Agrupamento com K-means sobre dados Médicos

Marcos Vinícius dos Santos Ferreira

2018-04-16

Contents

Database Diabetes

Diabetes 130-US hospitals for years 1999-2008 Data Set

Estes dados foram preparados para analisar os fatores relacionados à readmissão, bem como outros resultados referentes aos pacientes com diabetes.

Informação do dataset

O conjunto de dados representa 10 anos (1999-2008) de atendimento clínico em 130 hospitais dos EUA e redes de distribuição integradas. Inclui mais de 50 atributos multivariados que representam os resultados do paciente e do hospital. O dataset contém 100000 instâncias. Informações foram extraídas do banco de dados para encontros que satisfizeram os seguintes critérios.

```
1. É um encontro de internação (internação hospitalar).
2. É um encontro diabético.
3. O tempo de internação foi de no mínimo 1 dia e no máximo 14 dias.
4. Testes laboratoriais foram realizados durante o encontro.
5. Medicamentos foram administrados durante o encontro.
```

Os dados contêm atributos como número do paciente, raça, gênero, idade, tipo de internação, tempo no hospital, especialidade médica do médico admitido, número de exames laboratoriais realizados, resultado do exame de HbA1c, diagnóstico, número de medicamentos, medicamentos diabéticos, número de pacientes ambulatoriais , internação e visitas de emergência no ano anterior à hospitalização, etc.

Dificuldades Aprensentadas:

```
* Heterogêneas e difíceis em termos de valores ausentes
* Registros incompletos ou inconsistentes
* Alta dimensionalidade, entendida pelo número de características e  por sua complexidade.
```

Paper

Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records

- Impacto da Medida de HbA1c nas Taxas de Readmissão Hospitalar: Análise de 70.000 Registros de Pacientes com Base de Dados Clínicos.
- **Hipótesi:** Nossa hipótese é que a medida da HbA1c está associada a uma redução nas taxas de readmissão em indivíduos internados no hospital.

Minha Metodologia

Fazer o uso do aprendizado de máquina não supervisionado para identificar relação entre dados clínicos de diabetes e fornecer indícios de quais fatores influemciam mais na doença.

Objetivo

- Quais fatores influenciam e ou apontam indícios sobre o avanço ou cura da da diabete?

Metodos

- Usar o Aprendizado de Máquina não supervisionado com o algoritmo k-means para indentificar padrões que possam identificar padrões no dataset que evidenciam tais indícios.

Hipótese de quais fatores mais influenciam no diabetes(avanço da doença).

- **Atributo - Tipo - Valores Ausentes**
- Idade - Nominal - 0%
- Discharge disposition - Nominal - 0%
- Admission source - Nominal - 0%
- Time in hospital - Numeric - 0%
- Medical specialty - Nominal - 59% missing
- Number of lab procedures (Numeric)
- Number of procedures (Numeric)
- Number of medications (Numeric)
- Number of emergency visits (Numeric)

- Diagnosis 1 - Nominal - 0%
- Diagnosis 2 - Nominal - 0%
- Diagnosis 3 - Nominal - 1%
- Número de Diagnósticos - Numeric - 0%
- Glucose serum test result - Nominal - 0%

Carregando os dados

```
# lendo o dataset
data.diabetes <- read.csv('../../data/dataset_diabetes/diabetic_data.csv')

# Visualizando o dataset
str(data.diabetes)
```

```
## 'data.frame':    101766 obs. of  50 variables:
##  $ encounter_id            : int  2278392 149190 64410 500364 16680 35754 55842 63768 12522 15738 ..
##  $ patient_nbr             : int  8222157 55629189 86047875 82442376 42519267 82637451 84259809 11488
##  $ race                    : Factor w/ 6 levels "?","AfricanAmerican",..: 4 4 2 4 4 4 4 4 4 4 ...
##  $ gender                  : Factor w/ 3 levels "Female","Male",..: 1 1 1 2 2 2 2 2 1 1 ...
##  $ age                     : Factor w/ 10 levels "[0-10)","[10-20)",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ weight                  : Factor w/ 10 levels "?","[0-25)","[100-125)",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ admission_type_id       : int  6 1 1 1 1 2 3 1 2 3 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ payer_code              : Factor w/ 18 levels "?","BC","CH",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ medical_specialty       : Factor w/ 73 levels "?","AllergyandImmunology",..: 39 1 1 1 1 1 1 1 1 2
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ number_emergency        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ number_inpatient        : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ diag_1                  : Factor w/ 717 levels "?","10","11",..: 126 145 456 556 56 265 265 278 25
##  $ diag_2                  : Factor w/ 749 levels "?","11","110",..: 1 81 80 99 26 248 248 316 262 48
##  $ diag_3                  : Factor w/ 790 levels "?","11","110",..: 1 123 768 250 88 88 772 88 231 3
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
##  $ max_glu_serum           : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ A1Cresult               : Factor w/ 4 levels ">7",">8","None",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ metformin               : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 3 2 2 2 ...
##  $ repaglinide             : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ nateglinide             : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ chlorpropamide          : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ glimepiride             : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 3 2 2 2 ...
##  $ acetohexamide           : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ glipizide               : Factor w/ 4 levels "Down","No","Steady",..: 2 2 3 2 3 2 2 2 3 2 ...
##  $ glyburide               : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 3 2 2 ...
##  $ tolbutamide             : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pioglitazone            : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ rosiglitazone           : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 3 ...
##  $ acarbose                : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ miglitol                : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ troglitazone            : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ tolazamide              : Factor w/ 3 levels "No","Steady",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ examide                 : Factor w/ 1 level "No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ citoglipton             : Factor w/ 1 level "No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ insulin                 : Factor w/ 4 levels "Down","No","Steady",..: 2 4 2 4 3 3 3 2 3 3 ...
##  $ glyburide.metformin     : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ glipizide.metformin     : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ glimepiride.pioglitazone: Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ metformin.rosiglitazone : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ metformin.pioglitazone  : Factor w/ 2 levels "No","Steady": 1 1 1 1 1 1 1 1 1 1 ...
##  $ change                  : Factor w/ 2 levels "Ch","No": 2 1 2 1 1 2 1 2 1 1 ...
##  $ diabetesMed             : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 2 2 2 2 2 ...
##  $ readmitted              : Factor w/ 3 levels "<30",">30","NO": 3 2 3 3 3 2 3 2 3 3 ...
```

