Predicting Hospital Readmission for Diabetes patients

Rishabh Lavangad

26 December 2018

```
library(dplyr)
```

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

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

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

```
library(GGally)
```

```
## Loading required package: ggplot2

##
## Attaching package: 'GGally'

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

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(rpart)
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:psych':
##
##     outlier
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

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

```
library(nnet)
library(e1071)
library(naivebayes)
```

```
## naivebayes 0.9.6 loaded
```

Loading the data

```
raw.data <- read.csv("/home/sunbeam/R_loveeshbhat_imarticus/programs/R_project/github/provided_files/dia
head(raw.data)
```

```
##   encounter_id patient_nbr            race gender     age weight
## 1      2278392     8222157       Caucasian Female  [0-10)   <NA>
## 2       149190    55629189       Caucasian Female [10-20)   <NA>
## 3        64410    86047875 AfricanAmerican Female [20-30)   <NA>
## 4       500364    82442376       Caucasian   Male [30-40)   <NA>
## 5        16680    42519267       Caucasian   Male [40-50)   <NA>
## 6        35754    82637451       Caucasian   Male [50-60)   <NA>
##   admission_type_id discharge_disposition_id admission_source_id
## 1                 6                       25                   1
## 2                 1                        1                   7
## 3                 1                        1                   7
## 4                 1                        1                   7
## 5                 1                        1                   7
## 6                 2                        1                   2
##   time_in_hospital payer_code        medical_specialty num_lab_procedures
## 1                1       <NA> Pediatrics-Endocrinology                 41
## 2                3       <NA>                      <NA>                59
```

```
## 3               2       <NA>                    <NA>                11
## 4               2       <NA>                    <NA>                44
## 5               1       <NA>                    <NA>                51
## 6               3       <NA>                    <NA>                31
##   num_procedures num_medications number_outpatient number_emergency
## 1              0               1                 0                0
## 2              0              18                 0                0
## 3              5              13                 2                0
## 4              1              16                 0                0
## 5              0               8                 0                0
## 6              6              16                 0                0
##   number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
## 1                0 250.83   <NA>   <NA>                1          None
## 2                0    276 250.01    255                9          None
## 3                1    648    250    V27                6          None
## 4                0      8 250.43    403                7          None
## 5                0    197    157    250                5          None
## 6                0    414    411    250                9          None
##   A1Cresult metformin repaglinide nateglinide chlorpropamide glimepiride
## 1      None        No          No          No             No          No
## 2      None        No          No          No             No          No
## 3      None        No          No          No             No          No
## 4      None        No          No          No             No          No
## 5      None        No          No          No             No          No
## 6      None        No          No          No             No          No
##   acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone
## 1            No        No        No          No           No            No
## 2            No        No        No          No           No            No
## 3            No    Steady        No          No           No            No
## 4            No        No        No          No           No            No
## 5            No    Steady        No          No           No            No
## 6            No        No        No          No           No            No
##   acarbose miglitol troglitazone tolazamide examide citoglipton insulin
## 1       No       No           No         No      No          No      No
## 2       No       No           No         No      No          No      Up
## 3       No       No           No         No      No          No      No
## 4       No       No           No         No      No          No      Up
## 5       No       No           No         No      No          No  Steady
## 6       No       No           No         No      No          No  Steady
##   glyburide.metformin glipizide.metformin glimepiride.pioglitazone
## 1                  No                  No                       No
## 2                  No                  No                       No
## 3                  No                  No                       No
## 4                  No                  No                       No
## 5                  No                  No                       No
## 6                  No                  No                       No
##   metformin.rosiglitazone metformin.pioglitazone change diabetesMed
## 1                      No                     No     No          No
## 2                      No                     No     Ch         Yes
## 3                      No                     No     No         Yes
## 4                      No                     No     Ch         Yes
## 5                      No                     No     Ch         Yes
## 6                      No                     No     No         Yes
##   readmitted
```

```
## 1         NO
## 2        >30
## 3         NO
## 4         NO
## 5         NO
## 6        >30
```

```
summary(raw.data)
```

```
##   encounter_id        patient_nbr                    race
##  Min.   :    12522   Min.   :      135   AfricanAmerican:19210
##  1st Qu.: 84961194   1st Qu.: 23413221   Asian          :  641
##  Median :152388987   Median : 45505143   Caucasian      :76099
##  Mean   :165201646   Mean   : 54330401   Hispanic       : 2037
##  3rd Qu.:230270888   3rd Qu.: 87545950   Other          : 1506
##  Max.   :443867222   Max.   :189502619   NA's           : 2273
##
##              gender           age              weight
##  Female         :54708   [70-80):26068   [75-100) : 1336
##  Male           :47055   [60-70):22483   [50-75)  :  897
##  Unknown/Invalid:    3   [50-60):17256   [100-125):  625
##                           [80-90):17197   [125-150):  145
##                           [40-50): 9685   [25-50)  :   97
##                           [30-40): 3775   (Other)  :   97
##                           (Other): 5302   NA's     :98569
##  admission_type_id discharge_disposition_id admission_source_id
##  Min.   :1.000     Min.   : 1.000           Min.   : 1.000
##  1st Qu.:1.000     1st Qu.: 1.000           1st Qu.: 1.000
##  Median :1.000     Median : 1.000           Median : 7.000
##  Mean   :2.024     Mean   : 3.716           Mean   : 5.754
##  3rd Qu.:3.000     3rd Qu.: 4.000           3rd Qu.: 7.000
##  Max.   :8.000     Max.   :28.000           Max.   :25.000
##
##  time_in_hospital   payer_code                medical_specialty
##  Min.   : 1.000   MC     :32439   InternalMedicine      :14635
##  1st Qu.: 2.000   HM     : 6274   Emergency/Trauma      : 7565
##  Median : 4.000   SP     : 5007   Family/GeneralPractice: 7440
##  Mean   : 4.396   BC     : 4655   Cardiology            : 5352
##  3rd Qu.: 6.000   MD     : 3532   Surgery-General       : 3099
##  Max.   :14.000   (Other): 9603   (Other)               :13726
##                   NA's   :40256   NA's                  :49949
##  num_lab_procedures num_procedures num_medications number_outpatient
##  Min.   :  1.0      Min.   :0.00   Min.   : 1.00   Min.   : 0.0000
##  1st Qu.: 31.0      1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000
##  Median : 44.0      Median :1.00   Median :15.00   Median : 0.0000
##  Mean   : 43.1      Mean   :1.34   Mean   :16.02   Mean   : 0.3694
##  3rd Qu.: 57.0      3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000
##  Max.   :132.0      Max.   :6.00   Max.   :81.00   Max.   :42.0000
##
##  number_emergency  number_inpatient      diag_1          diag_2
##  Min.   : 0.0000   Min.   : 0.0000   428    : 6862   276    : 6752
##  1st Qu.: 0.0000   1st Qu.: 0.0000   414    : 6581   428    : 6662
##  Median : 0.0000   Median : 0.0000   786    : 4016   250    : 6071
##  Mean   : 0.1978   Mean   : 0.6356   410    : 3614   427    : 5036
```

```
##  3rd Qu.: 0.0000   3rd Qu.: 1.0000   486    : 3508   401    : 3736
##  Max.   :76.0000   Max.   :21.0000   (Other):77164   (Other):73151
##                                      NA's   :   21   NA's   :  358
##      diag_3      number_diagnoses max_glu_serum A1Cresult
##  250    :11555   Min.   : 1.000   >200: 1485    >7  : 3812
##  401    : 8289   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  276    : 5175   Median : 8.000   None:96420    None:84748
##  428    : 4577   Mean   : 7.423   Norm: 2597    Norm: 4990
##  427    : 3955   3rd Qu.: 9.000
##  (Other):66792   Max.   :16.000
##  NA's   : 1423
##   metformin     repaglinide     nateglinide     chlorpropamide
##  Down  :  575   Down  :    45   Down  :    11   Down  :     1
##  No    :81778   No    :100227   No    :101063   No    :101680
##  Steady:18346   Steady:  1384   Steady:   668   Steady:    79
##  Up    : 1067   Up    :   110   Up    :    24   Up    :     6
##
##
##
##  glimepiride    acetohexamide    glipizide      glyburide
##  Down  :  194   No    :101765   Down  :  560   Down  :  564
##  No    :96575   Steady:     1   No    :89080   No    :91116
##  Steady: 4670                   Steady:11356   Steady: 9274
##  Up    :  327                   Up    :  770   Up    :  812
##
##
##
##  tolbutamide     pioglitazone   rosiglitazone    acarbose
##  No    :101743   Down  :  118   Down  :   87   Down  :     3
##  Steady:    23   No    :94438   No    :95401   No    :101458
##                  Steady: 6976   Steady: 6100   Steady:   295
##                  Up    :  234   Up    :  178   Up    :    10
##
##
##
##    miglitol      troglitazone     tolazamide     examide     citoglipton
##  Down  :     5   No    :101763   No    :101727   No:101766   No:101766
##  No    :101728   Steady:     3   Steady:    38
##  Steady:    31                   Up    :     1
##  Up    :     2
##
##
##
##    insulin      glyburide.metformin glipizide.metformin
##  Down  :12218   Down  :     6       No    :101753
##  No    :47383   No    :101060       Steady:    13
##  Steady:30849   Steady:   692
##  Up    :11316   Up    :     8
##
##
##
##  glimepiride.pioglitazone metformin.rosiglitazone metformin.pioglitazone
##  No    :101765            No    :101764           No    :101765
##  Steady:     1            Steady:     2           Steady:     1
```

```
##
##
##
##
##
##  change     diabetesMed readmitted
##  Ch:47011   No :23403   <30:11357
##  No:54755   Yes:78363   >30:35545
##                         NO :54864
##
##
##
##
```

For getting the readmission prediction we dont need many columns hence we should be selecting only relevant values to make it simpler.Also we need to remove those columns which mainly has NA values.

```
data <- select(raw.data,  -encounter_id, -patient_nbr, -weight,-(25:41),-(43:47))
head(data)
```

```
##             race gender     age admission_type_id
## 1       Caucasian Female  [0-10)                 6
## 2       Caucasian Female [10-20)                 1
## 3 AfricanAmerican Female [20-30)                 1
## 4       Caucasian   Male [30-40)                 1
## 5       Caucasian   Male [40-50)                 1
## 6       Caucasian   Male [50-60)                 2
##   discharge_disposition_id admission_source_id time_in_hospital payer_code
## 1                       25                   1                1       <NA>
## 2                        1                   7                3       <NA>
## 3                        1                   7                2       <NA>
## 4                        1                   7                2       <NA>
## 5                        1                   7                1       <NA>
## 6                        1                   2                3       <NA>
##         medical_specialty num_lab_procedures num_procedures
## 1 Pediatrics-Endocrinology                41              0
## 2                    <NA>                59              0
## 3                    <NA>                11              5
## 4                    <NA>                44              1
## 5                    <NA>                51              0
## 6                    <NA>                31              6
##   num_medications number_outpatient number_emergency number_inpatient
## 1               1                 0                0                0
## 2              18                 0                0                0
## 3              13                 2                0                1
## 4              16                 0                0                0
## 5               8                 0                0                0
## 6              16                 0                0                0
##   diag_1 diag_2 diag_3 number_diagnoses max_glu_serum A1Cresult insulin
## 1 250.83   <NA>   <NA>                1          None      None      No
## 2    276 250.01    255                9          None      None      Up
## 3    648    250    V27                6          None      None      No
## 4      8 250.43    403                7          None      None      Up
```

