Research log

Dennis Scheper

11-9-2020

1 Week 1 - original dataset

Date: 11 September, 2020 (week 1)

1.1 Introduction

The article, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records", states that the management of hyperglycemia has a significant impact on the outcome and readmission rates of hospitalized patients. The authors based this conclusion on the comprehensive assessment of 70.000 diabetes patient records retrieved from 140 US hospitals. The results that depict the relationship between readmission rates and measurement of HbA1c levels, can aid in the enlargement of already existing diabetes tactics to reduce readmission rates further.

1.2 Dataset and Attributes Information

All information used in this article comes from a database, consisting of 41 tables and totals 117 features, such as demographics (like gender, race and age), inpatient or outpatient, and (in-hospital) mortality. Data came from 130 hospitals in the USA over a ten-year period (1998-2008) and contains around 74 million unique visits by 18 million unique patients. This research used information that need to accede to the following specifications:

1. is a hospital admission;
2. the encounter is a diagnosed with 'diabetes', any kind will satisfy;
3. the length of admission was at least one day up to eighteen days;
4. laboratory test results are available; and
5. medications were administered.

On these five criteria points, 101.000 encounters fulfilled all specifications. After some considerations with removing encounters based on incomplete (weight and medical specialty) or biased data (discharge to a hospice or death), 69.984 encounters remained in the final dataset.

The initial dataset consists of 55 attributes with the class attribute being an encounter of one patient. As there are way too many attributes to describe, please refer to the codebook for all descriptions; we only look at some important attributes and their type and possible valuations. The age of a patient is nominal and is grouped into ten-year intervals. The admission type or for what specific reason a patient was hospitalized, and comes in 9 distinct values while the type is nominal. Some attributes are numeric and count, for example, the amount of lab procedures, the number of medications and number of emergency visits. The database consists of three diagnosis attributes, which can have 848, 923 and 954 distinct values, respectively. The values are based on ICD9 three-letter codes and are of nominal type. Some other important attributes are whether a patient changed medications (with the values 'no change' and 'change'; nominal type) or had diabetes medication ('yes' or 'no' values; also of nominal typing). 24 other attributes depict whether a medicine is prescribed or not, if prescribed, then if the dosage was increased ('up'), decreased ('down'), or stayed the same ('steady') during the encounter. Readmission rates were calculated by looking at a nominal type ('Readmitted') with the possible valuations of '<30' for a patient that was readmitted within 30 days, '>30' for a patient that was readmitted after 30 days, and 'No' for patients that were not readmitted. The authors' goal was to determine whether a relationship between readmission rates and HbA1c measurement exists, therefore they introduced a new attribute 'HbA1c' with four different valuations, based on the information from the database: 1) no HbA1c test performed; 2) HbA1c performed and in normal range; 3) HbA1c performed and the result is greater than 8% with no change in diabetic medication; and 4) Hb1Ac performed, result is greater than 8% and diabetic medication was changed.

1.3 Research Question

Is it possible, using machine learning techniques, to predict whether a patient's time in hospital is linked to the results of a HbA1c measurement?

