R Notebook

The following is the head of the data.

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
train = read.csv('train.csv')
head(train)
```

```
##   Id Product_Info_1 Product_Info_2 Product_Info_3 Product_Info_4
## 1  2              1              D3             10     0.07692308
## 2  5              1              A1             26     0.07692308
## 3  6              1              E1             26     0.07692308
## 4  7              1              D4             10     0.48717949
## 5  8              1              D2             26     0.23076923
## 6 10              1              D2             26     0.23076923
##   Product_Info_5 Product_Info_6 Product_Info_7    Ins_Age        Ht
## 1              2              1              1 0.64179104 0.5818182
## 2              2              3              1 0.05970149 0.6000000
## 3              2              3              1 0.02985075 0.7454545
## 4              2              3              1 0.16417910 0.6727273
## 5              2              3              1 0.41791045 0.6545455
## 6              3              1              1 0.50746269 0.8363636
##          Wt       BMI Employment_Info_1 Employment_Info_2
## 1 0.1485356 0.3230080             0.028                12
## 2 0.1317992 0.2722877             0.000                 1
## 3 0.2887029 0.4287804             0.030                 9
## 4 0.2050209 0.3524377             0.042                 9
## 5 0.2343096 0.4240456             0.027                 9
## 6 0.2991632 0.3648867             0.325                15
##   Employment_Info_3 Employment_Info_4 Employment_Info_5 Employment_Info_6
## 1                 1                 0                 3                NA
## 2                 3                 0                 2            0.0018
## 3                 1                 0                 2            0.0300
## 4                 1                 0                 3            0.2000
## 5                 1                 0                 2            0.0500
## 6                 1                 0                 2            1.0000
##   InsuredInfo_1 InsuredInfo_2 InsuredInfo_3 InsuredInfo_4 InsuredInfo_5
## 1             1             2             6             3             1
## 2             1             2             6             3             1
## 3             1             2             8             3             1
## 4             2             2             8             3             1
## 5             1             2             6             3             1
## 6             1             2             8             3             1
##   InsuredInfo_6 InsuredInfo_7 Insurance_History_1 Insurance_History_2
## 1             2             1                   1                   1
## 2             2             1                   2                   1
## 3             1             1                   2                   1
## 4             2             1                   2                   1
```

```
## 5             2             1                   2                   1
## 6             1             1                   2                   1
##   Insurance_History_3 Insurance_History_4 Insurance_History_5
## 1                   3                   1         0.000666667
## 2                   3                   1         0.000133333
## 3                   1                   3                  NA
## 4                   1                   3                  NA
## 5                   1                   3                  NA
## 6                   3                   2         0.005000000
##   Insurance_History_7 Insurance_History_8 Insurance_History_9
## 1                   1                   1                   2
## 2                   1                   3                   2
## 3                   3                   2                   3
## 4                   3                   2                   3
## 5                   3                   2                   3
## 6                   1                   3                   2
##   Family_Hist_1 Family_Hist_2 Family_Hist_3 Family_Hist_4 Family_Hist_5
## 1             2            NA     0.5980392            NA     0.5267857
## 2             2     0.1884058            NA    0.08450704            NA
## 3             3     0.3043478            NA    0.22535211            NA
## 4             3     0.4202899            NA    0.35211268            NA
## 5             2     0.4637681            NA    0.40845070            NA
## 6             2            NA     0.2941176    0.50704225            NA
##   Medical_History_1 Medical_History_2 Medical_History_3 Medical_History_4
## 1                 4               112                 2                 1
## 2                 5               412                 2                 1
## 3                10                 3                 2                 2
## 4                 0               350                 2                 2
## 5                NA               162                 2                 2
## 6                 6               491                 2                 2
##   Medical_History_5 Medical_History_6 Medical_History_7 Medical_History_8
## 1                 1                 3                 2                 2
## 2                 1                 3                 2                 2
## 3                 1                 3                 2                 2
## 4                 1                 3                 2                 2
## 5                 1                 3                 2                 2
## 6                 1                 3                 2                 2
##   Medical_History_9 Medical_History_10 Medical_History_11
## 1                 1                 NA                 3
## 2                 1                 NA                 3
## 3                 2                 NA                 3
## 4                 2                 NA                 3
## 5                 2                 NA                 3
## 6                 2                 NA                 3
##   Medical_History_12 Medical_History_13 Medical_History_14
## 1                  2                  3                  3
## 2                  2                  3                  3
## 3                  2                  3                  3
## 4                  2                  3                  3
## 5                  2                  3                  3
## 6                  2                  3                  3
##   Medical_History_15 Medical_History_16 Medical_History_17
## 1                240                  3                  3
## 2                  0                  1                  3
```

