```
rm(list=ls()); gc()
setwd('/Users/shwetasaloni/Downloads/')
dat=read.csv('diabetic_data.csv', head=T, stringsAsFactors = F)
#View(dat)
dim(dat)
[1] 101766    48
```

```
> head(dat)
  encounter_id patient_nbr           race gender     age admission_type_id discharge_disposition_id admission_source_id time_in_hospital       medical_specialty num_lab_procedures
1      2278392     8222157      Caucasian Female  [0-10)               6                       25                  1                1 Pediatrics-Endocrinology                41
2       149190    55629189      Caucasian Female [10-20)               1                        1                  7                3                       ?                59
3        64410    86047875 AfricanAmerican Female [20-30)               1                        1                  7                2                       ?                11
4       500364    82442376      Caucasian   Male [30-40)               1                        1                  7                2                       ?                44
5        16680    42519267      Caucasian   Male [40-50)               1                        1                  7                1                       ?                51
6        35754    82637451      Caucasian   Male [50-60)               2                        1                  2                3                       ?                31
  num_procedures num_medications number_outpatient number_emergency number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum A1Cresult metformin repaglinide nateglinide
1              0                1                 0                 0                0 250.83      ?      ?                1         None       None        No          No          No
2              0               18                 0                 0                0    276 250.01    255                9         None       None        No          No          No
3              5               13                 2                 0                1    648    250    V27                6         None       None        No          No          No
4              1               16                 0                 0                0      8 250.43    403                7         None       None        No          No          No
5              0                8                 0                 0                0    197    157    250                5         None       None        No          No          No
6              6               16                 0                 0                0    414    411    250                9         None       None        No          No          No
  chlorpropamide glimepiride acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone acarbose miglitol troglitazone tolazamide examide citoglipton insulin
1             No          No            No        No       No          No           No            No       No       No           No         No      No          No      No
2             No          No            No        No       No          No           No            No       No       No           No         No      No          No      Up
3             No          No            No    Steady       No          No           No            No       No       No           No         No      No          No      No
4             No          No            No        No       No          No           No            No       No       No           No         No      No          No      Up
5             No          No            No    Steady       No          No           No            No       No       No           No         No      No          No  Steady
6             No          No            No        No       No          No           No            No       No       No           No         No      No          No  Steady
  glyburide.metformin glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone metformin.pioglitazone change diabetesMed readmitted
1                  No                  No                       No                     No                     No     No         No         NO
2                  No                  No                       No                     No                     No     Ch        Yes        >30
3                  No                  No                       No                     No                     No     No        Yes         NO
4                  No                  No                       No                     No                     No     Ch        Yes         NO
5                  No                  No                       No                     No                     No     Ch        Yes         NO
6                  No                  No                       No                     No                     No     No        Yes        >30
```

```
> dat[dat == '?'] <- NA
> dat1=na.omit(dat)
> dat[dat == '?'] <- NA
> dat1=na.omit(dat)
> dim(dat1)
[1] 49735    48
>
```

Removing the features that doesn't contribute to the model accuracy. The features are not adding significant information towards the objective. Hence getting rid of these.

```
> dat1=dat1[,setdiff(colnames(dat),c('examide', 'citoglipton', 'glimepiride.pioglitazone',     "metformin.rosiglitazone", 'acetohexamide', 'repaglinide',     'nateglinide',
'chlorpropamide', 'tolbutamide',
+                          'rosiglitazone',    'acarbose',     'miglitol', 'troglitazone',     'tolazamide', 'glyburide.metformin',    'glipizide.metformin',
'metformin.pioglitazone',
+                          'metformin'))]
>
> dim(dat1)
[1] 49735    30
>
```

```
> unique(dat1$gender)
[1] "Female"          "Male"            "Unknown/Invalid"
> #removing gender=='Unknown/Invalid' value from the model as gender= Unknown/Invalid doesn't make any sence.
> dat2=dat1[!(dat1$gender=='Unknown/Invalid'), ]
> unique(dat2$gender)
[1] "Female" "Male"
>
> #Category reduction: The % of Asian and Hispanic is less
> unique(dat2$race)
[1] "Caucasian"       "Other"           "AfricanAmerican" "Asian"           "Hispanic"
> dat2$race= as.factor(ifelse(dat2$race == 'Asian' | dat2$race == 'Hispanic', 'Other',dat2$race))
>
> #Category reduction: Dividing age feature into more distingushable term
> table(dat2$age)

  [0-10)  [10-20)  [20-30)  [30-40)  [40-50)  [50-60)  [60-70)  [70-80)  [80-90) [90-100)
      52      285      812     1909     4985     8693    11013    12709     7950     1326
> dat2$age=as.factor(ifelse(dat2$age=="[0-10)" | dat2$age== '[10-20)' | dat2$age== '[20-30)' ,'age_under30',
+                           ifelse(dat2$age == '[30-40)' | dat2$age == '[40-50)' | dat2$age == '[50-60)','age_30To60',
+                                  ifelse(dat2$age == '[60-70)' | dat2$age == '[70-80)' | dat2$age == '[80-90)' | dat2$age == '[90-100)','Above60', dat2$age))))
>
```

```
> ##Category reduction: The % of admission_type_id as 4,5,6,8 is less, hence merging all into admission_type_id=4.
> table(dat2$admission_type_id)

    1     2     3     4
19885 12299 11754  5796
> dat2$admission_type_id=as.factor(ifelse(dat2$admission_type_id == '4' |
+                                         dat2$admission_type_id == '5' |
+                                         dat2$admission_type_id == '6' |
+                                         dat2$admission_type_id == '8', '4', dat2$admission_type_id))
>
> table(dat2$discharge_disposition_id)

    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20    22    23    24    25    27    28
31367  1080  6528   407   698  5220   267    41     3     2   812     3   152   104    22    11    11   877     8     1  1126   207     8   757     1    21
>
```

```
> dat2$admission_source_id=as.factor(ifelse(dat2$admission_source_id %in% c(as.character(2:6), as.character(8:10), '13','14', '17','22'), '0',dat2$admission_source_id))
> #Category reduction: Set the value of discharge_disposition_id as 0 whose frequency is very less
> dat2$discharge_disposition_id=as.factor(ifelse(dat2$discharge_disposition_id ==  setdiff( unique(dat2$discharge_disposition_id), c('2', '4', '5', as.character(7:28))  ), dat2$discharge_dispo
sition_id, '0') )
>
> #Category reduction: Set the value of admission_source_id as 0 whose frequency is very less
> table(dat2$admission_source_id)

    1     2     3
 7837 17114 24783
> dat2$admission_source_id=as.factor(ifelse(dat2$admission_source_id %in% c(as.character(2:6), as.character(8:10), '13','14', '17','22'), '0',dat2$admission_source_id))
>
```

```
> table(dat2$medical_specialty)

          Cardiology       Emergency/Trauma Family/GeneralPractice       InternalMedicine                 Others       Surgery-General
                5217                   7472                   7140                  13967                  12969                   2969
> dat2$medical_specialty=as.factor(ifelse(dat2$medical_specialty %in% setdiff(unique(dat2$medical_specialty), c('Cardiology', 'InternalMedicine', 'Family/GeneralPractice', 'Emergency/Trauma',
'Surgery-General')),
+                                       'Others', dat2$medical_specialty))
>
```

```
> table(dat2$diag_1)
```