2 Week 2 - EDA (Exploratory Data Analysis)

Date: 14 September, 2020

In this section, we will perform an exploratory data analysis (EDA) to determine variations exists in our dataset, exposing any missing values, and if they exist,

```
library(formattable)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
encounter.data <- read.table(file = 'dataset_diabetes/diabetic_data.csv',
                   sep = ',', header = TRUE)

codebook <- read.table(file = 'codebook.csv', sep = ';',
                       header = T)
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
## EOF within quoted string
```

```
glimpse(encounter.data)
```

```
## Rows: 101,766
## Columns: 50
## $ encounter_id             <int> 2278392, 149190, 64410, 500364, 16680, 357...
## $ patient_nbr              <int> 8222157, 55629189, 86047875, 82442376, 425...
## $ race                     <fct> Caucasian, Caucasian, AfricanAmerican, Cau...
## $ gender                   <fct> Female, Female, Female, Male, Male, Male, ...
## $ age                      <fct> [0-10), [10-20), [20-30), [30-40), [40-50)...
## $ weight                   <fct> ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ...
## $ admission_type_id        <int> 6, 1, 1, 1, 1, 2, 3, 1, 2, 3, 1, 2, 1, 1, ...
## $ discharge_disposition_id <int> 25, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 3, 6,...
## $ admission_source_id      <int> 1, 7, 7, 7, 7, 2, 2, 7, 4, 4, 7, 4, 7, 7, ...
## $ time_in_hospital         <int> 1, 3, 2, 2, 1, 3, 4, 5, 13, 12, 9, 7, 7, 1...
## $ payer_code               <fct> ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ...
## $ medical_specialty        <fct> Pediatrics-Endocrinology, ?, ?, ?, ?, ?, ?...
## $ num_lab_procedures       <int> 41, 59, 11, 44, 51, 31, 70, 73, 68, 33, 47...
## $ num_procedures           <int> 0, 0, 5, 1, 0, 6, 1, 0, 2, 3, 2, 0, 0, 1, ...
## $ num_medications          <int> 1, 18, 13, 16, 8, 16, 21, 12, 28, 18, 17, ...
## $ number_outpatient        <int> 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ number_emergency         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ...
## $ number_inpatient         <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

```
## $ diag_1                   <fct> 250.83, 276, 648, 8, 197, 414, 414, 428, 3...
## $ diag_2                   <fct> ?, 250.01, 250, 250.43, 157, 411, 411, 492...
## $ diag_3                   <fct> ?, 255, V27, 403, 250, 250, V45, 250, 38, ...
## $ number_diagnoses         <int> 1, 9, 6, 7, 5, 9, 7, 8, 8, 8, 9, 7, 8, 8, ...
## $ max_glu_serum            <fct> None, None, None, None, None, None, None, ...
## $ A1Cresult                <fct> None, None, None, None, None, None, None, ...
## $ metformin                <fct> No, No, No, No, No, No, Steady, No, No, No...
## $ repaglinide              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ nateglinide              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ chlorpropamide           <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glimepiride              <fct> No, No, No, No, No, No, Steady, No, No, No...
## $ acetohexamide            <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glipizide                <fct> No, No, Steady, No, Steady, No, No, No, St...
## $ glyburide                <fct> No, No, No, No, No, No, No, Steady, No, No...
## $ tolbutamide              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ pioglitazone             <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ rosiglitazone            <fct> No, No, No, No, No, No, No, No, No, Steady...
## $ acarbose                 <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ miglitol                 <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ troglitazone             <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ tolazamide               <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ examide                  <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ citoglipton              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ insulin                  <fct> No, Up, No, Up, Steady, Steady, Steady, No...
## $ glyburide.metformin      <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glipizide.metformin      <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glimepiride.pioglitazone <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ metformin.rosiglitazone  <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ metformin.pioglitazone   <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ change                   <fct> No, Ch, No, Ch, Ch, No, Ch, No, Ch, Ch, No...
## $ diabetesMed              <fct> No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes...
## $ readmitted               <fct> NO, >30, NO, NO, NO, >30, NO, >30, NO, NO,...
```

