Predicting readmission probability for diabetes inpatients

STAT 471/571/701, Fall 2017

Due: April 2, 2017 at 11:59PM

Instructions

- This project is due at **11:59pm on Sunday, April. 2, 2017.**
- It is an individual project, amounting to 15% of your final grade. See the *Collaboration* section at the bottom of this document.
- There is no single correct answer. You will be graded on the general quality of your work.
- The entire write up should not be more than 10 pages. You may put any supporting documents, code, graphics, or other exhibits into an Appendix, which is not counted in the 10 page limit.

Introduction

Background

Diabetes is a chronic medical condition affecting millions of Americans, but if managed well, with good diet, exercise and medication, patients can lead relatively normal lives. However, if improperly managed, diabetes can lead to patients being continuously admitted and readmitted to hospitals. Readmissions are especially serious - they represent a failure of the health system to provide adequate support to the patient and are extremely costly to the system. As a result, the Centers for Medicare and Medicaid Services announced in 2012 that they would no longer reimburse hospitals for services rendered if a patient was readmitted with complications within 30 days of discharge.

Given these policy changes, being able to identify and predict those patients most at risk for costly readmissions has become a pressing priority for hospital administrators.

In this project, we shall explore how to use the techniques we have learned in order to help better manage diabetes patients who have been admitted to a hospital. Our goal is to avoid patients being readmitted within 30 days of discharge, which reduces costs for the hospital and improves outcomes for patients.

The original data is from the Center for Clinical and Translational Research at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are over 100,000 unique hospital admissions in this dataset, from ~70,000 unique patients. The data includes demographic elements, such as age, gender, and race, as well as clinical attributes such as tests conducted, emergency/inpatient visits, etc. Refer to the original documentation for more details on the dataset. Three former students Spencer Luster, Matthew Lesser and Mridul Ganesh, brought this data set into the class and did a wonderful final project. We will use a subset processed by the group but with a somewhat different objective.

Goals of the analysis

1. Identify the factors predicting whether or not the patient will be readmitted within 30 days.
2. Propose a classification rule to predict if a patient will be readmitted within 30 days.

Characteristics of the Data Set

All observations have five things in common:

1. They are all hospital admissions
2. Each patient had some form of diabetes
3. The patient stayed for between 1 and 14 days.
4. The patient had laboratory tests performed on him/her.
5. The patient was given some form of medication during the visit.

The data was collected during a ten-year period from 1999 to 2008. There are over 100,000 unique hospital admissions in the data set, with ~70,000 unique patients.

Description of variables

The dataset used covers ~50 different variables to describe every hospital diabetes admission. In this section we give an overview and brief description of the variables in this dataset.

a) Patient identifiers:

a. `encounter_id`: unique identifier for each admission
b. `patient_nbr`: unique identifier for each patient

b) Patient Demographics:

`race`, `age`, `gender`, `weight` cover the basic demographic information associated with each patient. `Payer_code` is an additional variable that identifies which health insurance (Medicare /Medicaid / Commercial) the patient holds.

c) Admission and discharge details:

a. `admission_source_id` and `admission_type_id` identify who referred the patient to the hospital (e.g. physician vs. emergency dept.) and what type of admission this was (Emergency vs. Elective vs. Urgent).
b. `discharge_disposition_id` indicates where the patient was discharged to after treatment.

d) Patient Medical History:

a. `num_outpatient`: number of outpatient visits by the patient in the year prior to the current encounter
b. `num_inpatient`: number of inpatient visits by the patient in the year prior to the current encounter
c. `num_emergency`: number of emergency visits by the patient in the year prior to the current encounter

e) Patient admission details:

a. `medical_specialty`: the specialty of the physician admitting the patient
b. `diag_1`, `diag_2`, `diag_3`: ICD9 codes for the primary, secondary and tertiary diagnoses of the patient. ICD9 are the universal codes that all physicians use to record diagnoses. There are various easy to use tools to lookup what individual codes mean (Wikipedia is pretty decent on its own)
c. `time_in_hospital`: the patient's length of stay in the hospital (in days)
d. `number_diagnoses`: Total no. of diagnosis entered for the patient
e. `num_lab_procedures`: No. of lab procedures performed in the current encounter
f. `num_procedures`: No. of non-lab procedures performed in the current encounter
g. `num_medications`: No. of distinct medications prescribed in the current encounter

f) Clinical Results:

a. `max_glu_serum`: indicates results of the glucose serum test
b. `A1Cresult`: indicates results of the A1c test

g) Medication Details:

a. `diabetesMed`: indicates if any diabetes medication was prescribed
b. `change`: indicates if there was a change in diabetes medication
c. `24 medication variables`: indicate whether the dosage of the medicines was changed in any manner during the encounter

h) Readmission indicator:

Indicates whether a patient was readmitted after a particular admission. There are 3 levels for this variable: "NO" = no readmission, "< 30" = readmission within 30 days and "> 30" = readmission after more than 30 days. The 30 day distinction is of practical importance to hospitals because federal regulations penalize hospitals for an excessive proportion of such readmissions.

To save your time we are going to use some data sets cleaned by the group. Thus, we provide two datasets:

`diabetic.data.csv` is the original data. You may use it for the purpose of summary if you wish. You will see that the original data can't be used directly for your analysis, yet.

`readmission.csv` is a cleaned version and they are modified in the following ways:

1) `Payer code`, `weight` and `Medical Specialty` are not included since they have a large number of missing values.

2) Variables such as `acetohexamide` (col 30), `glimepiride.pioglitazone` (45), `metformin.rosiglitazone`(46), `metformin.pioglitazone`(47) have little variability, and are as such excluded. This also includes the following variables: `chlorpropamide`(28), `acetohexamide`(30), `tolbutamide`(33), `acarbose`(36), `miglitor`(37), `troglitazone`(38), `tolazamide`(39), `examide`(40), `citoglipton`(41), `glyburide.metformin`(43), `glipizide.metformin`(44), and `glimepiride.pioglitazone`(45).

3) Some categorical variables have been regrouped. For example, `Diag1_mod` keeps some original levels with large number of patients and aggregates other patients as `others`. This process is known as 'binning.'

4) The event of interest is **readmitted within $<$ 30 days**. Note that you need to create this response first by regrouping **Readmission indicator**!