```
## 5    197    157    250                5          None      None  Steady
## 6    414    411    250                9          None      None  Steady
##   change diabetesMed readmitted
## 1     No          No         NO
## 2     Ch         Yes        >30
## 3     No         Yes         NO
## 4     Ch         Yes         NO
## 5     Ch         Yes         NO
## 6     No         Yes        >30
```

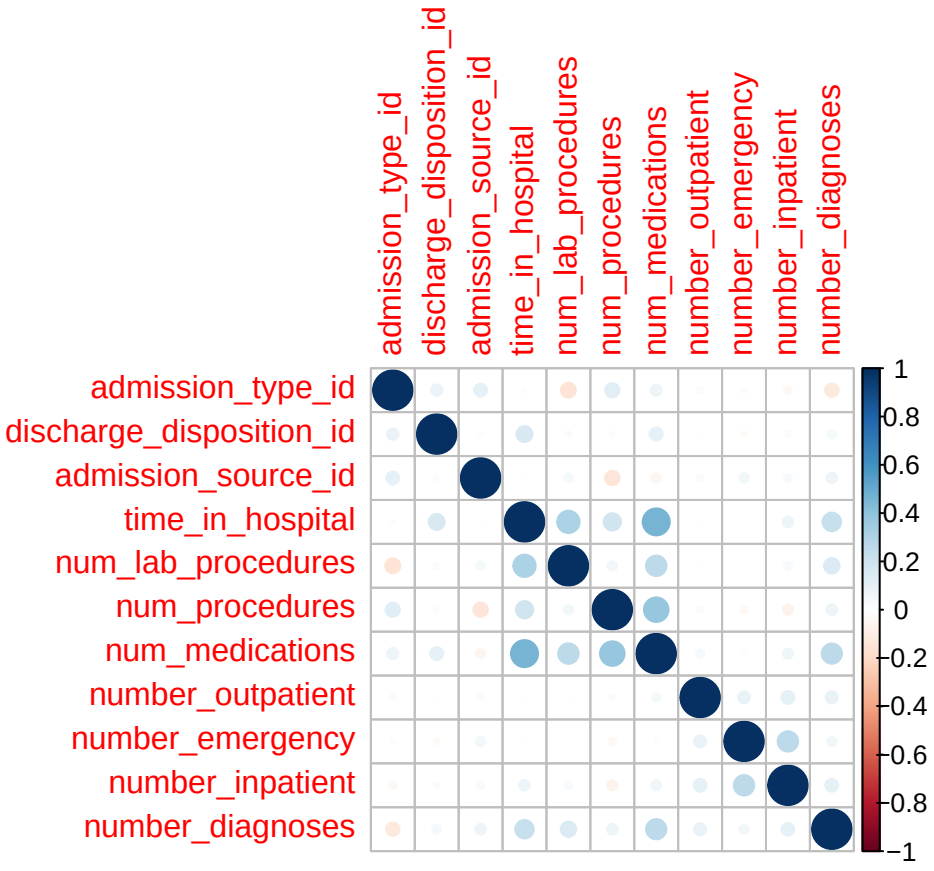
Basic summary of selected data and correlation plot of numeric data

```
summary(data)
```

```
##               race                   gender           age
##  AfricanAmerican:19210   Female         :54708   [70-80):26068
##  Asian          :  641   Male           :47055   [60-70):22483
##  Caucasian      :76099   Unknown/Invalid:    3   [50-60):17256
##  Hispanic       : 2037                           [80-90):17197
##  Other          : 1506                           [40-50): 9685
##  NA's           : 2273                           [30-40): 3775
##                                                  (Other): 5302
##  admission_type_id discharge_disposition_id admission_source_id
##  Min.   :1.000     Min.   : 1.000           Min.   : 1.000
##  1st Qu.:1.000     1st Qu.: 1.000           1st Qu.: 1.000
##  Median :1.000     Median : 1.000           Median : 7.000
##  Mean   :2.024     Mean   : 3.716           Mean   : 5.754
##  3rd Qu.:3.000     3rd Qu.: 4.000           3rd Qu.: 7.000
##  Max.   :8.000     Max.   :28.000           Max.   :25.000
##
##  time_in_hospital   payer_code                medical_specialty
##  Min.   : 1.000   MC     :32439   InternalMedicine      :14635
##  1st Qu.: 2.000   HM     : 6274   Emergency/Trauma      : 7565
##  Median : 4.000   SP     : 5007   Family/GeneralPractice: 7440
##  Mean   : 4.396   BC     : 4655   Cardiology            : 5352
##  3rd Qu.: 6.000   MD     : 3532   Surgery-General       : 3099
##  Max.   :14.000   (Other): 9603   (Other)               :13726
##                   NA's   :40256   NA's                  :49949
##  num_lab_procedures num_procedures num_medications number_outpatient
##  Min.   :  1.0      Min.   :0.00   Min.   : 1.00   Min.   : 0.0000
##  1st Qu.: 31.0      1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000
##  Median : 44.0      Median :1.00   Median :15.00   Median : 0.0000
##  Mean   : 43.1      Mean   :1.34   Mean   :16.02   Mean   : 0.3694
##  3rd Qu.: 57.0      3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000
##  Max.   :132.0      Max.   :6.00   Max.   :81.00   Max.   :42.0000
##
##  number_emergency  number_inpatient      diag_1          diag_2
##  Min.   : 0.0000   Min.   : 0.0000   428    : 6862   276    : 6752
##  1st Qu.: 0.0000   1st Qu.: 0.0000   414    : 6581   428    : 6662
##  Median : 0.0000   Median : 0.0000   786    : 4016   250    : 6071
##  Mean   : 0.1978   Mean   : 0.6356   410    : 3614   427    : 5036
##  3rd Qu.: 0.0000   3rd Qu.: 1.0000   486    : 3508   401    : 3736
##  Max.   :76.0000   Max.   :21.0000   (Other):77164   (Other):73151
```

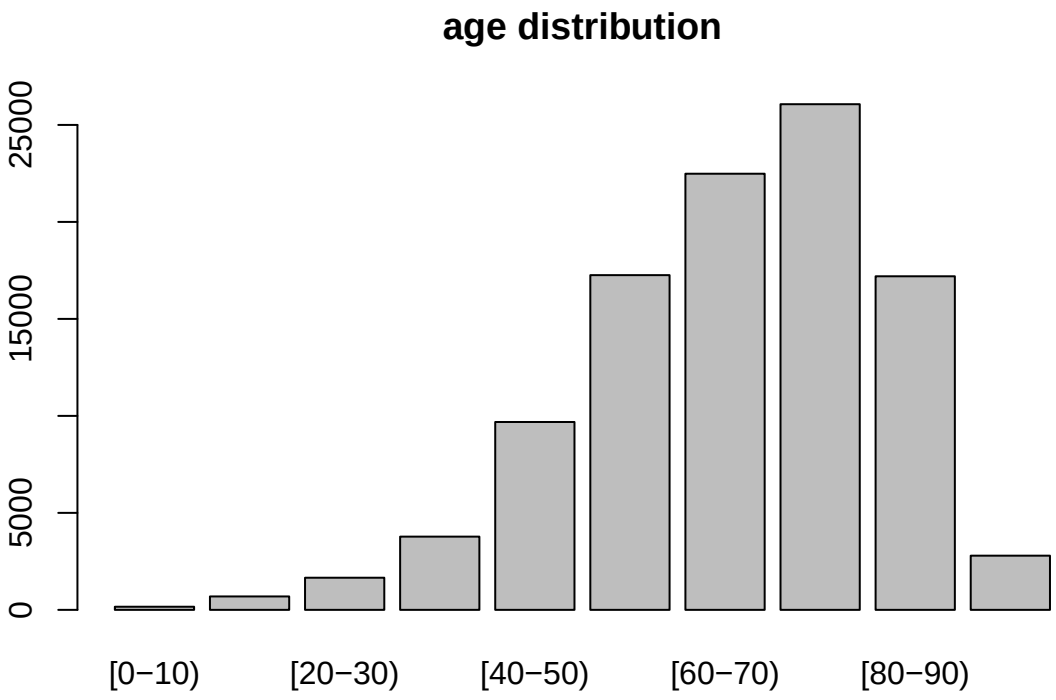
```
##                                        NA's   :   21   NA's   :  358
##      diag_3      number_diagnoses max_glu_serum A1Cresult
##  250    :11555   Min.   : 1.000   >200: 1485    >7  : 3812
##  401    : 8289   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  276    : 5175   Median : 8.000   None:96420    None:84748
##  428    : 4577   Mean   : 7.423   Norm: 2597    Norm: 4990
##  427    : 3955   3rd Qu.: 9.000
##  (Other):66792   Max.   :16.000
##  NA's   : 1423
##    insulin      change     diabetesMed readmitted
##  Down  :12218   Ch:47011   No :23403   <30:11357
##  No    :47383   No:54755   Yes:78363   >30:35545
##  Steady:30849                          NO :54864
##  Up    :11316
##
##
##
```

```
numeric_data <- select_if(data, is.numeric)
c <- cor(numeric_data, use= "pairwise.complete.obs")

