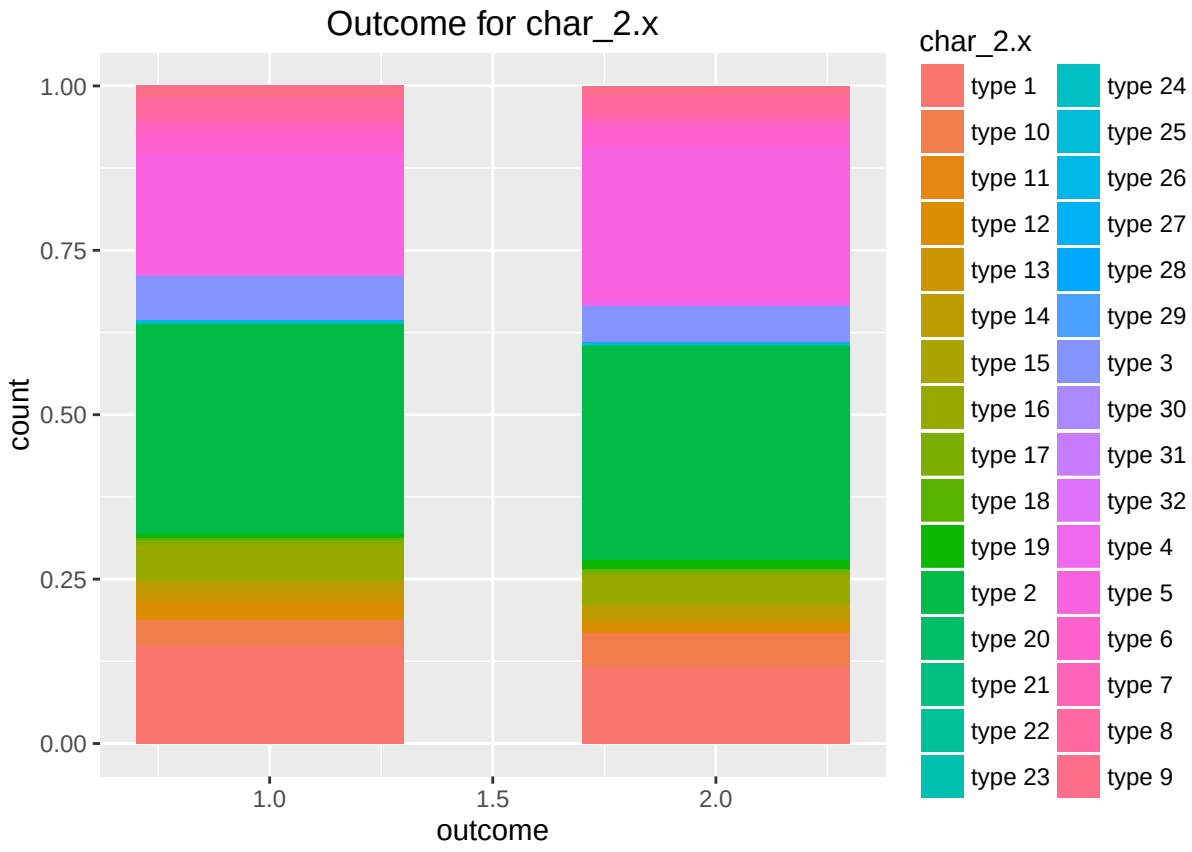

Top 5 Activity Categories for Char_1.x
Count per activity category
50000
40000
30000
20000
10000
0
type 1
type 12
type 2
type 3
type 5
Activity Category