```
## 3                 NA                  1                  3
## 4                 NA                  1                  3
## 5                 NA                  1                  3
## 6                 NA                  1                  3
##   Medical_History_18 Medical_History_19 Medical_History_20
## 1                  1                  1                  2
## 2                  1                  1                  2
## 3                  1                  1                  2
## 4                  1                  1                  2
## 5                  1                  1                  2
## 6                  2                  1                  2
##   Medical_History_21 Medical_History_22 Medical_History_23
## 1                  1                  2                  3
## 2                  1                  2                  3
## 3                  1                  2                  3
## 4                  2                  2                  3
## 5                  1                  2                  3
## 6                  2                  2                  3
##   Medical_History_24 Medical_History_25 Medical_History_26
## 1                 NA                  1                  3
## 2                 NA                  1                  3
## 3                 NA                  2                  2
## 4                 NA                  1                  3
## 5                 NA                  2                  2
## 6                 NA                  1                  3
##   Medical_History_27 Medical_History_28 Medical_History_29
## 1                  3                  1                  3
## 2                  3                  1                  3
## 3                  3                  1                  3
## 4                  3                  1                  3
## 5                  3                  1                  3
## 6                  3                  1                  3
##   Medical_History_30 Medical_History_31 Medical_History_32
## 1                  2                  3                 NA
## 2                  2                  3                 NA
## 3                  2                  3                 NA
## 4                  2                  3                 NA
## 5                  2                  3                 NA
## 6                  2                  3                 NA
##   Medical_History_33 Medical_History_34 Medical_History_35
## 1                  1                  3                  1
## 2                  3                  1                  1
## 3                  3                  3                  1
## 4                  3                  3                  1
## 5                  3                  3                  1
## 6                  3                  1                  1
##   Medical_History_36 Medical_History_37 Medical_History_38
## 1                  2                  2                  1
## 2                  2                  2                  1
## 3                  3                  2                  1
## 4                  2                  2                  1
## 5                  3                  2                  1
## 6                  2                  2                  1
##   Medical_History_39 Medical_History_40 Medical_History_41
```

```
## 1                  3                  3                  3
## 2                  3                  3                  1
## 3                  3                  3                  1
## 4                  3                  3                  1
## 5                  3                  3                  1
## 6                  3                  3                  3
##   Medical_Keyword_1 Medical_Keyword_2 Medical_Keyword_3 Medical_Keyword_4
## 1                 0                  0                  0                  0
## 2                 0                  0                  0                  0
## 3                 0                  0                  0                  0
## 4                 0                  0                  0                  0
## 5                 0                  0                  0                  0
## 6                 0                  0                  0                  0
##   Medical_Keyword_5 Medical_Keyword_6 Medical_Keyword_7 Medical_Keyword_8
## 1                 0                  0                  0                  0
## 2                 0                  0                  0                  0
## 3                 0                  0                  0                  0
## 4                 0                  0                  0                  0
## 5                 0                  0                  0                  0
## 6                 0                  0                  0                  0
##   Medical_Keyword_9 Medical_Keyword_10 Medical_Keyword_11
## 1                 0                  0                  0
## 2                 0                  0                  0
## 3                 0                  0                  0
## 4                 0                  0                  0
## 5                 0                  0                  0
## 6                 0                  0                  0
##   Medical_Keyword_12 Medical_Keyword_13 Medical_Keyword_14
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_15 Medical_Keyword_16 Medical_Keyword_17
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_18 Medical_Keyword_19 Medical_Keyword_20
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_21 Medical_Keyword_22 Medical_Keyword_23
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
```

```
## 6                  0                  1                  0
##   Medical_Keyword_24 Medical_Keyword_25 Medical_Keyword_26
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_27 Medical_Keyword_28 Medical_Keyword_29
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_30 Medical_Keyword_31 Medical_Keyword_32
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  1
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_33 Medical_Keyword_34 Medical_Keyword_35
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  1                  0
##   Medical_Keyword_36 Medical_Keyword_37 Medical_Keyword_38
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_39 Medical_Keyword_40 Medical_Keyword_41
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_42 Medical_Keyword_43 Medical_Keyword_44
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   Medical_Keyword_45 Medical_Keyword_46 Medical_Keyword_47
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
```

```
## 4                   0                   0                   0
## 5                   0                   0                   0
## 6                   0                   0                   0
##   Medical_Keyword_48 Response
## 1                  0        8
## 2                  0        4
## 3                  0        8
## 4                  0        8
## 5                  0        8
## 6                  0        8
```

In this report we focus on data cleaning and EDA. We have removed 12 features that had more than 6000 NA values. After the features were dropped, we also dropped rows that had NA values.

Removing the columns with NAs greater than 6000(more than 10% of data)