10	11	112	115	117	131	133	135	136	141	142	143	145	146	149	150	151	152	153	154	155	156	157	158	160	161
1	7	28	2	3	1	1	17	4	3	2	1	1	4	1	10	35	10	159	54	32	9	57	3	1	6
162	163	164	170	171	172	173	174	175	179	180	182	183	184	185	187	188	189	191	192	193	194	195	196	197	198
203	4	1	1	6	3	2	85	2	1	4	50	13	3	102	1	58	81	22	1	19	2	4	10	152	126
199	200	201	202	203	204	205	207	210	211	212	214	215	216	217	218	220	223	225	226	227	228	230	233	235	236
11	8	3	53	16	10	11	1	3	51	6	2	1	1	1	82	22	3	26	11	29	2	4	15	17	2
237	238	239	240	241	242	244	245	246	250	250.01	250.02	250.03	250.1	250.11	250.12	250.13	250.2	250.21	250.22	250.23	250.3	250.31	250.32	250.33	250.4
4	32	14	1	38	8	5	4	1	75	11	372	88	110	234	175	324	52	5	81	20	10	6	11	15	159
250.41	250.42	250.43	250.5	250.51	250.52	250.53	250.6	250.7	250.8	250.81	250.82	250.83	250.9	250.91	250.92	250.93	251	252	253	255	261	262	263	27	272
40	47	15	3	1	3	2	573	497	772	89	178	59	2	2	19	5	4	13	25	12	2	2	9	2	1
273	274	275	276	277	278	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299
2	46	23	941	13	120	158	10	29	14	34	157	22	34	43	5	44	57	48	32	67	308	558	7	40	1
3	300	301	303	304	305	306	307	308	309	31	310	311	312	314	318	320	323	324	327	331	332	333	334	335	337
8	39	1	46	23	9	4	11	2	25	2	6	27	13	1	1	3	4	2	9	109	22	13	2	6	5
338	34	340	341	342	344	345	346	348	349	35	350	351	353	354	355	356	357	358	359	360	361	362	365	366	368
10	4	26	2	5	3	35	40	64	16	7	6	15	4	1	12	1	8	3	4	1	5	3	1	1	6
369	370	373	374	375	376	377	378	379	38	380	382	383	384	385	386	388	389	39	391	394	396	397	398	401	402
1	2	1	2	1	7	2	8	4	716	11	2	3	1	1	58	3	1	2	1	3	9	2	76	175	265
403	404	405	41	410	411	412	413	414	415	416	417	42	420	421	423	424	425	426	427	428	429	430	431	432	433

```
> #Category Reduction: Based on ICD 9 codes for diabetic diagnosis
> dat2$diag_1= as.factor(ifelse((dat2$diag_1 >=390 & dat2$diag_1 <= 459) | dat2$diag_1 == 785, 'Circulatory',
+                               ifelse((dat2$diag_1 >=460 & dat2$diag_1 <=519) | dat2$diag_1 == 786, 'Respiratory',
+                                      ifelse((dat2$diag_1 >= 520 & dat2$diag_1 <= 579) | dat2$diag_1 == 787, 'Digestive',
+                                             ifelse(dat2$diag_1 >= 250 & dat2$diag_1 < 251, 'Diabetes',
+                                                    ifelse(dat2$diag_1 >= 800 & dat2$diag_1<=999, 'Injury',
+                                                           ifelse(dat2$diag_1>=710 & dat2$diag_1<= 739, 'Musculoskeletal',
+                                                                  ifelse((dat2$diag_1>= 580 & dat2$diag_1 <= 629) | dat2$diag_1 == 788, 'Genitourinary',
+                                                                         ifelse((dat2$diag_1>= 140 & dat2$diag_1 <= 239) | dat2$diag_1 == 780 | dat2$diag_1 == 781 | dat2$diag_1 == 784 |
+                                                                                  (dat2$diag_1>= 790 & dat2$diag_1 <= 799) | (dat2$diag_1>= 240 & dat2$diag_1 <= 279 & dat2$diag_1!= 250) |
+                                                                                  (dat2$diag_1>= 680 & dat2$diag_1 <= 709 | dat2$diag_1 == 782) | (dat2$diag_1>= 001 & dat2$diag_1<= 139),'Neop
lasms', 'Other')))))))))
> table(dat2$diag_2)
```