```
summary(encounter.data)
```

```
##   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      admission_type_id
##  Female         :54708   [70-80):26068   ?        :98569   Min.   :1.000
##  Male           :47055   [60-70):22483   [75-100) : 1336   1st Qu.:1.000
##  Unknown/Invalid:    3   [50-60):17256   [50-75)  :  897   Median :1.000
##                           [80-90):17197   [100-125):  625   Mean   :2.024
##                           [40-50): 9685   [125-150):  145   3rd Qu.:3.000
##                           [30-40): 3775   [25-50)  :   97   Max.   :8.000
##                           (Other): 5302   (Other)  :   97
##  discharge_disposition_id admission_source_id time_in_hospital   payer_code
##  Min.   : 1.000           Min.   : 1.000      Min.   : 1.000   ?      :40256
##  1st Qu.: 1.000           1st Qu.: 1.000      1st Qu.: 2.000   MC     :32439
##  Median : 1.000           Median : 7.000      Median : 4.000   HM     : 6274
##  Mean   : 3.716           Mean   : 5.754      Mean   : 4.396   SP     : 5007
##  3rd Qu.: 4.000           3rd Qu.: 7.000      3rd Qu.: 6.000   BC     : 4655
##  Max.   :28.000           Max.   :25.000      Max.   :14.000   MD     : 3532
##                                                                (Other): 9603
##               medical_specialty num_lab_procedures num_procedures
##  ?                     :49949   Min.   :  1.0      Min.   :0.00
##  InternalMedicine      :14635   1st Qu.: 31.0      1st Qu.:0.00
##  Emergency/Trauma      : 7565   Median : 44.0      Median :1.00
##  Family/GeneralPractice: 7440   Mean   : 43.1      Mean   :1.34
##  Cardiology            : 5352   3rd Qu.: 57.0      3rd Qu.:2.00
##  Surgery-General       : 3099   Max.   :132.0      Max.   :6.00
##  (Other)               :13726
##  num_medications number_outpatient number_emergency  number_inpatient
##  Min.   : 1.00   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:10.00   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :15.00   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   :16.02   Mean   : 0.3694   Mean   : 0.1978   Mean   : 0.6356
##  3rd Qu.:20.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000
##  Max.   :81.00   Max.   :42.0000   Max.   :76.0000   Max.   :21.0000
##
##      diag_1          diag_2          diag_3      number_diagnoses max_glu_serum
##  428    : 6862   276    : 6752   250    :11555   Min.   : 1.000   >200: 1485
##  414    : 6581   428    : 6662   401    : 8289   1st Qu.: 6.000   >300: 1264
##  786    : 4016   250    : 6071   276    : 5175   Median : 8.000   None:96420
##  410    : 3614   427    : 5036   428    : 4577   Mean   : 7.423   Norm: 2597
##  486    : 3508   401    : 3736   427    : 3955   3rd Qu.: 9.000
##  427    : 2766   496    : 3305   414    : 3664   Max.   :16.000
##  (Other):74419   (Other):70204   (Other):64551
##  A1Cresult     metformin     repaglinide     nateglinide     chlorpropamide
##  >7  : 3812   Down  :  575   Down  :    45   Down  :    11   Down  :     1
##  >8  : 8216   No    :81778   No    :100227   No    :101063   No    :101680
```

```
##  None:84748   Steady:18346   Steady:  1384   Steady:   668   Steady:    79
##  Norm: 4990   Up    : 1067   Up    :   110   Up    :    24   Up    :     6
##
##
##
##  glimepiride    acetohexamide    glipizide      glyburide     tolbutamide
##  Down  :  194   No    :101765   Down  :  560   Down  :  564   No    :101743
##  No    :96575   Steady:     1   No    :89080   No    :91116   Steady:    23
##  Steady: 4670                   Steady:11356   Steady: 9274
##  Up    :  327                   Up    :  770   Up    :  812
##
##
##
##  pioglitazone   rosiglitazone    acarbose        miglitol      troglitazone
##  Down  :  118   Down  :   87   Down  :     3   Down  :     5   No    :101763
##  No    :94438   No    :95401   No    :101458   No    :101728   Steady:     3
##  Steady: 6976   Steady: 6100   Steady:   295   Steady:    31
##  Up    :  234   Up    :  178   Up    :    10   Up    :     2
##
##
##
##   tolazamide     examide     citoglipton   insulin      glyburide.metformin
##  No    :101727   No:101766   No:101766   Down  :12218   Down  :     6
##  Steady:    38                           No    :47383   No    :101060
##  Up    :     1                           Steady:30849   Steady:   692
##                                           Up    :11316   Up    :     8
##
##
##
##  glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone
##  No    :101753       No    :101765            No    :101764
##  Steady:    13       Steady:     1            Steady:     2
##
##
##
##
##
##
##  metformin.pioglitazone change     diabetesMed readmitted
##  No    :101765          Ch:47011   No :23403   <30:11357
##  Steady:     1          No:54755   Yes:78363   >30:35545
##                                                 NO :54864
##
##
```

