DataAuto\_task

Variant 3

Evgenii Berdinskikh

2024-10-24

Table of Contents

# Чтение данных

В вашем варианте нужно использовать датасет framingham.

list.files("data/raw")

## [1] "framingham.csv"

data <- read\_csv("data/raw/framingham.csv", show\_col\_types = FALSE)

# Выведите общее описание данных

skim(data)

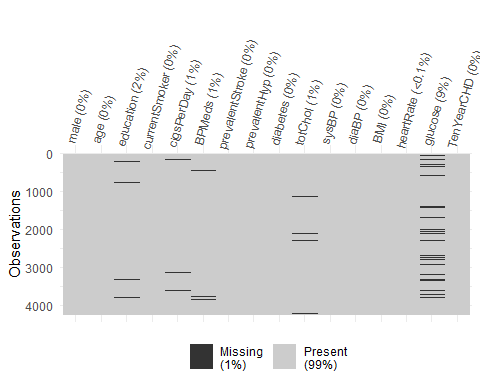
Data summary

|  |  |
| --- | --- |
| Name | data |
| Number of rows | 4238 |
| Number of columns | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| male | 0 | 1.00 | 0.43 | 0.50 | 0.00 | 0.00 | 0.0 | 1.00 | 1.0 | ▇▁▁▁▆ |
| age | 0 | 1.00 | 49.58 | 8.57 | 32.00 | 42.00 | 49.0 | 56.00 | 70.0 | ▃▇▆▆▂ |
| education | 105 | 0.98 | 1.98 | 1.02 | 1.00 | 1.00 | 2.0 | 3.00 | 4.0 | ▇▆▁▃▂ |
| currentSmoker | 0 | 1.00 | 0.49 | 0.50 | 0.00 | 0.00 | 0.0 | 1.00 | 1.0 | ▇▁▁▁▇ |
| cigsPerDay | 29 | 0.99 | 9.00 | 11.92 | 0.00 | 0.00 | 0.0 | 20.00 | 70.0 | ▇▃▁▁▁ |
| BPMeds | 53 | 0.99 | 0.03 | 0.17 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▁ |
| prevalentStroke | 0 | 1.00 | 0.01 | 0.08 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▁ |
| prevalentHyp | 0 | 1.00 | 0.31 | 0.46 | 0.00 | 0.00 | 0.0 | 1.00 | 1.0 | ▇▁▁▁▃ |
| diabetes | 0 | 1.00 | 0.03 | 0.16 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▁ |
| totChol | 50 | 0.99 | 236.72 | 44.59 | 107.00 | 206.00 | 234.0 | 263.00 | 696.0 | ▆▇▁▁▁ |
| sysBP | 0 | 1.00 | 132.35 | 22.04 | 83.50 | 117.00 | 128.0 | 144.00 | 295.0 | ▇▇▁▁▁ |
| diaBP | 0 | 1.00 | 82.89 | 11.91 | 48.00 | 75.00 | 82.0 | 89.88 | 142.5 | ▁▇▅▁▁ |
| BMI | 19 | 1.00 | 25.80 | 4.08 | 15.54 | 23.07 | 25.4 | 28.04 | 56.8 | ▅▇▁▁▁ |
| heartRate | 1 | 1.00 | 75.88 | 12.03 | 44.00 | 68.00 | 75.0 | 83.00 | 143.0 | ▂▇▃▁▁ |
| glucose | 388 | 0.91 | 81.97 | 23.96 | 40.00 | 71.00 | 78.0 | 87.00 | 394.0 | ▇▁▁▁▁ |
| TenYearCHD | 0 | 1.00 | 0.15 | 0.36 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▂ |

vis\_miss(data)+  
 theme(axis.text.x = element\_text(angle = 75, hjust = 0.1))



# Очистка данных

1. Уберите переменные, в которых пропущенных значений больше 20% или уберите субъектов со слишком большим количеством пропущенных значений. Или совместите оба варианта. Напишите обоснование, почему вы выбрали тот или иной вариант:

data %>%   
 rowwise() %>%   
 filter(  
 sum(is.na( across(everything()) )) > 1  
 )