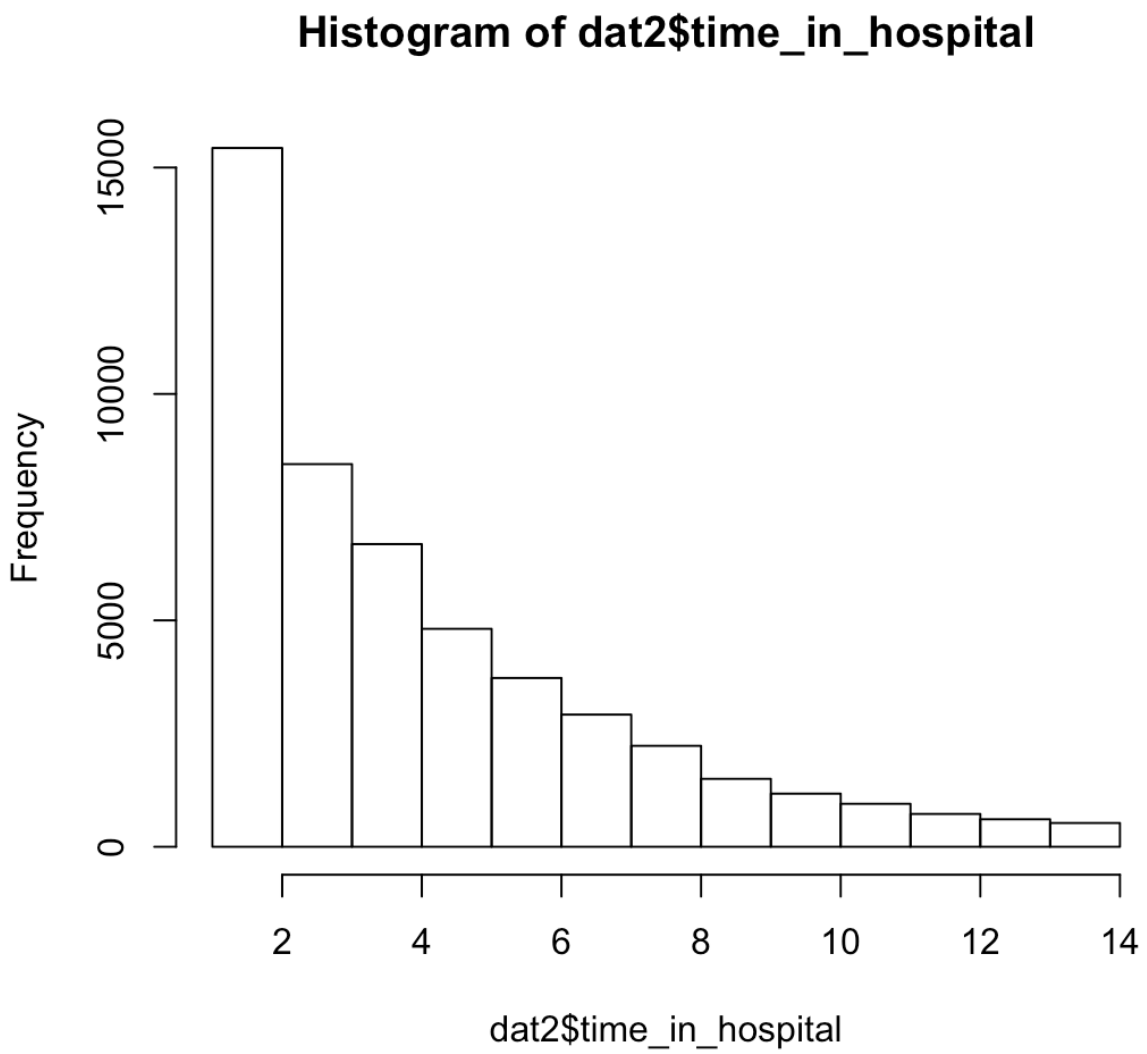
11	110	112	117	130	131	135	136	137	138	141	150	151	152	153	154	155	156	157	162	163	172	173	174	182	183
2	5	90	5	1	1	68	4	1	9	1	15	10	8	73	23	13	7	39	177	1	4	1	15	1	8
185	188	189	191	193	195	196	197	198	199	200	201	202	203	204	205	208	211	214	217	218	220	223	225	226	227
16	5	30	6	3	1	100	200	174	5	13	5	105	48	59	30	3	30	5	3	26	6	1	4	1	5
232	233	238	239	241	242	244	245	246	250	250.01	250.02	250.03	250.1	250.11	250.12	250.13	250.2	250.21	250.22	250.23	250.31	250.32	250.33	250.4	250.41
1	5	8	1	9	49	82	1	3	3034	969	1066	140	9	34	48	50	4	1	18	2	1	2	3	107	136
250.42	250.43	250.5	250.51	250.52	250.53	250.6	250.7	250.8	250.81	250.82	250.83	250.9	250.91	250.92	250.93	251	252	253	255	258	260	261	262	263	266
62	24	39	42	32	13	496	77	90	59	69	17	5	7	61	11	1	11	24	34	4	4	7	8	92	4
269	27	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293
1	2	1	1	212	2	14	20	3170	41	146	6	328	5	26	9	80	699	61	170	35	16	21	39	52	39
294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	31	310	311	312	314	317	318	319	320	322

```
> # Category Reduction: Based on ICD 9 codes for diabetic diagnosis
> dat2$diag_2= as.factor(ifelse((dat2$diag_2 >=390 & dat2$diag_2 <= 459) | dat2$diag_2 == 785, 'Circulatory',
+                               ifelse((dat2$diag_2 >=460 & dat2$diag_2 <= 519) | dat2$diag_2 == 786, 'Respiratory',
+                                      ifelse((dat2$diag_2 >= 520 & dat2$diag_2 <= 579) | dat2$diag_2 == 787, 'Digestive',
+                                             ifelse(dat2$diag_2 >= 250 & dat2$diag_2 < 251, 'Diabetes',
+                                                    ifelse(dat2$diag_2 >= 800 & dat2$diag_2 <=999, 'Injury',
+                                                           ifelse(dat2$diag_2>=710 & dat2$diag_2 <= 739, 'Musculoskeletal',
+                                                                  ifelse((dat2$diag_2>= 580 & dat2$diag_2 <= 629) | dat2$diag_2 == 788, 'Genitourinary',
+                                                                         ifelse((dat2$diag_2>= 140 & dat2$diag_2 <= 239) | dat2$diag_2 == 780 | dat2$diag_2 == 781 | dat2$diag_2 == 784 |
+                                                                                  (dat2$diag_2>= 790 & dat2$diag_2 <= 799) | (dat2$diag_2 >= 240 & dat2$diag_2 <= 279 & dat2$diag_2!= 250) |
+                                                                                  (dat2$diag_2>= 680 & dat2$diag_2 <= 709 | dat2$diag_2 == 782) | (dat2$diag_2>= 001 & dat2$diag_2<= 139),'Neop
lasms', 'Other')))))))))
> table(dat2$diag_3)
```