```
##
##
```

```
# number of medications, medical specialty, diabetes med, insulin, change,
```

```
myvars <- c('race', 'weight', 'payer_code', 'medical_specialty')
amountRecords <- length(rownames(encounter.data))
tableMissingValues <- encounter.data %>%
  summarise_each_(funs("TotalMissing" = sum(.=='?', na.rm = TRUE),
                       "MissingValuesPercentage" = round(sum(.=='?', na.rm = TRUE)/amour
  t()
```

```
## Warning: `summarise_each_()` is deprecated as of dplyr 0.7.0.
## Please use `across()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
tableMissingValues <- data.frame(missingValues = tableMissingValues[1:4,],
                                 missingValuesPer = tableMissingValues[5:8,])
rownames(tableMissingValues) <- myvars

formattable(tableMissingValues)
```

missingValues

missingValuesPer

race

2273

2.23

weight

98569

96.86

payer_code

40256

39.56

medical_specialty

49949

49.08

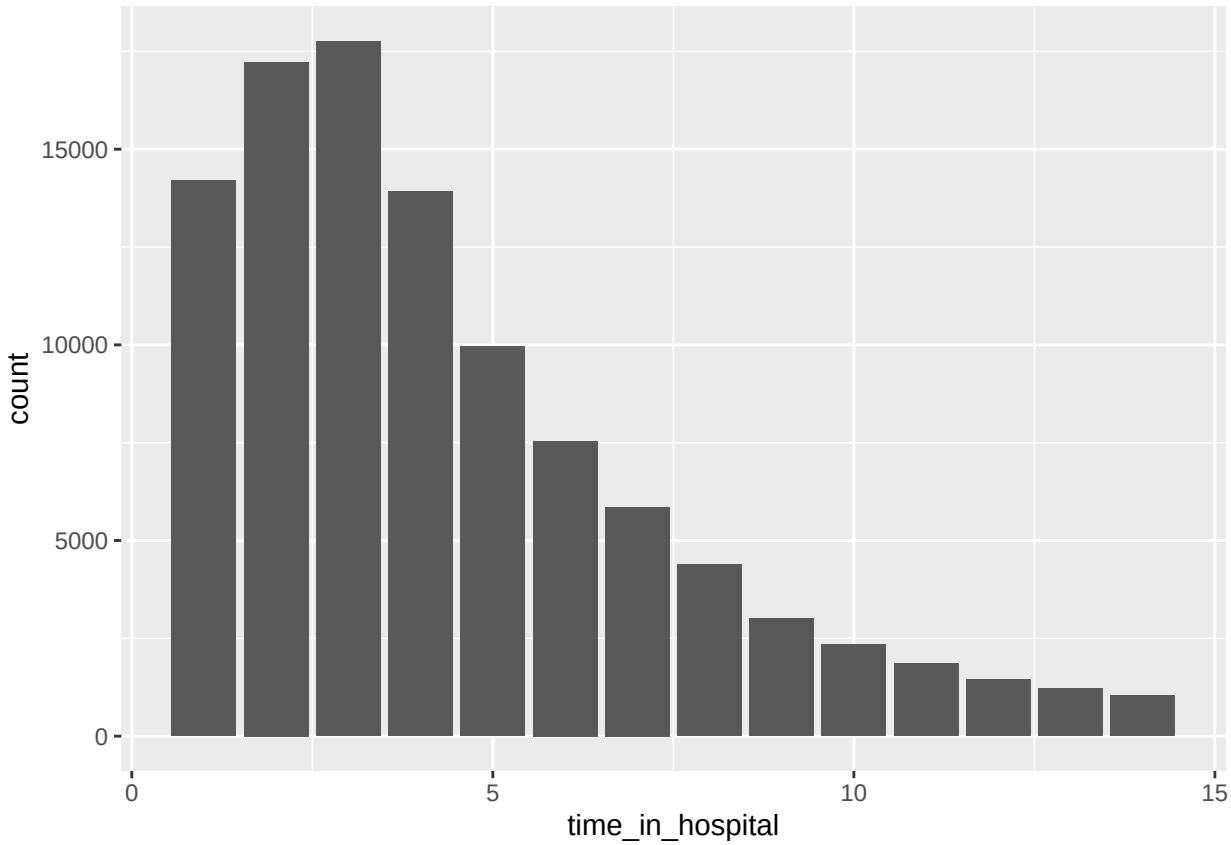
```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
```

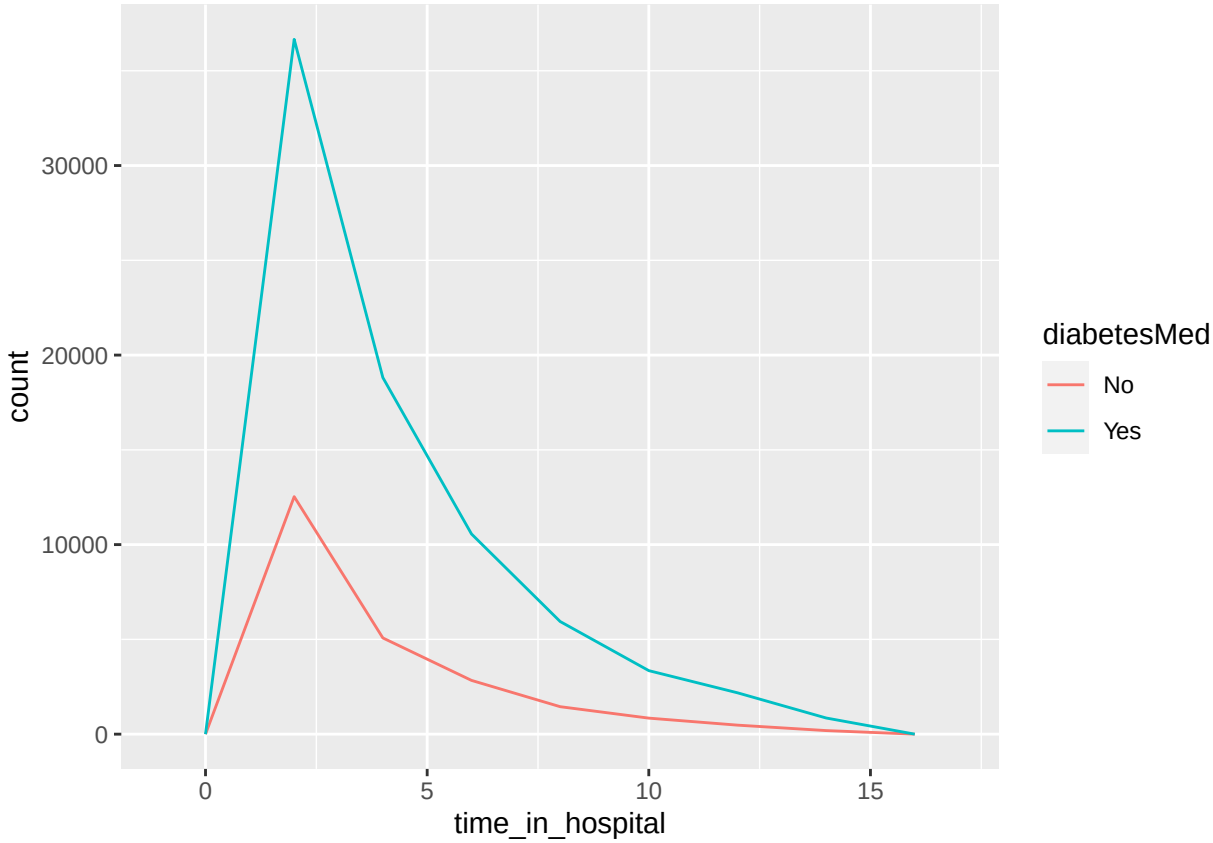
```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