Exploratory Data Analysis

```
## ============================== STANDARD EDA TECHNIQUES
## ==============================

## <<<< READING IN DATA >>> ===== FULL DATASET ===== bill.data.test <-
## read.csv('Bills.subset.test.csv', header=TRUE, sep=',', na.strings='') #
## accounts for header, CSV, and na strings
df.full <- read.csv("diabetic.data.csv", header = TRUE, sep = ",", na.strings = "")  # accounts for hea
dim(df.full)  # 101,766 observations x 50 variables
```

```
## [1] 101766     50
```

```
head(df.full, 30)
```

```
##    encounter_id patient_nbr            race gender     age weight
## 1       2278392     8222157       Caucasian Female  [0-10)      ?
## 2        149190    55629189       Caucasian Female [10-20)      ?
## 3         64410    86047875 AfricanAmerican Female [20-30)      ?
## 4        500364    82442376       Caucasian   Male [30-40)      ?
## 5         16680    42519267       Caucasian   Male [40-50)      ?
## 6         35754    82637451       Caucasian   Male [50-60)      ?
## 7         55842    84259809       Caucasian   Male [60-70)      ?
## 8         63768   114882984       Caucasian   Male  [70-80)      ?
## 9         12522    48330783       Caucasian Female  [80-90)      ?
## 10        15738    63555939       Caucasian Female [90-100)      ?
## 11        28236    89869032 AfricanAmerican Female  [40-50)      ?
## 12        36900    77391171 AfricanAmerican   Male  [60-70)      ?
## 13        40926    85504905       Caucasian Female  [40-50)      ?
## 14        42570    77586282       Caucasian   Male  [80-90)      ?
## 15        62256    49726791 AfricanAmerican Female  [60-70)      ?
## 16        73578    86328819 AfricanAmerican   Male  [60-70)      ?
## 17        77076    92519352 AfricanAmerican   Male  [50-60)      ?
## 18        84222   108662661       Caucasian Female  [50-60)      ?
## 19        89682   107389323 AfricanAmerican   Male  [70-80)      ?
## 20       148530    69422211               ?   Male  [70-80)      ?
## 21       150006    22864131               ? Female  [50-60)      ?
## 22       150048    21239181               ?   Male  [60-70)      ?
## 23       182796    63000108 AfricanAmerican Female  [70-80)      ?
## 24       183930   107400762       Caucasian Female  [80-90)      ?
## 25       216156    62718876 AfricanAmerican Female  [70-80)      ?
## 26       221634    21861756           Other Female  [50-60)      ?
## 27       236316    40523301       Caucasian   Male  [80-90)      ?
## 28       248916   115196778       Caucasian Female  [50-60)      ?
## 29       250872    41606064       Caucasian   Male  [20-30)      ?
## 30       252822    18196434       Caucasian Female  [80-90)      ?
##    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
## 7                  3                        1                   2
## 8                  1                        1                   7
## 9                  2                        1                   4
## 10                 3                        3                   4
## 11                 1                        1                   7
## 12                 2                        1                   4
## 13                 1                        3                   7
## 14                 1                        6                   7
## 15                 3                        1                   2
## 16                 1                        3                   7
## 17                 1                        1                   7
## 18                 1                        1                   7
## 19                 1                        1                   7
## 20                 3                        6                   2
## 21                 2                        1                   4
## 22                 2                        1                   4
## 23                 2                        1                   4
## 24                 2                        6                   1
## 25                 3                        1                   2
## 26                 1                        1                   7
## 27                 1                        3                   7
## 28                 1                        1                   1
## 29                 2                        1                   2
## 30                 1                        2                   7
##    time_in_hospital payer_code        medical_specialty num_lab_procedures
## 1                 1          ? Pediatrics-Endocrinology                 41
## 2                 3          ?                        ?                 59
## 3                 2          ?                        ?                 11
## 4                 2          ?                        ?                 44
## 5                 1          ?                        ?                 51
## 6                 3          ?                        ?                 31
## 7                 4          ?                        ?                 70
## 8                 5          ?                        ?                 73
## 9                13          ?                        ?                 68
## 10               12          ?         InternalMedicine                 33
## 11                9          ?                        ?                 47
## 12                7          ?                        ?                 62
## 13                7          ?   Family/GeneralPractice                 60
## 14               10          ?   Family/GeneralPractice                 55
## 15                1          ?                        ?                 49
## 16               12          ?                        ?                 75
## 17                4          ?                        ?                 45
## 18                3          ?               Cardiology                 29
## 19                5          ?                        ?                 35
## 20                6          ?                        ?                 42
## 21                2          ?                        ?                 66
## 22                2          ?                        ?                 36
## 23                2          ?                        ?                 47
## 24               11          ?                        ?                 42
## 25                3          ?                        ?                 19
## 26                1          ?                        ?                 33
## 27                6          ?               Cardiology                 64
## 28                2          ?          Surgery-General                 25
## 29               10          ?                        ?                 53
## 30                5          ?               Cardiology                 52
##    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
## 7               1              21                 0                0
## 8               0              12                 0                0
## 9               2              28                 0                0
## 10              3              18                 0                0
## 11              2              17                 0                0
## 12              0              11                 0                0
## 13              0              15                 0                1
## 14              1              31                 0                0
## 15              5               2                 0                0
## 16              5              13                 0                0
## 17              4              17                 0                0
## 18              0              11                 0                0
## 19              5              23                 0                0
## 20              2              23                 0                0
## 21              1              19                 0                0
## 22              2              11                 0                0
## 23              0              12                 0                0
## 24              2              19                 0                0
## 25              4              18                 0                0
## 26              0               7                 0                0
## 27              3              18                 0                0
## 28              2              11                 0                0
## 29              0              20                 0                0
## 30              0              14                 0                0
##    number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
## 1                 0 250.83      ?      ?                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
## 7                 0    414    411    V45                7          None
## 8                 0    428    492    250                8          None
## 9                 0    398    427     38                8          None
## 10                0    434    198    486                8          None
## 11                0  250.7    403    996                9          None
## 12                0    157    288    197                7          None
## 13                0    428 250.43  250.6                8          None
## 14                0    428    411    427                8          None
## 15                0    518    998    627                8          None
## 16                0    999    507    996                9          None
## 17                0    410    411    414                8          None
## 18                0    682    174    250                3          None
## 19                0    402    425    416                9          None
## 20                0    737    427    714                8          None
## 21                0    410    427    428                7          None
## 22                0    572    456    427                6          None
## 23                0    410    401    582                8          None
## 24                0    V57    715    V43                8          None
## 25                0    189    496    427                6          None
## 26                0    786    401    250                3          None
## 27                0    427    428    414                7          None
## 28                0    996    585 250.01                3          None
## 29                0    277 250.02    263                6          None
## 30                0    428    410    414                8          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
## 7       None    Steady          No          No             No      Steady
## 8       None        No          No          No             No          No
## 9       None        No          No          No             No          No
## 10      None        No          No          No             No          No
## 11      None        No          No          No             No          No
## 12      None        No          No          No             No          No
## 13      None    Steady          Up          No             No          No
## 14      None        No          No          No             No          No
## 15      None        No          No          No             No          No
## 16      None        No          No          No             No          No
## 17      None        No          No          No             No          No
## 18      None        No          No          No             No          No
## 19      None        No          No          No             No          No
## 20      None        No          No          No             No          No
## 21      None        No          No          No             No          No
## 22      None    Steady          No          No             No      Steady
## 23      None        No          No          No             No          No
## 24      None        No          No          No             No          No
## 25      None        No          No          No             No          No
## 26      None    Steady          No          No             No          No
## 27        >7    Steady          No          No             No          No
## 28      None        No          No          No             No          No
## 29      None        No          No          No             No          No
## 30      None    Steady          No          No             No          No
##    acetohexamide glipizide glyburide tolbutamide pioglitazone
## 1             No        No        No          No           No
## 2             No        No        No          No           No
## 3             No    Steady        No          No           No
## 4             No        No        No          No           No
## 5             No    Steady        No          No           No
## 6             No        No        No          No           No
## 7             No        No        No          No           No
## 8             No        No    Steady          No           No
## 9             No    Steady        No          No           No
## 10            No        No        No          No           No
## 11            No        No        No          No           No
## 12            No        No        Up          No           No
## 13            No        No        No          No           No
## 14            No        No        No          No           No
## 15            No        No        No          No           No
## 16            No        No        No          No           No
## 17            No    Steady        No          No           No
## 18            No        No    Steady          No           No
## 19            No        No        No          No           No
## 20            No        No      Down          No           No
## 21            No        No        No          No           No
## 22            No        No        No          No           No
## 23            No        No        No          No           No
## 24            No        No        No          No           No
## 25            No    Steady        No          No           No
## 