Final Project

CUNY SPS Masters in Data Science - DATA 698

Ali Harb
Dilip Ganesan
Raghunathan Ramnath

May 10, 2019

Contents

1 Load Data

```
#Load data set
dset <- read.csv("C:/cuny/2019/698/dataset_diabetes/dataset_diabetes/diabetic_data.csv")

head(dset)
```

encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_
2278392	8222157	Caucasian	Female	[0-10)	?	6	
149190	55629189	Caucasian	Female	[10-20)	?	1	
64410	86047875	AfricanAmerican	Female	[20-30)	?	1	
500364	82442376	Caucasian	Male	[30-40)	?	1	
16680	42519267	Caucasian	Male	[40-50)	?	1	
35754	82637451	Caucasian	Male	[50-60)	?	2	

2 Summary Statistics

The current data set is composed of 101766 records and 50 variables.

```
##   encounter_id        patient_nbr                   race
##  Min.   :    12522   Min.   :      135   ?              : 2273
##  1st Qu.: 84961194   1st Qu.: 23413221   AfricanAmerican:19210
##  Median :152388987   Median : 45505143   Asian          :  641
##  Mean   :165201646   Mean   : 54330401   Caucasian      :76099
##  3rd Qu.:230270888   3rd Qu.: 87545950   Hispanic       : 2037
##  Max.   :443867222   Max.   :189502619   Other          : 1506
##
##              gender           age              weight
##  Female          :54708   [70-80):26068   ?        :98569
##  Male            :47055   [60-70):22483   [75-100) : 1336
##  Unknown/Invalid:    3   [50-60):17256   [50-75)  :  897
##                           [80-90):17197   [100-125):  625
##                           [40-50): 9685   [125-150):  145
##                           [30-40): 3775   [25-50)  :   97
##                           (Other): 5302   (Other)  :   97
##  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   ?      :40256   ?                     :49949
##  1st Qu.: 2.000   MC     :32439   InternalMedicine      :14635
##  Median : 4.000   HM     : 6274   Emergency/Trauma      : 7565
##  Mean   : 4.396   SP     : 5007   Family/GeneralPractice: 7440
##  3rd Qu.: 6.000   BC     : 4655   Cardiology            : 5352
##  Max.   :14.000   MD     : 3532   Surgery-General       : 3099
##                   (Other): 9603   (Other)               :13726
##  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   427    : 2766   496    : 3305
##                                      (Other):74419   (Other):70204
##      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
##  414    : 3664   Max.   :16.000
##  (Other):64551
##   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
##
##
##
##
```

3 Data Analysis

encounter_id

patient_nbr

Race: 75% of the population are Caucasian. Remaining 25% includes other 4 races that include others and unknown.

Gender: Almost 54% of the population is female and 46% of the population is male.

Age: [70-80] has the highest patient population.

admission_type_id: Almost 54% of the patient population were admitted in "Emergency"

discharge_disposition_id:60% of the patient population where discharged to home

admission_source_id: "Emergency Room" was the source for 57% of patient population

payer_code:40% of the patient population does not have payer_code.

medical_specialty:50% of the population does not have the specialty.

num_lab_procedures:varies..

num_procedures:47% patient population did not have any procedures. 21% had 1 procedure and 13% had 2 procedures and the remaining had between 3 and 6.

num_medications: varies..

number_outpatient:85% its 0

number_emergency:90% of time its not emergency

number_inpatient:67% its 0.

number_diagnoses:50% of the population have 9 diagnoses

max_glue_serum:96% of the patient population does not have.

A1Cresult:85% of the patient population did not have A1C.

metformin:81% of the patient population did not have metformin.

insulin:varies

change:47% its ch.

diabetesMed:78% its yes.

readmitted:46% admitted.

diag_1

diag_2

diag_3

3.1 Following can be ignored.

weight: this can be ignored as for 98% of the patient population its unknown.

repaglinide:Mostly its no.

naateglinide:Mostly its no.

chlorpropamide:Mostly its no.

glimepiride:Mostly its no.

acetohexamide:No except one.

glipizide:No except one.

glyburide:91% its no.

miglitol:91% its no.

troglitazone:Mostly its no.

tolazamide:Mostly its no.

examide:Mostly its no.

citoglipton:Mostly its no.

glyburide.metformin:Mostly its no.

glipizide.metformin:Mostly its no.

glimepiride.pioglitazone:Mostly its no.