Pré-Processamento

```
# selecionar os atributos que por hipótese podem ser relevantes
data.diabetes.rel <- data.diabetes[,c('age', 'discharge_disposition_id', 'admission_source_id',
                                      'time_in_hospital', 'medical_specialty', 'num_lab_procedures', 'nu
                                      'num_medications', 'number_outpatient', 'diag_1', 'diag_2', 'diag_
                                      'number_diagnoses', 'max_glu_serum')]

# estrutura dos novos dados
str(data.diabetes.rel)
```

```
## 'data.frame':    101766 obs. of  14 variables:
##  $ age                     : Factor w/ 10 levels "[0-10)","[10-20)",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ medical_specialty       : Factor w/ 73 levels "?","AllergyandImmunology",..: 39 1 1 1 1 1 1 1 1 20
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ diag_1                  : Factor w/ 717 levels "?","10","11",..: 126 145 456 556 56 265 265 278 25
##  $ diag_2                  : Factor w/ 749 levels "?","11","110",..: 1 81 80 99 26 248 248 316 262 48
##  $ diag_3                  : Factor w/ 790 levels "?","11","110",..: 1 123 768 250 88 88 772 88 231 3
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
##  $ max_glu_serum           : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
# selecionar as 200 primeiras instancias

data.diabetes.rel.sel <- data.diabetes.rel[1:200,]

str(data.diabetes.rel.sel)
```

```
## 'data.frame':    200 obs. of  14 variables:
##  $ age                     : Factor w/ 10 levels "[0-10)","[10-20)",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ medical_specialty       : Factor w/ 73 levels "?","AllergyandImmunology",..: 39 1 1 1 1 1 1 1 1 20
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
```

```
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ diag_1                  : Factor w/ 717 levels "?","10","11",..: 126 145 456 556 56 265 265 278 25
##  $ diag_2                  : Factor w/ 749 levels "?","11","110",..: 1 81 80 99 26 248 248 316 262 48
##  $ diag_3                  : Factor w/ 790 levels "?","11","110",..: 1 123 768 250 88 88 772 88 231 3
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
##  $ max_glu_serum           : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
# tranformar as colunas de palavras em numeros
med.esp <-as.numeric(data.diabetes.rel.sel$medical_specialty)

data.diabetes.rel.sel['medical_specialty'] <- as.integer(med.esp)

str(data.diabetes.rel.sel)
```

```
## 'data.frame':    200 obs. of  14 variables:
##  $ age                     : Factor w/ 10 levels "[0-10)","[10-20)",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ medical_specialty       : int  39 1 1 1 1 1 1 1 1 20 ...
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ diag_1                  : Factor w/ 717 levels "?","10","11",..: 126 145 456 556 56 265 265 278 25
##  $ diag_2                  : Factor w/ 749 levels "?","11","110",..: 1 81 80 99 26 248 248 316 262 48
##  $ diag_3                  : Factor w/ 790 levels "?","11","110",..: 1 123 768 250 88 88 772 88 231 3
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
##  $ max_glu_serum           : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
# dados de idade

data.diabetes.rel.sel$age
```

```
##   [1] [0-10)   [10-20)  [20-30)  [30-40)  [40-50)  [50-60)  [60-70)
##   [8] [70-80)  [80-90)  [90-100) [40-50)  [60-70)  [40-50)  [80-90)
##  [15] [60-70)  [60-70)  [50-60)  [50-60)  [70-80)  [70-80)  [50-60)
##  [22] [60-70)  [70-80)  [80-90)  [70-80)  [50-60)  [80-90)  [50-60)
##  [29] [20-30)  [80-90)  [60-70)  [70-80)  [70-80)  [60-70)  [70-80)
##  [36] [60-70)  [70-80)  [60-70)  [70-80)  [50-60)  [70-80)  [40-50)
##  [43] [70-80)  [50-60)  [80-90)  [40-50)  [70-80)  [70-80)  [50-60)
##  [50] [60-70)  [50-60)  [70-80)  [40-50)  [50-60)  [60-70)  [60-70)
##  [57] [50-60)  [40-50)  [80-90)  [70-80)  [70-80)  [50-60)  [40-50)
##  [64] [80-90)  [50-60)  [90-100) [10-20)  [80-90)  [50-60)  [50-60)
##  [71] [70-80)  [50-60)  [60-70)  [70-80)  [70-80)  [70-80)  [60-70)
##  [78] [60-70)  [50-60)  [50-60)  [70-80)  [50-60)  [50-60)  [60-70)
##  [85] [60-70)  [40-50)  [40-50)  [60-70)  [60-70)  [40-50)  [70-80)
##  [92] [70-80)  [40-50)  [50-60)  [60-70)  [70-80)  [70-80)  [70-80)
##  [99] [50-60)  [30-40)  [70-80)  [60-70)  [30-40)  [60-70)  [70-80)
## [106] [80-90)  [50-60)  [80-90)  [60-70)  [50-60)  [50-60)  [60-70)
## [113] [40-50)  [70-80)  [70-80)  [30-40)  [60-70)  [70-80)  [60-70)
## [120] [60-70)  [70-80)  [40-50)  [40-50)  [70-80)  [50-60)  [30-40)
## [127] [80-90)  [30-40)  [20-30)  [60-70)  [50-60)  [60-70)  [60-70)
## [134] [70-80)  [90-100) [70-80)  [60-70)  [60-70)  [50-60)  [80-90)
## [141] [30-40)  [60-70)  [80-90)  [20-30)  [90-100) [50-60)  [50-60)
```

```
## [148] [50-60)  [50-60)  [70-80)  [60-70)  [40-50)  [50-60)  [70-80)
## [155] [50-60)  [60-70)  [60-70)  [50-60)  [60-70)  [80-90)  [80-90)
## [162] [50-60)  [80-90)  [50-60)  [80-90)  [40-50)  [80-90)  [30-40)
## [169] [50-60)  [50-60)  [60-70)  [60-70)  [70-80)  [50-60)  [70-80)
## [176] [70-80)  [80-90)  [40-50)  [70-80)  [40-50)  [40-50)  [70-80)
## [183] [50-60)  [50-60)  [60-70)  [70-80)  [80-90)  [40-50)  [40-50)
## [190] [70-80)  [70-80)  [20-30)  [40-50)  [60-70)  [20-30)  [60-70)
## [197] [60-70)  [40-50)  [30-40)  [20-30)
## 10 Levels: [0-10) [10-20) [20-30) [30-40) [40-50) [50-60) ... [90-100)
```

Tratando os valores de diagnostico.

```
# função que troca de valores
troca.valor<-function(estrutura, valor, troca){

  temp <- as.vector(estrutura)

  temp[which(temp==valor)]=troca

  temp
}

data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[0-10)", "10")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[10-20)", "15")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[20-30)", "25")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[30-40)", "35")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[40-50)", "45")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[50-60)", "55")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[60-70)", "65")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[70-80)", "75")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[80-90)", "85")
data.diabetes.rel.sel$age <- troca.valor(data.diabetes.rel.sel$age,"[90-100)", "95")

# vendoos valores convertidos pela media de idades
data.diabetes.rel.sel$age
```

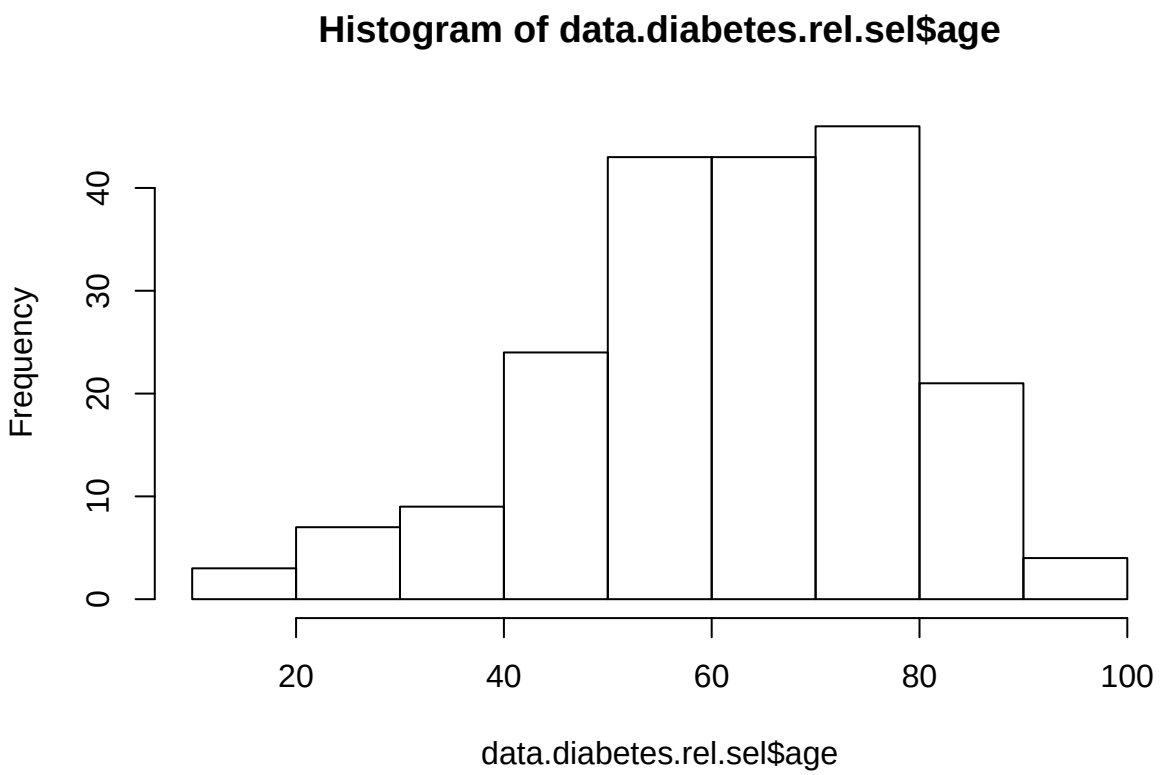
```
##   [1] "10" "15" "25" "35" "45" "55" "65" "75" "85" "95" "45" "65" "45" "85"
##  [15] "65" "65" "55" "55" "75" "75" "55" "65" "75" "85" "75" "55" "85" "55"
##  [29] "25" "85" "65" "75" "75" "65" "75" "65" "75" "65" "75" "55" "75" "45"
##  [43] "75" "55" "85" "45" "75" "75" "55" "65" "55" "75" "45" "55" "65" "65"
##  [57] "55" "45" "85" "75" "75" "55" "45" "85" "55" "95" "15" "85" "55" "55"
##  [71] "75" "55" "65" "75" "75" "75" "65" "65" "55" "55" "75" "55" "55" "65"
##  [85] "65" "45" "45" "65" "65" "45" "75" "75" "45" "55" "65" "75" "75" "75"
##  [99] "55" "35" "75" "65" "35" "65" "75" "85" "55" "85" "65" "55" "55" "65"
## [113] "45" "75" "75" "35" "65" "75" "65" "65" "75" "45" "45" "75" "55" "35"
## [127] "85" "35" "25" "65" "55" "65" "65" "75" "95" "75" "65" "65" "55" "85"
## [141] "35" "65" "85" "25" "95" "55" "55" "55" "55" "75" "65" "45" "55" "75"
## [155] "55" "65" "65" "55" "65" "85" "85" "55" "85" "55" "85" "45" "85" "35"
## [169] "55" "55" "65" "65" "75" "55" "75" "75" "85" "45" "75" "45" "45" "75"
## [183] "55" "55" "65" "75" "85" "45" "45" "75" "75" "25" "45" "65" "25" "65"
## [197] "65" "45" "35" "25"
```

```
# convertendo os nuvemos para interios
data.diabetes.rel.sel$age <- as.integer(data.diabetes.rel.sel$age)

data.diabetes.rel.sel$age
```