corrplot(c)
```

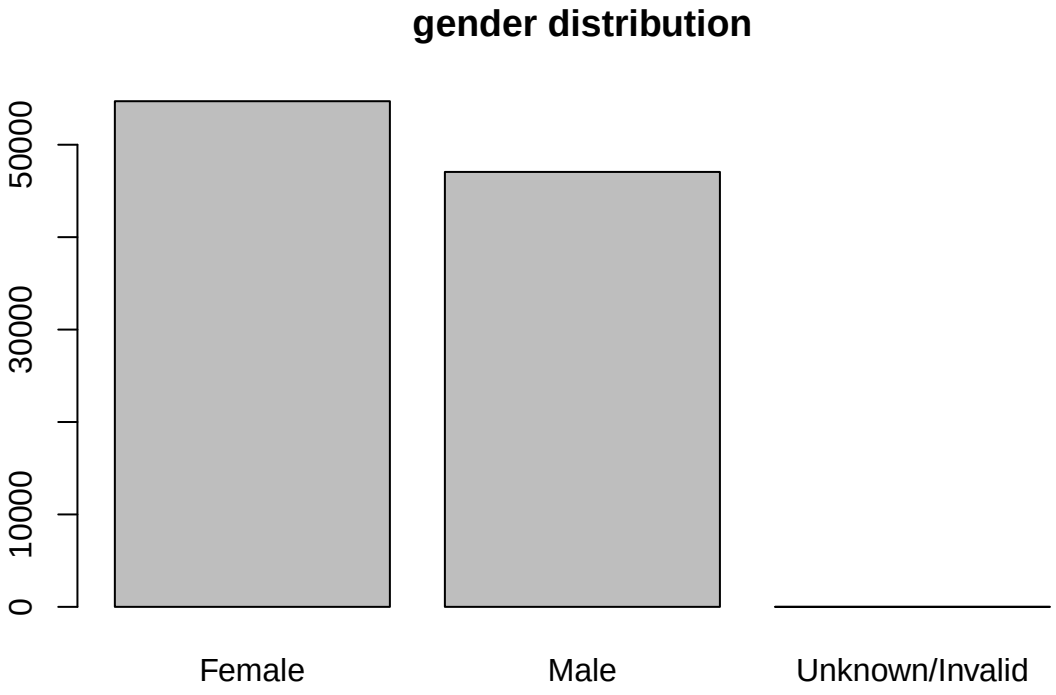



```
data$race[is.na(data$race)] <- "Other"
```

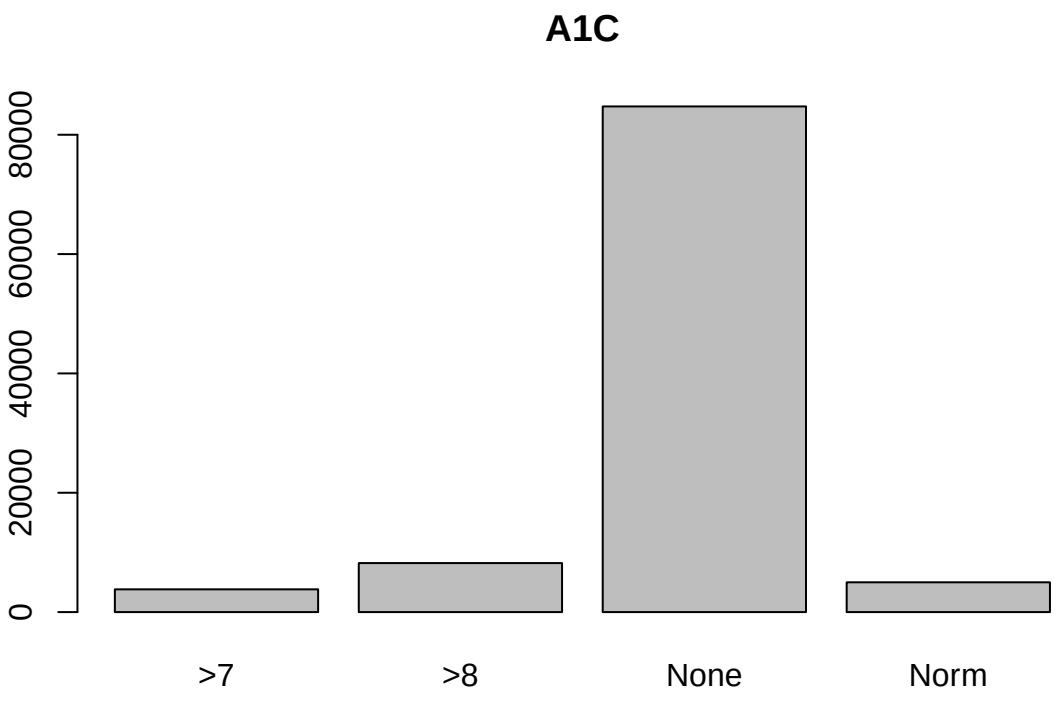
```r
plot(data$age, main = "age distribution") # age: mode 70-80yrs normal distribution, left skewed
```



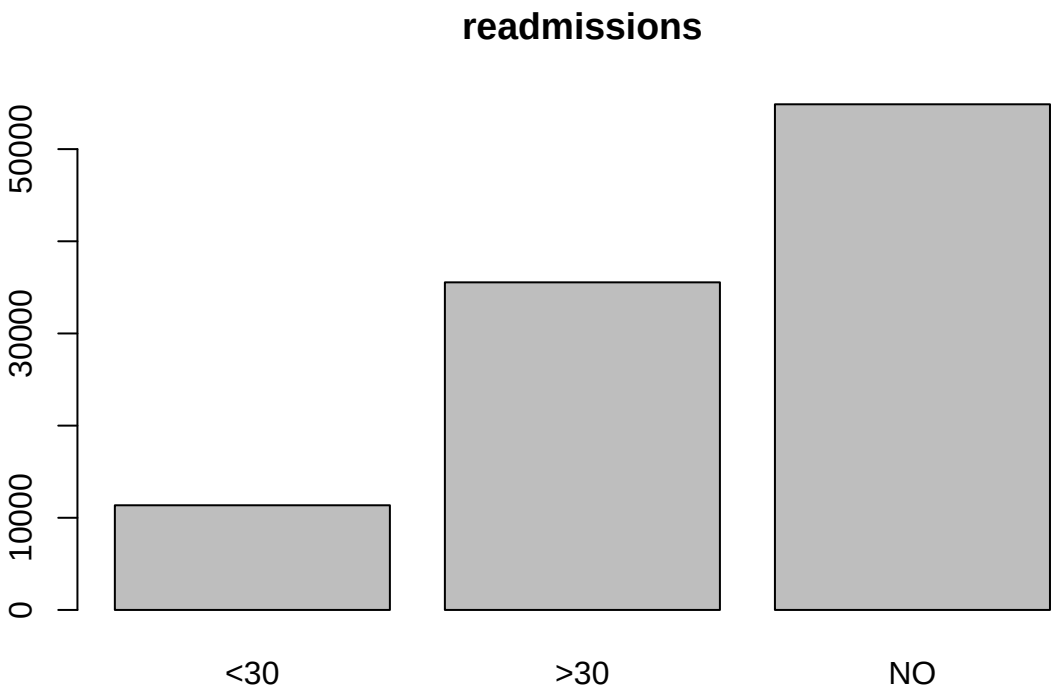

```r
plot(data$gender, main = "gender distribution") # gender: female 53% male 47%
```



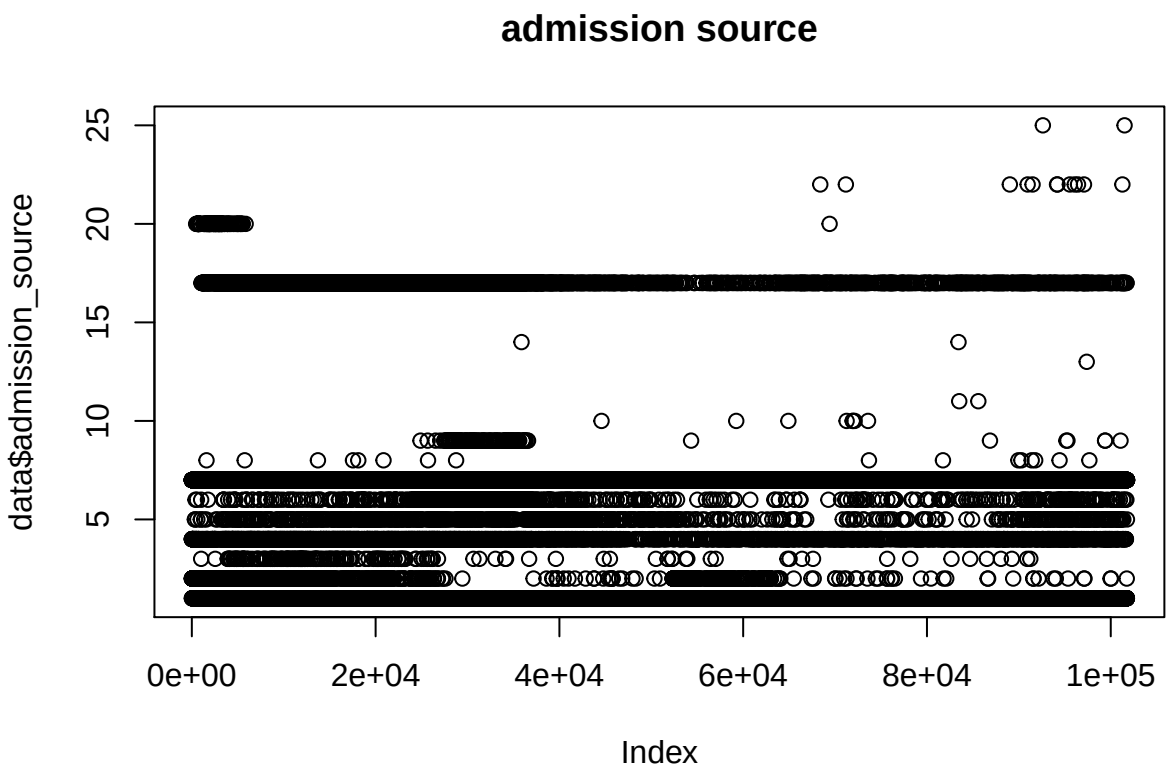

```r
plot(data$A1Cresult, main = "A1C") # A1Cresult: 84% no A1c results, 8% >8
```




```
plot(data$readmitted, main = "readmissions") # readmission: >50% no readmission
```




```
plot(data$admission_source, main = "admission source") # emergency 60%
```



Changing Categorical data

Changing categorical vriables

```
data2 <- data
data2$diag_1 <- as.numeric(levels(data2$diag_1)[data2$diag_1])
```

```
## Warning: NAs introduced by coercion
```

```
data2$diag_2 <- as.numeric(levels(data2$diag_2)[data2$diag_2])
```

```
## Warning: NAs introduced by coercion
```

```
data2$diag_3 <- as.numeric(levels(data2$diag_3)[data2$diag_3])
```

```
## Warning: NAs introduced by coercion
```

diagnosis1

```
data2$diagnosis_group <- factor( rep("other",nrow(data2)),ordered = F,
                                 levels =     c("circulatory","respiratory","Digestive","Diabetes","Inj
                                   "Musculoskeletal","Genitourinary","Neoplasms","other"))
```

```
data2$diagnosis_group[data2$diag_1 >= 390 &
                        data2$diag_1 <= 459 | data2$diag_1 == 785] <- "circulatory"

data2$diagnosis_group[data2$diag_1 >= 460 &
                        data2$diag_1 <= 519 | data2$diag_1 == 786] <- "respiratory"

data2$diagnosis_group[data2$diag_1 >= 520 &
                        data2$diag_1 <= 579 | data2$diag_1==787] <- "Digestive"

data2$diagnosis_group[data2$diag_1>=250 & data2$diag_1 < 251] <- "Diabetes"

data2$diagnosis_group[data2$diag_1>800 & data2$diag_1 <= 999] <- "Injury"

data2$diagnosis_group[data2$diag_1>=710 & data2$diag_1 <= 739] <- "Musculoskeletal"

data2$diagnosis_group[data2$diag_1>=580 &
                        data2$diag_1 <= 629 | data2$diag_1==788] <- "Genitourinary"

data2$diagnosis_group[data2$diag_1>=140 & data2$diag_1 <= 239 | data2$diag_1>=790 &
                        data2$diag_1 <= 799 | data2$diag_1==780 | data2$diag_1>=240 & data2$diag_1 < 25
                        data2$diag_1>=251 & data2$diag_1 <= 279 | data2$diag_1>=680 & data2$diag_1 <= 7
                        data2$diag_1>=001 & data2$diag_1 <= 139 | data2$diag_1==781 |
                      data2$diag_1==782 | data2$diag_1==784] <- "Neoplasms"
```

diagnosis_2

```
data2$diagnosis_2 <- factor( rep("other",nrow(data2)),ordered = F,
                                levels = c("circulatory","respiratory","Digestive","Diabetes","Injury"
                                           "Musculoskeletal","Genitourinary","Neoplasms","other"))
data2$diagnosis_2[data2$diag_2>=390 & data2$diag_2 <= 459 | data2$diag_2==785] <- "circulatory"
data2$diagnosis_2[data2$diag_2>=460 & data2$diag_2 <= 519 | data2$diag_2==786] <- "respiratory"
data2$diagnosis_2[data2$diag_2>=520 & data2$diag_2 <= 579 | data2$diag_2==787] <- "Digestive"
data2$diagnosis_2[data2$diag_2>=250 & data2$diag_2 < 251] <- "Diabetes"
data2$diagnosis_2[data2$diag_2>800 & data2$diag_2 <= 999] <- "Injury"
data2$diagnosis_2[data2$diag_2>=710 & data2$diag_2 <= 739] <- "Musculoskeletal"
data2$diagnosis_2[data2$diag_2>=580 & data2$diag_2 <= 629 | data2$diag_2==788] <- "Genitourinary"
data2$diagnosis_2[data2$diag_2>=140 & data2$diag_2 <= 239 | data2$diag_2>=790 &
                        data2$diag_2 <= 799 | data2$diag_2==780 | data2$diag_2>=240 & data2$diag_2 < 25
                        data2$diag_2>=251 & data2$diag_2 <= 279 | data2$diag_2>=680 & data2$diag_2 <= 7
                        data2$diag_2>=001 & data2$diag_2 <= 139 | data2$diag_2==781 |
                        data2$diag_2==782 | data2$diag_2==784] <- "Neoplasms"
```

diagnosis_3

```
data2$diagnosis_3 <- factor( rep("other",nrow(data2)),ordered = F,
                                levels = c("circulatory","respiratory","Digestive","Diabetes","Injury"
                                           "Musculoskeletal","Genitourinary","Neoplasms","other"))
data2$diagnosis_3[data2$diag_3>=390 & data2$diag_3 <= 459 | data2$diag_3==785] <- "circulatory"
```

```
data2$diagnosis_3[data2$diag_3>=460 & data2$diag_3 <= 519 | data2$diag_3==786] <- "respiratory"
data2$diagnosis_3[data2$diag_3>=520 & data2$diag_3 <= 579 | data2$diag_3==787] <- "Digestive"
data2$diagnosis_3[data2$diag_3>=250 & data2$diag_3 < 251] <- "Diabetes"
data2$diagnosis_3[data2$diag_3>800 & data2$diag_3 <= 999] <- "Injury"
data2$diagnosis_3[data2$diag_3>=710 & data2$diag_3 <= 739] <- "Musculoskeletal"
data2$diagnosis_3[data2$diag_3>=580 & data2$diag_3 <= 629 | data2$diag_3==788] <- "Genitourinary"
data2$diagnosis_3[data2$diag_3>=140 & data2$diag_3 <= 239 | data2$diag_3>=790 &
                        data2$diag_3 <= 799 | data2$diag_3==780 | data2$diag_3>=240 & data2$diag_3 < 25
                        data2$diag_3>=251 & data2$diag_3 <= 279 | data2$diag_3>=680 & data2$diag_3 <= 7
                        data2$diag_3>=001 & data2$diag_3 <= 139 | data2$diag_3==781 |
                        data2$diag_3==782 | data2$diag_3==784] <- "Neoplasms"
```