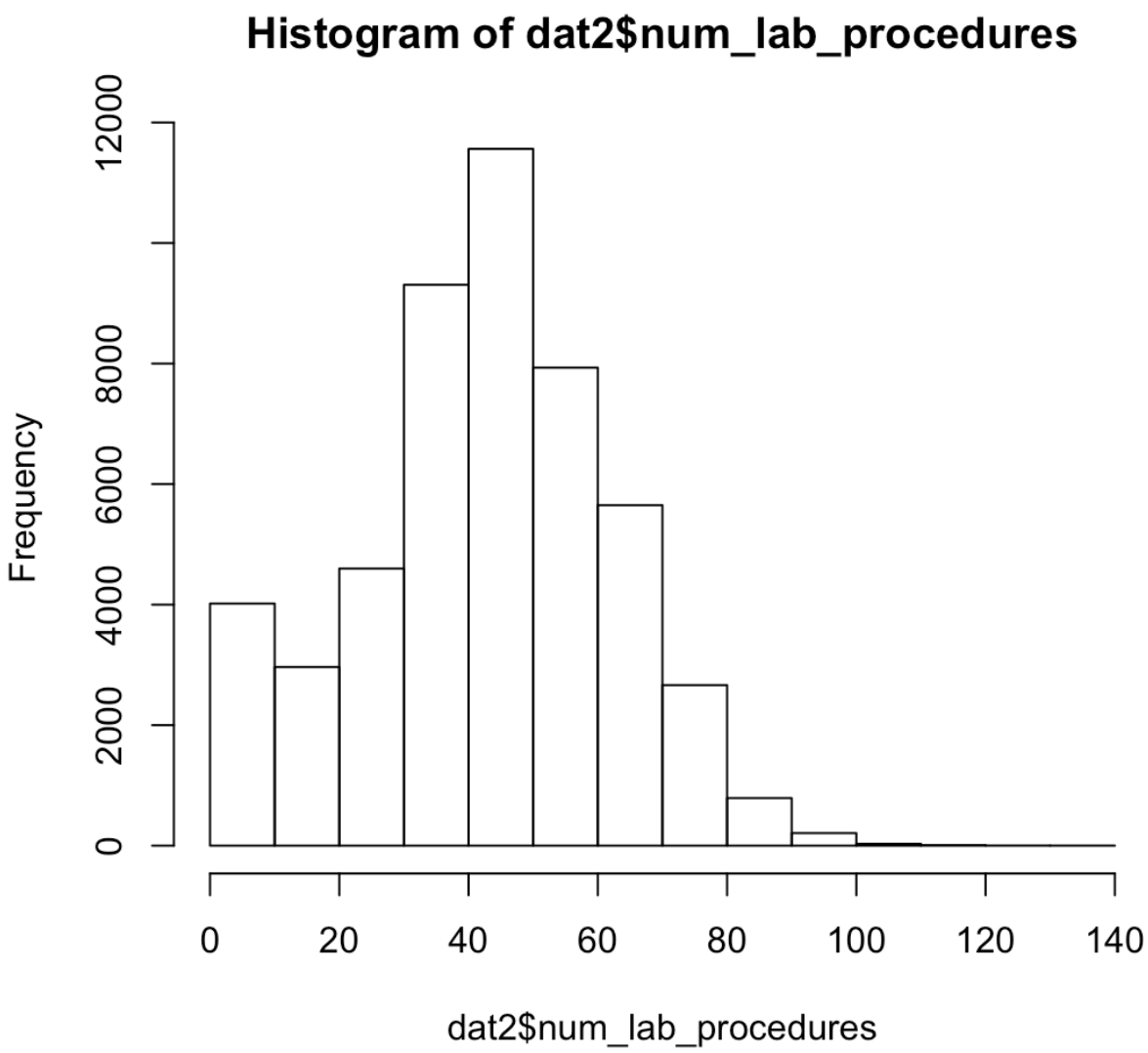
11	110	112	115	117	122	123	131	135	138	139	14	141	146	150	151	152	153	154	155	156	157	158	161	162	163
1	12	108	1	1	1	1	3	38	10	1	1	4	1	4	5	1	30	6	4	2	11	1	1	77	4
164	17	170	171	172	173	174	175	179	182	183	185	188	189	191	192	193	195	196	197	198	199	200	201	202	203
1	1	3	1	2	1	20	1	3	3	3	27	9	12	4	2	1	1	45	207	112	28	6	8	67	25
204	205	208	211	214	215	216	218	220	223	225	226	227	228	233	235	236	238	239	240	241	242	243	244	245	246
27	12	3	53	1	1	2	34	2	1	2	1	4	2	4	1	1	14	3	5	7	41	1	270	2	2
250	250.01	250.02	250.03	250.1	250.11	250.12	250.13	250.2	250.22	250.23	250.3	250.4	250.41	250.42	250.43	250.5	250.51	250.52	250.53	250.6	250.7	250.8	250.81	250.82	250.83
6123	574	639	79	9	11	9	8	6	8	1	1	272	116	77	42	62	68	33	13	561	83	124	25	40	13
250.9	250.91	250.92	250.93	251	252	253	255	256	258	259	260	261	262	263	265	266	270	271	272	273	274	275	276	277	278
8	3	40	10	2	12	16	28	3	1	3	1	12	5	132	1	4	3	1	1013	3	36	39	2424	22	362
279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	303	304	305

```
> # Category Reduction: Based on ICD 9 codes for diabetic diagnosis
> dat2$diag_3= as.factor(ifelse((dat2$diag_3 >=390 & dat2$diag_3 <= 459) | dat2$diag_3 == 785, 'Circulatory',
+                               ifelse((dat2$diag_3 >=460 & dat2$diag_3 <= 519) | dat2$diag_3 == 786, 'Respiratory',
+                                      ifelse((dat2$diag_3 >= 520 & dat2$diag_3 <= 579) | dat2$diag_3 == 787, 'Digestive',
+                                             ifelse(dat2$diag_3 >= 250 & dat2$diag_3 < 251, 'Diabetes',
+                                                    ifelse(dat2$diag_3 >= 800 & dat2$diag_3 <=999, 'Injury',
+                                                           ifelse(dat2$diag_3 >=710 & dat2$diag_3 <= 739, 'Musculoskeletal',
+                                                                  ifelse((dat2$diag_3 >= 580 & dat2$diag_3 <= 629) | dat2$diag_3 == 788, 'Genitourinary',
+                                                                         ifelse((dat2$diag_3 >= 140 & dat2$diag_3 <= 239) | dat2$diag_3 == 780 | dat2$diag_3 == 781 | dat2$diag_3 == 784 |
+                                                                                  (dat2$diag_3 >= 790 & dat2$diag_3 <= 799) | (dat2$diag_3 >= 240 & dat2$diag_3 <= 279 & dat2$diag_3 != 250) |
+                                                                                  (dat2$diag_3 >= 680 & dat2$diag_3 <= 709 | dat2$diag_3 == 782) | (dat2$diag_3 >= 001 & dat2$diag_3 <= 139),'N
eoplasms', 'Other'))))))))))
> table(dat2$diag_3)

   Circulatory       Diabetes      Digestive   Genitourinary         Injury Musculoskeletal      Neoplasms          Other     Respiratory
         15588           9058           1910           2912            936           1038           8131           6670           3491
>
```

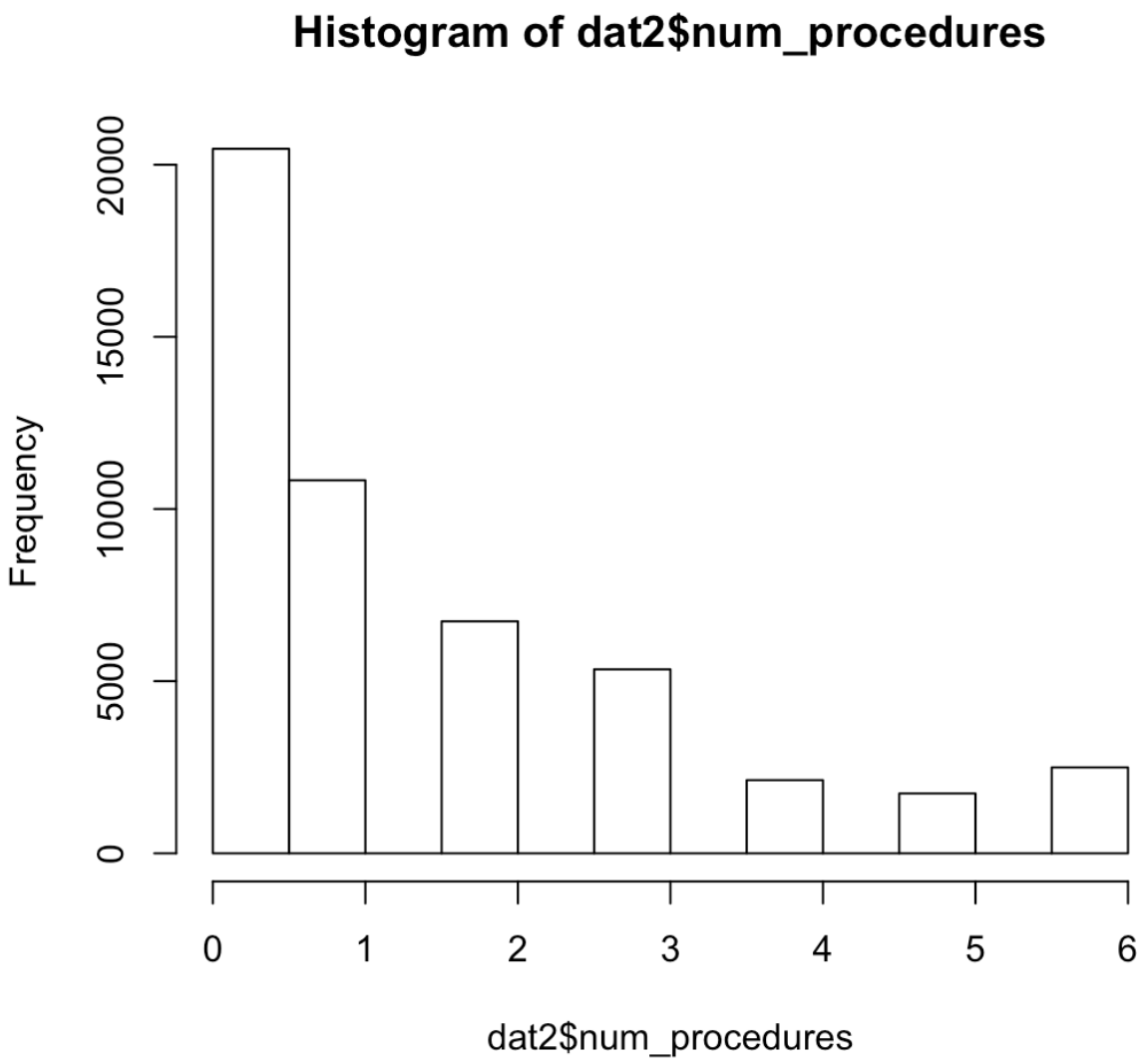
```
> #Category Reduction: Based on the distribution of values of feature time_in_hospital
> hist(dat2$time_in_hospital)
```