26            No        No        No          No           No
## 27            No        No    Steady          No           No
## 28            No        No        No          No           No
## 29            No        No        No          No           No
## 30            No        No    Steady          No           No
##    rosiglitazone acarbose miglitol troglitazone tolazamide examide
## 1             No       No       No           No         No      No
## 2             No       No       No           No         No      No
## 3             No       No       No           No         No      No
## 4             No       No       No           No         No      No
## 5             No       No       No           No         No      No
## 6             No       No       No           No         No      No
## 7             No       No       No           No         No      No
## 8             No       No       No           No         No      No
## 9             No       No       No           No         No      No
## 10        Steady       No       No           No         No      No
## 11            No       No       No           No         No      No
## 12            No       No       No           No         No      No
## 13            No       No       No           No         No      No
## 14            No       No       No           No         No      No
## 15            No       No       No           No         No      No
## 16            No       No       No           No         No      No
## 17            No       No       No           No         No      No
## 18            No       No       No           No         No      No
## 19            No       No       No           No         No      No
## 20            No       No       No           No         No      No
## 21            No       No       No           No         No      No
## 22            No       No       No           No         No      No
## 23            No       No       No           No         No      No
## 24            No       No       No           No         No      No
## 25            No       No       No           No         No      No
## 26            No       No       No           No         No      No
## 27            No       No       No           No         No      No
## 28            No       No       No           No         No      No
## 29            No       No       No           No         No      No
## 30            No       No       No           No         No      No
##    citoglipton insulin glyburide.metformin glipizide.metformin
## 1           No      No                  No                  No
## 2           No      Up                  No                  No
## 3           No      No                  No                  No
## 4           No      Up                  No                  No
## 5           No  Steady                  No                  No
## 6           No  Steady                  No                  No
## 7           No  Steady                  No                  No
## 8           No      No                  No                  No
## 9           No  Steady                  No                  No
## 10          No  Steady                  No                  No
## 11          No  Steady                  No                  No
## 12          No  Steady                  No                  No
## 13          No    Down                  No                  No
## 14          No  Steady                  No                  No
## 15          No  Steady                  No                  No
## 16          No      Up                  No                  No
## 17          No  Steady                  No                  No
## 18          No      No                  No                  No
## 19          No  Steady                  No                  No
## 20          No  Steady                  No                  No
## 21          No    Down                  No                  No
## 22          No  Steady                  No                  No
## 23          No      No                  No                  No
## 24          No      No                  No                  No
## 25          No  Steady                  No                  No
## 26          No      No                  No                  No
## 27          No      No                  No                  No
## 28          No  Steady                  No                  No
## 29          No    Down                  No                  No
## 30          No      No                  No                  No
##    glimepiride.pioglitazone metformin.rosiglitazone metformin.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
## 7                        No                      No                     No
## 8                        No                      No                     No
## 9                        No                      No                     No
## 10                       No                      No                     No
## 11                       No                      No                     No
## 12                       No                      No                     No
## 13                       No                      No                     No
## 14                       No                      No                     No
## 15                       No                      No                     No
## 16                       No                      No                     No
## 17                       No                      No                     No
## 18                       No                      No                     No
## 19                       No                      No                     No
## 20                       No                      No                     No
## 21                       No                      No                     No
## 22                       No                      No                     No
## 23                       No                      No                     No
## 24                       No                      No                     No
## 25                       No                      No                     No
## 26                       No                      No                     No
## 27                       No                      No                     No
## 28                       No                      No                     No
## 29                       No                      No                     No
## 30                       No                      No                     No
##    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
## 7      Ch         Yes         NO
## 8      No         Yes        >30
## 9      Ch         Yes         NO
## 10     Ch         Yes         NO
## 11     No         Yes        >30
## 12     Ch         Yes        <30
## 13     Ch         Yes        <30
## 14     No         Yes         NO
## 15     No         Yes        >30
## 16     Ch         Yes         NO
## 17     Ch         Yes        <30
## 18     No         Yes         NO
## 19     No         Yes        >30
## 20     Ch         Yes         NO
## 21     Ch         Yes         NO
## 22     Ch         Yes         NO
## 23     No          No         NO
## 24     No          No        >30
## 25     Ch         Yes         NO
## 26     No         Yes         NO
## 27     Ch         Yes         NO
## 28     No         Yes        >30
## 29     Ch         Yes        >30
## 30     Ch         Yes        >30
```

```
View(df.full)
```

```
## Warning: running command ''/usr/bin/otool' -L '/Library/Frameworks/
## R.framework/Resources/modules/R_de.so'' had status 1
```

```
summary(df.full)
```

```
##   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  
##                                    
##                                    
##                                    
##                                    
```

```
summary(df.full$readmitted)
```

```
##   <30   >30    NO 
## 11357 35545 54864
```

```
# ====== CLEANED DATASET ====
data1 <- read.csv("diabetic.data.csv", header = TRUE, sep = ",", na.strings = "")  # accounts for heade
dim(data1)  #101766 observations x 50 variables
```

```
## [1] 101766     50
```

```
tail(data1, 20)
```

```
##        encounter_id patient_nbr            race gender     age weight
## 101747    443797298    89955270       Caucasian   Male [70-80)      ?
## 101748    443804570    33230016       Caucasian Female [70-80)      ?
## 101749    443811536   189481478       Caucasian Female [40-50)      ?
## 101750    443816024   106392411       Caucasian Female [70-80)      ?
## 101751    443824292   138784172       Caucasian Female [80-90)      ?
## 101752    443835140   175326800       Caucasian   Male [70-80)      ?
## 101753    443835512   139605341           Other Female [40-50)      ?
## 101754    443841992   184875899           Other   Male [40-50)      ?
## 101755    443842016   183087545       Caucasian Female [70-80)      ?
## 101756    443842022   188574944           Other Female [40-50)      ?
## 101757    443842070   140199494           Other Female [60-70)      ?
## 101758    443842136   181593374       Caucasian Female [70-80)      ?
## 101759    443842340   120975314       Caucasian Female [80-90)      ?
## 101760    443842778    86472243       Caucasian   Male [80-90)      ?
## 101761    443847176    50375628 AfricanAmerican Female [60-70)      ?
## 101762    443847548   100162476 AfricanAmerican   Male [70-80)      ?
## 101763    443847782    74694222 AfricanAmerican Female [80-90)      ?
## 101764    443854148    41088789       Caucasian   Male [70-80)      ?
## 101765    443857166    31693671       Caucasian Female [80-90)      ?
## 101766    443867222   175429310       Caucasian   Male [70-80)      ?
##        admission_type_id discharge_disposition_id admission_source_id
## 101747                 1                        1                   7
## 101748                 1                       22                   7
## 101749                 1                        4                   7
## 101750                 3                        6                   1
## 101751                 3                        1                   1
## 101752                 3                        6                   1
## 101753                 3                        1                   1
## 101754                 1                        1                   7
## 101755                 1                        1                   7
## 101756                 1                        1                   7
## 101757                 1                        1                   7
## 101758                 1                        1                   7
## 101759                 1                        1                   7
## 101760                 1                        1                   7
## 101761                 1                        1                   7
## 101762                 1                        3                   7
## 101763                 1                        4                   5
## 101764                 1                        1                   7
## 101765                 2                        3                   7
## 101766                 1                        1                   7
##        time_in_hospital payer_code medical_specialty num_lab_procedures
## 101747                4         MC                 ?                  2
## 101748                8         MC  InternalMedicine                 51
## 101749               14         MD                 ?                 69
## 101750                3         MC       Orthopedics                 27
## 101751                3         MD                 ?                 31
## 101752               13         MC                 ?                 77
## 101753                3         HM                 ?                 13
## 101754               13          ?                 ?                 51
## 101755                9          ?                 ?                 50
## 101756               14         MD                 ?                 73
## 101757                2         MD                 ?                 46
## 101758                5          ?                 ?                 21
## 101759                5         MC                 ?                 76
## 101760                1         MC                 ?                  1
## 101761                6         DM                 ?                 45
## 101762                3         MC                 ?                 51
## 101763                5         MC                 ?                 33
## 101764                1         MC                 ?                 53
## 101765               10         MC   Surgery-General                 45
## 101766                6          ?                 ?                 