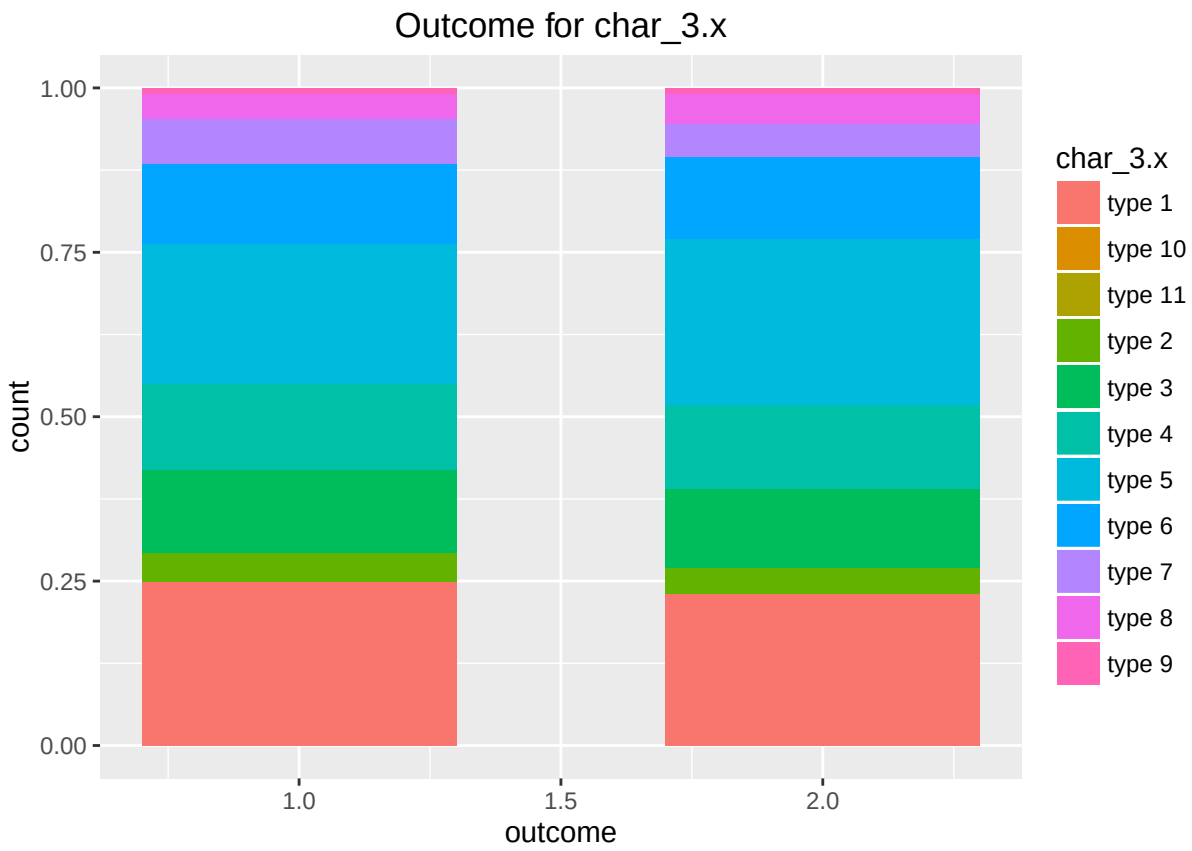

Top 5 Activity Categories for Char_2.x
Count per activity category
60000
40000
20000
0
type 1
type 16
type 2
type 3
type 5
Activity Category