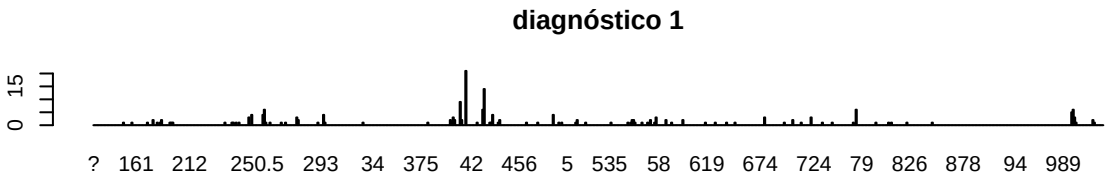
```
##   [1] 10 15 25 35 45 55 65 75 85 95 45 65 45 85 65 65 55 55 75 75 55 65 75
##  [24] 85 75 55 85 55 25 85 65 75 75 65 75 65 75 65 75 55 75 45 75 55 85 45
##  [47] 75 75 55 65 55 75 45 55 65 65 55 45 85 75 75 55 45 85 55 95 15 85 55
##  [70] 55 75 55 65 75 75 75 65 65 55 55 75 55 55 65 65 45 45 65 65 45 75 75
##  [93] 45 55 65 75 75 75 55 35 75 65 35 65 75 85 55 85 65 55 55 65 45 75 75
## [116] 35 65 75 65 65 75 45 45 75 55 35 85 35 25 65 55 65 65 75 95 75 65 65
## [139] 55 85 35 65 85 25 95 55 55 55 55 75 65 45 55 75 55 65 65 55 65 85 85
## [162] 55 85 55 85 45 85 35 55 55 65 65 75 55 75 75 85 45 75 45 45 75 55 55
## [185] 65 75 85 45 45 75 75 25 45 65 25 65 65 45 35 25
```

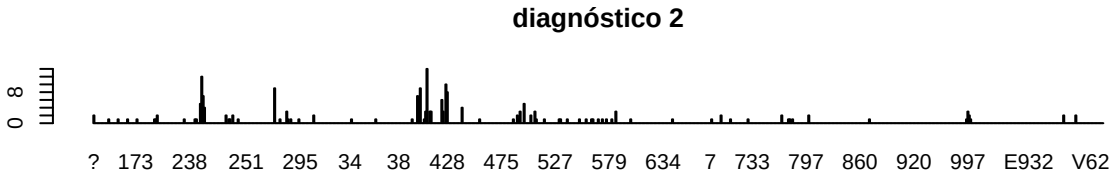
vendo a distribuição dos valores com o histograma.

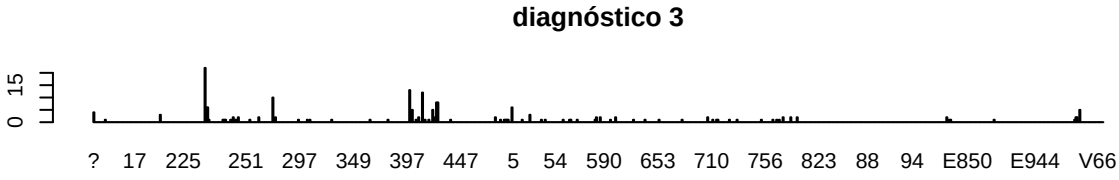
```
hist(data.diabetes.rel.sel$age)
```




```
# tratar valores nulos dos diagnóstico - solução - média.

# visualizando os dados dos 3 dianosticos
par(mfrow=c(3,1))
plot(data.diabetes.rel.sel$diag_1, main='diagnóstico 1')
plot(data.diabetes.rel.sel$diag_2, main='diagnóstico 2')
plot(data.diabetes.rel.sel$diag_3, main='diagnóstico 3')
```