13
##        num_procedures num_medications number_outpatient number_emergency
## 101747              0               7                 1                0
## 101748              6              19                 0                0
## 101749              0              16                 0                0
## 101750              1              29                 0                1
## 101751              2              24                 0                0
## 101752              6              65                 0                0
## 101753              1               5                 0                0
## 101754              2              13                 0                0
## 101755              2              33                 0                0
## 101756              6              26                 0                1
## 101757              6              17                 1                1
## 101758              1              16                 0                0
## 101759              1              22                 0                1
## 101760              0              15                 3                0
## 101761              1              25                 3                1
## 101762              0              16                 0                0
## 101763              3              18                 0                0
## 101764              0               9                 1                0
## 101765              2              21                 0                0
## 101766              3               3                 0                0
##        number_inpatient diag_1 diag_2 diag_3 number_diagnoses
## 101747                0    427    427    250                5
## 101748                0    410    311    250                9
## 101749                0    295    305    250                5
## 101750                0    715    401    250                9
## 101751                0    574    574    250                9
## 101752                0    424    429    486               16
## 101753                0    348    784    782                8
## 101754                0  250.8    730    731                9
## 101755                0    574    574 250.02                9
## 101756                0    592    599    518                9
## 101757                1    996    585    403                9
## 101758                1    491    518    511                9
## 101759                0    292      8    304                9
## 101760                0    435    784    250                7
## 101761                2    345    438    412                9
## 101762                0 250.13    291    458                9
## 101763                1    560    276    787                9
## 101764                0     38    590    296               13
## 101765                1    996    285    998                9
## 101766                0    530    530    787                9
##        max_glu_serum A1Cresult metformin repaglinide nateglinide
## 101747          None      None        No          No          No
## 101748          None        >7        No          No          No
## 101749          None        >7        Up          No          No
## 101750          None      Norm    Steady          No          No
## 101751          None      None        No          No          No
## 101752          None      Norm        No          No          No
## 101753          None      None    Steady          No          No
## 101754          None      None    Steady          No          No
## 101755          None        >7        No          No          No
## 101756          None        >8        No          No          No
## 101757          None      None        No          No          No
## 101758          None      None        No          No          No
## 101759          None      None        No          No          No
## 101760          None      None        No          No          No
## 101761          None      None        No          No          No
## 101762          None        >8    Steady          No          No
## 101763          None      None        No          No          No
## 101764          None      None    Steady          No          No
## 101765          None      None        No          No          No
## 101766          None      None        No          No          No
##        chlorpropamide glimepiride acetohexamide glipizide glyburide
## 101747             No          No            No    Steady        No
## 101748             No          No            No        No        No
## 101749             No          No            No        No    Steady
## 101750             No          No            No    Steady        No
## 101751             No          No            No        No        No
## 101752             No          No            No        No        No
## 101753             No          No            No        No    Steady
## 101754             No          No            No        No        No
## 101755             No          No            No        No        Up
## 101756             No          No            No    Steady        No
## 101757             No          No            No        No        No
## 101758             No          No            No        No        No
## 101759             No          No            No        No        No
## 101760             No          No            No        No        No
## 101761             No          No            No        No        No
## 101762             No          No            No        No        No
## 101763             No          No            No        No        No
## 101764             No          No            No        No        No
## 101765             No          No            No    Steady        No
## 101766             No          No            No        No        No
##        tolbutamide pioglitazone rosiglitazone acarbose miglitol
## 101747          No           No            No       No       No
## 101748          No           No            No       No       No
## 101749          No           No            No       No       No
## 101750          No           No            No       No       No
## 101751          No           No            No       No       No
## 101752          No           No            No       No       No
## 101753          No           No            No       No       No
## 101754          No           No            No       No       No
## 101755          No           No            No       No       No
## 101756          No           No            No       No       No
## 101757          No           No            No       No       No
## 101758          No           No            No       No       No
## 101759          No           No            No       No       No
## 101760          No           No            No       No       No
## 101761          No           No        Steady       No       No
## 101762          No           No            No       No       No
## 101763          No           No            No       No       No
## 101764          No           No            No       No       No
## 101765          No       Steady            No       No       No
## 101766          No           No            No       No       No
##        troglitazone tolazamide examide citoglipton insulin
## 101747           No         No      No          No      No
## 101748           No         No      No          No  Steady
## 101749           No         No      No          No    Down
## 101750           No         No      No          No  Steady
## 101751           No         No      No          No    Down
## 101752           No         No      No          No      Up
## 101753           No         No      No          No  Steady
## 101754           No         No      No          No    Down
## 101755           No         No      No          No  Steady
## 101756           No         No      No          No      Up
## 101757           No         No      No          No  Steady
## 101758           No         No      No          No  Steady
## 101759           No         No      No          No      Up
## 101760           No         No      No          No      Up
## 101761           No         No      No          No    Down
## 101762           No         No      No          No    Down
## 101763           No         No      No          No  Steady
## 101764           No         No      No          No    Down
## 101765           No         No      No          No      Up
## 101766           No         No      No          No      No
##        glyburide.metformin glipizide.metformin glimepiride.pioglitazone
## 101747                  No                  No                       No
## 101748                  No                  No                       No
## 101749                  No                  No                       No
## 101750                  No                  No                       No
## 101751                  No                  No                       No
## 101752                  No                  No                       No
## 101753                  No                  No                       No
## 101754                  No                  No                       No
## 101755                  No                  No                       No
## 101756                  No                  No                       No
## 101757                  No                  No                       No
## 101758                  No                  No                       No
## 101759                  No                  No                       No
## 101760                  No                  No                       No
## 101761                  No                  No                       No
## 101762                  No                  No                       No
## 101763                  No                  No                       No
## 101764                  No                  No                       No
## 101765                  No                  No                       No
## 101766                  No                  No                       No
##        metformin.rosiglitazone metformin.pioglitazone change diabetesMed
## 101747                      No                     No     No         Yes
## 101748                      No                     No     No         Yes
## 101749                      No                     No     Ch         Yes
## 101750                      No                     No     Ch         Yes
## 101751                      No                     No     Ch         Yes
## 101752                      No                     No     Ch         Yes
## 101753                      No                     No     Ch         Yes
## 101754                      No                     No     Ch         Yes
## 101755                      No                     No     Ch         Yes
## 101756                      No                     No     Ch         Yes
## 101757                      No                     No     No         Yes
## 101758                      No                     No     No         Yes
## 101759                      No                     No     Ch         Yes
## 101760                      No                     No     Ch         Yes
## 101761                      No                     No     Ch         Yes
## 101762                      No                     No     Ch         Yes
## 101763                      No                     No     No         Yes
## 101764                      No                     No     Ch         Yes
## 101765                      No                     No     Ch         Yes
## 101766                      No                     No     No          No
##        readmitted
## 101747        <30
## 101748        >30
## 101749        >30
## 101750         NO
## 101751        <30
## 101752         NO
## 101753         NO
## 101754         NO
## 101755        >30
## 101756        >30
## 101757        >30
## 101758         NO
## 101759         NO
## 101760         NO
## 101761        >30
## 101762        >30
## 101763         NO
## 101764         NO
## 101765         NO
## 101766         NO
```

```
head(data1, 20)
```

```
##    encounter_id patient_nbr            race gender      age weight
## 1       2278392     8222157       Caucasian Female   [0-10)      ?
## 2        149190    55629189       Caucasian Female  [10-20)      ?
## 3         64410    86047875 AfricanAmerican Female  [20-30)      ?
## 4        500364    82442376       Caucasian   Male  [30-40)      ?
## 5         16680    42519267       Caucasian   Male  [40-50)      ?
## 6         35754    82637451       Caucasian   Male  [50-60)      ?
## 7         55842    84259809       Caucasian   Male  [60-70)      ?
## 8         63768   114882984       Caucasian   Male  [70-80)      ?
## 9         12522    48330783       Caucasian Female  [80-90)      ?
## 10        15738    63555939       Caucasian Female [90-100)      ?
## 11        28236    89869032 AfricanAmerican Female  [40-50)      ?
## 12        36900    77391171 AfricanAmerican   Male  [60-70)      ?
## 13        40926    85504905       Caucasian Female  [40-50)      ?
## 14        42570    77586282       Caucasian   Male  [80-90)      ?
## 15        62256    49726791 AfricanAmerican Female  [60-70)      ?
## 16        73578    86328819 AfricanAmerican   Male  [60-70)      ?
## 17        77076    92519352 AfricanAmerican   Male  [50-60)      ?
## 18        84222   108662661       Caucasian Female  [50-60)      ?
## 19        89682   107389323 AfricanAmerican   Male  [70-80)      ?
## 20       148530    69422211               ?   Male  [70-80)      ?
##    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
## 7                  3                        1                   2
## 8                  1                        1                   7
## 9                  2                        1                   4
## 10                 3                        3                   4
## 11                 1                        1                   7
## 12                 2                        1                   4
## 13                 1                        3                   7
## 14                 1                        6                   7
## 15                 3                        1                   2
## 16                 1                        3                   7
## 17                 1                       1                  7
## 18                 1                       1                  7
## 19                 1                       1                  7
## 20                 3                       6                  2
##    time_in_hospital payer_code        medical_specialty num_lab_procedures
## 1                 1          ? Pediatrics-Endocrinology                 41
## 2                 3          ?                        ?                 59
## 3                 2          ?                        ?                 11
## 4                 2          ?                        ?                 44
## 5                 1          ?                        ?                 51
## 6                 3          ?                        ?                 31
## 7                 4          ?                        ?                 70
## 8                 5          ?                        ?                 73
## 9                13          ?                        ?                 68
## 10               12          ?         