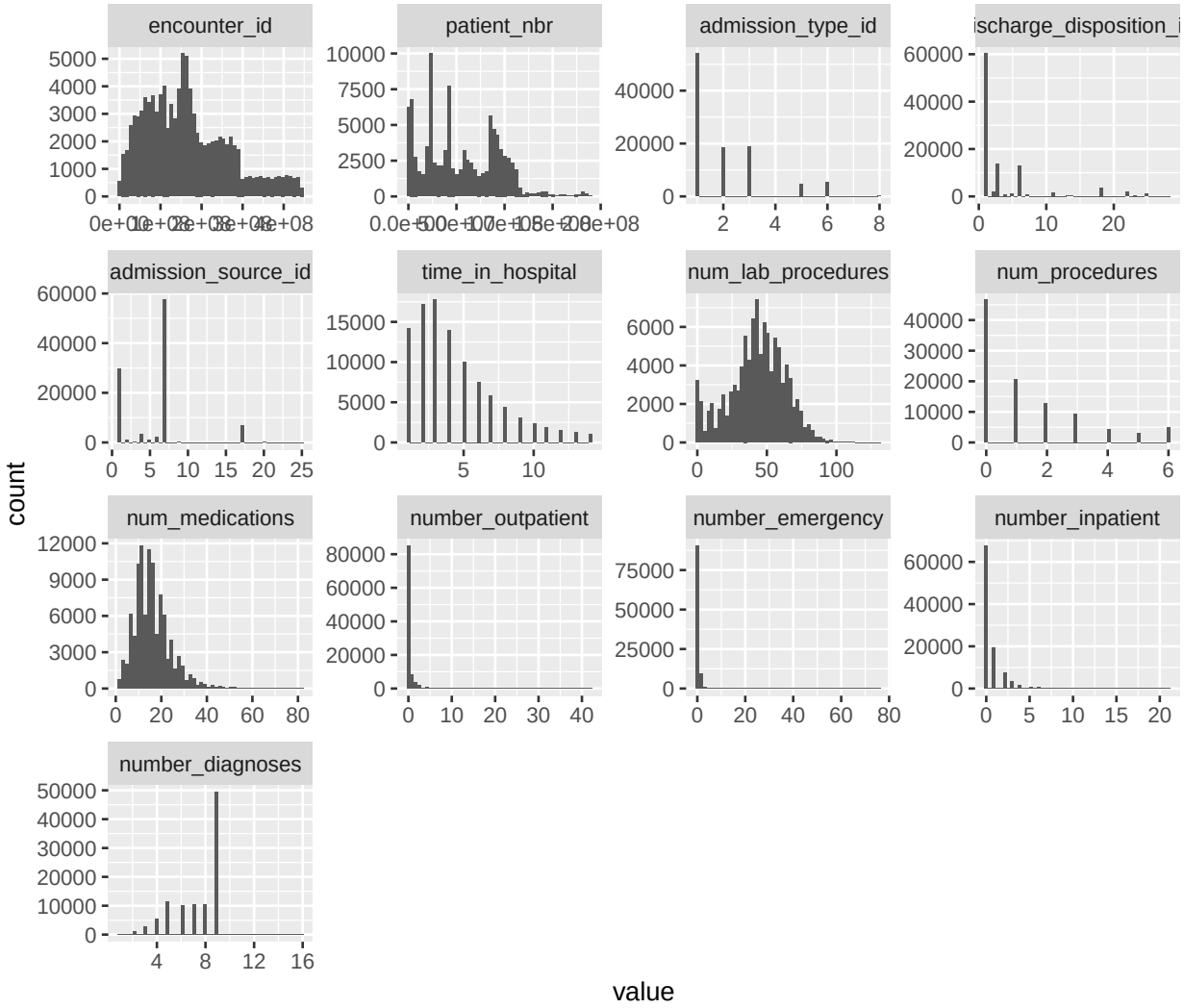
```
## 'data.frame':    101766 obs. of  50 variables:
##  $ encounter_id            : int  2278392 149190 64410 500364 16680 35754 55842 63768 12522 1
##  $ patient_nbr             : int  8222157 55629189 86047875 82442376 42519267 82637451 842598
##  $ race                    : Factor w/ 6 levels "?","AfricanAmerican",..: 4 4 2 4 4 4 4 4 4 4
##  $ gender                  : Factor w/ 3 levels "Female","Male",..: 1 1 1 2 2 2 2 2 1 1 ...
##  $ age                     : Factor w/ 10 levels "[0-10)","[10-20)",..: 1 2 3 4 5 6 7 8 9 10
##  $ weight                  : Factor w/ 10 levels "?","[0-25)","[100-125)",..: 1 1 1 1 1 1 1 1
##  $ admission_type_id       : int  6 1 1 1 1 2 3 1 2 3 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ payer_code              : Factor w/ 18 levels "?","BC","CH",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ medical_specialty       : Factor w/ 73 levels "?","AllergyandImmunology",..: 39 1 1 1 1 1
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ number_emergency        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ number_inpatient        : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ diag_1                  : Factor w/ 717 levels "?","10","11",..: 126 145 456 556 56 265 26
##  $ diag_2                  : Factor w/ 749 levels "?","11","110",..: 1 81 80 99 26 248 248 31
##  $ diag_3                  : Factor w/ 790 levels "?","11","110",..: 1 123 768 250 88 88 772
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
##  $ max_glu_serum           : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ A1Cresult               : Factor w/ 4 levels ">7",">8","None",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ metformin               : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 3 2 2 2
##  $ repaglinide             : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ nateglinide             : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ chlorpropamide          : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ glimepiride             : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 3 2 2 2
##  $ acetohexamide           : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ glipizide               : Factor w/ 4 levels "Down","No","Steady",..: 2 2 3 2 3 2 2 2 3 2
##  $ glyburide               : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 3 2 2
##  $ tolbutamide             : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pioglitazone            : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ rosiglitazone           : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 3
##  $ acarbose                : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ miglitol                : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ troglitazone            : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tolazamide              : Factor w/ 3 levels "No","Steady",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ examide                 : Factor w/ 1 level "No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ citoglipton             : Factor w/ 1 level "No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ insulin                 : Factor w/ 4 levels "Down","No","Steady",..: 2 4 2 4 3 3 3 2 3 3
##  $ glyburide.metformin     : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2
##  $ glipizide.metformin     : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ glimepiride.pioglitazone: Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ metformin.rosiglitazone : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ metformin.pioglitazone  : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ change                  : Factor w/ 2 levels "Ch","No": 2 1 2 1 1 2 1 2 1 1 ...
##  $ diabetesMed             : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 2 2 2 2 2 ...
##  $ readmitted              : Factor w/ 3 levels "<30",">30","NO": 3 2 3 3 3 2 3 2 3 3 ...
```