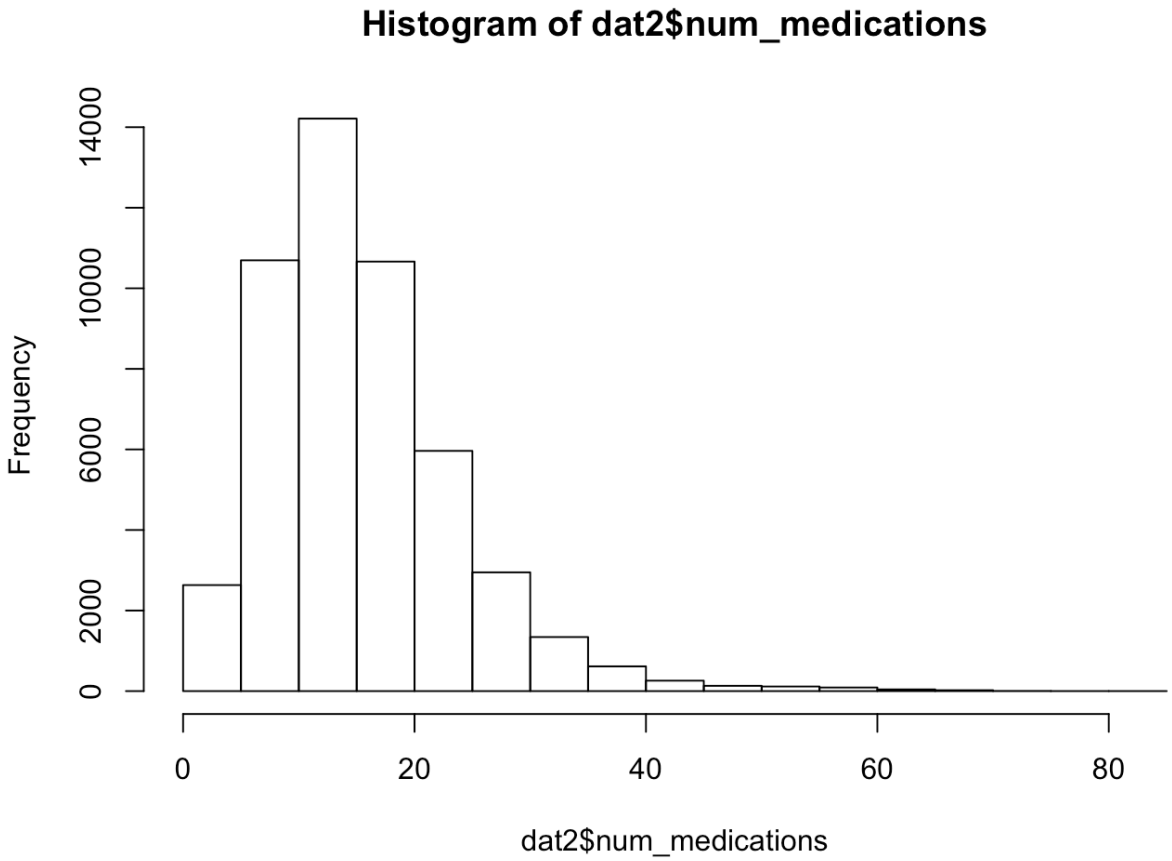

```
> #Category Reduction: Based on the distribution of values of feature time_in_hospital
> hist(dat2$time_in_hospital)
> dat2$time_in_hospital=(ifelse(dat2$time_in_hospital>= 8,'GreaterThan8',
+                               ifelse(dat2$time_in_hospital>=3 & dat2$time_in_hospital < 8, 'between4to8Days
',
+                                      ifelse(dat2$time_in_hospital>0 & dat2$time_in_hospital<=2, 'lessthan2d
ays',dat2$time_in_hospital))))
> table(dat2$time_in_hospital)

between4to8Days    GreaterThan8   lessthan2days
          26593            7709           15432
> #Category Reduction: Based on the distribution of values of feature num_lab_procedures
> hist(dat2$num_lab_procedures)
>
```



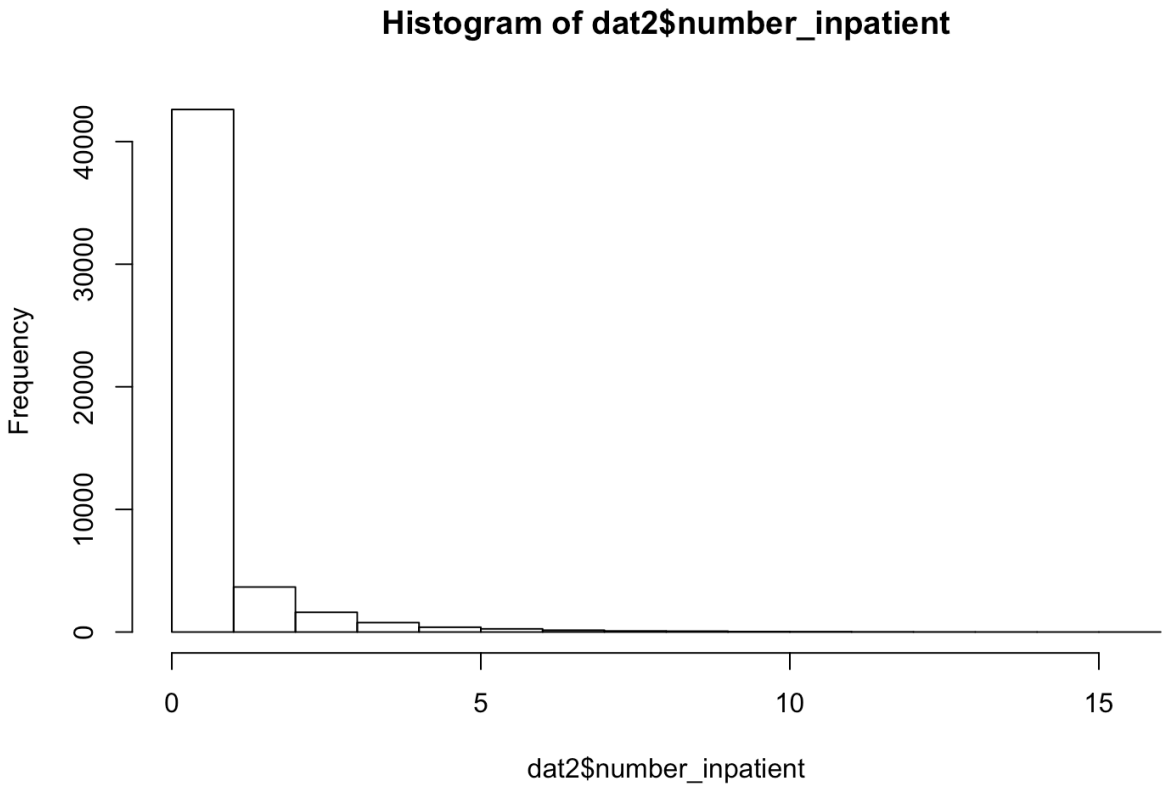

```
> dat2$num_lab_procedures=(ifelse(dat2$num_lab_procedures>= 70,'Long',
+                                 ifelse(dat2$num_lab_procedures>=30 & dat2$num_lab_procedures < 70, 'Medium',
+                                        ifelse(dat2$num_lab_procedures>=0 & dat2$num_lab_procedures<30, 'less',dat2$n
um_lab_procedures))))
> table(dat2$num_lab_procedures)

  less   Long Medium
 11041   4087  34606
> #Category Reduction: Based on the distribution of values of feature num_procedures
> hist(dat2$num_procedures)
>
```