```
summary(data2)
```

```
##               race                   gender           age
##  AfricanAmerican:19210   Female         :54708   [70-80):26068
##  Asian          :  641   Male           :47055   [60-70):22483
##  Caucasian      :76099   Unknown/Invalid:    3   [50-60):17256
##  Hispanic       : 2037                           [80-90):17197
##  Other          : 3779                           [40-50): 9685
##                                                  [30-40): 3775
##                                                  (Other): 5302
##  admission_type_id discharge_disposition_id admission_source_id
##  Min.   :1.000     Min.   : 1.000           Min.   : 1.000
##  1st Qu.:1.000     1st Qu.: 1.000           1st Qu.: 1.000
##  Median :1.000     Median : 1.000           Median : 7.000
##  Mean   :2.024     Mean   : 3.716           Mean   : 5.754
##  3rd Qu.:3.000     3rd Qu.: 4.000           3rd Qu.: 7.000
##  Max.   :8.000     Max.   :28.000           Max.   :25.000
##
##  time_in_hospital   payer_code                medical_specialty
##  Min.   : 1.000   MC     :32439   InternalMedicine      :14635
##  1st Qu.: 2.000   HM     : 6274   Emergency/Trauma      : 7565
##  Median : 4.000   SP     : 5007   Family/GeneralPractice: 7440
##  Mean   : 4.396   BC     : 4655   Cardiology            : 5352
##  3rd Qu.: 6.000   MD     : 3532   Surgery-General       : 3099
##  Max.   :14.000   (Other): 9603   (Other)               :13726
##                   NA's   :40256   NA's                  :49949
##  num_lab_procedures num_procedures num_medications number_outpatient
##  Min.   :  1.0      Min.   :0.00   Min.   : 1.00   Min.   : 0.0000
##  1st Qu.: 31.0      1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000
##  Median : 44.0      Median :1.00   Median :15.00   Median : 0.0000
##  Mean   : 43.1      Mean   :1.34   Mean   :16.02   Mean   : 0.3694
##  3rd Qu.: 57.0      3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000
##  Max.   :132.0      Max.   :6.00   Max.   :81.00   Max.   :42.0000
##
##  number_emergency  number_inpatient      diag_1          diag_2
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :  3.0   Min.   :  5.0
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:410.0   1st Qu.:276.0
##  Median : 0.0000   Median : 0.0000   Median :440.0   Median :425.0
##  Mean   : 0.1978   Mean   : 0.6356   Mean   :493.6   Mean   :438.7
##  3rd Qu.: 0.0000   3rd Qu.: 1.0000   3rd Qu.:599.0   3rd Qu.:530.0
##  Max.   :76.0000   Max.   :21.0000   Max.   :999.0   Max.   :999.0
```

```
##                                      NA's   :1666    NA's   :2894
##      diag_3      number_diagnoses max_glu_serum A1Cresult
##  Min.   :  3.0   Min.   : 1.000   >200: 1485    >7  : 3812
##  1st Qu.:272.0   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  Median :403.0   Median : 8.000   None:96420    None:84748
##  Mean   :418.2   Mean   : 7.423   Norm: 2597    Norm: 4990
##  3rd Qu.:496.0   3rd Qu.: 9.000
##  Max.   :999.0   Max.   :16.000
##  NA's   :6481
##    insulin      change     diabetesMed readmitted     diagnosis_group
##  Down  :12218   Ch:47011   No :23403   <30:11357   circulatory:30437
##  No    :47383   No:54755   Yes:78363   >30:35545   respiratory:14423
##  Steady:30849                           NO :54864   Neoplasms  :14056
##  Up    :11316                                       Digestive  : 9475
##                                                     Diabetes   : 8757
##                                                     other      : 7576
##                                                     (Other)    :17042
##         diagnosis_2           diagnosis_3
##  circulatory  :31881   circulatory  :30306
##  Neoplasms    :18805   Neoplasms    :17849
##  Diabetes     :12794   Diabetes     :17157
##  respiratory  :10895   other        :14627
##  other        :10654   respiratory  : 7358
##  Genitourinary: 8376   Genitourinary: 6680
##  (Other)      : 8361   (Other)      : 7789
```

admission_source

```
data2$admission_source <- factor( rep("other", nrow(data2)), ordered = F,
                            levels = c("clinic_referral", "emergency","other"))

data2$admission_source[data2$admission_source_id == c(1,2,3)]<- "clinic_referral"

data2$admission_source[data2$admission_source_id == 7]<- "emergency"

head(data2$admission_source)
```

```
## [1] clinic_referral emergency       emergency       emergency
## [5] emergency       other
## Levels: clinic_referral emergency other
```

discharged_to

```
data2$discharged_to <- factor( rep("transferred", nrow(data2)), ordered = F,
                                levels = c("home", "transferred", "left_AMA"))

data2$discharged_to[data2$discharge_disposition_id==c(1,6,8)]<- "home"
```

```r
data2$discharged_to[data2$discharge_disposition_id==7]<- "left_AMA"

data2 <- select(data2, -diag_1, -diag_2, -diag_3, -admission_type_id, -discharge_disposition_id)

data2 <- select(data2, -medical_specialty)

data2 <- rename(data2, diag1 = diagnosis_group, diag2=diagnosis_2, diag3 = diagnosis_3)
summary(data2)
```

```
##               race                   gender           age
##  AfricanAmerican:19210   Female         :54708   [70-80):26068
##  Asian          :  641   Male           :47055   [60-70):22483
##  Caucasian      :76099   Unknown/Invalid:    3   [50-60):17256
##  Hispanic       : 2037                           [80-90):17197
##  Other          : 3779                           [40-50): 9685
##                                                  [30-40): 3775
##                                                  (Other): 5302
##  admission_source_id time_in_hospital   payer_code    num_lab_procedures
##  Min.   : 1.000      Min.   : 1.000   MC     :32439   Min.   :  1.0
##  1st Qu.: 1.000      1st Qu.: 2.000   HM     : 6274   1st Qu.: 31.0
##  Median : 7.000      Median : 4.000   SP     : 5007   Median : 44.0
##  Mean   : 5.754      Mean   : 4.396   BC     : 4655   Mean   : 43.1
##  3rd Qu.: 7.000      3rd Qu.: 6.000   MD     : 3532   3rd Qu.: 57.0
##  Max.   :25.000      Max.   :14.000   (Other): 9603   Max.   :132.0
##                                       NA's   :40256
##  num_procedures num_medications number_outpatient number_emergency
##  Min.   :0.00   Min.   : 1.00   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :1.00   Median :15.00   Median : 0.0000   Median : 0.0000
##  Mean   :1.34   Mean   :16.02   Mean   : 0.3694   Mean   : 0.1978
##  3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :6.00   Max.   :81.00   Max.   :42.0000   Max.   :76.0000
##
##  number_inpatient  number_diagnoses max_glu_serum A1Cresult
##  Min.   : 0.0000   Min.   : 1.000   >200: 1485    >7  : 3812
##  1st Qu.: 0.0000   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  Median : 0.0000   Median : 8.000   None:96420    None:84748
##  Mean   : 0.6356   Mean   : 7.423   Norm: 2597    Norm: 4990
##  3rd Qu.: 1.0000   3rd Qu.: 9.000
##  Max.   :21.0000   Max.   :16.000
##
##    insulin      change     diabetesMed readmitted          diag1
##  Down  :12218   Ch:47011   No :23403   <30:11357   circulatory:30437
##  No    :47383   No:54755   Yes:78363   >30:35545   respiratory:14423
##  Steady:30849                           NO :54864   Neoplasms  :14056
##  Up    :11316                                       Digestive  : 9475
##                                                     Diabetes   : 8757
##                                                     other      : 7576
##                                                     (Other)    :17042
##           diag2                diag3              admission_source
##  circulatory  :31881   circulatory  :30306   clinic_referral:10233
##  Neoplasms    :18805   Neoplasms    :17849   emergency      :57494
##  Diabetes     :12794   Diabetes     :17157   other          :34039
```

```
##  respiratory  :10895   other        :14627
##  other        :10654   respiratory  : 7358
##  Genitourinary: 8376   Genitourinary: 6680
##  (Other)      : 8361   (Other)      : 7789
##      discharged_to
##  home       :24520
##  transferred:76623
##  left_AMA   :  623
##
##
##
##
```

payer_code

```
data2$payer_code2 <- factor( rep("other", nrow(data2)),
                             ordered = F, levels = c("other", "self_pay"))

data2$payer_code2[data2$payer_code=="SP"]<- "self_pay"

data2 <- select(data2, -payer_code)
data2 <- select(data2, -admission_source_id)
data2 <- rename(data2, payer_code=payer_code2)
summary(data2)
```

```
##               race                    gender           age
##  AfricanAmerican:19210   Female         :54708   [70-80):26068
##  Asian          :  641   Male           :47055   [60-70):22483
##  Caucasian      :76099   Unknown/Invalid:    3   [50-60):17256
##  Hispanic       : 2037                            [80-90):17197
##  Other          : 3779                            [40-50): 9685
##                                                   [30-40): 3775
##                                                   (Other): 5302
##  time_in_hospital num_lab_procedures num_procedures num_medications
##  Min.   : 1.000   Min.   :  1.0      Min.   :0.00   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 31.0      1st Qu.:0.00   1st Qu.:10.00
##  Median : 4.000   Median : 44.0      Median :1.00   Median :15.00
##  Mean   : 4.396   Mean   : 43.1      Mean   :1.34   Mean   :16.02
##  3rd Qu.: 6.000   3rd Qu.: 57.0      3rd Qu.:2.00   3rd Qu.:20.00
##  Max.   :14.000   Max.   :132.0      Max.   :6.00   Max.   :81.00
##
##  number_outpatient number_emergency  number_inpatient  number_diagnoses
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 1.000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 6.000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 8.000
##  Mean   : 0.3694   Mean   : 0.1978   Mean   : 0.6356   Mean   : 7.423
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000   3rd Qu.: 9.000
##  Max.   :42.0000   Max.   :76.0000   Max.   :21.0000   Max.   :16.000
##
##  max_glu_serum A1Cresult      insulin      change     diabetesMed
##  >200: 1485    >7  : 3812   Down  :12218   Ch:47011   No :23403
```

```
##  >300: 1264    >8  : 8216   No    :47383   No:54755   Yes:78363
##  None:96420    None:84748   Steady:30849
##  Norm: 2597    Norm: 4990   Up    :11316
##
##
##
##  readmitted          diag1                diag2
##  <30:11357   circulatory:30437   circulatory  :31881
##  >30:35545   respiratory:14423   Neoplasms    :18805
##  NO :54864   Neoplasms  :14056   Diabetes     :12794
##              Digestive  : 9475   respiratory  :10895
##              Diabetes   : 8757   other        :10654
##              other      : 7576   Genitourinary: 8376
##              (Other)    :17042   (Other)      : 8361
##           diag3              admission_source     discharged_to
##  circulatory  :30306   clinic_referral:10233   home       :24520
##  Neoplasms    :17849   emergency      :57494   transferred:76623
##  Diabetes     :17157   other          :34039   left_AMA   :  623
##  other        :14627
##  respiratory  : 7358
##  Genitourinary: 6680
##  (Other)      : 7789
##     payer_code
##  other   :96759
##  self_pay: 5007
##
##
##
##
##
```

```
head(data2)
```

```
##              race gender     age time_in_hospital num_lab_procedures
## 1       Caucasian Female  [0-10)                1                 41
## 2       Caucasian Female [10-20)                3                 59
## 3 AfricanAmerican Female [20-30)                2                 11
## 4       Caucasian   Male [30-40)                2                 44
## 5       Caucasian   Male [40-50)                1                 51
## 6       Caucasian   Male [50-60)                3                 31
##   num_procedures num_medications number_outpatient number_emergency
## 1              0               1                 0                 0
## 2              0              18                 0                 0
## 3              5              13                 2                 0
## 4              1              16                 0                 0
## 5              0               8                 0                 0
## 6              6              16                 0                 0
##   number_inpatient number_diagnoses max_glu_serum A1Cresult insulin change
## 1                0                1          None      None      No     No
## 2                0                9          None      None      Up     Ch
## 3                1                6          None      None      No     No
## 4                0                7          None      None      Up     Ch
## 5                0                5          None      None  Steady     Ch
## 6                0                9          None      None  Steady     No
```