4 Histogram

```
ggplot(melt(dset), aes(x=value)) + facet_wrap(~variable, scale="free") + geom_histogram(bins=50)
```

```
## Using race, gender, age, weight, payer_code, medical_specialty, diag_1, diag_2, diag_3, max_g
```

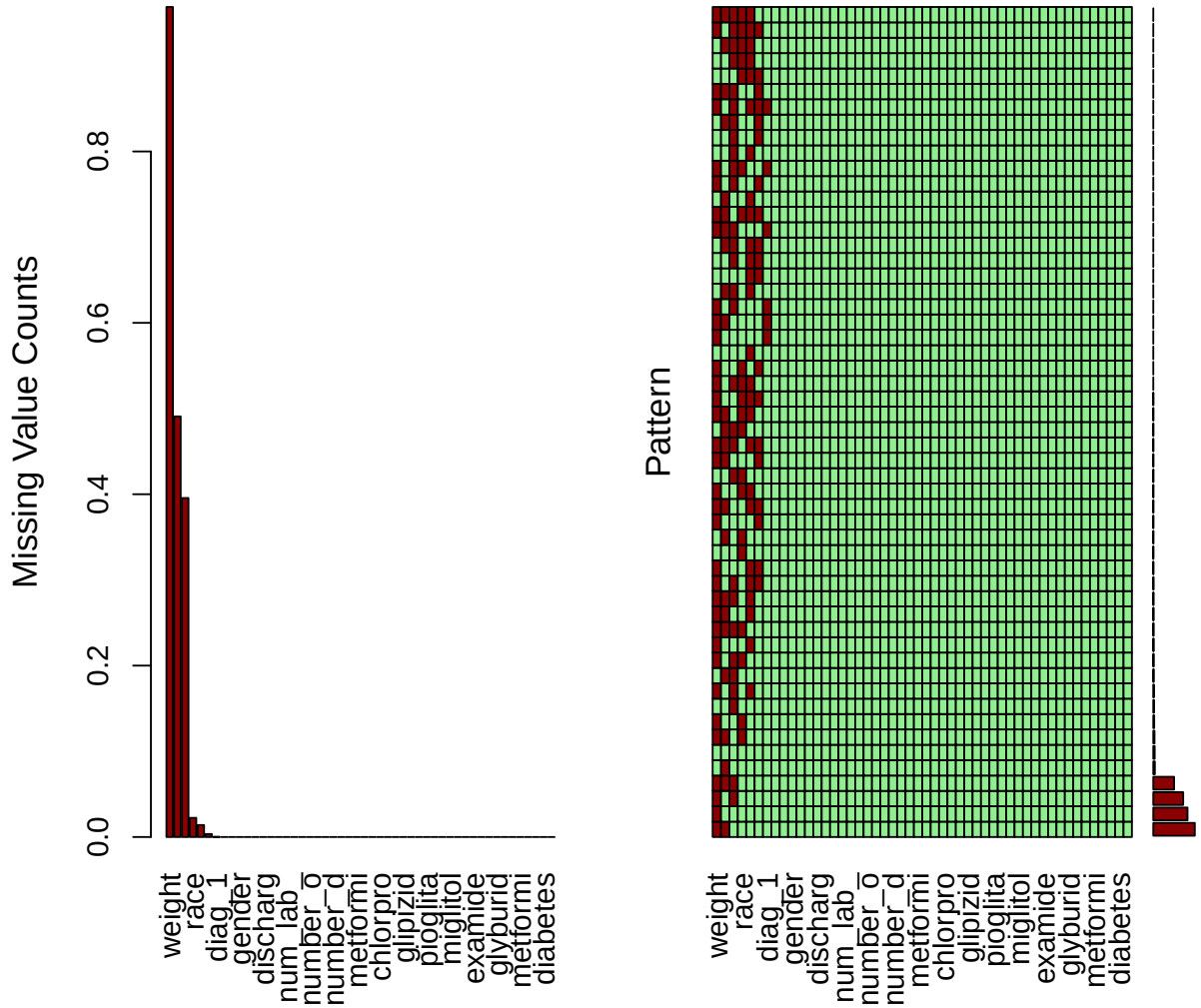


5 Missing Values:

```
#Replace ? with NA
dset[dset=="?"]<-NA
```

```
## Missing Values
options(scipen = 999)
missing_plot <- VIM::aggr(dset,
                          numbers = T,
                          sortVars = T,
                          col = c("lightgreen", "darkred", "orange"),
                          labels=str_sub(names(dset), 1, 8),
                          ylab=c("Missing Value Counts"
                                 , "Pattern"))
```

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```




```
##
##  Variables sorted by number of missings:
##  Variable        Count
##    weight 0.9685847926
```

```
##  medical_ 0.4908220820
##  payer_co 0.3955741603
##      race 0.0223355541
##    diag_3 0.0139830592
##    diag_2 0.0035178743
##    diag_1 0.0002063558
##  encounte 0.0000000000
##  patient_ 0.0000000000
##    gender 0.0000000000
##       age 0.0000000000
##  admissio 0.0000000000
##  discharg 0.0000000000
##  admissio 0.0000000000
##  time_in_ 0.0000000000
##  num_lab_ 0.0000000000
##  num_proc 0.0000000000
##  num_medi 0.0000000000
##  number_o 0.0000000000
##  number_e 0.0000000000
##  number_i 0.0000000000
##  number_d 0.0000000000
##  max_glu_ 0.0000000000
##  A1Cresul 0.0000000000
##  metformi 0.0000000000
##  repaglin 0.0000000000
##  nateglin 0.0000000000
##  chlorpro 0.0000000000
##  glimepir 0.0000000000
##  acetohex 0.0000000000
##  glipizid 0.0000000000
##  glyburid 0.0000000000
##  tolbutam 0.0000000000
##  pioglita 0.0000000000
##  rosiglit 0.0000000000
##  acarbose 0.0000000000
##  miglitol 0.0000000000
##  troglita 0.0000000000
##  tolazami 0.0000000000
##   examide 0.0000000000
##  citoglip 0.0000000000
##   insulin 0.0000000000
##  glyburid 0.0000000000
##  glipizid 0.0000000000
##  glimepir 0.0000000000
##  metformi 0.0000000000
##  metformi 0.0000000000
##    change 0.0000000000
##  diabetes 0.0000000000
```