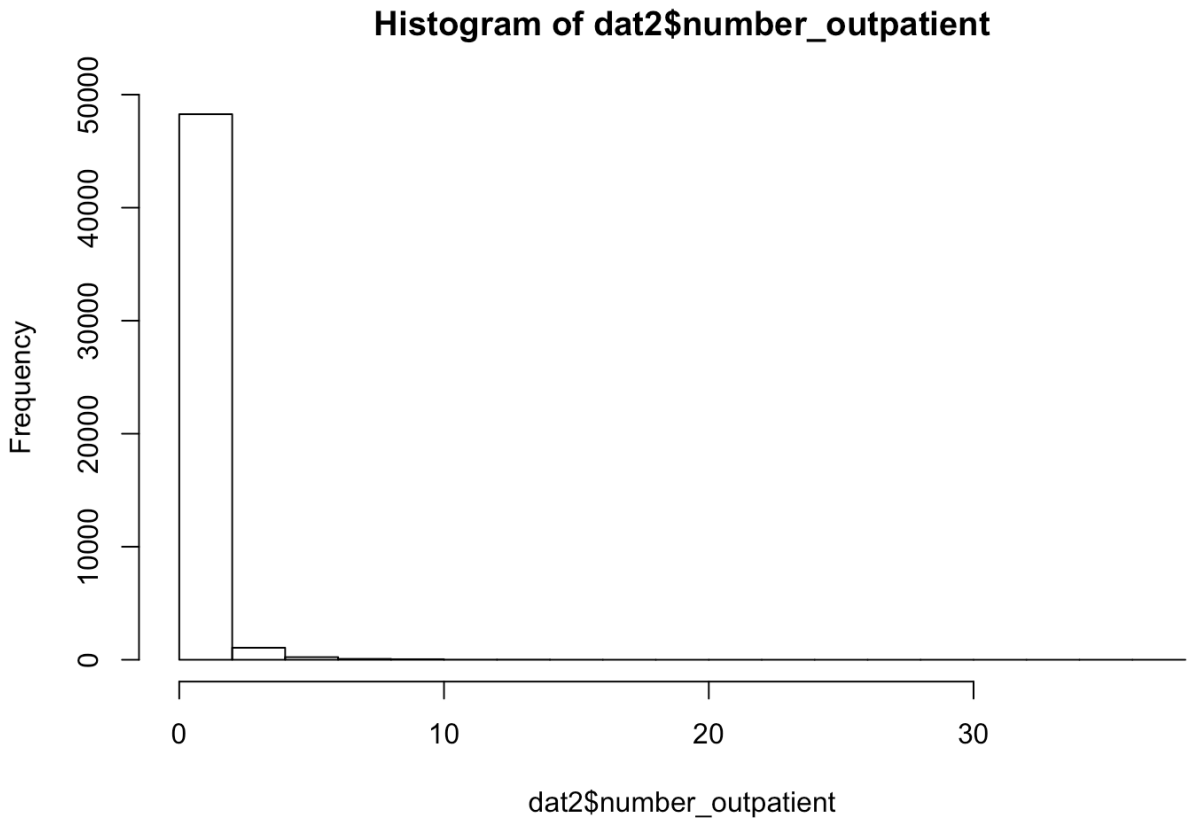

```
> dat2$num_procedures=(ifelse(dat2$num_procedures>= 4,'Long',
+                             ifelse(dat2$num_procedures>=2 & dat2$num_procedures < 4, 'Medium',
+                                    ifelse(dat2$num_procedures>=0 & dat2$num_procedures<2, 'less',dat2$num_procedures))))
> table(dat2$num_procedures)

  less   Long Medium
 31296   6357  12081
> #Category Reduction: Based on the distribution of values of feature num_medications
> hist(dat2$num_medications)
```



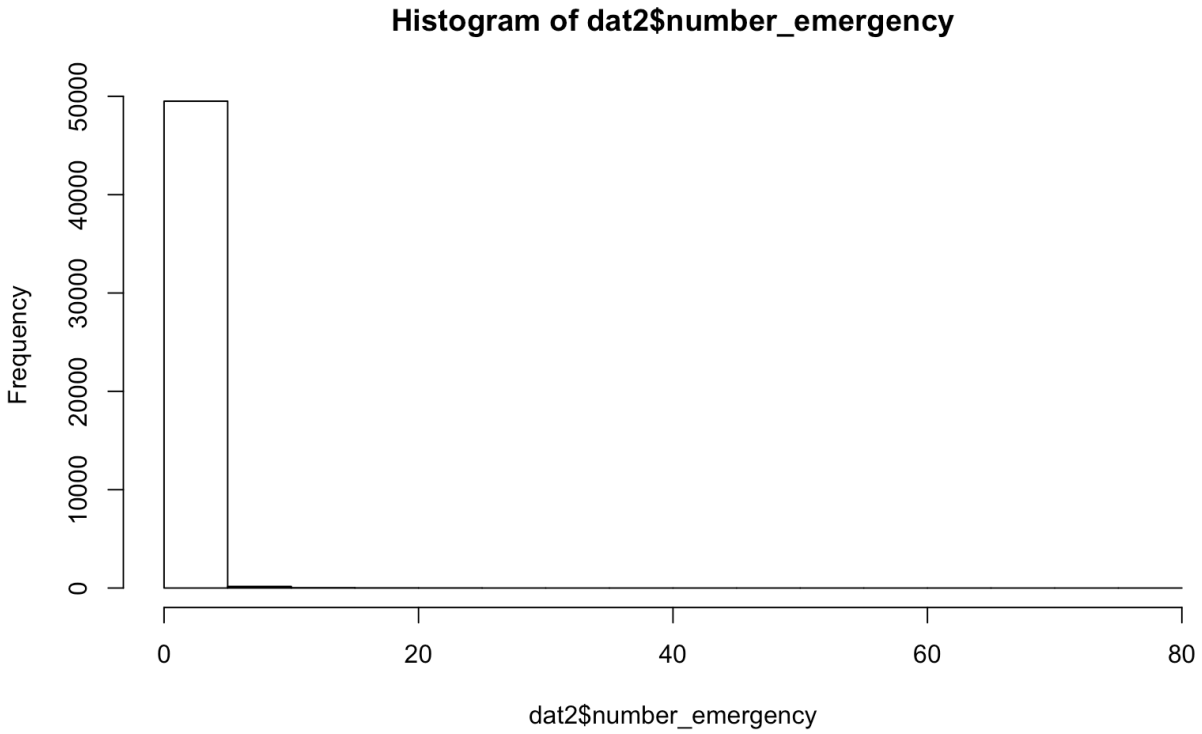

```
> dat2$num_procedures=(ifelse(dat2$num_procedures>= 4,'Long',
+                             ifelse(dat2$num_procedures>=2 & dat2$num_procedures < 4, 'Medium',
+                                    ifelse(dat2$num_procedures>=0 & dat2$num_procedures<2, 'less',dat2$num_procedures))))
> table(dat2$num_procedures)

  less   Long Medium
 31296   6357  12081
> #Category Reduction: Based on the distribution of values of feature num_medications
> hist(dat2$num_medications)
> dat2$num_medications=(ifelse(dat2$num_medications>= 40,'High',
+                             ifelse(dat2$num_medications>20 & dat2$num_medications < 40, 'Medium',
+                                    ifelse(dat2$num_medications>=0 & dat2$num_medications<=20, 'less',dat2$num_medications))))
> table(dat2$num_medications)

  High   less Medium
   763  38203  10768
> #Category Reduction: Based on the distribution of values of feature number_inpatient
> hist(dat2$number_inpatient)
```



