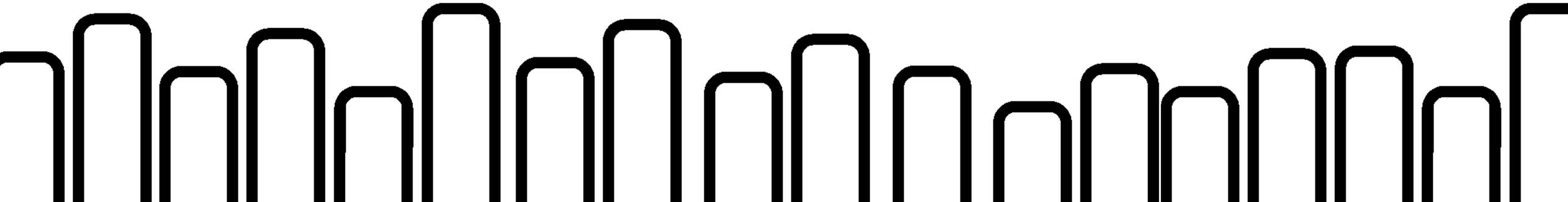
C19-109-M-DATA-BI

EDA

Exploratory data analysis




DATOS BRUTOS
CONCLUSIONES
ANÁLISIS EXPLORATORIO DE DATOS (EDA)
ANÁLISIS DESCRIPTIVO
AJUSTE DE TIPOS DE VARIABLES
DETECCIÓN Y TRATAMIENTO DE DATOS AUSENTES
IDENTIFICACIÓN DE DATOS ATÍPICOS
CORRELACIÓN DE VARIABLES

1.- Pregunta

¿Qué tipo de persona tiene probabilidad de reingreso?

2.- Generalidades del dataset

```
> dim(database)
[1] 101766     50
```

- 101,767 Registros(filas).
- 50 Características o variables(columnas).
- Algunas variables importantes: race, gender, age, time_in_hospital, num_lab_procedures, num_medications, readmitted, entre otras.

3.- Tipo de datos encontrados.