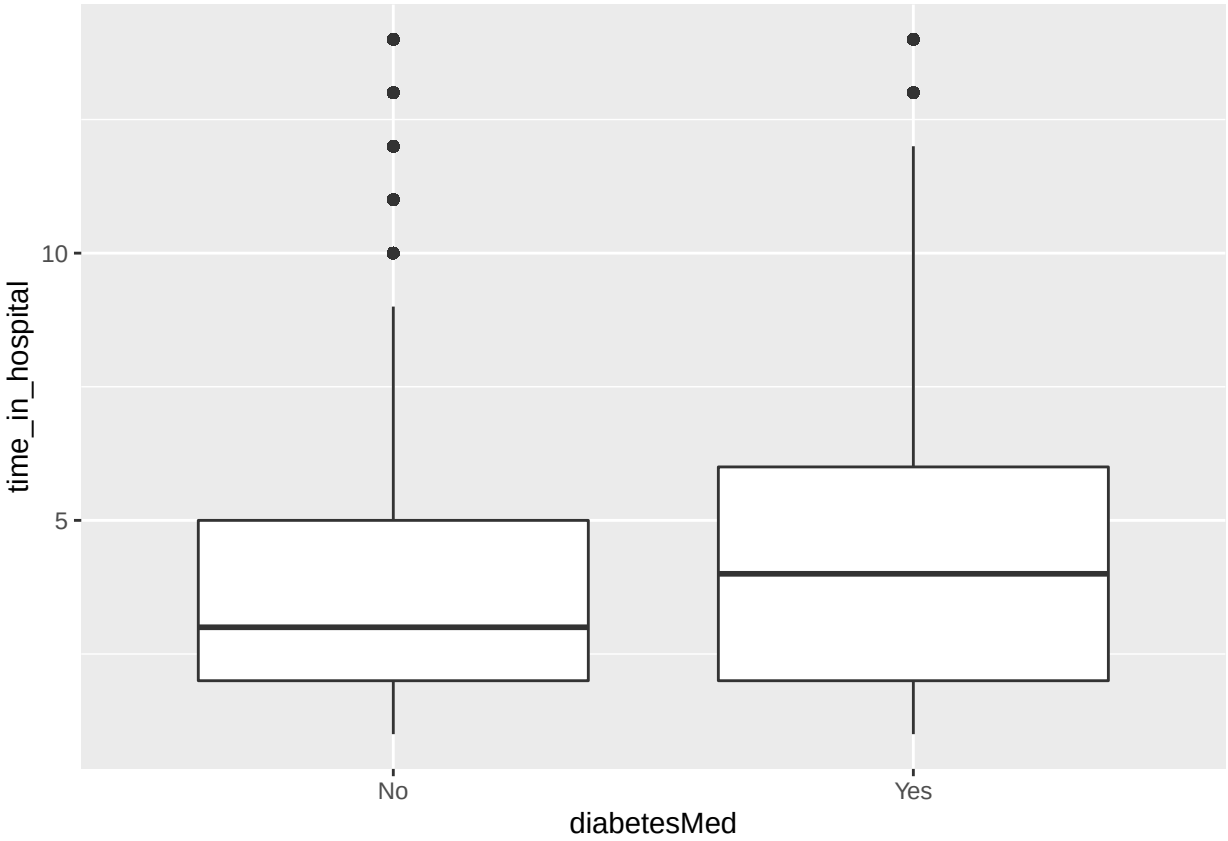
```
# We see a sharp decline in the number of days stayed in hospital
ggplot(data = encounter.data) +
  geom_bar(mapping = aes(x = time_in_hospital))
```




```
ggplot(data = encounter.data, mapping = aes(x = time_in_hospital, color = diabetesMed))
  geom_freqpoly(binwidth = 2)
```




```
ggplot(data = encounter.data, mapping = aes(x = diabetesMed, y = time_in_hospital)) +
  geom_boxplot()
```


time_in_hospital
10
5
No
Yes
diabetesMed