Exercise-4

Matthew McAvoy

September 2, 2016

Exercise 4

To begin to answer if its possible to predict discharge status, we should look at the quality of the data. According to Pipino and colleagues (Pipino, Lee, Wang 2002), there are 16 dimensions of data we should be concerned about before we can trust any derived solutions based on the data.

Let us begin by loading the data into a usable form.

```
setwd("C:/Users/homur/OneDrive/New College/EDA/Week 2/dataset_diabetes/dataset_diabetes/")
DiabetesData <- data.frame(read.csv(file="diabetic_data.csv",head=TRUE,sep=","))
dd <- DiabetesData
```

Since we were able to load the data into a statistical package, we can say a few things about it. First, it is accessible since we can access and retrieve it quickly. And it's manipulatable since R gives us that power.

We can't say much about objectivity, reputation, security, or timeliness, but can say a few things about its appropriateness, completeness, free-of-error, and maybe even it's value-added to predicting discharge status.

To determine completeness, we can measure column integrity.

The following counts the number of NA entries in each column

```
apply(dd, 2, function(z) sum(is.na(z)))
```

```
##             encounter_id              patient_nbr                     race
##                        0                        0                        0
##                   gender                      age                   weight
##                        0                        0                        0
##        admission_type_id discharge_disposition_id      admission_source_id
##                        0                        0                        0
##         time_in_hospital               payer_code        medical_specialty
##                        0                        0                        0
##       num_lab_procedures           num_procedures          num_medications
##                        0                        0                        0
##        number_outpatient         number_emergency         number_inpatient
##                        0                        0                        0
##                   diag_1                   diag_2                   diag_3
##                        0                        0                        0
##         number_diagnoses            max_glu_serum                A1Cresult
##                        0                        0                        0
##                metformin              repaglinide              nateglinide
##                        0                        0                        0
##           chlorpropamide              glimepiride            acetohexamide
##                        0                        0                        0
##                glipizide                glyburide              tolbutamide
##                        0                        0                        0
##             pioglitazone            rosiglitazone                 acarbose
##                        0                        0                        0
##                 miglitol             troglitazone               tolazamide
##                        0                        0                        0
##                  examide              citoglipton                  insulin
##                        0                        0                        0
##      glyburide.metformin      glipizide.metformin glimepiride.pioglitazone
##                        0                        0                        0
##  metformin.rosiglitazone   metformin.pioglitazone                   change
##                        0                        0                        0
##              diabetesMed               readmitted
##                        0                        0
```

It looks like this is a very clean file without any empty entries.

To determine how much error might be in it, we can calculate the variance of each item. We might be suspicious of erroneous data if the variance is very large. We can ignore the NA cases since these are categorical and not numerical.

```
apply(dd, 2, function(z) sum(var(z)))
```

```
## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion
## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion

## Warning in var(z): NAs introduced by coercion
```

```
##             encounter_id              patient_nbr                     race
##             1.053503e+16             1.497408e+15                       NA
##                   gender                      age                   weight
##                       NA                       NA                       NA
##        admission_type_id discharge_disposition_id      admission_source_id
##             2.089189e+00             2.788015e+01             1.651675e+01
##         time_in_hospital               payer_code        medical_specialty
##             8.910868e+00                       NA                       NA
##       num_lab_procedures           num_procedures          num_medications
##             3.870805e+02             2.909777e+00             6.605733e+01
##        number_outpatient         number_emergency         number_inpatient
##             1.605961e+00             8.657786e-01             1.594824e+00
##                   diag_1                   diag_2                   diag_3
##                       NA                       NA                       NA
##         number_diagnoses            max_glu_serum                A1Cresult
##             3.738810e+00                       NA                       NA
##                metformin              repaglinide              nateglinide
##                       NA                       NA                       NA
##           chlorpropamide              glimepiride            acetohexamide
##                       NA                       NA                       NA
##                glipizide                glyburide              tolbutamide
##                       NA                       NA                       NA
##             pioglitazone            rosiglitazone                 acarbose
##                       NA                       NA                       NA
##                 miglitol             troglitazone               tolazamide
##                       NA                       NA                       NA
##                  examide              citoglipton                  insulin
##                       NA                       NA                       NA
##      glyburide.metformin      glipizide.metformin glimepiride.pioglitazone
##                       NA                       NA                       NA
##  metformin.rosiglitazone   metformin.pioglitazone                   change
##                       NA                       NA                       NA
##              diabetesMed               readmitted
##                       NA                       NA
```

It looks like most acceptible columns have a small variance with the exception of num_lab_procedures which varies by quite a bit. This makes sense as the min and max are [1.0, 132.0]

```
summary(dd$num_lab_procedures)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     1.0    31.0    44.0    43.1    57.0   132.0
```

So far, it looks like the data is quite good and should be usuable for predicting discharge status.