## # A tibble: 61 × 16  
## # Rowwise:   
## male age education currentSmoker cigsPerDay BPMeds prevalentStroke  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 52 1 0 0 1 0  
## 2 1 43 4 1 NA 0 0  
## 3 0 38 3 1 3 1 0  
## 4 0 40 2 1 20 0 0  
## 5 0 47 1 1 25 0 0  
## 6 1 65 NA 0 0 0 0  
## 7 0 60 NA 0 0 0 0  
## 8 0 66 2 1 15 0 0  
## 9 0 37 2 1 30 0 0  
## 10 0 41 1 1 30 0 0  
## # ℹ 51 more rows  
## # ℹ 9 more variables: prevalentHyp <dbl>, diabetes <dbl>, totChol <dbl>,  
## # sysBP <dbl>, diaBP <dbl>, BMI <dbl>, heartRate <dbl>, glucose <dbl>,  
## # TenYearCHD <dbl>

**Обоснование**: больше всего пропусков в переменной glucose (9%),

1. Переименуйте переменные в человекочитаемый вид (что делать с пробелами в названиях?);
2. В соответствии с описанием данных приведите переменные к нужному типу (numeric или factor);
3. Отсортируйте данные по возрасту по убыванию;
4. Сохраните в файл outliers.csv субъектов, которые являются выбросами (например, по правилу трёх сигм) — это необязательное задание со звёздочкой;
5. Присвойте получившийся датасет переменной “cleaned\_data”.

# Сколько осталось переменных?

# Сколько осталось случаев?

# Есть ли в данных идентичные строки?

# Сколько всего переменных с пропущенными значениями в данных и сколько пропущенных точек в каждой такой переменной?

# Описательные статистики

## Количественные переменные

1. Рассчитайте для всех количественных переменных для каждой группы (TenYearCHD):

1.1) Количество значений;

1.2) Количество пропущенных значений;

1.3) Среднее;

1.4) Медиану;

1.5) Стандартное отклонение;

1.6) 25% квантиль и 75% квантиль;

1.7) Интерквартильный размах;

1.8) Минимум;

1.9) Максимум;

1.10) 95% ДИ для среднего - задание со звёздочкой.

## Категориальные переменные

1. Рассчитайте для всех категориальных переменных для каждой группы (TenYearCHD):

1.1) Абсолютное количество;

1.2) Относительное количество внутри группы;

1.3) 95% ДИ для доли внутри группы - задание со звёздочкой.

# Визуализация

## Количественные переменные

1. Для каждой количественной переменной сделайте боксплоты по группам. Расположите их либо на отдельных рисунках, либо на одном, но читаемо;
2. Наложите на боксплоты beeplots - задание со звёздочкой.
3. Раскрасьте боксплоты с помощью библиотеки RColorBrewer.

## Категориальные переменные

1. Сделайте подходящие визуализации категориальных переменных. Обоснуйте, почему выбрали именно этот тип.

# Статистические оценки

## Проверка на нормальность

1. Оцените каждую переменную на соответствие нормальному распределению с помощью теста Шапиро-Уилка. Какие из переменных являются нормальными и как как вы это поняли?
2. Постройте для каждой количественной переменной QQ-плот. Отличаются ли выводы от теста Шапиро-Уилка? Какой метод вы бы предпочли и почему?
3. Ниже напишите, какие ещё методы проверки на нормальность вы знаете и какие у них есть ограничения.

**Напишите текст здесь**

## Сравнение групп

1. Сравните группы (переменная **TenYearCHD**) по каждой переменной (как количественной, так и категориальной). Для каждой переменной выберите нужный критерий и кратко обоснуйте его выбор в комментариях.

# Далее идут **необязательные** дополнительные задания, которые могут принести вам дополнительные баллы в том числе в случае ошибок в предыдущих

## Корреляционный анализ

1. Создайте корреляционную матрицу с визуализацией и поправкой на множественные сравнения. Объясните, когда лучше использовать корреляционные матрицы и в чём минусы и плюсы корреляционных исследований.

## Моделирование

1. Постройте регрессионную модель для переменной **TenYearCHD**. Опишите процесс построения