Table 1: Variables Missing Values

	Variable	Count	pct_missing
1	weight	98569	0.969
2	medical_specialty	49949	0.491
3	payer_code	40256	0.396
4	race	2273	0.022
5	diag_3	1423	0.014
6	diag_2	358	0.004
7	diag_1	21	0.000

```
##  readmitt 0.0000000000
```

```
missing_plot$missings %>%
  mutate(
    pct_missing = Count / nrow(dset)
    ) %>%
  arrange(-pct_missing) %>%
  filter(pct_missing > 0) %>%
  kable(digits = 3, row.names = T, caption = "Variables Missing Values")
```

```
options(scipen=0, digits=7)
```

6 Drop Missing Values:

Weight is missing in over 98% records. Owing to the poor interpretability of missing values and little predictive generalizability to other patients, best thing is to just drop it.

Payer code and Medical Specialty of treating physician also have 40-50% missing values. We decided to drop these.

```
dset$weight<-NULL
dset$payer_code<-NULL
dset$medical_specialty<-NULL
dset$citoglipton<-NULL
dset$examide<-NULL

#deletes columns 'weight','payer_code','medical_specialty'

#We also noticed that for two variables (drugs named citoglipton and examide), all records have
```

7 Addition of variables:

Service utilization: The data contains variables for number of inpatient (admissions), emergency room visits and outpatient visits for a given patient in the previous one year. These are (crude)

measures of how much hospital/clinic services a person has used in the past year.

```
dset['service_utilization'] <- dset['number_outpatient'] +
   dset['number_emergency'] + dset['number_inpatient']
```

Number of medication changes: The dataset contains 23 features for 23 drugs (or combos) which indicate for each of these, whether a change in that medication was made or not during the current hospital stay of patient. Medication change for diabetics upon admission has been shown by previous research to be associated with lower readmission rates. We decided to count how many changes were made in total for each patient, and declared that a new feature. The reasoning here was to both simplify the model and possibly discover a relationship with number of changes regardless of which drug was changed.

```
keys = list ('metformin', 'repaglinide', 'nateglinide', 'chlorpropamide',
'glimepiride', 'glipizide', 'glyburide', 'pioglitazone', 'rosiglitazone',
 'acarbose', 'miglitol', 'insulin', 'glyburide.metformin', 'tolazamide',
'metformin.pioglitazone','metformin.rosiglitazone', 'glimepiride.pioglitazone',
'glipizide.metformin', 'troglitazone', 'tolbutamide', 'acetohexamide')


bv <- function(x) {
    if (x =='No'|x=='Steady'){s=0}
      else {s=1}
    return(s)
}

#dset$metformin <- dset$metformin.apply(lambda x: 0 if (x =='No'|x=='Steady') else 1)
#dset['metformin']<- sapply(dset['metformin'], bv)

for (i in keys)
  { dset[i]<- sapply(dset[i], bv)
}
```

```
## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
```

```
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used

## Warning in if (x == "No" | x == "Steady") {: the condition has length > 1
## and only the first element will be used
```

```
dset$numchange<- dset['metformin'] + dset['repaglinide'] + dset['nateglinide'] + dset['chlorprop

#for (i in keys)
#    { dset['numchange'] = dset['numchange'] + dset[i] }
```

8 Split Data:

```
    train<-sample_frac(dset, 0.7)
    sid<-as.numeric(rownames(train)) # because rownames() returns character
    test<-dset[-sid,]
```