InternalMedicine                 33
## 11                9          ?                        ?                 47
## 12                7          ?                        ?                 62
## 13                7          ?   Family/GeneralPractice                 60
## 14               10          ?   Family/GeneralPractice                 55
## 15                1          ?                        ?                 49
## 16               12          ?                        ?                 75
## 17                4          ?                        ?                 45
## 18                3          ?               Cardiology                 29
## 19                5          ?                        ?                 35
## 20                6          ?                        ?                 42
##    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
## 7               1              21                 0                0
## 8               0              12                 0                0
## 9               2              28                 0                0
## 10              3              18                 0                0
## 11              2              17                 0                0
## 12              0              11                 0                0
## 13              0              15                 0                1
## 14              1              31                 0                0
## 15              5               2                 0                0
## 16              5              13                 0                0
## 17              4              17                 0                0
## 18              0              11                 0                0
## 19              5              23                 0                0
## 20              2              23                 0                0
##    number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
## 1                 0 250.83      ?      ?                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
## 7                 0    414    411    V45                7          None
## 8                 0    428    492    250                8          None
## 9                 0    398    427     38                8          None
## 10                0    434    198    486                8          None
## 11                0  250.7    403    996                9          None
## 12                0    157    288    197                7          None
## 13                0    428 250.43  250.6                8          None
## 14                0    428    411    427                8          None
## 15                0    518    998    627                8          None
## 16                0    999    507    996                9          None
## 17                0    410    411    414                8          None
## 18                0    682    174    250                3          None
## 19                0    402    425    416                9          None
## 20                0    737    427    714                8          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
## 7       None    Steady          No          No             No      Steady
## 8       None        No          No          No             No          No
## 9       None        No          No          No             No          No
## 10      None        No          No          No             No          No
## 11      None        No          No          No             No          No
## 12      None        No          No          No             No          No
## 13      None    Steady          Up          No             No          No
## 14      None        No          No          No             No          No
## 15      None        No          No          No             No          No
## 16      None        No          No          No             No          No
## 17      None        No          No          No             No          No
## 18      None        No          No          No             No          No
## 19      None        No          No          No             No          No
## 20      None        No          No          No             No          No
##    acetohexamide glipizide glyburide tolbutamide pioglitazone
## 1             No        No        No          No           No
## 2             No        No        No          No           No
## 3             No    Steady        No          No           No
## 4             No        No        No          No           No
## 5             No    Steady        No          No           No
## 6             No        No        No          No           No
## 7             No        No        No          No           No
## 8             No        No    Steady          No           No
## 9             No    Steady        No          No           No
## 10            No        No        No          No           No
## 11            No        No        No          No           No
## 12            No        No        Up          No           No
## 13            No        No        No          No           No
## 14            No        No        No          No           No
## 15            No        No        No          No           No
## 16            No        No        No          No           No
## 17            No    Steady        No          No           No
## 18            No        No    Steady          No           No
## 19            No        No        No          No           No
## 20            No        No      Down          No            No
##    rosiglitazone acarbose miglitol troglitazone tolazamide examide
## 1             No       No       No           No         No      No
## 2             No       No       No           No         No      No
## 3             No       No       No           No         No      No
## 4             No       No       No           No         No      No
## 5             No       No       No           No         No      No
## 6             No       No       No           No         No      No
## 7             No       No       No           No         No      No
## 8             No       No       No           No         No      No
## 9             No       No       No           No         No      No
## 10        Steady       No       No           No         No      No
## 11            No       No       No           No         No      No
## 12            No       No       No           No         No      No
## 13            No       No       No           No         No      No
## 14            No       No       No           No         No      No
## 15            No       No       No           No         No      No
## 16            No       No       No           No         No      No
## 17            No       No       No           No         No      No
## 18            No       No       No           No         No      No
## 19            No       No       No           No         No      No
## 20            No       No       No           No         No      No
##    citoglipton insulin glyburide.metformin glipizide.metformin
## 1           No      No                  No                  No
## 2           No      Up                  No                  No
## 3           No      No                  No                  No
## 4           No      Up                  No                  No
## 5           No  Steady                  No                  No
## 6           No  Steady                  No                  No
## 7           No  Steady                  No                  No
## 8           No      No                  No                  No
## 9           No  Steady                  No                  No
## 10          No  Steady                  No                  No
## 11          No  Steady                  No                  No
## 12          No  Steady                  No                  No
## 13          No    Down                  No                  No
## 14          No  Steady                  No                  No
## 15          No  Steady                  No                  No
## 16          No      Up                  No                  No
## 17          No  Steady                  No                  No
## 18          No      No                  No                  No
## 19          No  Steady                  No                  No
## 20          No  Steady                  No                  No
##    glimepiride.pioglitazone metformin.rosiglitazone metformin.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
## 7                        No                      No                     No
## 8                        No                      No                     No
## 9                        No                      No                     No
## 10                       No                      No                     No
## 11                       No                      No                     No
## 12                       No                      No                     No
## 13                       No                      No                     No
## 14                       No                      No                     No
## 15                       No                      No                     No
## 16                       No                      No                     No
## 17                       No                      No                     No
## 18                       No                      No                     No
## 19                       No                      No                     No
## 20                       No                      No                     No
##    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
## 7      Ch         Yes         NO
## 8      No         Yes        >30
## 9      Ch         Yes         NO
## 10     Ch         Yes         NO
## 11     No         Yes        >30
## 12     Ch         Yes        <30
## 13     Ch         Yes        <30
## 14     No         Yes         NO
## 15     No         Yes        >30
## 16     Ch         Yes         NO
## 17     Ch         Yes        <30
## 18     No         Yes         NO
## 19     No         Yes        >30
## 20     Ch         Yes         NO
```

```
# View(data1)

data1 <- data1[-c(6, 11:12, 28, 30, 33, 36:41, 43:47)]  # getting rid of unhelpful vars
names(data1)
```

```
##  [1] "encounter_id"             "patient_nbr"
##  [3] "race"                     "gender"
##  [5] "age"                      "admission_type_id"
##  [7] "discharge_disposition_id" "admission_source_id"
##  [9] "time_in_hospital"         "num_lab_procedures"
## [11] "num_procedures"           "num_medications"
## [13] "number_outpatient"        "number_emergency"
## [15] "number_inpatient"         "diag_1"
## [17] "diag_2"                   "diag_3"
## [19] "number_diagnoses"         "max_glu_serum"
## [21] "A1Cresult"                "metformin"
## [23] "repaglinide"              "nateglinide"
## [25] "glimepiride"              "glipizide"
## [27] "glyburide"                "pioglitazone"
## [29] "rosiglitazone"            "insulin"
## [31] "change"                   "diabetesMed"
## [33] "readmitted"
```

```r
dim(data1)  # 101766 x 33
```

```
## [1] 101766     33
```

```r
summary(data1)
```

```
##   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        admission_type_id
##  Female         :54708   [70-80):26068   Min.   :1.000    
##  Male           :47055   [60-70):22483   1st Qu.:1.000    
##  Unknown/Invalid:    3   [50-60):17256   Median :1.000    
##                          [80-90):17197   Mean   :2.024    
##                          [40-50): 9685   3rd Qu.:3.000    
##                          [30-40): 3775   Max.   :8.000    
##                          (Other): 5302                    
##  discharge_disposition_id admission_source_id time_in_hospital
##  Min.   : 1.000           Min.   : 1.000      Min.   : 1.000  
##  1st Qu.: 1.000           1st Qu.: 1.000      1st Qu.: 2.000  
##  Median : 1.000           Median : 7.000      Median : 4.000  
##  Mean   : 3.716           Mean   : 5.754      Mean   : 4.396  
##  3rd Qu.: 4.000           3rd Qu.: 7.000      3rd Qu.: 6.000  
##  Max.   :28.000           Max.   :25.000      Max.   :14.000  
##                                                               
##  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     glimepiride
##  Down  :  575   Down  :    45   Down  :    11   Down  :  194
##  No    :81778   No    :100227   No    :101063   No    :96575
##  Steady:18346   Steady:  1384   Steady:   668   Steady: 4670
##  Up    : 1067   Up    :   110   Up    :    24   Up    :  327
##
##
##
##   glipizide      glyburide     pioglitazone   rosiglitazone
##  Down  :  560   Down  :  564   Down  :  118   Down  :   87
##  No    :89080   No    :91116   No    :94438   No    :95401
##  Steady:11356   Steady: 9274   Steady: 6976   Steady: 6100
##  Up    :  770   Up    :  812   Up    :  234   Up    :  178
##
##
##
##    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
##
##
##
```

```r
# <<<<<<<<<< NA VALUES >>>>>>>>>
sum(is.na(data1))
```

```
## [1] 0
```

```r
# show how many NA values in each column
sapply(data1, function(x) sum(is.na(x)))  # no 0 values
```

```
##             encounter_id              patient_nbr                     race
##                        0                        0                        0
##                   gender                      age        admission_type_id
##                        0                        0                        0
## discharge_disposition_id      admission_source_id         time_in_hospital
##                        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
##              glimepiride                glipizide                glyburide
##                        0                        0                        0
##             pioglitazone            rosiglitazone                  insulin
##                        0                        0                        0
##                   change              diabetesMed               readmitted
##                        0                        0                        0
```

