automatization\_notebook\_04

2023-10-03

# Чтение данных

В вашем варианте нужно использовать датасет healthcare-dataset-stroke-data.

setwd('/Users/user/Desktop/BioStat\_2023/biostat\_homework/data/raw')  
df<-readr::read\_csv('/Users/user/Desktop/BioStat\_2023/biostat\_homework/data/raw/healthcare-dataset-stroke-data.csv')

## Rows: 5110 Columns: 12  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (6): gender, ever\_married, work\_type, Residence\_type, bmi, smoking\_status  
## dbl (6): id, age, hypertension, heart\_disease, avg\_glucose\_level, stroke  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

df

## # A tibble: 5,110 × 12  
## id gender age hypertension heart\_disease ever\_married work\_type   
## <dbl> <chr> <dbl> <dbl> <dbl> <chr> <chr>   
## 1 9046 Male 67 0 1 Yes Private   
## 2 51676 Female 61 0 0 Yes Self-employed  
## 3 31112 Male 80 0 1 Yes Private   
## 4 60182 Female 49 0 0 Yes Private   
## 5 1665 Female 79 1 0 Yes Self-employed  
## 6 56669 Male 81 0 0 Yes Private   
## 7 53882 Male 74 1 1 Yes Private   
## 8 10434 Female 69 0 0 No Private   
## 9 27419 Female 59 0 0 Yes Private   
## 10 60491 Female 78 0 0 Yes Private   
## # ℹ 5,100 more rows  
## # ℹ 5 more variables: Residence\_type <chr>, avg\_glucose\_level <dbl>, bmi <chr>,  
## # smoking\_status <chr>, stroke <dbl>

# Выведите общее описание данных

View(df)

# 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:

**Обоснование**:

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 (stroke):

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 (stroke):

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 **stroke**) 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 **stroke**. Опишите процесс построения