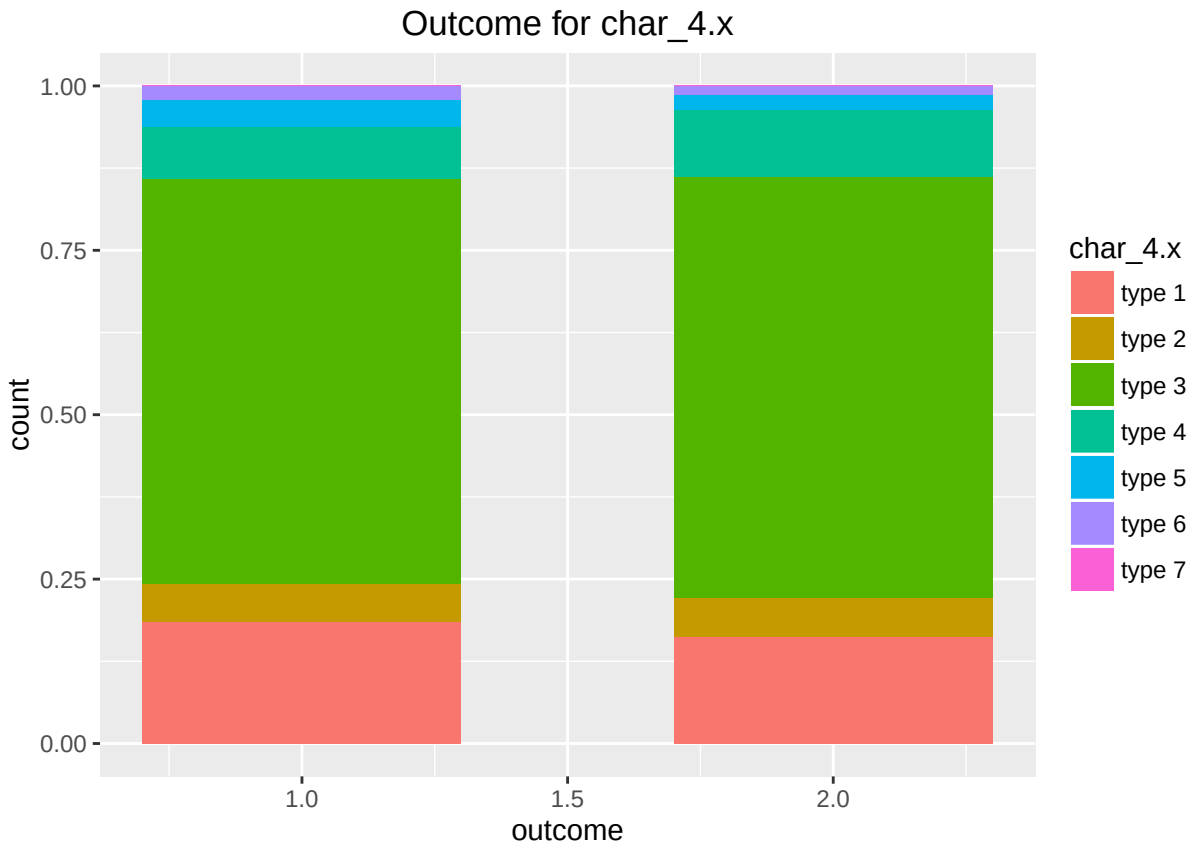

Top 5 Activity Categories for Char_3.x
Count per activity category
50000
40000
30000
20000
10000
0
type 1
type 3
type 4
type 5
type 6
Activity Category

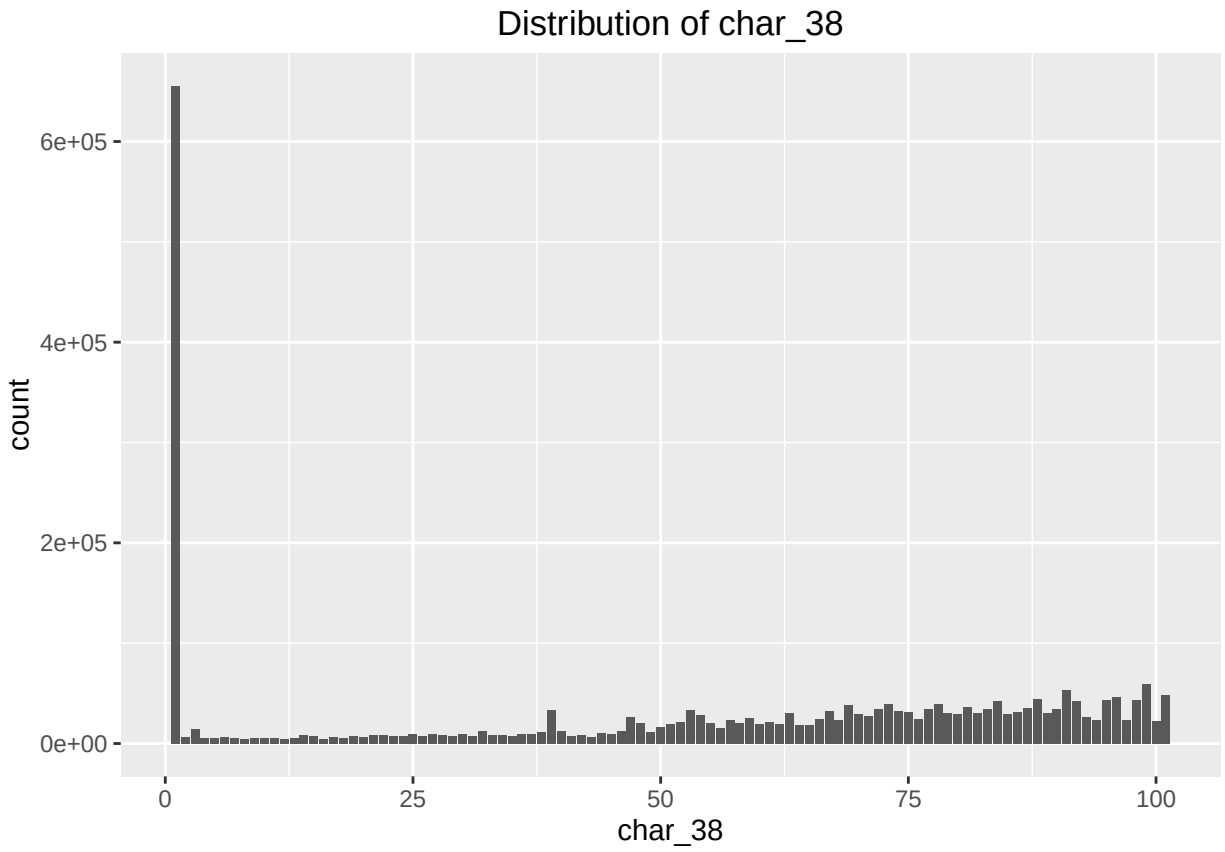

Top 5 Activity Categories for Char_4.x
Count per activity category
125000
100000
75000
50000
25000
0
type 1
type 2
type 3
type 4
type 5
Activity Category