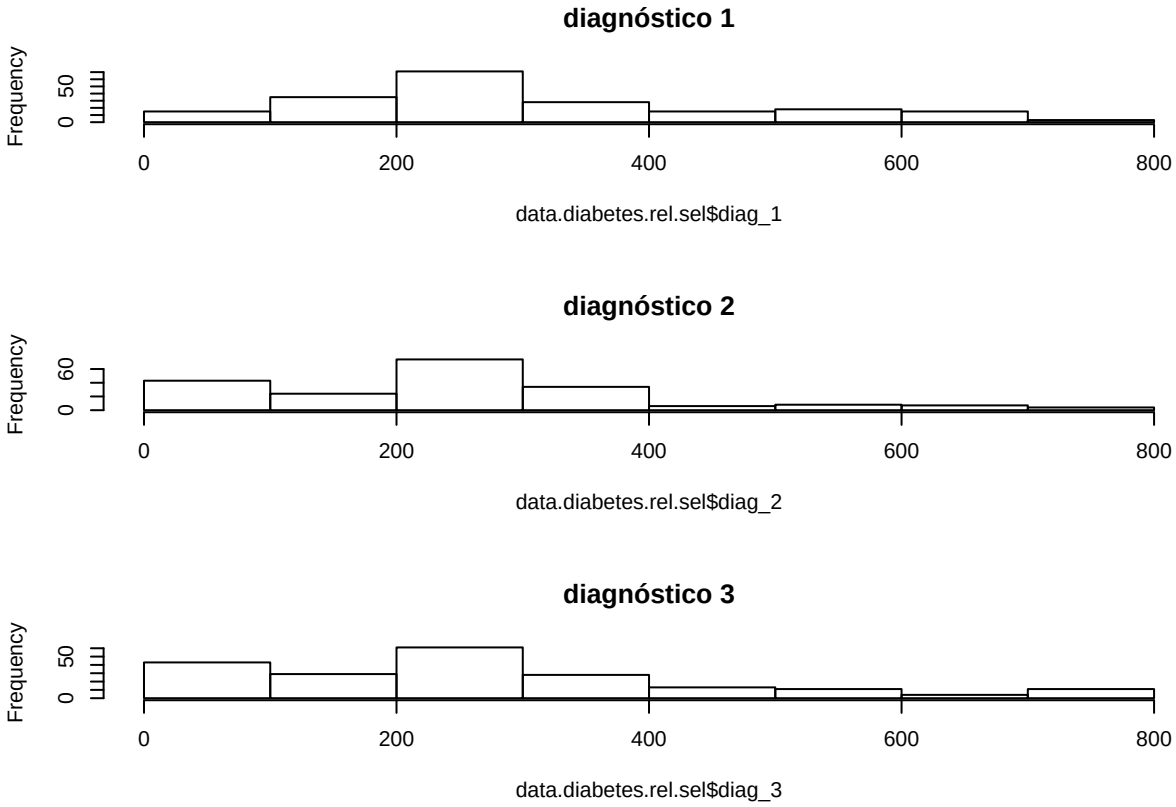



```
#tratar valores desconhecidos - ?
## possso remover esses valores ou aplicar um media
head(data.diabetes.rel.sel$diag_1)
```

```
## [1] 250.83 276    648    8      197    414
## 717 Levels: ? 10 11 110 112 114 115 117 131 133 135 136 141 142 143 ... V71
```

```
# solution 1
# data.diabetes.rel.sel$diag_1 <- troca.valor(data.diabetes.rel.sel$diag_1,'?', 414)
# data.diabetes.rel.sel$diag_1 <- as.integer(data.diabetes.rel.sel$diag_1)
# data.diabetes.rel.sel$diag_1[is.na(data.diabetes.rel.sel$diag_1)] <- 0
# data.diabetes.rel.sel$diag_1

# solution 2
# temp <- troca.valor(data.diabetes.rel.sel$diag_1,'?', 414)
# temp <- as.integer(temp)
# temp[is.na(temp)] <- 0
# hist(temp)

# solution 3
# diagnostico 1
data.diabetes.rel.sel$diag_1 <- as.integer(data.diabetes.rel.sel$diag_1)
data.diabetes.rel.sel$diag_1[is.na(data.diabetes.rel.sel$diag_1)] <- 0

# diagnostico 2
data.diabetes.rel.sel$diag_2 <- as.integer(data.diabetes.rel.sel$diag_2)
data.diabetes.rel.sel$diag_2[is.na(data.diabetes.rel.sel$diag_2)] <- 0

# diagnostico 3
data.diabetes.rel.sel$diag_3 <- as.integer(data.diabetes.rel.sel$diag_3)
data.diabetes.rel.sel$diag_3[is.na(data.diabetes.rel.sel$diag_3)] <- 0
```

```r
# visualizando os dados
par(mfrow=c(3,1))
hist(data.diabetes.rel.sel$diag_1, main='diagnóstico 1')
hist(data.diabetes.rel.sel$diag_2, main='diagnóstico 2')
hist(data.diabetes.rel.sel$diag_3, main='diagnóstico 3')
```




```r
# conferindo a nova estrutura do dataset
str(data.diabetes.rel.sel)
```

```
## 'data.frame':    200 obs. of  14 variables:
##  $ age                     : int  10 15 25 35 45 55 65 75 85 95 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ medical_specialty       : int  39 1 1 1 1 1 1 1 1 20 ...
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ diag_1                  : num  126 145 456 556 56 265 265 278 254 284 ...
##  $ diag_2                  : num  1 81 80 99 26 248 248 316 262 48 ...
##  $ diag_3                  : num  1 123 768 250 88 88 772 88 231 319 ...
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
##  $ max_glu_serum           : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```r
# removendo a coluna max_glu_serum por ter muitos valores discrepantes
data.diabetes.rel.sel = subset(data.diabetes.rel.sel, select = -c(max_glu_serum))
```

Análise dos dados

Primeiro temos que explorar e visualizar os dados.

```
# estrutura dos meus dados
str(data.diabetes.rel.sel)
```

```
## 'data.frame':    200 obs. of  13 variables:
##  $ age                     : int  10 15 25 35 45 55 65 75 85 95 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
##  $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
##  $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
##  $ medical_specialty       : int  39 1 1 1 1 1 1 1 1 20 ...
##  $ num_lab_procedures      : int  41 59 11 44 51 31 70 73 68 33 ...
##  $ num_procedures          : int  0 0 5 1 0 6 1 0 2 3 ...
##  $ num_medications         : int  1 18 13 16 8 16 21 12 28 18 ...
##  $ number_outpatient       : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ diag_1                  : num  126 145 456 556 56 265 265 278 254 284 ...
##  $ diag_2                  : num  1 81 80 99 26 248 248 316 262 48 ...
##  $ diag_3                  : num  1 123 768 250 88 88 772 88 231 319 ...
##  $ number_diagnoses        : int  1 9 6 7 5 9 7 8 8 8 ...
```

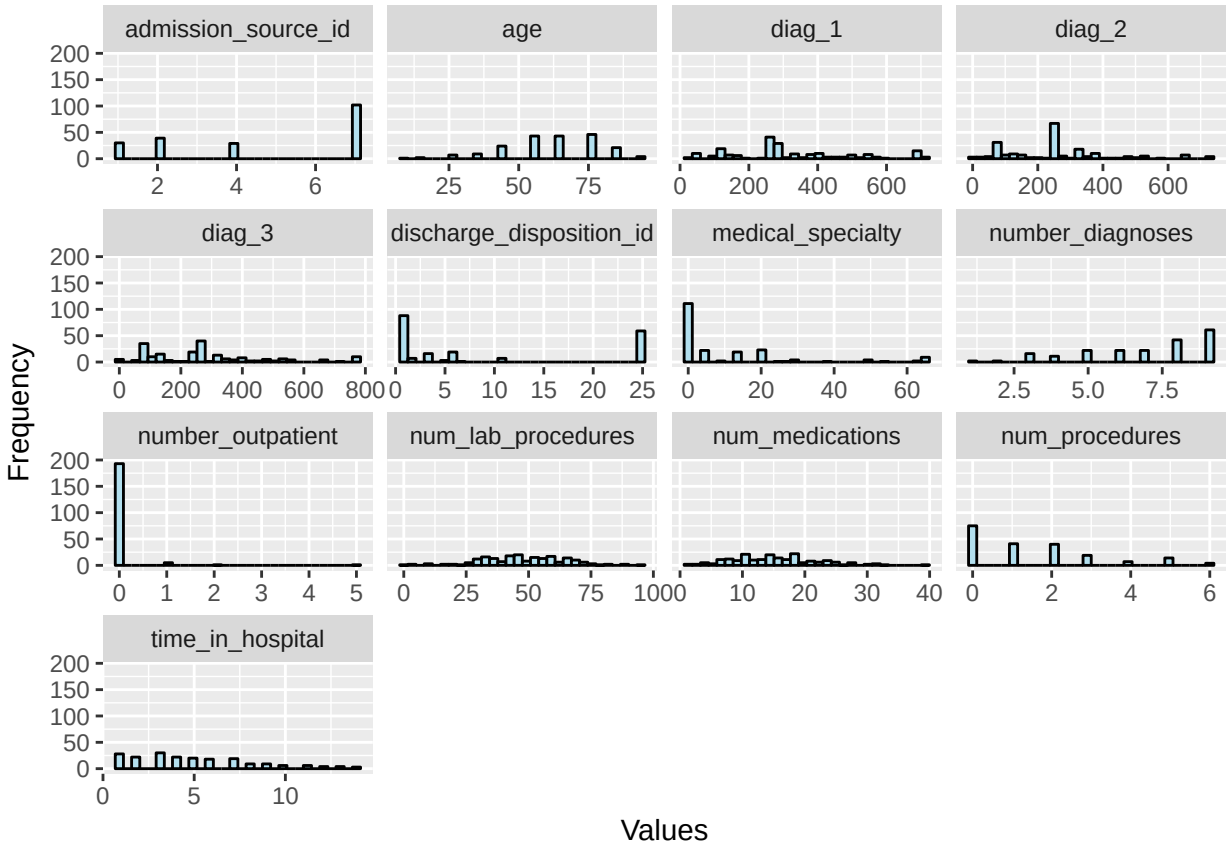
Todas as colunas são expressas como numéricas ou inteiras. E quanto à distribuição estatística?

```
summary(data.diabetes.rel.sel)
```

```
##       age        discharge_disposition_id admission_source_id
##  Min.   :10.00   Min.   : 1.00            Min.   :1.00
##  1st Qu.:55.00   1st Qu.: 1.00            1st Qu.:2.00
##  Median :65.00   Median : 3.00            Median :7.00
##  Mean   :61.92   Mean   : 9.19            Mean   :4.69
##  3rd Qu.:75.00   3rd Qu.:25.00            3rd Qu.:7.00
##  Max.   :95.00   Max.   :25.00            Max.   :7.00
##  time_in_hospital medical_specialty num_lab_procedures num_procedures
##  Min.   : 1.00    Min.   : 1.00     Min.   : 1.00      Min.   :0.0
##  1st Qu.: 2.75    1st Qu.: 1.00     1st Qu.:36.00      1st Qu.:0.0
##  Median : 4.00    Median : 1.00     Median :47.00      Median :1.0
##  Mean   : 5.06    Mean   :10.58     Mean   :48.17      Mean   :1.5
##  3rd Qu.: 7.00    3rd Qu.:13.25     3rd Qu.:59.00      3rd Qu.:2.0
##  Max.   :14.00    Max.   :66.00     Max.   :96.00      Max.   :6.0
##  num_medications number_outpatient     diag_1          diag_2
##  Min.   : 1.00   Min.   :0.00      Min.   : 22.0   Min.   :  1.0
##  1st Qu.:10.00   1st Qu.:0.00      1st Qu.:226.5   1st Qu.:135.0
##  Median :15.00   Median :0.00      Median :278.0   Median :248.0
##  Mean   :15.19   Mean   :0.06      Mean   :319.2   Mean   :255.6
##  3rd Qu.:19.00   3rd Qu.:0.00      3rd Qu.:407.0   3rd Qu.:320.0
##  Max.   :39.00   Max.   :5.00      Max.   :711.0   Max.   :729.0
##      diag_3      number_diagnoses
##  Min.   :  1.0   Min.   :1.000
##  1st Qu.:111.8   1st Qu.:5.000
##  Median :258.0   Median :8.000
##  Mean   :277.7   Mean   :6.895
##  3rd Qu.:344.2   3rd Qu.:9.000
##  Max.   :772.0   Max.   :9.000
```