```
##   diabetesMed readmitted       diag1       diag2       diag3
## 1          No         NO    Diabetes       other       other
## 2         Yes        >30   Neoplasms    Diabetes   Neoplasms
## 3         Yes         NO       other    Diabetes       other
## 4         Yes         NO   Neoplasms    Diabetes circulatory
## 5         Yes         NO   Neoplasms   Neoplasms    Diabetes
## 6         Yes        >30 circulatory circulatory    Diabetes
##   admission_source discharged_to payer_code
## 1  clinic_referral   transferred      other
## 2        emergency   transferred      other
## 3        emergency   transferred      other
## 4        emergency          home      other
## 5        emergency   transferred      other
## 6            other   transferred      other
```

PCA Principal Component Analysis In order to reduce 23 dimensions or knowing the significant impact of our reduction we will do PCA.

PCA section

```
numeric_data <- scale(numeric_data)
pcaObj <- princomp(numeric_data, cor = TRUE, scores = TRUE, covmat = NULL)
summary(pcaObj)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.4359984 1.1952241 1.0978608 1.04782538 0.98879294
## Proportion of Variance 0.1874629 0.1298692 0.1095726 0.09981255 0.08888286
## Cumulative Proportion  0.1874629 0.3173320 0.4269046 0.52671712 0.61559999
##                           Comp.6    Comp.7     Comp.8     Comp.9
## Standard deviation     0.97134088 0.9110183 0.85581637 0.80392090
## Proportion of Variance 0.08577301 0.0754504 0.06658379 0.05875353
## Cumulative Proportion  0.70137300 0.7768234 0.84340718 0.90216071
##                          Comp.10    Comp.11
## Standard deviation     0.7810781 0.68275115
## Proportion of Variance 0.0554621 0.04237719
## Cumulative Proportion  0.9576228 1.00000000
```

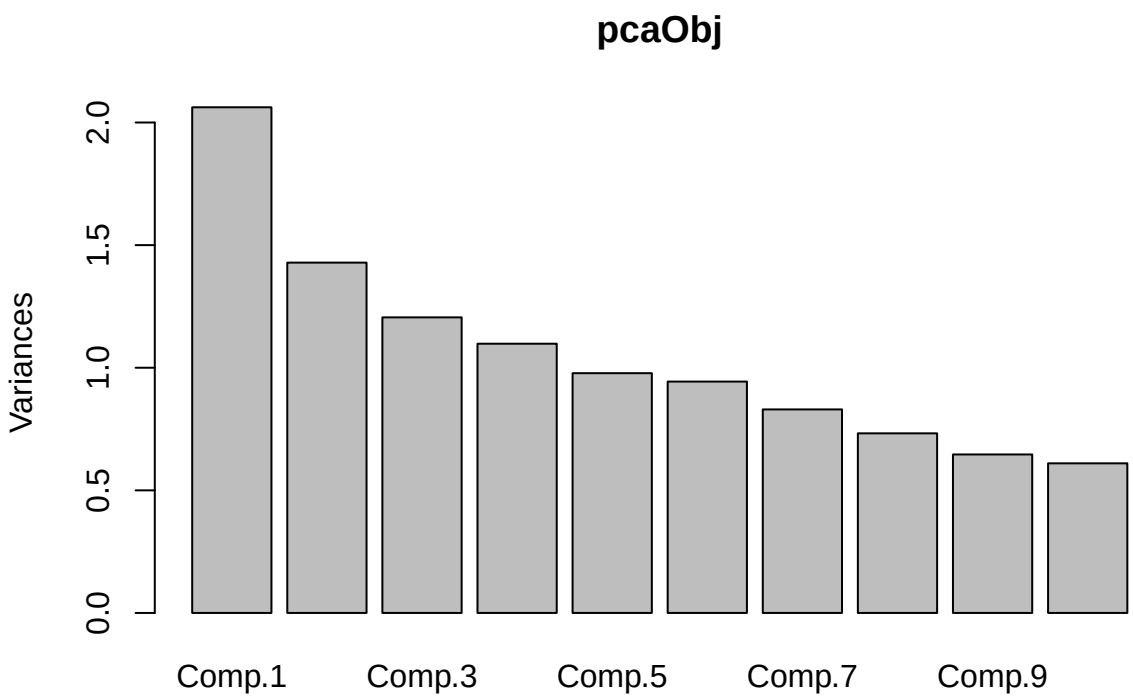
```
print(pcaObj)
```

```
## Call:
## princomp(x = numeric_data, cor = TRUE, scores = TRUE, covmat = NULL)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 1.4359984 1.1952241 1.0978608 1.0478254 0.9887929 0.9713409 0.9110183
##    Comp.8    Comp.9   Comp.10   Comp.11
## 0.8558164 0.8039209 0.7810781 0.6827512
##
##  11  variables and  101766 observations.
```

```r
names(pcaObj)
```

```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
## [7] "call"
```

```r
plot(pcaObj)
```

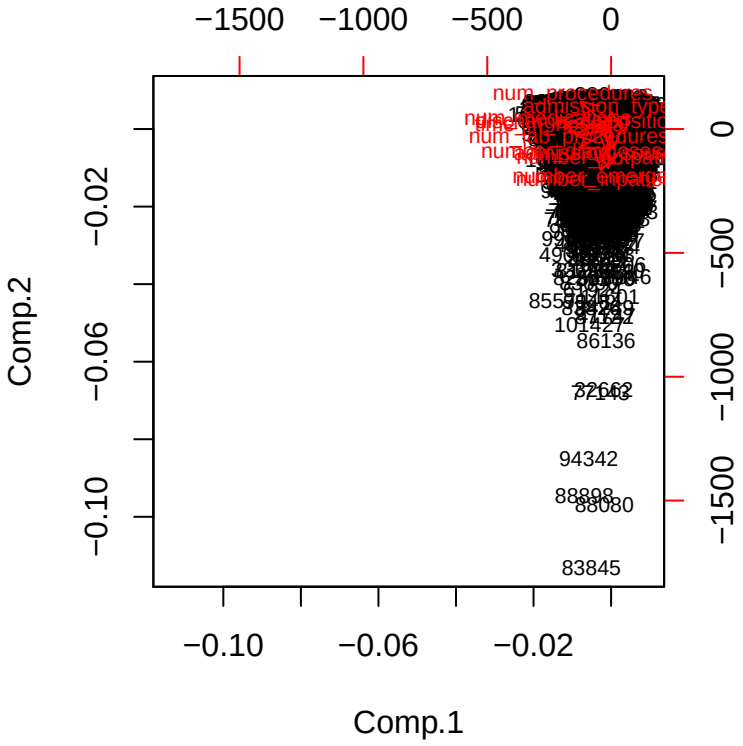



```r
pcaObj$loadings
```

```
##
## Loadings:
##                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## admission_type_id                0.229  0.723  0.127 -0.109 -0.167  0.121
## discharge_disposition_id -0.166         0.254  0.407  0.621  0.479
## admission_source_id              -0.260  0.215  0.667 -0.367 -0.307
## time_in_hospital         -0.522                0.137  0.118         0.162
## num_lab_procedures       -0.372        -0.375  0.222        -0.182  0.523
## num_procedures           -0.337  0.362  0.225 -0.380 -0.123 -0.168 -0.154
## num_medications          -0.558  0.112  0.117 -0.109
## number_outpatient                -0.314  0.253 -0.190 -0.466  0.641  0.380
## number_emergency                 -0.516  0.227 -0.266  0.212 -0.331
## number_inpatient         -0.113 -0.543  0.163 -0.193  0.317 -0.118
## number_diagnoses         -0.346 -0.253 -0.129        -0.259  0.219 -0.708
##                          Comp.8 Comp.9 Comp.10 Comp.11
## admission_type_id        -0.201  0.464 -0.296  -0.103
## discharge_disposition_id  0.238 -0.194 -0.156
## admission_source_id              -0.398  0.195
## time_in_hospital         -0.109  0.251  0.605  -0.472
## num_lab_procedures        0.117        -0.577
```

```
## num_procedures            0.168 -0.569 -0.149  -0.346
## num_medications                        0.156   0.786
## number_outpatient         0.104
## number_emergency          0.638  0.212
## number_inpatient         -0.649 -0.275 -0.116
## number_diagnoses                 0.269 -0.290  -0.135
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.091  0.091  0.091  0.091  0.091  0.091  0.091  0.091
## Cumulative Var  0.091  0.182  0.273  0.364  0.455  0.545  0.636  0.727
##                Comp.9 Comp.10 Comp.11
## SS loadings     1.000   1.000   1.000
## Proportion Var  0.091   0.091   0.091
## Cumulative Var  0.818   0.909   1.000
```

```
biplot(pcaObj, cex = 0.7)
```




```
final_data <- as.data.frame(pcaObj$scores)
```

According to PCA, The columns that are of utmost importance are as follows:

age discharged_to time_in_hospital num_lab_procedures num_procedures num_medications number_outpatient number_emergency number_inpatient number_diagnoses insulin change diabetesMed diag1 diag2 diag3 A1Cresult

Converting the dataset into training data and validation data.

```
set.seed(123)
inTrain <- createDataPartition(y = data2$readmitted, p = .67,list = FALSE)
train <- data2[ inTrain,]
test <- data2[-inTrain,]
nrow(train)
```

```
## [1] 68185
```

```
nrow(test)
```

```
## [1] 33581
```

```
summary(train)
```

```
##               race                   gender           age
##  AfricanAmerican:12921   Female         :36818   [70-80):17494
##  Asian          :  442   Male           :31366   [60-70):14994
##  Caucasian      :50925   Unknown/Invalid:    1   [50-60):11606
##  Hispanic       : 1370                           [80-90):11462
##  Other          : 2527                           [40-50): 6538
##                                                  [30-40): 2551
##                                                  (Other): 3540
##  time_in_hospital num_lab_procedures num_procedures  num_medications
##  Min.   : 1.000   Min.   :  1.00     Min.   :0.000   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 31.00     1st Qu.:0.000   1st Qu.:10.00
##  Median : 4.000   Median : 44.00     Median :1.000   Median :15.00
##  Mean   : 4.394   Mean   : 43.11     Mean   :1.335   Mean   :16.01
##  3rd Qu.: 6.000   3rd Qu.: 57.00     3rd Qu.:2.000   3rd Qu.:20.00
##  Max.   :14.000   Max.   :132.00     Max.   :6.000   Max.   :81.00
##
##  number_outpatient number_emergency  number_inpatient number_diagnoses
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 6.000
##  Median : 0.0000   Median : 0.0000   Median : 0.000   Median : 8.000
##  Mean   : 0.3674   Mean   : 0.1952   Mean   : 0.634   Mean   : 7.417
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.000   3rd Qu.: 9.000
##  Max.   :42.0000   Max.   :76.0000   Max.   :21.000   Max.   :16.000
##
##  max_glu_serum A1Cresult      insulin      change     diabetesMed
##  >200:  981    >7  : 2552   Down  : 8231   Ch:31572   No :15634
##  >300:  867    >8  : 5518   No    :31742   No:36613   Yes:52551
##  None:64591    None:56795   Steady:20612
##  Norm: 1746    Norm: 3320   Up    : 7600
##
##
##
##  readmitted          diag1                diag2
##  <30: 7610   circulatory:20231   circulatory  :21288
##  >30:23816   respiratory: 9701   Neoplasms    :12595
```

```
##  NO :36759   Neoplasms  : 9513   Diabetes     : 8530
##              Digestive  : 6394   respiratory  : 7271
##              Diabetes   : 5841   other        : 7184
##              other      : 5130   Genitourinary: 5700
##              (Other)    :11375   (Other)      : 5617
##           diag3              admission_source     discharged_to
##  circulatory  :20183   clinic_referral: 6796   home       :16459
##  Neoplasms    :12013   emergency      :38553   transferred:51313
##  Diabetes     :11486   other          :22836   left_AMA   :  413
##  other        : 9884
##  respiratory  : 4912
##  Genitourinary: 4496
##  (Other)      : 5211
##     payer_code
##  other   :64828
##  self_pay: 3357
##
##
##
##
##
```