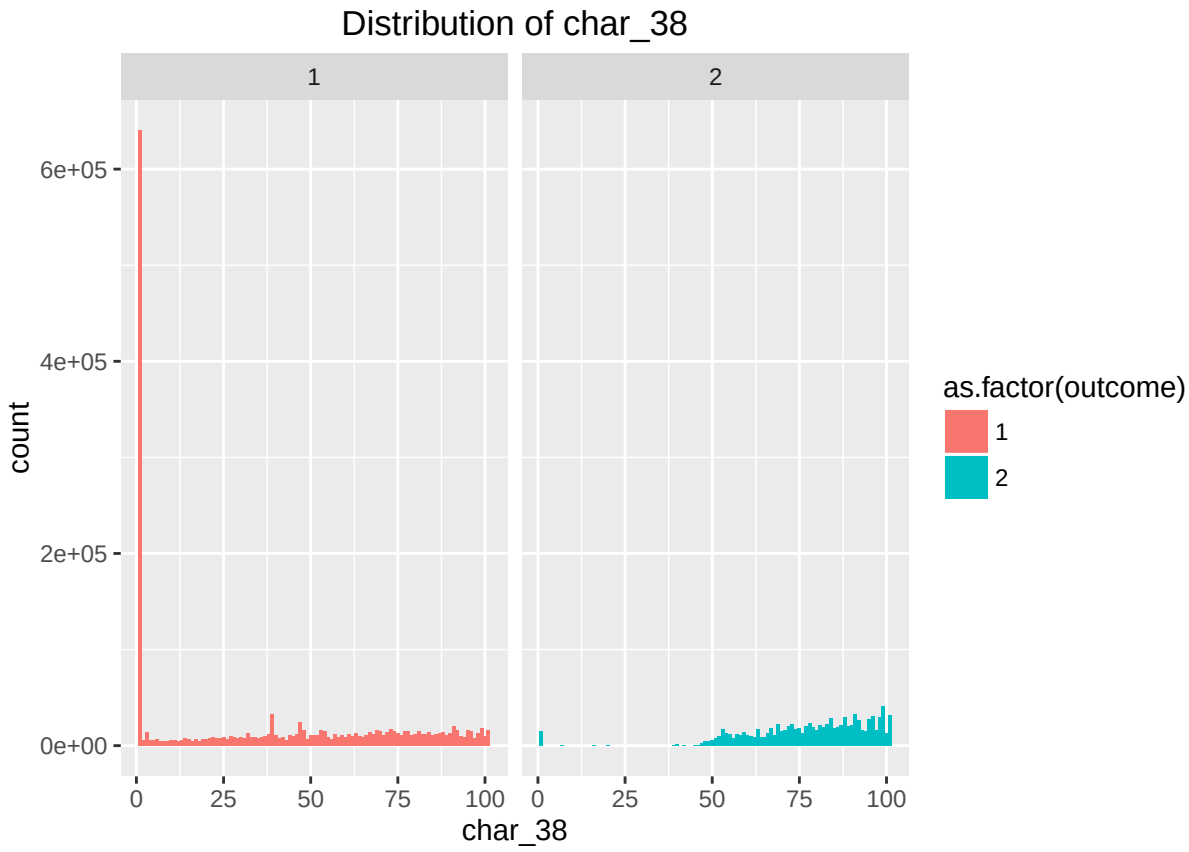

Outcome for char_1.x
count
1.00
0.75
0.50
0.25
0.00
1.0
1.5
2.0
outcome
char_1.x
type 1
type 10
type 11
type 12
type 13
type 14
type 15
type 16
type 17
type 18
type 19
type 2
type 20
type 21
type 22
type 23
type 24
type 25
type 26
type 27
type 28
type 29
type 3
type 30
type 31
type 32
type 33
type 34
type 35
type 36
type 37
type 38
type 39
type 4
type 40
type 41
type 42
type 43
type 44
type 45
type 46
type 47
type 48
type 49
type 5
type 50
type 52
type 6
type 7
type 8
type 9