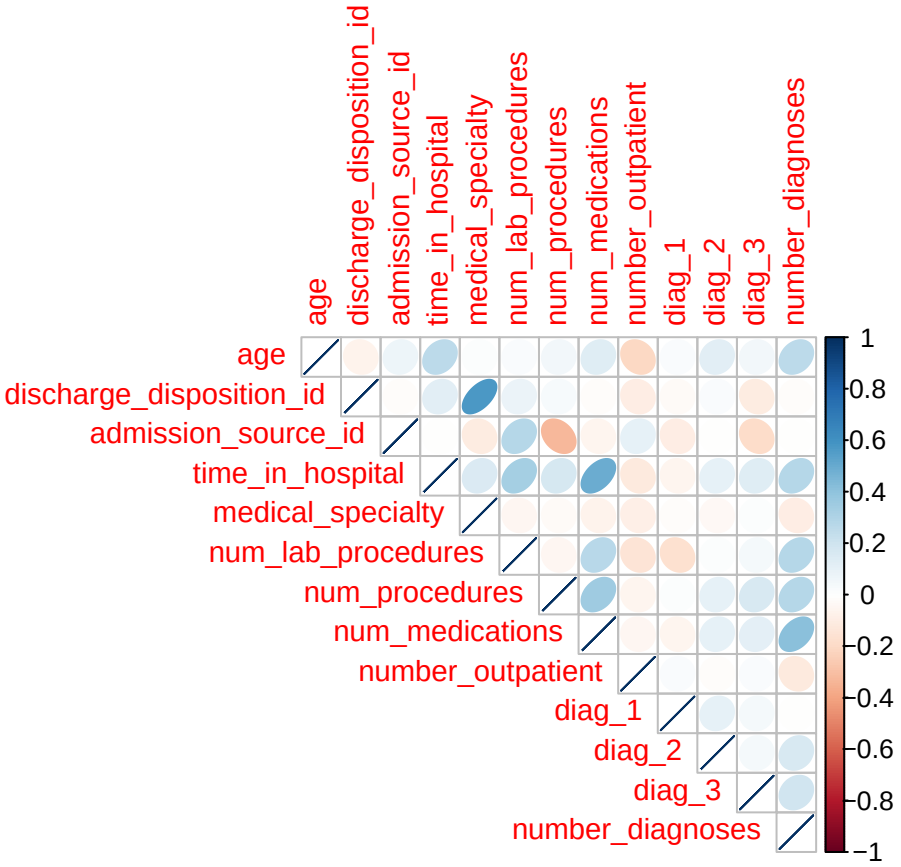
```
#load library
library(tidyverse)
```

```
library(corrplot)
library(gridExtra)
library(GGally)

# Histograma de cada atributo

data.diabetes.rel.sel %>%
  gather(Attributes, value, 1:13) %>%
  ggplot(aes(x=value)) +
  geom_histogram(fill="lightblue2", colour="black") +
  facet_wrap(~Attributes, scales="free_x") +
  labs(x="Values", y="Frequency")
```



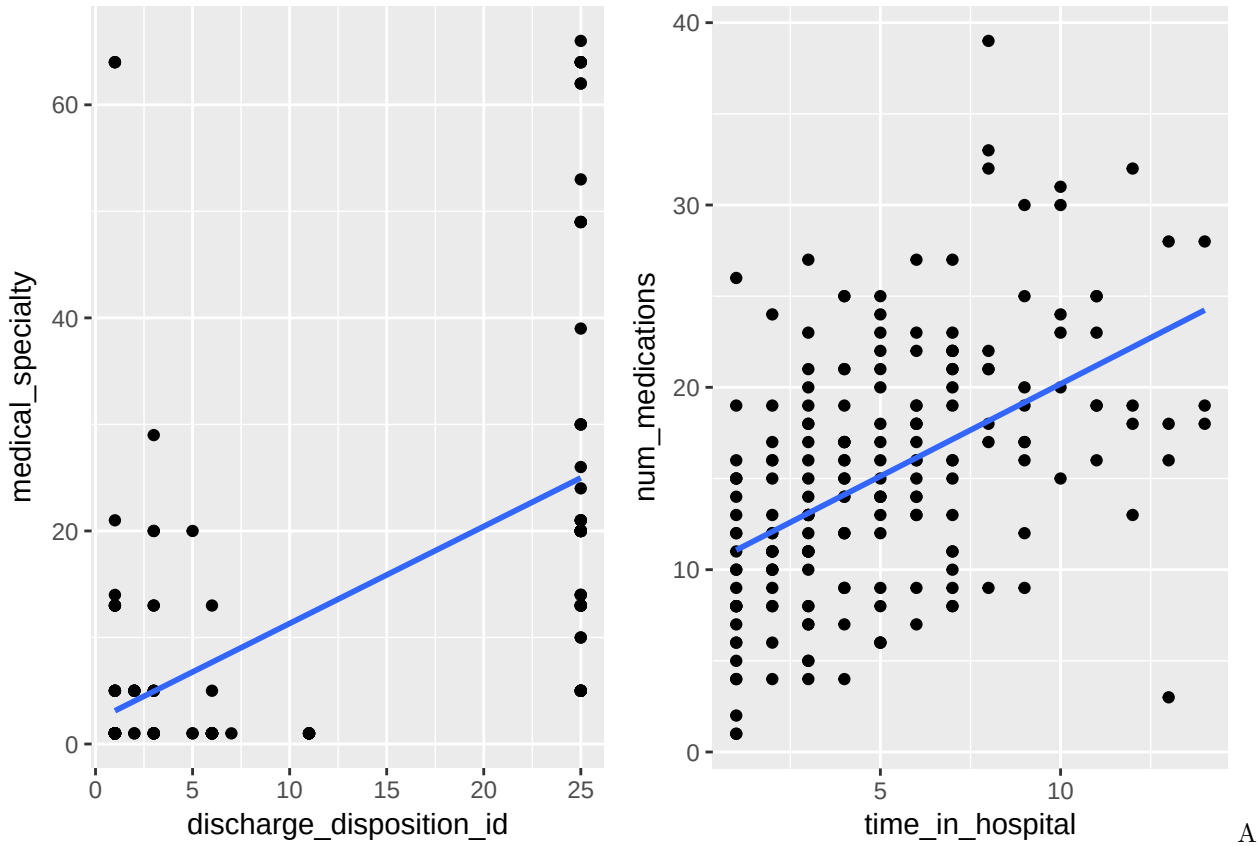
Qual é a relação entre os diferentes atributos? Podemos usar a função **corrplot()** para criar uma exibição gráfica de uma matriz de correlação.

```
# Matrix de correlação
corrplot(cor(data.diabetes.rel.sel), type="upper", method="ellipse", tl.cex=0.9)
```