```
summary(test)
```

```
##               race                    gender           age
##  AfricanAmerican: 6289   Female         :17890   [70-80):8574
##  Asian          :  199   Male           :15689   [60-70):7489
##  Caucasian      :25174   Unknown/Invalid:    2   [80-90):5735
##  Hispanic       :  667                            [50-60):5650
##  Other          : 1252                            [40-50):3147
##                                                   [30-40):1224
##                                                   (Other):1762
##  time_in_hospital num_lab_procedures num_procedures  num_medications
##  Min.   : 1.0     Min.   :  1.00     Min.   :0.000   Min.   : 1.00
##  1st Qu.: 2.0     1st Qu.: 31.00     1st Qu.:0.000   1st Qu.:10.00
##  Median : 4.0     Median : 44.00     Median :1.000   Median :15.00
##  Mean   : 4.4     Mean   : 43.06     Mean   :1.349   Mean   :16.04
##  3rd Qu.: 6.0     3rd Qu.: 57.00     3rd Qu.:2.000   3rd Qu.:20.00
##  Max.   :14.0     Max.   :126.00     Max.   :6.000   Max.   :79.00
##
##  number_outpatient number_emergency  number_inpatient  number_diagnoses
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 1.000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 6.000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 8.000
##  Mean   : 0.3733   Mean   : 0.2031   Mean   : 0.6388   Mean   : 7.434
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000   3rd Qu.: 9.000
##  Max.   :38.0000   Max.   :64.0000   Max.   :19.0000   Max.   :16.000
##
##  max_glu_serum A1Cresult      insulin      change     diabetesMed
##  >200:  504    >7  : 1260   Down  : 3987   Ch:15439   No : 7769
##  >300:  397    >8  : 2698   No    :15641   No:18142   Yes:25812
##  None:31829    None:27953   Steady:10237
##  Norm:  851    Norm: 1670   Up    : 3716
##
```

```
##
##
##  readmitted          diag1                diag2
##  <30: 3747   circulatory:10206   circulatory  :10593
##  >30:11729   respiratory: 4722   Neoplasms    : 6210
##  NO :18105   Neoplasms  : 4543   Diabetes     : 4264
##              Digestive  : 3081   respiratory  : 3624
##              Diabetes   : 2916   other        : 3470
##              other      : 2446   Genitourinary: 2676
##              (Other)    : 5667   (Other)      : 2744
##           diag3              admission_source     discharged_to
##  circulatory  :10123   clinic_referral: 3437   home       : 8061
##  Neoplasms    : 5836   emergency      :18941   transferred:25310
##  Diabetes     : 5671   other          :11203   left_AMA   :  210
##  other        : 4743
##  respiratory  : 2446
##  Genitourinary: 2184
##  (Other)      : 2578
##     payer_code
##  other   :31931
##  self_pay: 1650
##
##
##
##
##
```

Regression

Since the output variable readmitted is a categorical variable we need to find logistic regression instead of linear Regression

First by taking all the dimensions, in order to get the significance.

```
train_nonbinary <- train
test_nonbinary <- test

train$readmitted <- ifelse(train$readmitted == train$readmitted[1], 0 , 1)
test$readmitted <- ifelse(test$readmitted == test$readmitted[1], 0 , 1)
head(test)
```

```
##               race gender     age time_in_hospital num_lab_procedures
## 1        Caucasian Female  [0-10)                1                 41
## 3  AfricanAmerican Female [20-30)                2                 11
## 4        Caucasian   Male [30-40)                2                 44
## 7        Caucasian   Male [60-70)                4                 70
## 8        Caucasian   Male [70-80)                5                 73
## 13       Caucasian Female [40-50)                7                 60
##    num_procedures num_medications number_outpatient number_emergency
## 1               0               1                 0                0
## 3               5              13                 2                0
## 4               1              16                 0                0
## 7               1              21                 0                0
```

```
## 8               0              12                 0                0
## 13              0              15                 0                1
##    number_inpatient number_diagnoses max_glu_serum A1Cresult insulin
## 1                 0                1          None      None      No
## 3                 1                6          None      None      No
## 4                 0                7          None      None      Up
## 7                 0                7          None      None  Steady
## 8                 0                8          None      None      No
## 13                0                8          None      None    Down
##    change diabetesMed readmitted       diag1       diag2       diag3
## 1      No          No          0    Diabetes       other       other
## 3      No         Yes          0       other    Diabetes       other
## 4      Ch         Yes          0   Neoplasms    Diabetes circulatory
## 7      Ch         Yes          0 circulatory circulatory       other
## 8      No         Yes          1 circulatory respiratory    Diabetes
## 13     Ch         Yes          1 circulatory    Diabetes    Diabetes
##    admission_source discharged_to payer_code
## 1   clinic_referral   transferred      other
## 3         emergency   transferred      other
## 4         emergency          home      other
## 7             other          home      other
## 8         emergency   transferred      other
## 13        emergency   transferred      other
```

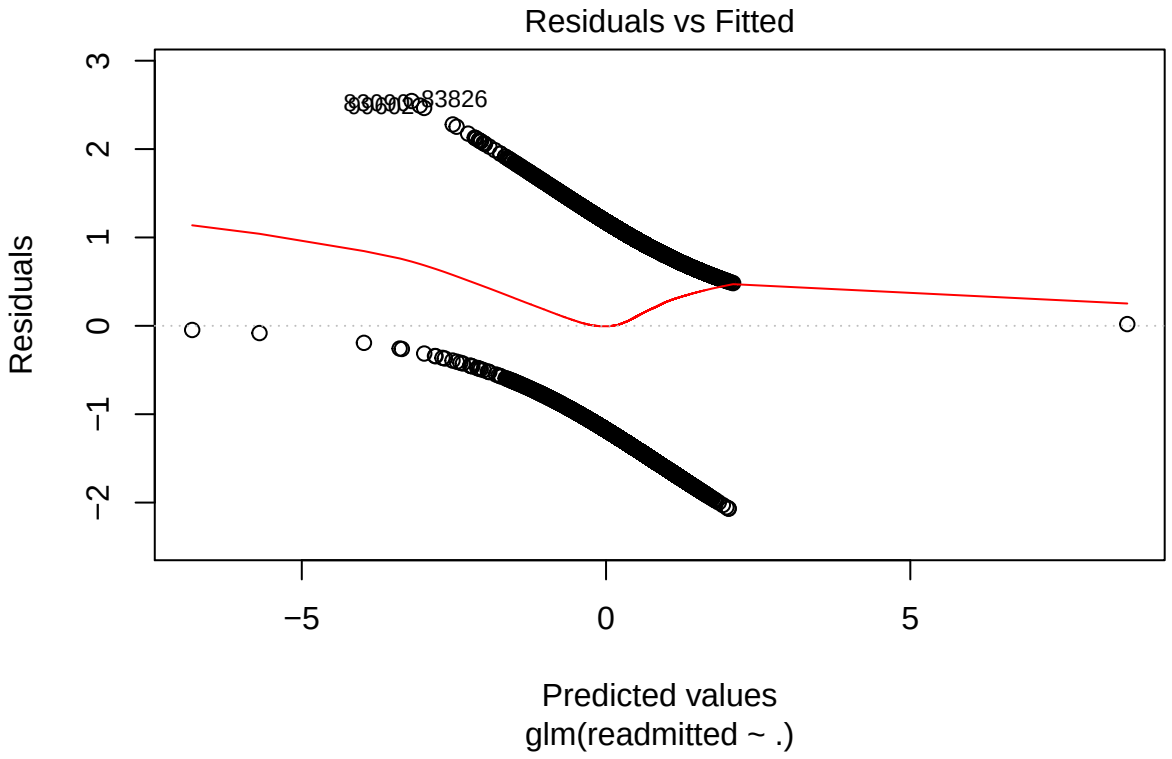
```
logit_model <- glm(readmitted ~., data = train, family = binomial(link = 'logit'))