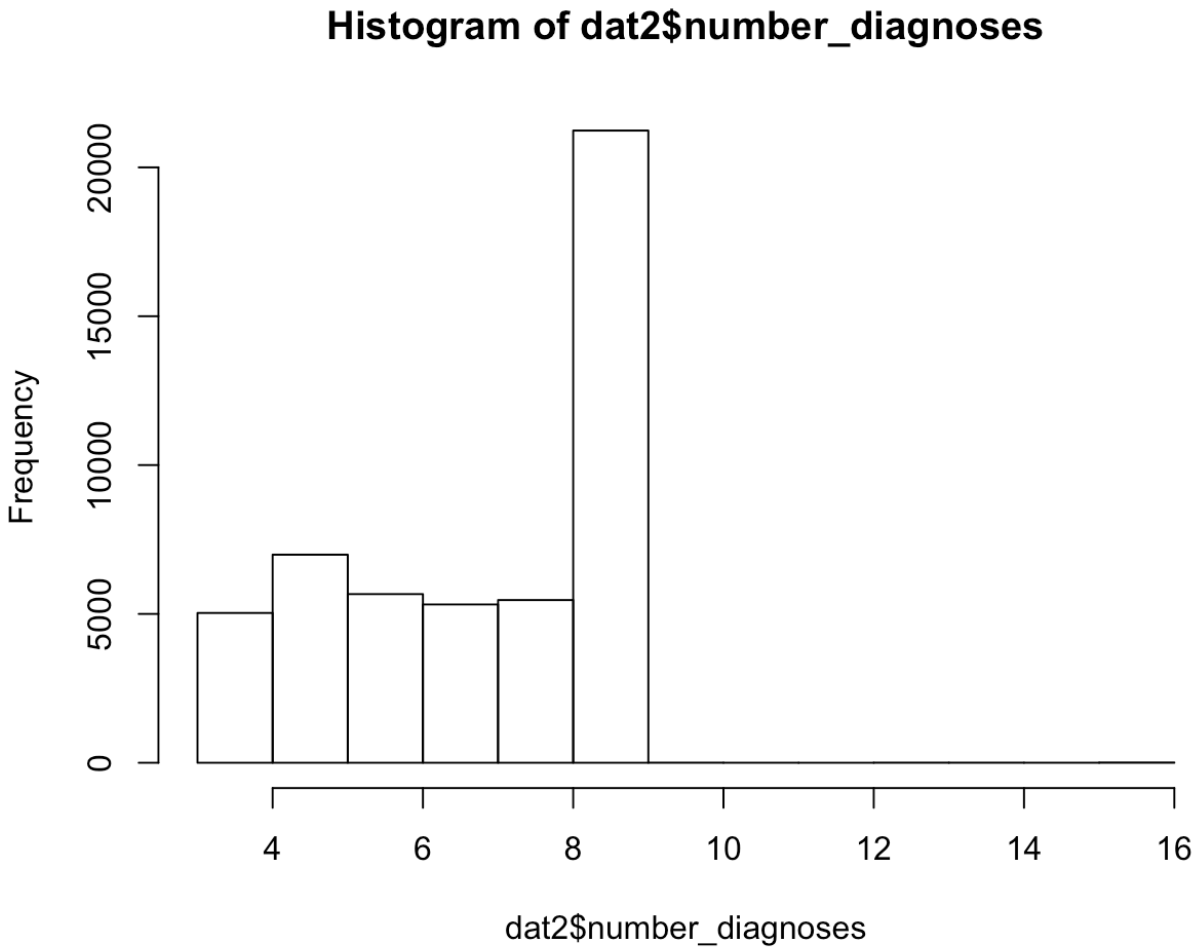

```
> dat2$number_inpatient=(ifelse(dat2$number_inpatient>4,'High',
+                               ifelse(dat2$number_inpatient>=2 & dat2$number_inpatient <= 4, 'Medium',
+                                      ifelse(dat2$number_inpatient>=1 & dat2$number_inpatient<=2, 'less',
+                                          'None'))))
> table(dat2$number_inpatient)

  High   less Medium   None
  1066   9678   6050  32940
> #Category Reduction: Based on the distribution of values of feature number_outpatient
> hist(dat2$number_outpatient)
```




```
> dat2$number_outpatient=(ifelse(dat2$number_outpatient>= 4,'High',
+                                ifelse(dat2$number_outpatient>=3 & dat2$number_outpatient <= 4, 'Medium',
+                                       ifelse(dat2$number_outpatient>=1 & dat2$number_outpatient<=2, 'less',
+                                              ifelse(dat2$number_outpatient==0, 'veryLow',dat2$number_outpatient)))))
> table(dat2$number_outpatient)

   High    less  Medium veryLow
    767    4905     695   43367
> #Category Reduction: Based on the distribution of values of feature number_emergency
> hist(dat2$number_emergency)
```



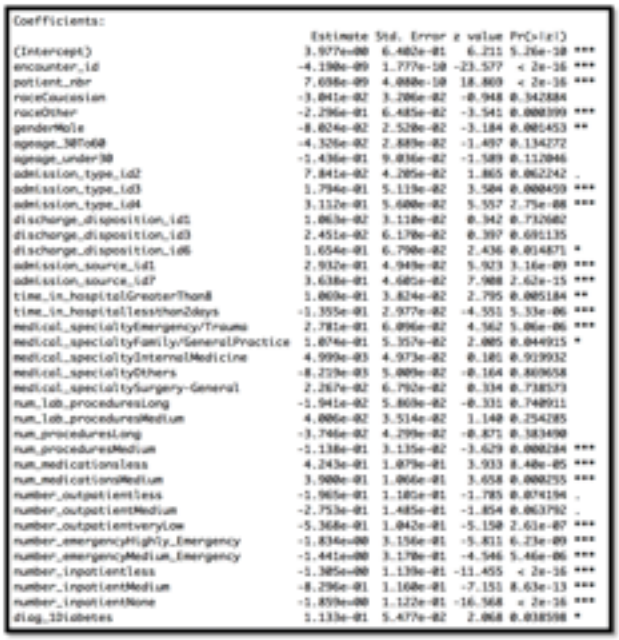

```
> dat2$number_emergency=(ifelse(dat2$number_emergency== 0,'Highly_Emergency',
+                               ifelse(dat2$number_emergency>=1 & dat2$number_emergency <= 5, 'Medium_Emergency','Emergency')))
> table(dat2$number_emergency)

       Emergency Highly_Emergency Medium_Emergency
             228            44214             5292
> #Category Reduction: Based on the distribution of values of feature number_diagnoses
> hist(dat2$number_diagnoses)
>
```