Existe uma forte correlação linear entre os atributos: *discharge-disposition-id* e *medical-specialty*, *time-in-hospital* e *num-medications* . Podemos modelar a relação entre essas duas variáveis ajustando uma equação linear.

```
# Relationship entre as variaveis que mais tem correlação

# discrarge disposition id and medical specialty
plt1 <- ggplot(data.diabetes.rel.sel, aes(x=discharge_disposition_id, y=medical_specialty)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)

#
plt2 <- ggplot(data.diabetes.rel.sel, aes(x=time_in_hospital, y=num_medications)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)

# plt3 <- ggplot(data.diabetes.rel.sel, aes(x=num_medications, y=number_diagnoses)) +
#   geom_point() +
#   geom_smooth(method="lm", se=FALSE)

grid.arrange(plt1, plt2, ncol=2)
```

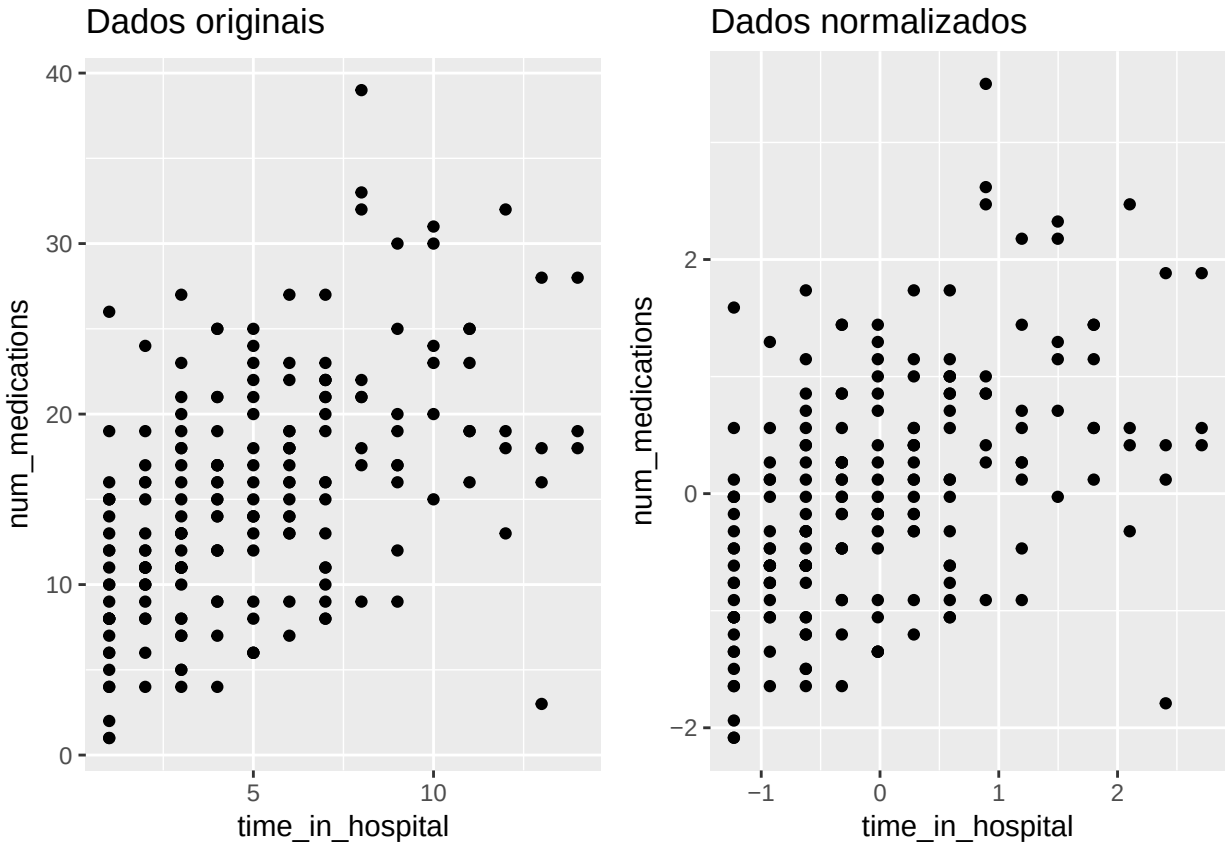


A correlação linear se aplica mais entre as variáveis *num-medications* e *time-in-hospital*. Agora que fizemos uma análise de dados exploratória, podemos preparar os dados para executar o algoritmo k-means.

Preparação dos dados

Temos que normalizar as variáveis para expressá-las no mesmo intervalo de valores. Em outras palavras, normalização significa ajustar os valores medidos em diferentes escalas para uma escala comum.

```
# Normalização
data.diabetes.rel.sel.norm <- as.data.frame(scale(data.diabetes.rel.sel))

# dados originais
data <- ggplot(data.diabetes.rel.sel, aes(x=time_in_hospital, y=num_medications)) +
  geom_point() +
  labs(title="Dados originais")

# dados normalizados
data.norm <- ggplot(data.diabetes.rel.sel.norm, aes(x=time_in_hospital, y=num_medications)) +
  geom_point() +
  labs(title="Dados normalizados")

# subplot
grid.arrange(data,data.norm, ncol=2)
```



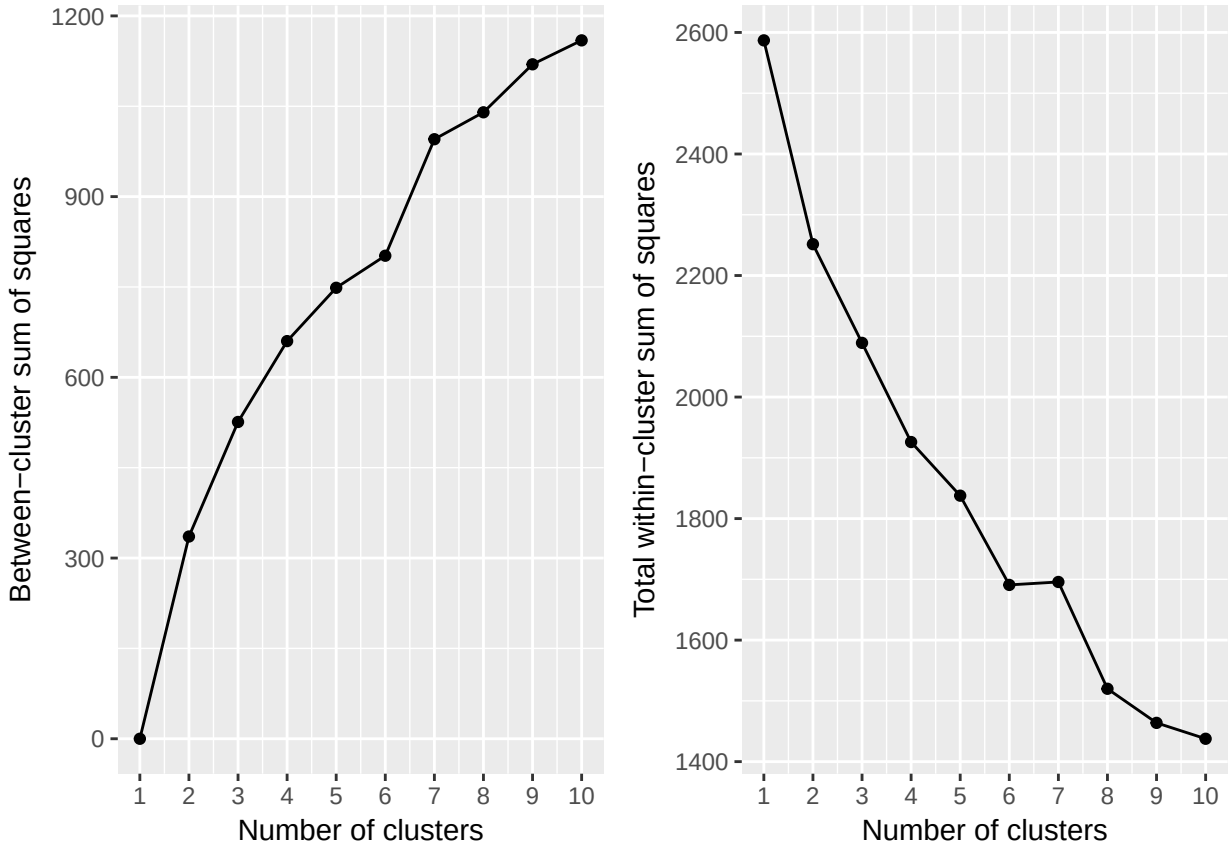
Quantos Clustes ?