summary(logit_model)
```

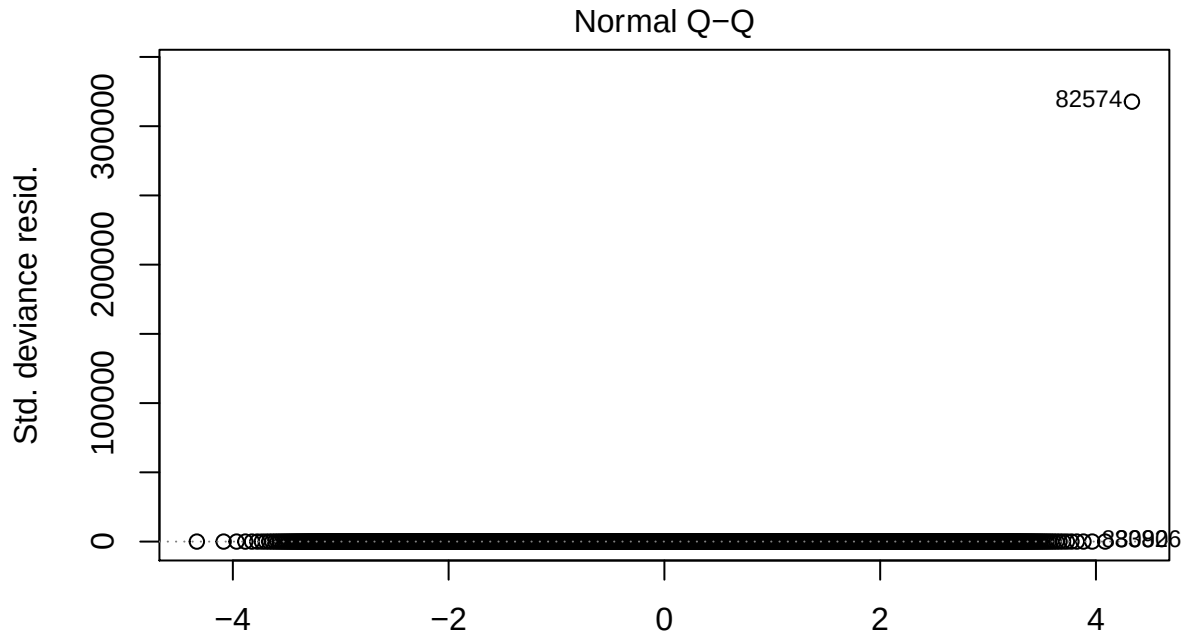
```
##
## Call:
## glm(formula = readmitted ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0694  -1.3345   0.8016   0.9330   2.5437
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.0218618  0.2801781   7.216 5.34e-13 ***
## raceAsian                  0.3608051  0.1117365   3.229 0.001242 **
## raceCaucasian             -0.0481757  0.0218377  -2.206 0.027379 *
## raceHispanic               0.1314631  0.0624489   2.105 0.035280 *
## raceOther                  0.3287038  0.0502299   6.544 5.99e-11 ***
## genderMale                 0.0662760  0.0166866   3.972 7.13e-05 ***
## genderUnknown/Invalid      7.2912357 43.9540602   0.166 0.868249
## age[10-20)                -0.6983308  0.2780125  -2.512 0.012009 *
## age[20-30)                -0.3485297  0.2691085  -1.295 0.195276
## age[30-40)                -0.5218984  0.2643436  -1.974 0.048346 *
## age[40-50)                -0.5420441  0.2626062  -2.064 0.039009 *
## age[50-60)                -0.5966177  0.2622610  -2.275 0.022912 *
## age[60-70)                -0.5930861  0.2622980  -2.261 0.023752 *
## age[70-80)                -0.6279053  0.2622953  -2.394 0.016671 *
## age[80-90)                -0.5845175  0.2626231  -2.226 0.026035 *
```

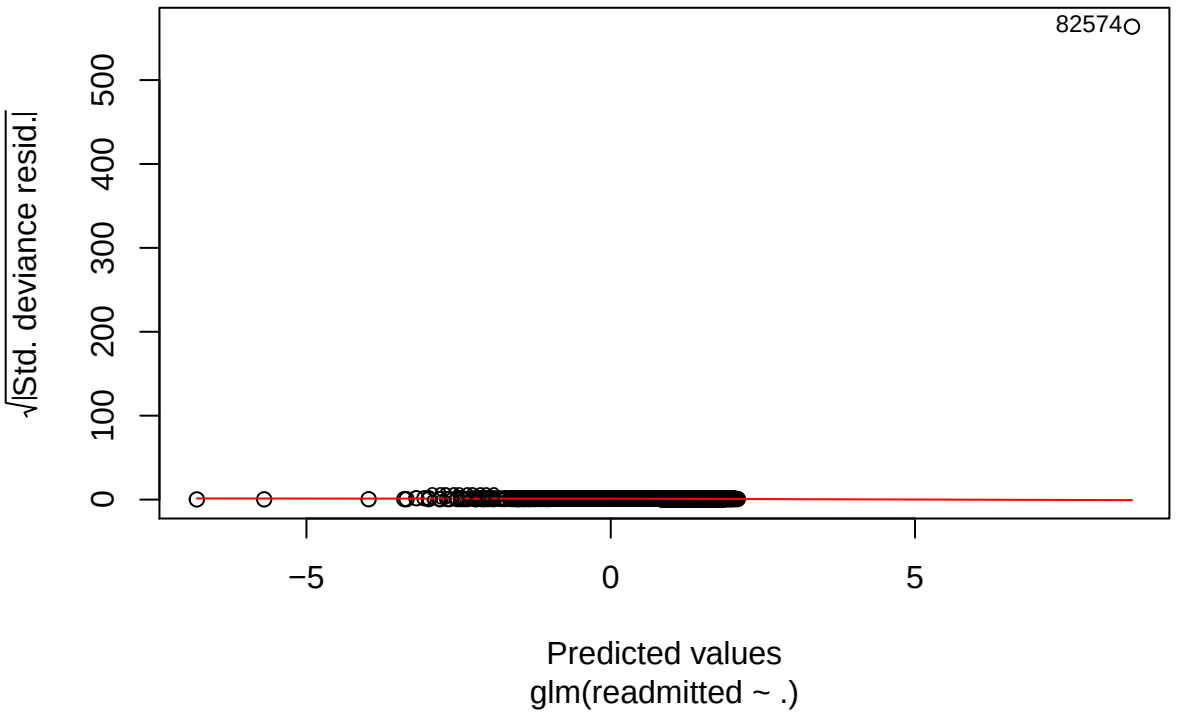
```
## age[90-100)               -0.2576538  0.2669355  -0.965 0.334430
## time_in_hospital          -0.0120905  0.0033183  -3.644 0.000269 ***
## num_lab_procedures        -0.0009467  0.0004866  -1.945 0.051724 .
## num_procedures             0.0359802  0.0057088   6.303 2.93e-10 ***
## num_medications            0.0018215  0.0013678   1.332 0.182948
## number_outpatient         -0.0792897  0.0068252 -11.617  < 2e-16 ***
## number_emergency          -0.0910685  0.0113445  -8.028 9.95e-16 ***
## number_inpatient          -0.1391976  0.0068584 -20.296  < 2e-16 ***
## number_diagnoses          -0.0747675  0.0051367 -14.555  < 2e-16 ***
## max_glu_serum>300         -0.1218766  0.0980782  -1.243 0.213998
## max_glu_serumNone          0.0900503  0.0694114   1.297 0.194513
## max_glu_serumNorm          0.0052957  0.0848788   0.062 0.950251
## A1Cresult>8               -0.1289469  0.0522154  -2.470 0.013529 *
## A1CresultNone             -0.1504887  0.0442873  -3.398 0.000679 ***
## A1CresultNorm             -0.0216940  0.0571035  -0.380 0.704015
## insulinNo                 -0.0245170  0.0328201  -0.747 0.455058
## insulinSteady              0.0955029  0.0303737   3.144 0.001665 **
## insulinUp                  0.0061281  0.0333627   0.184 0.854263
## changeNo                   0.0051303  0.0228705   0.224 0.822510
## diabetesMedYes            -0.2585747  0.0258637  -9.998  < 2e-16 ***
## diag1respiratory          -0.0461086  0.0271323  -1.699 0.089244 .
## diag1Digestive            -0.0176880  0.0325885  -0.543 0.587289
## diag1Diabetes             -0.0810701  0.0359829  -2.253 0.024258 *
## diag1Injury                0.1237662  0.0364720   3.393 0.000690 ***
## diag1Musculoskeletal       0.0926694  0.0437205   2.120 0.034041 *
## diag1Genitourinary         0.1219762  0.0408642   2.985 0.002837 **
## diag1Neoplasms             0.1379217  0.0281649   4.897 9.73e-07 ***
## diag1other                 0.0997307  0.0363829   2.741 0.006123 **
## diag2respiratory           0.0394568  0.0292339   1.350 0.177116
## diag2Digestive             0.1723829  0.0455261   3.786 0.000153 ***
## diag2Diabetes             -0.0037684  0.0301800  -0.125 0.900632
## diag2Injury                0.2431738  0.0590532   4.118 3.82e-05 ***
## diag2Musculoskeletal       0.0537948  0.0658835   0.817 0.414207
## diag2Genitourinary         0.0607008  0.0323864   1.874 0.060893 .
## diag2Neoplasms             0.0740163  0.0254566   2.908 0.003643 **
## diag2other                 0.1402409  0.0310455   4.517 6.26e-06 ***
## diag3respiratory           0.0343640  0.0337971   1.017 0.309261
## diag3Digestive             0.0355708  0.0454615   0.782 0.433957
## diag3Diabetes             -0.0027078  0.0263616  -0.103 0.918188
## diag3Injury                0.2849620  0.0651462   4.374 1.22e-05 ***
## diag3Musculoskeletal       0.0285607  0.0625148   0.457 0.647770
## diag3Genitourinary         0.0103034  0.0351788   0.293 0.769608
## diag3Neoplasms             0.1285461  0.0253255   5.076 3.86e-07 ***
## diag3other                 0.0501877  0.0273841   1.833 0.066843 .
## admission_sourceemergency -0.1573996  0.0305174  -5.158 2.50e-07 ***
## admission_sourceother      0.0384241  0.0320006   1.201 0.229856
## discharged_totransferred   0.1161715  0.0201843   5.756 8.64e-09 ***
## discharged_toleft_AMA      0.0933921  0.1064687   0.877 0.380390
## payer_codeself_pay        -0.1479700  0.0379324  -3.901 9.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 88232  on 68184  degrees of freedom
## Residual deviance: 85847  on 68121  degrees of freedom
## AIC: 85975
##
## Number of Fisher Scoring iterations: 7
```

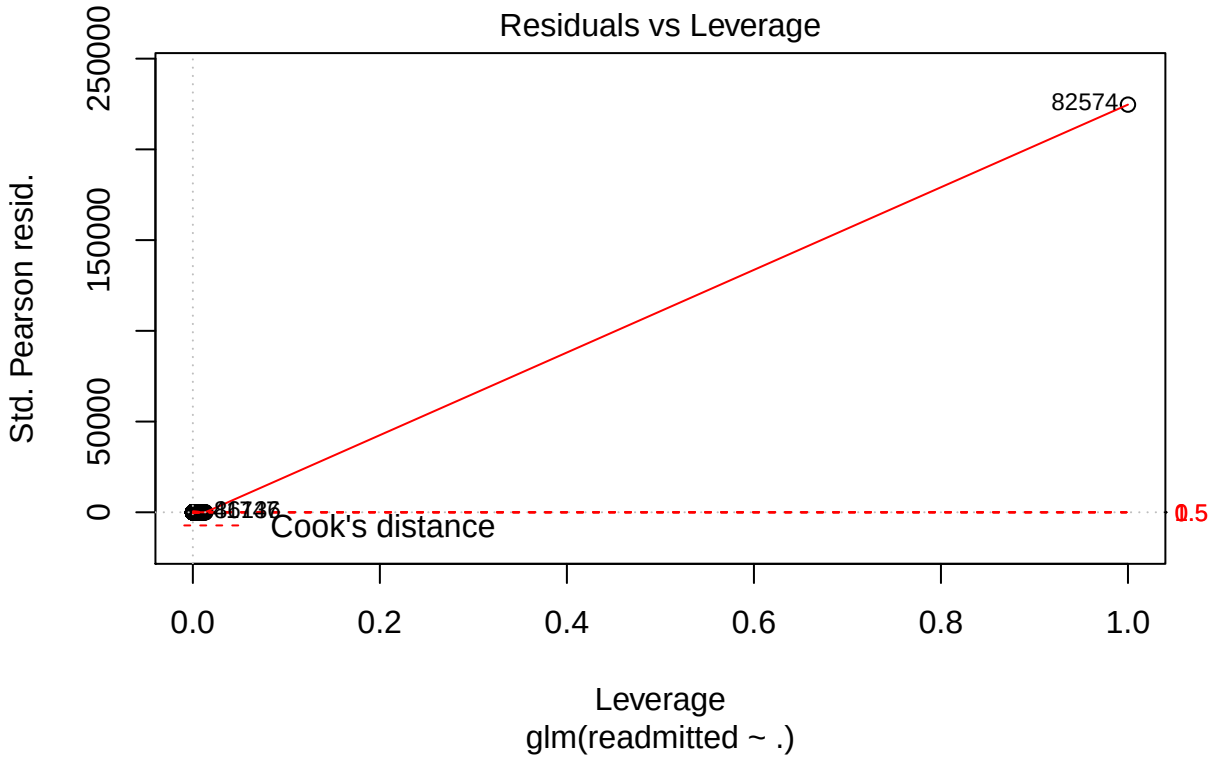
```
plot(logit_model)
```








```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
pred_logit <- predict(logit_model,test, type = "response")
#pred_logit

pred_logit <- ifelse(pred_logit > 0.5, 1, 0)
#pred_logit

result <- as.data.frame(table(pred_logit,test$readmitted))
#result

CorrectlyPredicted <- result[1,3]+result[4,3]

accuracy <- CorrectlyPredicted/nrow(test)
accuracy
```

```
## [1] 0.430422
```

```
senstivity_result <- result[4,3]/(result[2,3]+result[4,3])
senstivity_result
```

```
## [1] 0.4425347
```

```
specificity_result <- result[1,3]/(result[3,3]+result[1,3])
specificity_result
```

```
## [1] 0.2183021
```

Adding dimension which are most significant and then doing logistic regression

```r
normal_fit <- glm(readmitted ~ race + age + discharged_to + time_in_hospital +
             num_lab_procedures + num_procedures + num_medications + number_outpatient +
             number_emergency + number_inpatient + number_diagnoses +
             insulin + change + diabetesMed + diag1 + diag2 + diag3 + A1Cresult,
             data = train, family = binomial(link = 'logit'))

summary(normal_fit)
```

```
##
## Call:
## glm(formula = readmitted ~ race + age + discharged_to + time_in_hospital +
##     num_lab_procedures + num_procedures + num_medications + number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     insulin + change + diabetesMed + diag1 + diag2 + diag3 +
##     A1Cresult, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0513  -1.3427   0.8087   0.9316   2.5787
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.0460526  0.2705792   7.562 3.98e-14 ***
## raceAsian                  0.3443754  0.1116327   3.085 0.002036 **
## raceCaucasian             -0.0391613  0.0216363  -1.810 0.070298 .
## raceHispanic               0.1232683  0.0623490   1.977 0.048034 *
## raceOther                  0.3457385  0.0499631   6.920 4.52e-12 ***
## age[10-20)                -0.6949698  0.2778110  -2.502 0.012364 *
## age[20-30)                -0.3426753  0.2688467  -1.275 0.202447
## age[30-40)                -0.5118830  0.2641129  -1.938 0.052608 .
## age[40-50)                -0.5276386  0.2623599  -2.011 0.044312 *
## age[50-60)                -0.5784163  0.2620193  -2.208 0.027277 *
## age[60-70)                -0.5691966  0.2620549  -2.172 0.029852 *
## age[70-80)                -0.6063944  0.2620544  -2.314 0.020668 *
## age[80-90)                -0.5740896  0.2623821  -2.188 0.028670 *
## age[90-100)               -0.2599922  0.2666771  -0.975 0.329594
## discharged_totransferred   0.1213195  0.0191603   6.332 2.42e-10 ***
## discharged_toleft_AMA      0.0807748  0.1063483   0.760 0.447536
## time_in_hospital          -0.0122766  0.0033055  -3.714 0.000204 ***
## num_lab_procedures        -0.0017897  0.0004674  -3.829 0.000129 ***
## num_procedures             0.0465565  0.0056126   8.295  < 2e-16 ***
## num_medications            0.0031748  0.0013517   2.349 0.018841 *
## number_outpatient         -0.0773435  0.0067844 -11.400  < 2e-16 ***
## number_emergency          -0.0976367  0.0114084  -8.558  < 2e-16 ***
## number_inpatient          -0.1414956  0.0068468 -20.666  < 2e-16 ***
## number_diagnoses          -0.0801890  0.0050762 -15.797  < 2e-16 ***
## insulinNo                 -0.0026700  0.0326626  -0.082 0.934849
## insulinSteady              0.1182338  0.0302075   3.914 9.08e-05 ***
## insulinUp                  0.0097506  0.0333331   0.293 0.769889
## changeNo                   0.0044865  0.0228332   0.196 0.844228
## diabetesMedYes            -0.2561936  0.0258152  -9.924  < 2e-16 ***
## diag1respiratory          -0.0615840  0.0270612  -2.276 0.022862 *
## diag1Digestive            -0.0290236  0.0325419  -0.892 0.372456
```