- 37 chr
- 13 int

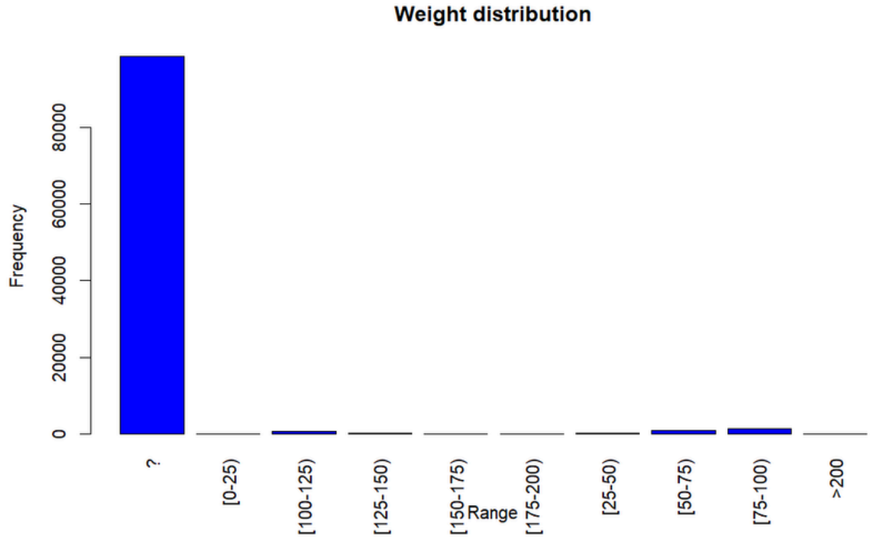
4.- Estadística descriptiva

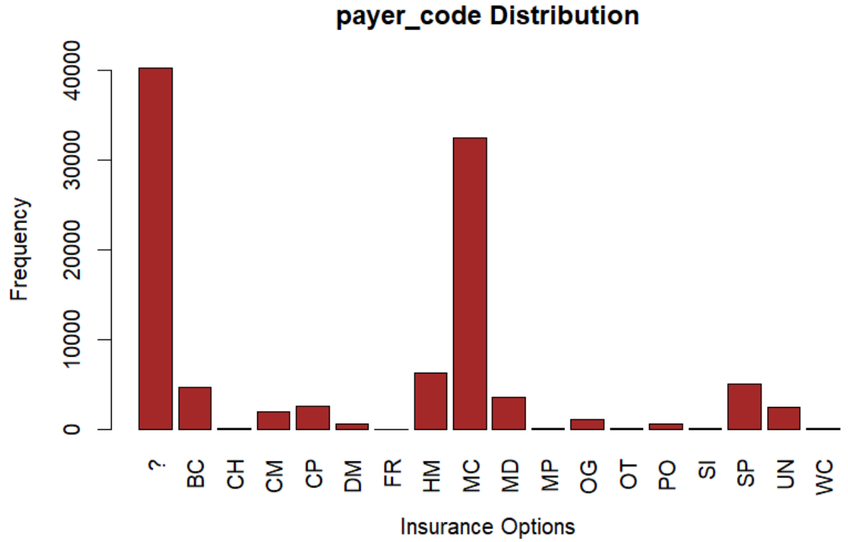
Estadística Descriptiva

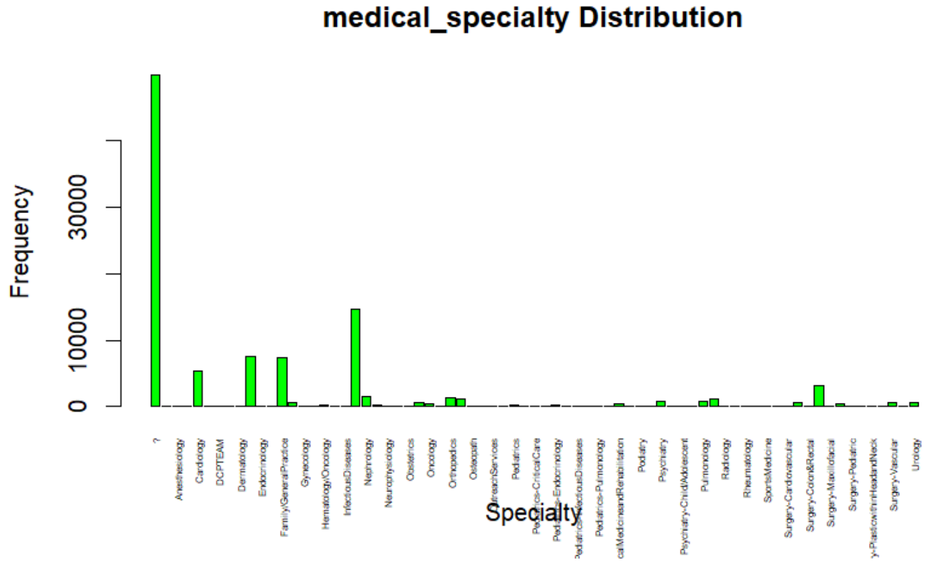
- Estadística Descriptiva
  - Medidas
    - Tendencia Central
      - Media
      - Mediana
      - Moda
    - Dispersión
      - Rango
      - Varianza
      - Desviación Estándar
      - Coeficiente de Variación
    - Posición
      - Deciles
      - Percentiles
      - Cuartiles
    - Forma
      - Sesgo
      - Curtosis
  - Gráficos
    - No agrupados
      - Diagrama de puntos
      - Diagrama de cajas y bigotes
      - Diagrama de Dispersión
    - Agrupados
      - Histograma
      - Polígono de Frecuencias
      - Ojiva
      - Diagrama de sectores o "diagrama de pastel"

5.-Visualización

Frecuencia







7. Sumarización (conclusiones).

DATOS NULOS DEL >5% O > 20%

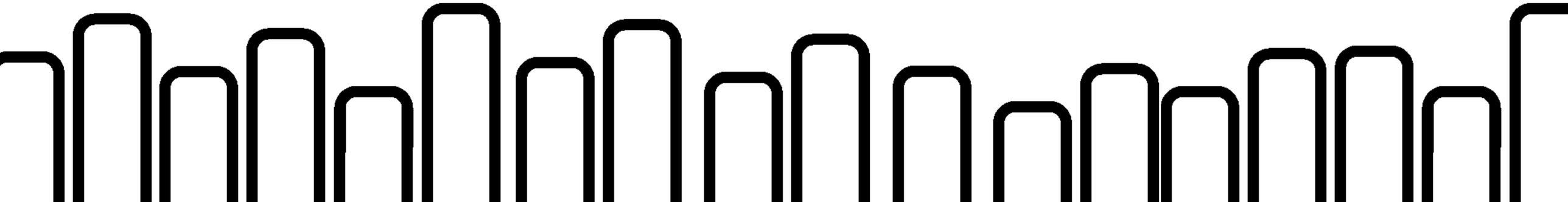
Descartar:

- Weight ? >90%
- payer_code ? >30%
- medical_specialty ? > %45
- max_glu_serum ? > 80%
- A1Cresult ? > 80%
- examide
- citoglipton

Considerar

- diag_1
- diag_2
- diag_3

PARTE II
–
LIMPIEZA



1. Eliminación de variables

```
> dim(database)
[1] 101766     50
```

Datos nulos 30%

1. Weight ? >90%
2. payer_code ? >30%
3. medical_specialty ? > %45
4. max_glu_serum ? > 80%
5. A1Cresult ? > 80%

1 valor

1. examide (1)
2. citoglipton (1)
3. metformin.rosiglitazone(1)

2 valores

1. acetohexamide(frecuencia 1)
2. tolbutamide (22)
3. troglitazone(3)
4. glimepiride.pioglitazone(1)
5. metformin.pioglitazone (1)

****Valores***

1. *diag_1 (713)*
2. *diag_2 (740)*
3. *diag_3 (786)*

****valores < 100***

1. glipizide.metformin
2. chlorpropamide
3. miglitol
4. tolazamide

2. Convertir “?” a NA

```
> print(columns_with_na)
  race diag_1 diag_2 diag_3
  2273     21    358   1423
> 
```

PORCENTAJE

```
      race      diag_1                    diag_2                    diag_3
2.23355541  0.02063558                0.35178743                1.39830592
```

3. Eliminar filas/registros con NA

Porcentaje de eliminar registros con valores nulos

```
> print(porcentaje_eliminado)
[1] 3.648566
```

```
> dim(clean
[1] 98053
```

	gender	Freq
1	Female	52833
2	Male	45219
3	Unknown/Invalid	1

4. Clasificar variables

Convertir a numéricos: Variables categóricas con orden intrínseco.
(Ej. primaria<sec<uni).
Crear variables dummy: Variables categóricas sin orden intrínseco.
(Ej. rojo, verde, azul,).

```
> print(numeric_columns)
 [1] "encounter_id"          "patient_nbr"           "admission_type_id"       "discharge_disposition_id"
 [5] "admission_source_id"   "time_in_hospital"      "num_lab_procedures"      "num_procedures"
 [9] "num_medications"       "number_outpatient"     "number_emergency"        "number_inpatient"
[13] "number_diagnoses"
```

Dummy

1. race
2. gender
3. age
4. metformin
5. repaglinide
6. nateglinide
7. glimepiride
8. glipizide
9. glyburide
10. pioglitazone
11. rosiglotazone
12. acarbose
13. insulin
14. gliyburide.metformin
15. readmitted

numéricos

1. change (2)
2. diabetesMed(2)