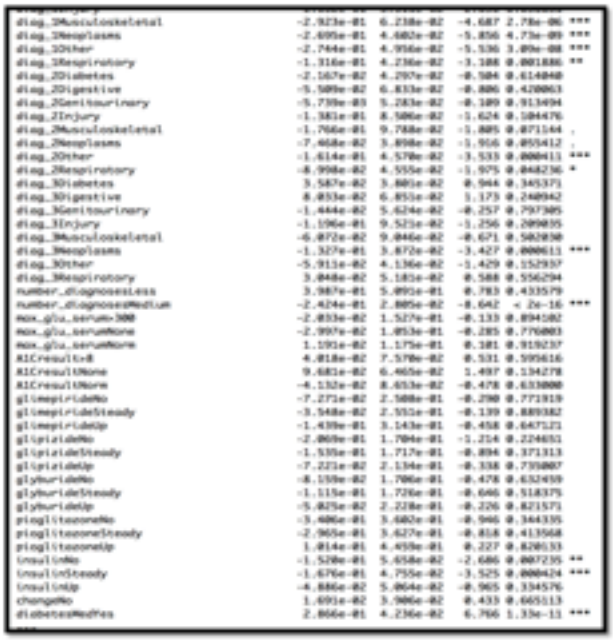



```
> dat2$number_diagnoses=ifelse(dat2$number_diagnoses>=10 & dat2$number_diagnoses <= 16 ,'Less',
+                              ifelse(dat2$number_diagnoses>=3 & dat2$number_diagnoses <= 8, 'Medium',
+                                     ifelse(dat2$number_diagnoses==9, 'High', dat2$number_diagnoses)))
> table(dat2$number_diagnoses)

  High   Less Medium
 21237     25  28472
> #Category reduction: as per the research patient admitted before 30 days is called the returning patient
> dat2$readmitted=(ifelse(dat2$readmitted<=30, 1,0))
> table(dat2$readmitted)

    0     1
27417 22317
```

```
> #Data Partition: 60% for Training and rest for validation
> set.seed(1)
> id.train = sample(1:nrow(dat2), nrow(dat2)*.6)
> id.test=setdiff(1:nrow(dat2), id.train)
> dat2.train = dat2[id.train,]
> dat2.test = dat2[id.test,]
> #Trained minimum model with only one variable called intercept
> min.model = glm(readmitted ~ 1, data = dat2.train, family = 'binomial')
> #measuring % of dependent variable is explained by variation in independent variable
> library(pscl)
> pR2(min.model)# r2 of the model
      llh    llhNull         G2   McFadden       r2ML       r2CU
-20515.67 -20515.67       0.00       0.00       0.00       0.00
>
```

```
> summary(obj) # it will give you the final model

Call:
glm(formula = readmitted ~ number_inpatient + patient_nbr + encounter_id +
    number_emergency + diabetesMed + number_diagnoses + diag_1 +
    number_outpatient + medical_specialty + admission_source_id +
    admission_type_id + time_in_hospital + num_procedures + race +
    gender + num_medications + insulin + diag_3 + diag_2 + A1Cresult +
    age, family = "binomial", data = dat2.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7515  -0.9984  -0.7330   1.1447   2.2293

Coefficients:
```





```
    Null deviance: 41031  on 29839  degrees of freedom
Residual deviance: 37634  on 29754  degrees of freedom
AIC: 37806

Number of Fisher Scoring iterations: 4
```

```
> require(MASS)
> exp(cbind((coef(obj))) )
```

encounter_id	1.0000000
patient_nbr	1.0000000
raceCaucasian	0.9700519
raceOther	0.7948569
genderMale	0.9228958
ageage_30To60	0.9576616
ageage_under30	0.8662461
admission_type_id2	1.0815648
admission_type_id3	1.1964542
admission_type_id4	1.3650353
discharge_disposition_id1	1.0106841
discharge_disposition_id3	1.0248169
discharge_disposition_id6	1.1798397
admission_source_id1	1.3406572
admission_source_id7	1.4388244
time_in_hospitalGreaterThan8	1.1128145
time_in_hospitallessthan2days	0.8732864
medical_specialtyEmergency/Trauma	1.3206228
medical_specialtyFamily/GeneralPractice	1.1134193
medical_specialtyInternalMedicine	1.0050117
medical_specialtyOthers	0.9918147
medical_specialtySurgery-General	1.0229252
num_lab_proceduresLong	0.9807807
num_lab_proceduresMedium	1.0408729
num_proceduresLong	0.9632292
num_proceduresMedium	0.8924684
num_medicationsless	1.5285002
num_medicationsMedium	1.4769782
number_outpatientless	0.8215876
number_outpatientMedium	0.7593065
number_outpatientveryLow	0.5846068
number_emergencyHighly_Emergency	0.1597776
number_emergencyMedium_Emergency	0.2366607
number_inpatientless	0.2712408
number_inpatientMedium	0.4362228

number_inpatientNone	0.1558417
diag_1Diabetes	1.1199566
diag_1Digestive	0.9101868
diag_1Genitourinary	0.7935029
diag_1Injury	0.8912172
diag_1Musculoskeletal	0.7465244
diag_1Neoplasms	0.7637530
diag_1Other	0.7600559
diag_1Respiratory	0.8766496
diag_2Diabetes	0.9785631
diag_2Digestive	0.9463971
diag_2Genitourinary	0.9942774
diag_2Injury	0.8710145
diag_2Musculoskeletal	0.8380877
diag_2Neoplasms	0.9280431
diag_2Other	0.8509250
diag_2Respiratory	0.9139527
diag_3Diabetes	1.0365213
diag_3Digestive	1.0836493
diag_3Genitourinary	0.9856595
diag_3Injury	0.8872701
diag_3Musculoskeletal	0.9410826
diag_3Neoplasms	0.8757358
diag_3Other	0.9426066
diag_3Respiratory	1.0309514
number_diagnosesLess	1.4898537
number_diagnosesMedium	0.7847762
max_glu_serum>300	0.9798738
max_glu_serumNone	0.9704772
max_glu_serumNorm	1.0119829
A1Cresult>8	1.0409932
A1CresultNone	1.1016514
A1CresultNorm	0.9595217
glimepirideNo	0.9298751
glimepirideSteady	0.9651449
glimepirideUp	0.8659918
glipizideNo	0.8130660

glyburideSteady	0.8945234
glyburideUp	0.9509870
pioglitazoneNo	0.7113319
pioglitazoneSteady	0.7433795
pioglitazoneUp	1.1067067
insulinNo	0.8590298
insulinSteady	0.8456958
insulinUp	0.9523100
changeNo	1.0170495
diabetesMedYes	1.3318968

```
>
> # to check the fit of the model McFaddenR2 is used to
> #check the fit of the model, it is same as linear regression R2
> library(pscl)
> pR2(obj)
         llh       llhNull            G2      McFadden          r2ML          r2CU
-1.882734e+04 -2.051567e+04  3.376650e+03  8.229442e-02  1.069909e-01  1.431945e-01
>
> #Auc & ROC
> library(ROCR)
> p <- predict(obj, newdata=subset(dat2.test,select=setdiff(colnames(dat2.test),c('readmitted')), type="response"))
> pr <- prediction(p, dat2.test$readmitted)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
>
> table(dat2.test$readmitted, p>0.5)

    FALSE  TRUE
  0 10017   899
  1  6824  2154
> ER<-(6842+892)/(6842+892+10024+2136)
> 1-ER
[1] 0.6112396
> plot(prf, col='red')
> #As a rule of thumb, a model with good predictive ability
> #should have an AUC closer to 1 (1 is ideal) than to 0.5
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.6836654
```