Variables of interest

Readmitted

```
summary(data1$readmitted)
```

```
##   <30   >30    NO
## 11357 35545 54864
```

```
# <30 >30 NO 11357 35545 54864
```

Race

```
# variables of interest
summary(data1$race)  # boxplot readmit by race
```

```
##               ? AfricanAmerican           Asian       Caucasian
##            2273           19210             641           76099
##        Hispanic           Other
##            2037            1506
```

```
# filter by race (AfricanAmerican, Asian, Caucasian, Hispanic, Other) &&
# ------ AfricanAmerican ----
readmit_less30.afamer <- filter(data1, race == "AfricanAmerican", readmitted ==
    "<30")
dim(readmit_less30.afamer)  # 2155
```

```
## [1] 2155   33
```

```
readmit_more30.afamer <- filter(data1, race == "AfricanAmerican", readmitted ==
    ">30")
dim(readmit_more30.afamer)  # 6634
```

```
## [1] 6634   33
```

```
readmit_none.afamer <- filter(data1, race == "AfricanAmerican", readmitted ==
    "NO")
dim(readmit_none.afamer)  # 10421
```

```
## [1] 10421    33
```

```
slices.afamer <- c(2155, 6634, 10421)
lbls.afamer <- c("<30", ">30", "none")
pct.afamer <- round(slices.afamer/sum(slices.afamer) * 100)
lbls.afamer <- paste(lbls.afamer, "-(", pct.afamer, ")")  # add percents to labels
lbls.afamer <- paste(lbls.afamer, "%", sep = "")  # ad % to labels

# ---- ASIAN ----
readmit_less30.asian <- filter(data1, race == "Asian", readmitted == "<30")
dim(readmit_less30.asian)  # 65
```

```
## [1] 65 33
```

```
readmit_more30.asian <- filter(data1, race == "Asian", readmitted == ">30")
dim(readmit_more30.asian)  # 161
```

```
## [1] 161  33
```

```
readmit_none.asian <- filter(data1, race == "Asian", readmitted == "NO")
dim(readmit_none.asian)  # 415
```

```
## [1] 415  33
```

```
slices.asian <- c(65, 161, 415)
lbls.asian <- c("<30", ">30", "none")
pct.asian <- round(slices.asian/sum(slices.asian) * 100)
lbls.asian <- paste(lbls.asian, "-(", pct.asian, ")")  # add percents to labels
lbls.asian <- paste(lbls.asian, "%", sep = "")  # ad % to labels

# ---- CAUCASIAN ----
readmit_less30.cau <- filter(data1, race == "Caucasian", readmitted == "<30")
dim(readmit_less30.cau)  # 8592
```

```
## [1] 8592   33
```

```
readmit_more30.cau <- filter(data1, race == "Caucasian", readmitted == ">30")
dim(readmit_more30.cau)  # 27124
```

```
## [1] 27124    33
```

```
readmit_none.cau <- filter(data1, race == "Caucasian", readmitted == "NO")
dim(readmit_none.cau)  # 40383
```

```
## [1] 40383    33
```

```
slices.cau <- c(8592, 27124, 40383)  #76099 total
lbls.cau <- c("<30", ">30", "none")
pct.cau <- round(slices.cau/sum(slices.cau) * 100)
lbls.cau <- paste(lbls.cau, "-(", pct.cau, ")")  # add percents to labels
lbls.cau <- paste(lbls.cau, "%", sep = "")  # ad % to labels

# ---- HISPANIC ----
readmit_less30.hisp <- filter(data1, race == "Hispanic", readmitted == "<30")
dim(readmit_less30.hisp)  # 212
```

```
## [1] 212  33
```

```
readmit_more30.hisp <- filter(data1, race == "Hispanic", readmitted == ">30")
dim(readmit_more30.hisp)  # 27124
```

```
## [1] 642  33
```

```
readmit_none.hisp <- filter(data1, race == "Hispanic", readmitted == "NO")
dim(readmit_none.hisp)  # 40383
```

```
## [1] 1183   33
```

```
slices.hisp <- c(212, 642, 1183)  #76099 total
lbls.hisp <- c("<30", ">30", "none")
pct.hisp <- round(slices.hisp/sum(slices.hisp) * 100)
lbls.hisp <- paste(lbls.hisp, "-(", pct.hisp, ")")  # add percents to labels
lbls.hisp <- paste(lbls.hisp, "%", sep = "")  # ad % to labels

# ---- OTHER ----
readmit_less30.oth <- filter(data1, race == "Other", readmitted == "<30")
dim(readmit_less30.oth)  # 145
```

```
## [1] 145  33
```

```
readmit_more30.oth <- filter(data1, race == "Other", readmitted == ">30")
dim(readmit_more30.oth)  # 446
```

```
## [1] 446  33
```