4. Conversion numérica/dummy

```
> str(clean_data)
'data.frame':   98052 obs. of  79 variables:
 $ encounter_id             : int  149190 64410 500364 16680 35754 55842 8
 $ patient_nbr              : int  55629189 86047875 82442376 42519267 8
 $ admission_type_id        : int  1 1 1 1 2 3 1 2 3 1 ...
 $ discharge_disposition_id: int  1 1 1 1 1 1 1 1 3 1 ...
 $ admission_source_id      : int  7 7 7 7 2 2 7 4 4 7 ...
 $ time_in_hospital         : int  3 2 2 1 3 4 5 13 12 9 ...
 $ num_lab_procedures       : int  59 11 44 51 31 70 73 68 33 47 ...
 $ num_procedures           : int  0 5 1 0 6 1 0 2 3 2 ...
 $ num_medications          : int  18 13 16 8 16 21 12 28 18 17 ...
 $ number_outpatient        : int  0 2 0 0 0 0 0 0 0 0 ...
 $ number_emergency         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ number_inpatient         : int  0 1 0 0 0 0 0 0 0 0 ...
 $ number_diagnoses         : int  9 6 7 5 9 7 8 8 8 9 ...
 $ change                   : num  1 0 1 1 0 1 0 1 1 0 ...
 $ diabetesMed              : num  1 1 1 1 1 1 1 1 1 1 ...
 $ raceAfricanAmerican      : num  0 1 0 0 0 0 0 0 0 1 ...
 $ raceAsian                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ raceCaucasian            : num  1 0 1 1 1 1 1 1 1 0 ...
 $ raceHispanic             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ raceOther                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ genderFemale             : num  1 1 0 0 0 0 0 1 1 1 ...
 $ genderMale               : num  0 0 1 1 1 1 1 0 0 0 ...
 $ age[0-10)                : num  0 0 0 0 0 0 0 0 0 0 ...
 $ age[10-20)               : num  1 0 0 0 0 0 0 0 0 0 ...
 $ age[20-30)               : num  0 1 0 0 0 0 0 0 0 0 ...
 $ age[30-40)               : num  0 0 1 0 0 0 0 0 0 0 ...
 $ age[40-50)               : num  0 0 0 1 0 0 0 0 0 1 ...
 $ age[50-60)               : num  0 0 0 0 1 0 0 0 0 0 ...
 $ age[60-70)               : num  0 0 0 0 0 1 0 0 0 0 ...
 $ age[70-80)               : num  0 0 0 0 0 0 1 0 0 0 ...
 $ age[80-90)               : num  0 0 0 0 0 0 0 1 0 0 ...
 $ age[90-100)              : num  0 0 0 0 0 0 0 0 1 0 ...
 $ metforminDown            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ metforminNo              : num  1 1 1 1 1 0 1 1 1 1 ...
 $ metforminSteady          : num  0 0 0 0 0 1 0 0 0 0 ...
 $ metforminUp              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ repaglinideDown          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ repaglinideNo            : num  1 1 1 1 1 1 1 1 1 1 ...
 $ repaglinideSteady        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ repaglinideUp            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ nateglinideDown          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ nateglinideNo            : num  1 1 1 1 1 1 1 1 1 1 ...
 $ nateglinideSteady        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ nateglinideUp            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ glimepirideDown          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ glimepirideNo            : num  1 1 1 1 1 0 1 1 1 1 ...
 $ glimepirideSteady        : num  0 0 0 0 0 1 0 0 0 0 ...
 $ glimepirideUp            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ glipizideDown            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ glipizideNo              : num  1 0 1 0 1 1 1 0 1 1 ...
 $ glipizideSteady          : num  0 1 0 1 0 0 0 1 0 0 ...
 $ glipizideUp              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ glyburideDown            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ glyburideNo              : num  1 1 1 1 1 1 0 1 1 1 ...
 $ glyburideSteady          : num  0 0 0 0 0 0 1 0 0 0 ...
 $ glyburideUp              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ pioglitazoneDown         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ pioglitazoneNo           : num  1 1 1 1 1 1 1 1 1 1 ...
 $ pioglitazoneSteady       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ pioglitazoneUp           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ rosiglitazoneDown        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ rosiglitazoneNo          : num  1 1 1 1 1 1 1 1 0 1 ...
 $ rosiglitazoneSteady      : num  0 0 0 0 0 0 0 0 1 0 ...
 $ rosiglitazoneUp          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ acarboseDown             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ acarboseNo               : num  1 1 1 1 1 1 1 1 1 1 ...
 $ acarboseSteady           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ acarboseUp               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ insulinDown              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ insulinNo                : num  0 1 0 0 0 0 1 0 0 0 ...
 $ insulinSteady            : num  0 0 0 1 1 1 0 1 1 1 ...
 $ insulinUp                : num  1 0 1 0 0 0 0 0 0 0 ...
 $ metforminDown.1          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ metforminNo.1            : num  1 1 1 1 1 0 1 1 1 1 ...
 $ metforminSteady.1        : num  0 0 0 0 0 1 0 0 0 0 ...
 $ metforminUp.1            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ readmitted<30            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ readmitted>30            : num  1 0 0 0 1 0 1 0 0 1 ...
 $ readmittedNO             : num  0 1 1 1 0 1 0 1 1 0 ...
```

```
> dim(clean_data)
[1] 98052    79
```