O algoritmo K-means para encontrar similiaridade entre os grupos, precisa do parâmetro **k**, que é a quantidade de grupos a serem escolhidos.

Qual é o valor ideal para k? Deve-se escolher um número de clusters para que adicionar outro cluster não forneça uma partição muito melhor dos dados. Em algum momento, o ganho cairá, dando um ângulo no gráfico (critério do cotovelo). O número de clusters é escolhido neste momento. No nosso caso, é claro que 3 é o valor apropriado para k. Para estudar graficamente qual valor de k nos dá a melhor partição, podemos traçar entre e tot.withinss vs Choice de k.

```
bss <- numeric()
wss <- numeric()

# rodar o algoritmo com diferentes valores de K
set.seed(1234)

for(i in 1:10){

  # para casa k, calcula betweenss e tot.withinss
  bss[i] <- kmeans(data.diabetes.rel.sel.norm, centers=i)$betweenss
  wss[i] <- kmeans(data.diabetes.rel.sel.norm, centers=i)$tot.withinss
}

# Soma entre os quadrados dos quadrados vs Escolha de k

d3 <- qplot(1:10, bss, geom=c("point", "line"),
```

```
            xlab="Number of clusters", ylab="Between-cluster sum of squares") +
  scale_x_continuous(breaks=seq(0, 10, 1))

# Soma total de quadrados dentro do cluster vs Escolha de k

d4 <- qplot(1:10, wss, geom=c("point", "line"),
            xlab="Number of clusters", ylab="Total within-cluster sum of squares") +
  scale_x_continuous(breaks=seq(0, 10, 1))

# subplot
grid.arrange(d3, d4, ncol=2)
```



Execução do k-means

Com o algoritmo k-means identificamos a quantidade de grupos que meus dados formam e com isso podemos trazer semântica aos dados e tirar conclusões. De acordo com o método do cotovelo, a quantidade de clusters é igual 6, ou seja, o meu parâmetro **k**.

```
# selecionar somente os atributos de clusters
data.diabetes.rel.sel.norm.atts <- data.diabetes.rel.sel.norm[,c('time_in_hospital', 'num_medications')]

# Execução do K-means com k  = 6
set.seed(1234)
kmenas.diabetes <- kmeans(data.diabetes.rel.sel.norm.atts, centers=6)
```

*Vetor de inteiros indicando o custer ao qual cada ponto é alocado.

```
kmenas.diabetes$cluster
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   1   5   1   5   1   5   3   5   6   2   2   4   5   6   1   2   5   1
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   3   3   5   1   1   2   5   1   5   1   2   5   5   1   5   5   3   1
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   1   5   3   1   1   4   2   1   3   1   2   4   3   6   5   5   2   1
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   1   5   2   1   3   4   5   5   5   5   5   2   1   1   1   4   3   6
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   5   6   5   1   1   3   1   5   1   5   5   1   5   5   3   5   5   5
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   5   3   5   1   5   1   3   3   1   5   6   3   5   1   1   1   2   1
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
##   3   5   3   3   5   5   2   1   5   6   6   5   3   6   5   2   3   3
## 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
##   4   1   1   2   5   3   3   5   4   2   2   5   6   1   1   3   4   1
## 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
##   2   1   1   1   3   6   1   2   5   3   1   2   3   1   3   5   1   3
## 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   4   1   5   4   4   5   5   5   1   4   3   1   4   2   4   6   4   1
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
##   4   1   1   1   5   2   1   1   5   4   1   4   1   1   1   5   5   3
## 199 200
##   3   6
```

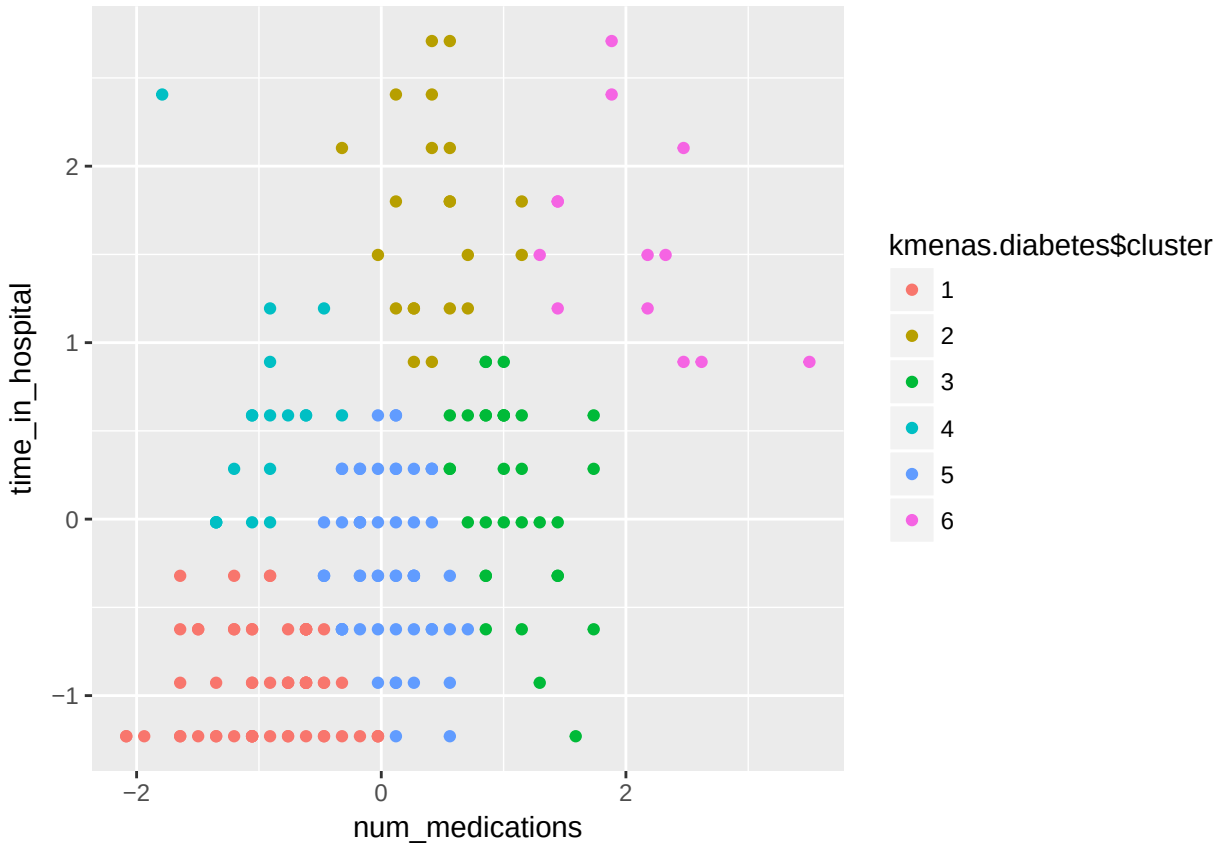
A matriz com o centro dos clusters.

```
kmenas.diabetes$centers
```

```
##   time_in_hospital num_medications
## 1       -0.9373370     -0.93354148
## 2        1.7133527      0.42773466
## 3        0.1238588      1.07053520
## 4        0.5710214     -0.97436810
## 5       -0.2562670      0.05679566
## 6        1.5668383      2.08700025
```