```
#sapply(train, function(x) sum(is.na(x)))
#rm(train_imp)
#removing the columns with NAs greater than 6000(more than 10% of data)
train_imp=train
x = list()
count =1
for (i in 1:128){
  if (sum(is.na(train[,i]))>6000){
    x[count] = i
    count = count+1
  }
}
x = unlist(x)
train_imp = train[,-x]
train_imp = na.omit(train_imp)
train_imp$Product_Info_2 = as.numeric(train_imp$Product_Info_2)

write.csv(train_imp, "Final-Train.csv")
```

Dim after cleaning

```
dim(train_imp)
```

```
## [1] 59362   116
```

Let us now find out the distribution of classes in the data.

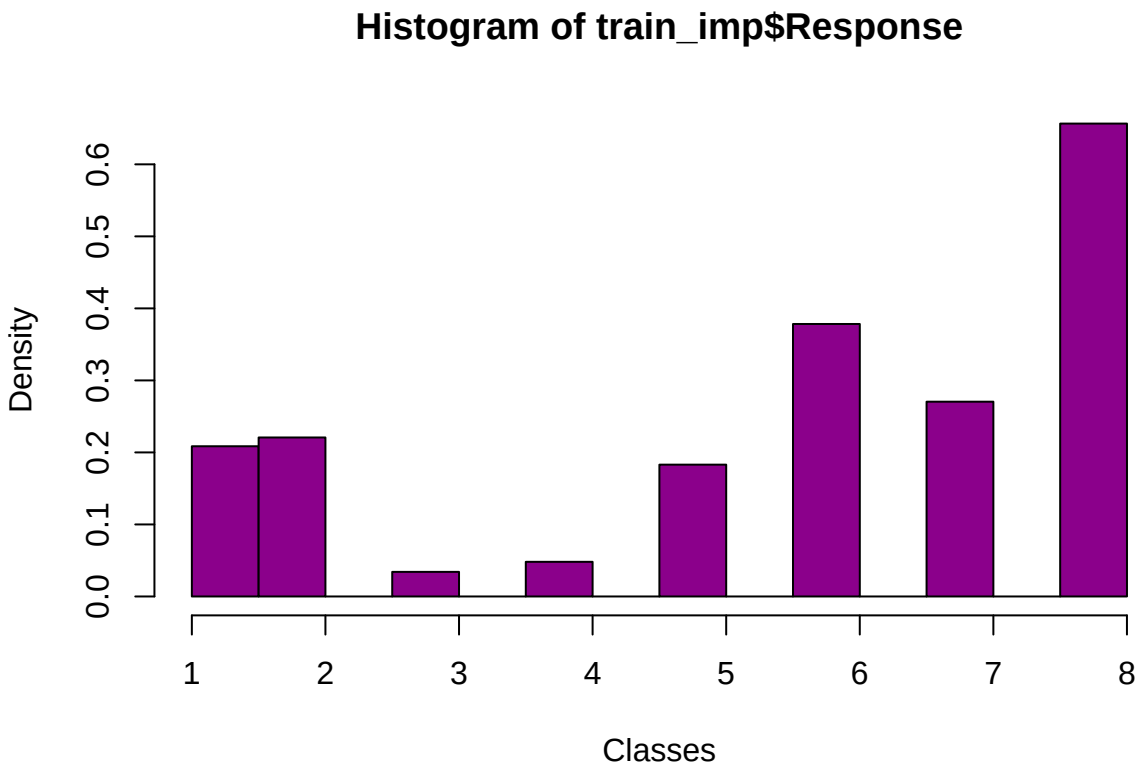
```
hist(train_imp$Response, xlab="Classes", xlim=c(1,8), col="darkmagenta", freq=FALSE, title="Class distri
```

```
## Warning in plot.window(xlim, ylim, "", ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "title" is not a graphical parameter

## Warning in axis(1, ...): "title" is not a graphical parameter
```

```
## Warning in axis(2, ...): "title" is not a graphical parameter
```




```
classCount = rep(0, 8)
for(i in 1:8){
  classCount[i] = length(train_imp$Response[train_imp$Response == i])
}
classes = seq(1,8)
classCount
```

```
## [1]  6191  6552  1013  1428  5432 11230  8027 19489
```

```
data.frame(classes, classCount)
```

```
##   classes classCount
## 1       1       6191
## 2       2       6552
## 3       3       1013
## 4       4       1428
## 5       5       5432
## 6       6      11230
## 7       7       8027
## 8       8      19489
```

It can be seen that there aren't as many samples for class 3(1013) and class 4(1428), as there are for other classes. Out assumption is that they still should be enough.