Outcome for char_2.x
count
1.00
0.75
0.50
0.25
0.00
1.0
1.5
2.0
outcome
char_2.x
type 1
type 10
type 11
type 12
type 13
type 14
type 15
type 16
type 17
type 18
type 19
type 2
type 20
type 21
type 22
type 23
type 24
type 25
type 26
type 27
type 28
type 29
type 3
type 30
type 31
type 32
type 4
type 5
type 6
type 7
type 8
type 9


Outcome for char_3.x
1.00
0.75
0.50
0.25
0.00
count
1.0
1.5
2.0
outcome
char_3.x
type 1
type 10
type 11
type 2
type 3
type 4
type 5
type 6
type 7
type 8
type 9

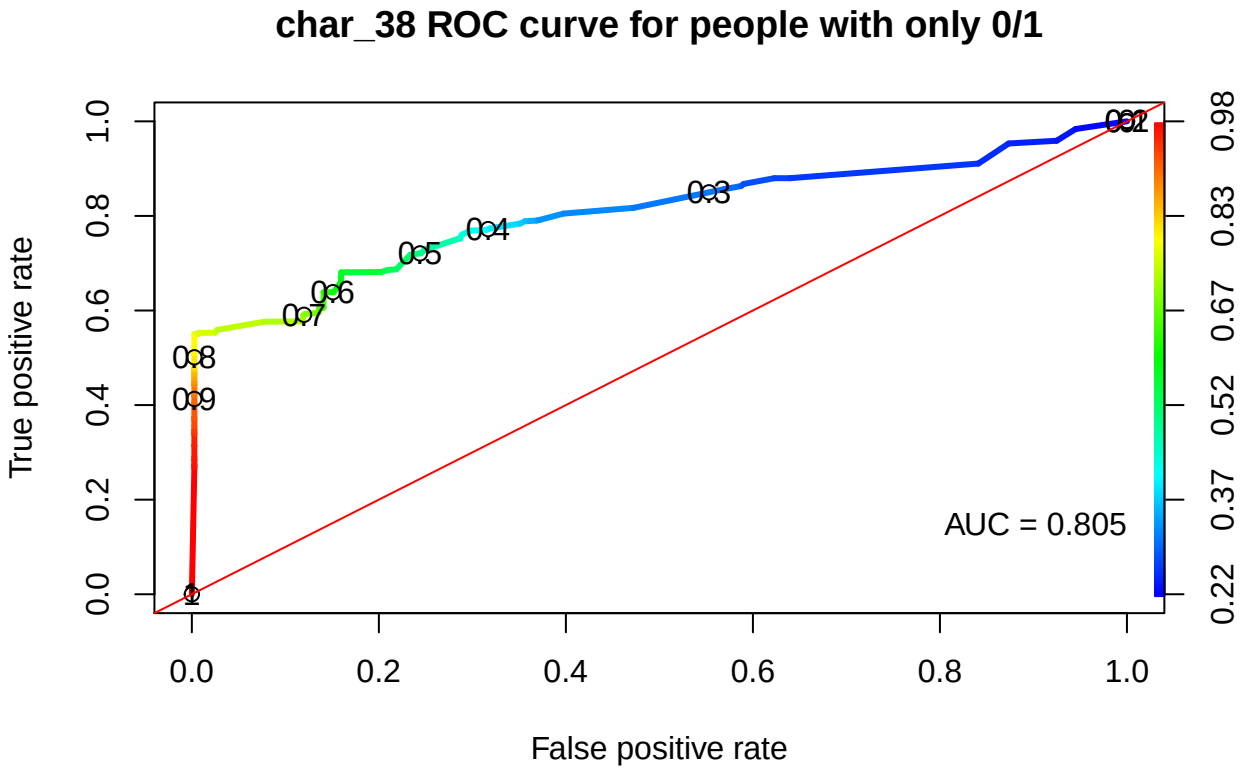

Outcome for char_4.x
1.00
0.75
0.50
0.25
0.00
count
1.0
1.5
2.0
outcome
char_4.x
type 1
type 2
type 3
type 4
type 5
type 6
type 7


Distribution of char_38
6e+05
4e+05
2e+05
0e+00
count
0
25
50
75
100
char_38



Receiver Operating Characteristics

```
## numeric(0)
```



Conclusion

The ROC curve for char_38 was plotted with a sample size of 5000 records from the combined data set comprising of people, act_train and act_test. The AUC value comes to 0.805 which shows that the prediction is correct as it is very near to AUC value of 1.