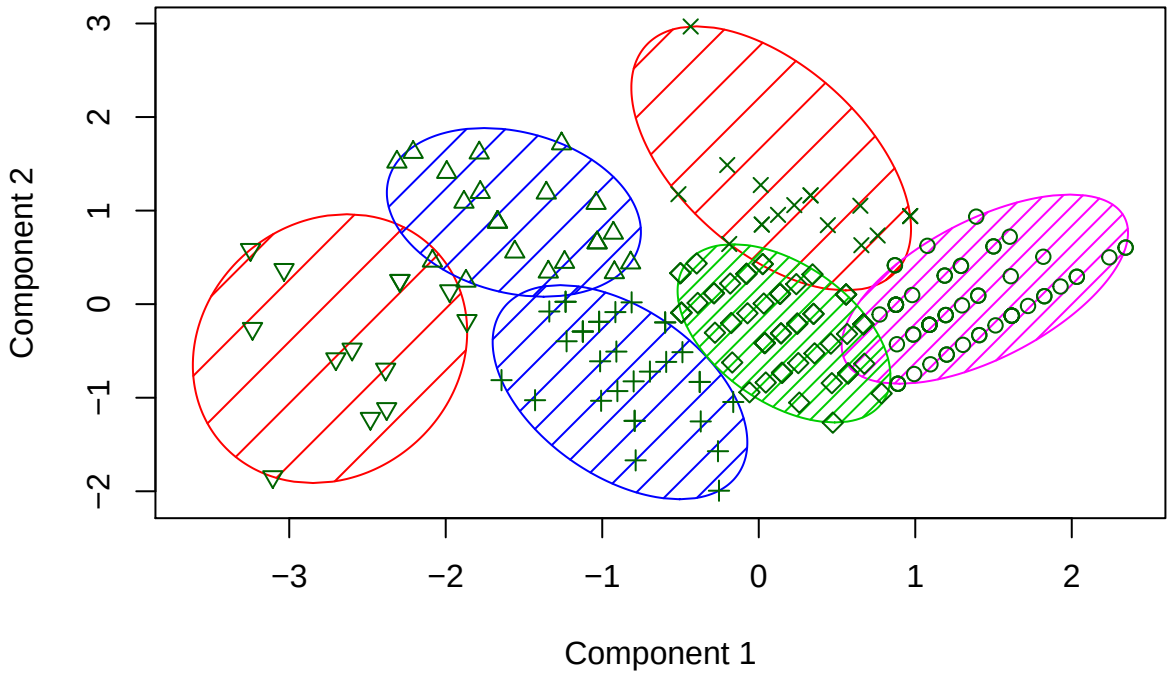
Visualizando o agrupamento.

```
# melhorar os labels do meu grafico
kmenas.diabetes$cluster <- as.factor(kmenas.diabetes$cluster)

p1 <- ggplot(data.diabetes.rel.sel.norm, aes(num_medications, time_in_hospital, color = kmenas.diabetes$

p1
```

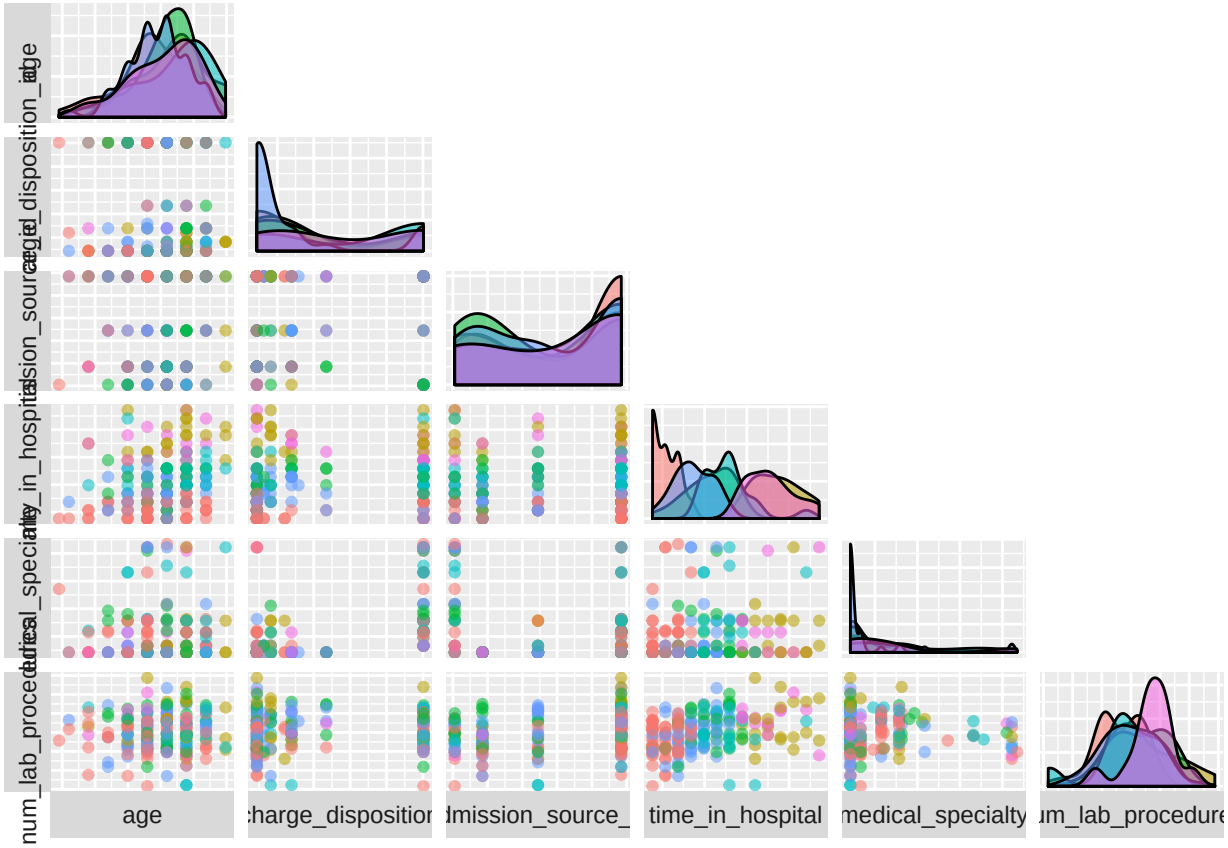


Visualizado com o clusplot com todos as 5 variáveis do grupo

```
library("cluster")
clusplot(data.diabetes.rel.sel.norm.atts, kmenas.diabetes$cluster, color = TRUE, shade= TRUE, labels = 
```




```
# Clustering
ggpairs(cbind(data.diabetes.rel.sel.norm, Cluster=as.factor(kmenas.diabetes$cluster)),
        columns=1:6, aes(colour=Cluster, alpha=0.5),
        lower=list(continuous="points"),
        upper=list(continuous="blank"),
        axisLabels="none", switch="both")
```



Validação

Aqui validamos o quão o método conseguio agrupar conforme um índice de validação. Validar com critérios internos, pois vai medir a qualidade do agrupamento com base nos dados originais, ja que, os dados não posuem rótulos ou estruturas definidas.

- **Critério Interno**
  - Mede o grau que uma partição obtida representa a estrutura presente nos dados;

Consusão

- Os grupos podem indicar a variedade do estado de saúde das pessoas com diabétes, ou seja, com diferentes graus, leve, moderado, normal, grave e diabete melitus.
- Os vastos grupos indicam que o tratamento merece mais cuidados.
- Pode-se concluir que diante das diversas características dos paciente, a readmissão dos pacientes acontecem nos mais diversos casos da diabete, é uma doença severa e que merece uma atenção e tratamento adequado, sendo grande parte resposável o própio paciente a seus limites.

Referências

- Origem do Dataset
- Descrição dos Atributos

- Silhueta
- Indroduction Data Mining
- Chapter 8