```
## diag1Diabetes            -0.0764618  0.0358919  -2.130 0.033144 *
## diag1Injury               0.1182402  0.0364082   3.248 0.001164 **
## diag1Musculoskeletal      0.1387815  0.0433325   3.203 0.001361 **
## diag1Genitourinary        0.1292029  0.0407808   3.168 0.001534 **
## diag1Neoplasms            0.1469129  0.0280982   5.229 1.71e-07 ***
## diag1other                0.1376419  0.0361199   3.811 0.000139 ***
## diag2respiratory          0.0327373  0.0291906   1.122 0.262075
## diag2Digestive            0.1617472  0.0454570   3.558 0.000373 ***
## diag2Diabetes            -0.0115543  0.0301155  -0.384 0.701225
## diag2Injury               0.2460349  0.0589642   4.173 3.01e-05 ***
## diag2Musculoskeletal      0.0560565  0.0658273   0.852 0.394453
## diag2Genitourinary        0.0537248  0.0323312   1.662 0.096573 .
## diag2Neoplasms            0.0712138  0.0254189   2.802 0.005085 **
## diag2other                0.1397573  0.0310044   4.508 6.55e-06 ***
## diag3respiratory          0.0318279  0.0337551   0.943 0.345728
## diag3Digestive            0.0291063  0.0454091   0.641 0.521536
## diag3Diabetes            -0.0074476  0.0263207  -0.283 0.777211
## diag3Injury               0.2823331  0.0650920   4.337 1.44e-05 ***
## diag3Musculoskeletal      0.0235742  0.0624272   0.378 0.705708
## diag3Genitourinary        0.0097317  0.0351307   0.277 0.781769
## diag3Neoplasms            0.1242859  0.0252897   4.914 8.90e-07 ***
## diag3other                0.0479913  0.0273508   1.755 0.079318 .
## A1Cresult>8              -0.1262972  0.0521724  -2.421 0.015488 *
## A1CresultNone            -0.1461778  0.0442470  -3.304 0.000954 ***
## A1CresultNorm            -0.0230205  0.0570591  -0.403 0.686617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88232  on 68184  degrees of freedom
## Residual deviance: 85992  on 68129  degrees of freedom
## AIC: 86104
##
## Number of Fisher Scoring iterations: 4
```

```
pred_logit <- predict(normal_fit,test, type = "response")
#pred_logit

pred_logit <- ifelse(pred_logit > 0.5, 1, 0)
#pred_logit

result <- as.data.frame(table ( pred_logit , test$readmitted ))
result
```

```
##   pred_logit Var2  Freq
## 1          0    0   382
## 2          1    0 17723
## 3          0    1  1334
## 4          1    1 14142
```

```
CorrectlyPredicted <- result[1, 3] + result[4, 3]
```

```r
accuracy <- CorrectlyPredicted / nrow(test)
accuracy
```

```
## [1] 0.4325065
```

```r
senstivity_result <- result[4, 3] / (result[2, 3] + result[4, 3])
senstivity_result
```

```
## [1] 0.4438098
```

```r
specificity_result <- result[1, 3] / (result[3, 3] + result[1, 3])
specificity_result
```

```
## [1] 0.2226107
```

Rpart Decision Tree

R part decision Trees Prediction

Prediction using three level category

Training

```r
rpart_tree <- rpart(formula = readmitted ~ age + discharged_to + time_in_hospital +
                      num_lab_procedures+num_procedures+num_medications  +         number_outpatient + n
                      insulin + change + diabetesMed + diag1 + diag2 + diag3 + A1Cresult,
                    data=train_nonbinary, method = 'class')
summary(rpart_tree)
```

```
## Call:
## rpart(formula = readmitted ~ age + discharged_to + time_in_hospital +
##     num_lab_procedures + num_procedures + num_medications + number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     insulin + change + diabetesMed + diag1 + diag2 + diag3 +
##     A1Cresult, data = train_nonbinary, method = "class")
##   n= 68185
##
##           CP nsplit rel error    xerror        xstd
## 1 0.03777127      0 1.0000000 1.0000000 0.004141836
## 2 0.01759689      1 0.9622287 0.9622287 0.004127938
## 3 0.01000000      2 0.9446318 0.9446318 0.004119714
##
## Variable importance
## number_inpatient number_emergency
##               94                6
##
## Node number 1: 68185 observations,    complexity param=0.03777127
##   predicted class=NO   expected loss=0.4608932  P(node) =1
##     class counts:  7610 23816 36759
##    probabilities: 0.112 0.349 0.539
##   left son=2 (22809 obs) right son=3 (45376 obs)
```

```
##   Primary splits:
##       number_inpatient  < 0.5  to the right, improve=1148.6830, (0 missing)
##       number_emergency  < 0.5  to the right, improve= 399.5199, (0 missing)
##       number_outpatient < 0.5  to the right, improve= 362.4015, (0 missing)
##       number_diagnoses  < 5.5  to the right, improve= 338.4749, (0 missing)
##       num_medications   < 10.5 to the right, improve= 161.7848, (0 missing)
##   Surrogate splits:
##       number_emergency  < 0.5  to the right, agree=0.685, adj=0.058, (0 split)
##       number_outpatient < 4.5  to the right, agree=0.667, adj=0.004, (0 split)
##
## Node number 2: 22809 observations,    complexity param=0.01759689
##   predicted class=>30  expected loss=0.556447  P(node) =0.3345164
##     class counts:  3762 10117  8930
##    probabilities: 0.165 0.444 0.392
##   left son=4 (9769 obs) right son=5 (13040 obs)
##   Primary splits:
##       number_inpatient  < 1.5  to the right,   improve=197.61520, (0 missing)
##       number_emergency  < 0.5  to the right,   improve= 79.77849, (0 missing)
##       number_outpatient < 0.5  to the right,   improve= 74.30586, (0 missing)
##       age               splits as  RLLLLLLRRR, improve= 32.12547, (0 missing)
##       num_medications   < 12.5 to the right,   improve= 30.18158, (0 missing)
##   Surrogate splits:
##       number_emergency  < 0.5  to the right,   agree=0.599, adj=0.065, (0 split)
##       number_outpatient < 0.5  to the right,   agree=0.578, adj=0.014, (0 split)
##       diag1             splits as  RRRLRRRRR,  agree=0.577, adj=0.012, (0 split)
##       age               splits as  RRLLRRRRRR, agree=0.576, adj=0.011, (0 split)
##       discharged_to     splits as  RRL,        agree=0.573, adj=0.003, (0 split)
##
## Node number 3: 45376 observations
##   predicted class=NO   expected loss=0.3867022  P(node) =0.6654836
##     class counts:  3848 13699 27829
##    probabilities: 0.085 0.302 0.613
##
## Node number 4: 9769 observations
##   predicted class=>30  expected loss=0.517965  P(node) =0.143272
##     class counts:  2091  4709  2969
##    probabilities: 0.214 0.482 0.304
##
## Node number 5: 13040 observations
##   predicted class=NO   expected loss=0.5428681  P(node) =0.1912444
##     class counts:  1671  5408  5961
##    probabilities: 0.128 0.415 0.457
```

Prediction

```
pred_tree <- predict(rpart_tree, test_nonbinary, type="class")
head(pred_tree)
```

```
##  1  3  4  7  8 13
## NO NO NO NO NO NO
## Levels: <30 >30 NO
```

Performance

```
table(predict(rpart_tree, test_nonbinary, type="class"), test_nonbinary$readmitted)
```

```
##
##         <30   >30    NO
##   <30     0     0     0
##   >30  1037  2341  1468
##   NO   2710  9388 16637
```

```
result <- as.data.frame(table(predict(rpart_tree, test_nonbinary, type="class"),
                              test_nonbinary$readmitted))
result
```

```
##   Var1 Var2  Freq
## 1  <30  <30     0
## 2  >30  <30  1037
## 3   NO  <30  2710
## 4  <30  >30     0
## 5  >30  >30  2341
## 6   NO  >30  9388
## 7  <30   NO     0
## 8  >30   NO  1468
## 9   NO   NO 16637
```

```
confusionMatrix(pred_tree, test_nonbinary$readmitted)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   <30   >30    NO
##        <30     0     0     0
##        >30  1037  2341  1468
##        NO   2710  9388 16637
##
## Overall Statistics
##
##                Accuracy : 0.5651
##                  95% CI : (0.5598, 0.5705)
##     No Information Rate : 0.5391
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1094
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity              0.0000    0.19959    0.9189
## Specificity              1.0000    0.88537    0.2183
## Pos Pred Value              NaN    0.48308    0.5790
## Neg Pred Value           0.8884    0.67329    0.6971
## Prevalence               0.1116    0.34927    0.5391
```

```
## Detection Rate           0.0000    0.06971    0.4954
## Detection Prevalence     0.0000    0.14431    0.8557
## Balanced Accuracy        0.5000    0.54248    0.5686
```

```
prop.table(table(test_nonbinary$readmitted, pred_tree),1)
```

```
##      pred_tree
##               <30        >30         NO
##   <30 0.00000000 0.27675474 0.72324526
##   >30 0.00000000 0.19959076 0.80040924
##   NO  0.00000000 0.08108257 0.91891743
```

Prediction in binary format

Training

```
train$readmitted <- ifelse(train$readmitted == train$readmitted[1], 0, 1)
test$readmitted <- ifelse(test$readmitted == test$readmitted[1], 0, 1)

rpart_tree <- rpart(formula = readmitted ~ age + discharged_to + time_in_hospital +
                      num_lab_procedures+num_procedures+num_medications  +         number_outpatient + n
                      insulin + change + diabetesMed + diag1 + diag2 + diag3 + A1Cresult,
                    data=train, method = 'class')

summary(rpart_tree)
```

```
## Call:
## rpart(formula = readmitted ~ age + discharged_to + time_in_hospital +
##     num_lab_procedures + num_procedures + num_medications + number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     insulin + change + diabetesMed + diag1 + diag2 + diag3 +
##     A1Cresult, data = train, method = "class")
##   n= 68185
##
##   CP nsplit rel error xerror xstd
## 1  0      0         1      0    0
##
## Node number 1: 68185 observations
##   predicted class=1  expected loss=0.349285  P(node) =1
##     class counts: 23816 44369
##    probabilities: 0.349 0.651
```

Prediction

```
pred_tree <- predict(rpart_tree, test, type="class")
```

Performance

```
table(predict(rpart_tree, test, type = "class"), test$readmitted)
```

```
##
##         0     1
##   0     0     0
##   1 18105 15476
```

```r
result <- as.data.frame(table(predict(rpart_tree, test, type = "class"), test$readmitted))
result
```

```
##   Var1 Var2  Freq
## 1    0    0     0
## 2    1    0 18105
## 3    0    1     0
## 4    1    1 15476
```

```r
CorrectlyPredicted <- result[1, 3] + result[4, 3]

accuracy <- CorrectlyPredicted / nrow(test)
accuracy
```

```
## [1] 0.4608558
```

```r
senstivity_result <- result[4, 3] / (result[2, 3] + result[4, 3])
senstivity_result
```

```
## [1] 0.4608558
```

```r
specificity_result <- result[1, 3] / (result[3, 3] + result[1, 3])
specificity_result
```

```
## [1] NaN
```

```r
prop.table(table(test$readmitted, pred_tree), 1)
```

```
##    pred_tree
##     0 1
##   0 0 1
##   1 0 1
```