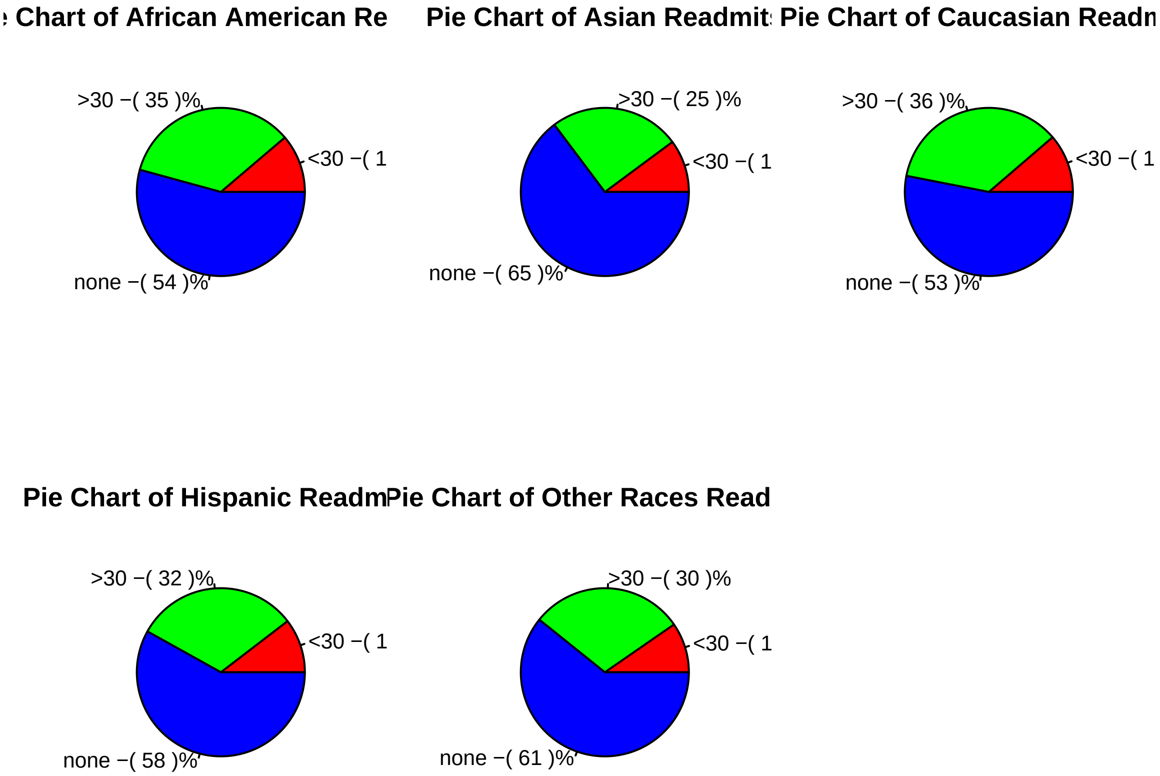
```
readmit_none.oth <- filter(data1, race == "Other", readmitted == "NO")
dim(readmit_none.oth)  # 915
```

```
## [1] 915  33
```

```
slices.oth <- c(145, 446, 915)
lbls.oth <- c("<30", ">30", "none")
pct.oth <- round(slices.oth/sum(slices.oth) * 100)
lbls.oth <- paste(lbls.oth, "-(", pct.oth, ")")  # add percents to labels
lbls.oth <- paste(lbls.oth, "%", sep = "")  # ad % to labels



par(mfrow = c(2, 3))
pie(slices.afamer, labels = lbls.afamer, col = rainbow(length(lbls.afamer)),
    main = "Pie Chart of African American Readmits")
pie(slices.asian, labels = lbls.asian, col = rainbow(length(lbls.asian)), main = "Pie Chart of Asian Rea
pie(slices.cau, labels = lbls.cau, col = rainbow(length(lbls.cau)), main = "Pie Chart of Caucasian Readm
pie(slices.hisp, labels = lbls.hisp, col = rainbow(length(lbls.hisp)), main = "Pie Chart of Hispanic Rea
pie(slices.oth, labels = lbls.oth, col = rainbow(length(lbls.hisp)), main = "Pie Chart of Other Races Re
```



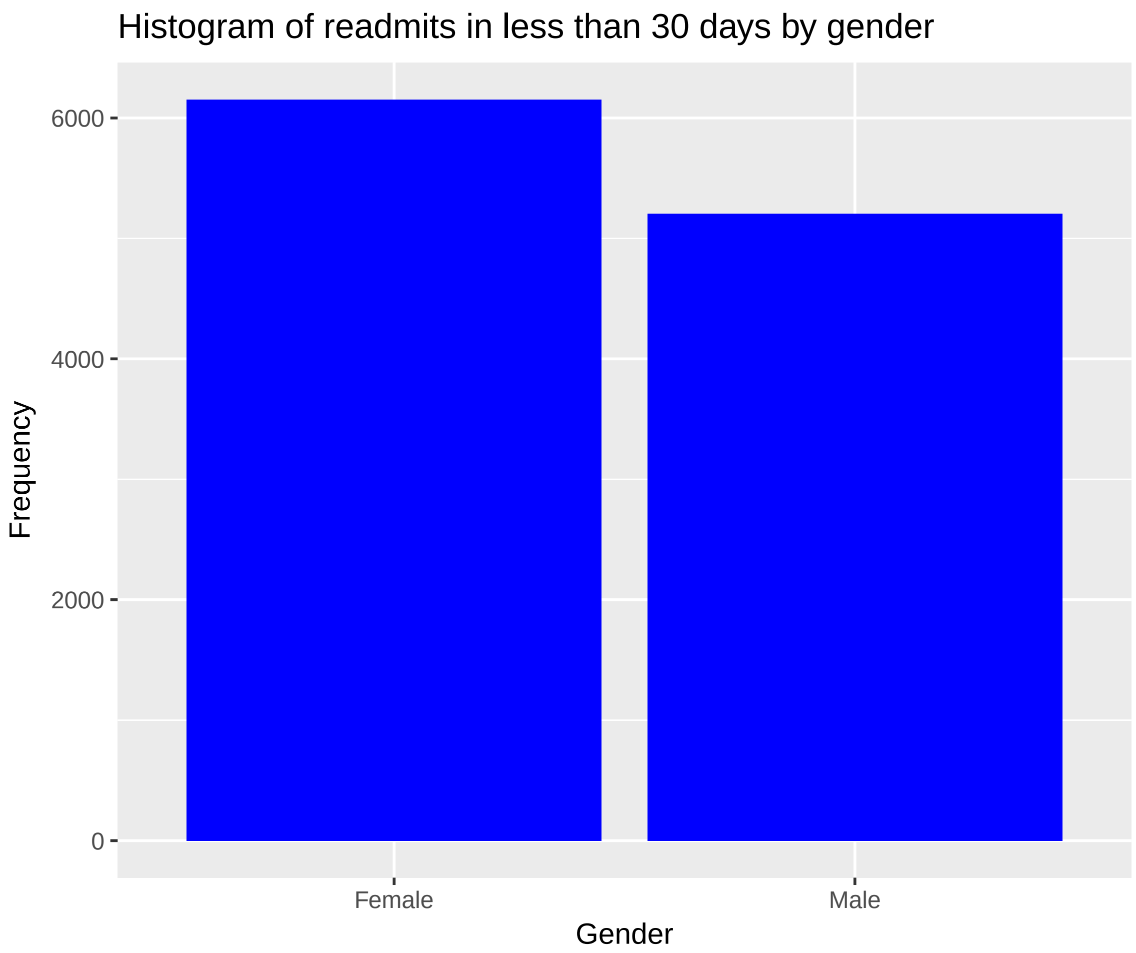
Gender

```
summary(data1$gender)   #boxplot
```

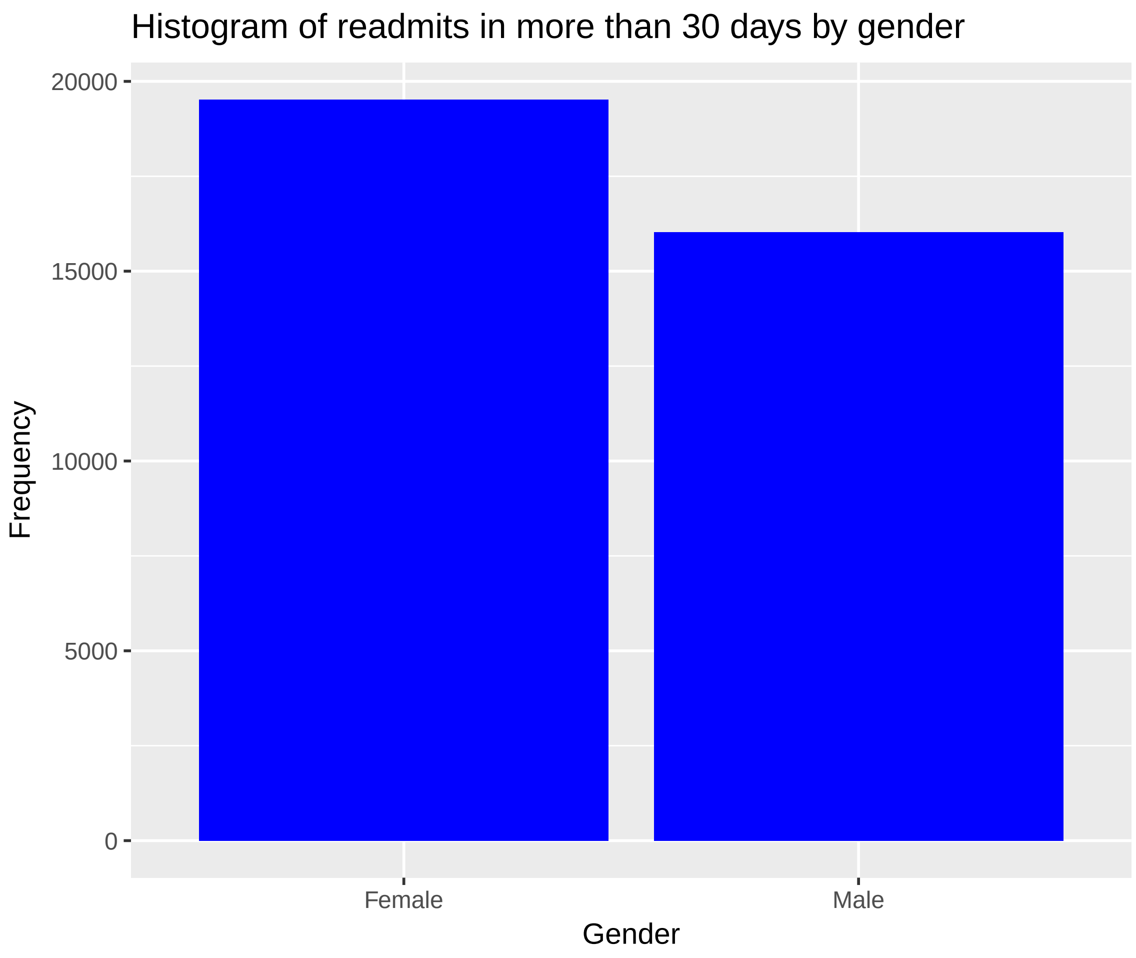
```
##          Female             Male Unknown/Invalid
##           54708            47055               3
```

```
readmit_less30.gender <- filter(data1, readmitted == "<30")
readmit_more30.gender <- filter(data1, readmitted == ">30")

par(mfrow = c(2, 2))
ggplot(readmit_less30.gender) + geom_bar(aes(x = gender), fill = "blue") + labs(title = "Histogram of r
    x = "Gender", y = "Frequency")
```



```
ggplot(readmit_more30.gender) + geom_bar(aes(x = gender), fill = "blue") + labs(title = "Histogram of re
    x = "Gender", y = "Frequency")
```



Age

```
summary(data1$age)   #scatterplot
```

```
##    [0-10)  [10-20)  [20-30)  [30-40)  [40-50)  [50-60)  [60-70)  [70-80)
##       161      691     1657     3775     9685    17256    22483    26068
##   [80-90) [90-100)
##     17197     2793
```

Change (in diabetes medication)

```
summary(data1$change)   #boxplot - change in diabetes medication
```

```
##    Ch    No
## 47011 54755
```

Number of diagnosis

```
summary(data1$number_diagnoses)   #bar plot
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000   8.000   7.423   9.000  16.000
```

```
# #<<<<< SUMMARY STATS >>>>>>>> dim(county_data) #sum.isna(county_data) #
# str(county_data) summary(county_data) tail(county_data, 10)
# head(county_data, 20) View(county_data) #class(county_data)
# names(county_data) county_data[1,'adult_smoking'] # row 1 of column
# 'adult_smoking' # <<<<LEVELS of a categorical variable>>>
# levels(county_data$state) # <<<<<<setting the BASE LEVEL >>>>>> #
# levels(county_data$median_income) # 'HIGH', 'LOW', 'MEDIUM' (default is
# alphabetical order) # county_data$median_income <-
# factor(county_data$median_income, levels = c('LOW', 'MEDIUM', 'HIGH'))
# #<<<<< RENAME VARIABLES >>>>>>>> #county_data <- county_data %>%
# rename(smoking_adults = adult_smoking ) # new = old name
# names(county_data) # <<<<<<<<<< NA VALUES >>>>>>>>>
# sum(is.na(county_data)) # show how many NA values in each column
# sapply(county_data, function(x) sum(is.na(x))) # <<<<<< REMOVE A COLUMN
# >>>>>>>> #remove a column # county_data.mod <- county_data[,-1]
# #head(county_data.mod) #summary(county_data.mod) #View(county_data.mod) #
# <<<<<<<<<<<< GETTING DISTINCT VALUES IN A COLUMN >>>>>>>>>>>> #county_data
# %>% county_data(,-1) county_data %>% select(state) %>% distinct() # get
# distinct values for state column county_data %>% select(median_income) %>%
# distinct() # get distince values for median_income column #county_data %>%
# select(median_income) #base R way #unique(county_data$state) # <<<<<<<
# MIN/MAX VALUES FOR A COLUMN >>>>>>>>> min(county_data$uninsured)
# max(county_data$uninsured) min(county_data$poverty_frac)
# max(county_data$poverty_frac)
```

Research approach

From the *Goals* section above, your study should respond to the following:

1) Identify important factors that capture the chance of a readmission within 30 days.

The set of available predictors is not limited to the raw variables in the data set. You may engineer any factors using the data, that you think will improve your model's quality.

2) For the purpose of classification, propose a model that can be used to predict whether a patient will be a readmit within 30 days. Justify your choice. Hint: use a decision criterion, such as AUC, to choose among a few candidate models.

Based on a quick and somewhat arbitrary guess, we estimate it costs twice as much to mislabel a readmission than it does to mislabel a non-readmission. Based on this risk ratio, propose a specific classification rule to minimize the cost. If you find any information that could provide a better cost estimate, please justify it in your write-up and use the better estimate in your answer.

Suggestion: You may use any of the methods covered so far in parts 1) and 2), and they need not be the same. Also keep in mind that a training/testing data split may be necessary.

Suggested outline

As you all know, it is very important to present your findings well. To achieve the best possible results you need to understand your audience.

Your target audience is a manager within the hospital organization. They hold an MBA, are familiar with medical terminology (though you do not need any previous medical knowledge), and have gone through a similar course to our Modern Data Mining with someone like your professor. You can assume thus some level of technical familiarity, but should not let the paper be bogged down with code or other difficult to understand output.

Note then that the most important elements of your report are the clarity of your analysis and the quality of your proposals.

A suggested outline of the report would include the following components:

1) Executive Summary

- This section should be accessible by people with very little statistical background (avoid using technical words and no direct R output is allowed)
- Give a background of the study. You may check the original website or other sources to fill in some details, such as to why the questions we address here are important.
- A quick summary about the data.
- Methods used and the main findings.
- You may use clearly labelled and explained visualizations.
- Issues, concerns, limitations of the conclusions. This is an especially important section to be honest in - we might be Penn students, but we are statisticians today.

2) Detailed process of the analysis

i) Data Summary

- Nature of the data, origin
- Necessary quantitative and graphical summaries
- Are there any problems with the data?
- Which variables are considered as input

ii) Analyses

- Various appropriate statistical methods: e.g. glmnet
- Comparisons various models
- Final model(s)

iii) Conclusion

- Summarize results and the final model
- Final recommendations

Maintain a good descriptive flow in the text of your report. Use Appendices to display lengthy output.

iii) Appendix

- All your R code (code without comments is no good!) if you are not using `rmd` format.
- Any thing necessary to keep but for which you don't want them to be in the main report.

Collaboration

This is an **individual** assignment. We will only allow private Piazza posts for questions. If there are questions that are generally useful, we will release that information.