automatization\_notebook\_03

2023-10-09

# Чтение данных

framingham <- read\_csv ("framingham.csv")

# Общее описание данных

Используемые данные находятся в открытом доступе на сайте [kaggle.com](https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/). Анализируемый датасет содержит информацию о более чем 4 тысячах жителей города Framingham (Massachusetts, USA), включающую различные потенциальные факторы риска и исход (диагностирование или нет ишемической болезни сердца в течение 10 лет наблюдения).

framingham %>% glimpse()

## Rows: 4,238  
## Columns: 16  
## $ male <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, …  
## $ age <dbl> 39, 46, 48, 61, 46, 43, 63, 45, 52, 43, 50, 43, 46, 41…  
## $ education <dbl> 4, 2, 1, 3, 3, 2, 1, 2, 1, 1, 1, 2, 1, 3, 2, 2, 3, 2, …  
## $ currentSmoker <dbl> 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, …  
## $ cigsPerDay <dbl> 0, 0, 20, 30, 23, 0, 0, 20, 0, 30, 0, 0, 15, 0, 9, 20,…  
## $ BPMeds <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, …  
## $ prevalentStroke <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ prevalentHyp <dbl> 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, …  
## $ diabetes <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ totChol <dbl> 195, 250, 245, 225, 285, 228, 205, 313, 260, 225, 254,…  
## $ sysBP <dbl> 106.0, 121.0, 127.5, 150.0, 130.0, 180.0, 138.0, 100.0…  
## $ diaBP <dbl> 70.0, 81.0, 80.0, 95.0, 84.0, 110.0, 71.0, 71.0, 89.0,…  
## $ BMI <dbl> 26.97, 28.73, 25.34, 28.58, 23.10, 30.30, 33.11, 21.68…  
## $ heartRate <dbl> 80, 95, 75, 65, 85, 77, 60, 79, 76, 93, 75, 72, 98, 65…  
## $ glucose <dbl> 77, 76, 70, 103, 85, 99, 85, 78, 79, 88, 76, 61, 64, 8…  
## $ TenYearCHD <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, …

# Очистка данных

## Поиск переменных с долей пропущенных значений более 20%

framingham\_more\_than\_20\_cols <- framingham %>%   
 select (where (function(x) sum (is.na(x))/ nrow(framingham) \* 100 > 20))  
 ifelse (ncol (framingham\_more\_than\_20\_cols) > 0,   
 framingham\_more\_than\_20\_cols,   
 "Переменных с долей пропущенных значений более 20% не выявлено")

## [1] "Переменных с долей пропущенных значений более 20% не выявлено"

## Поиск субъектов со слишком большим количеством пропущенных значений

framingham\_shortened <- framingham [apply (framingham, 1, function(x) sum (is.na(x))/ ncol (framingham) \* 100 < 10), ]

*Выбраны субъекты с долей пропущенных значений менее 10%*

## Переименовывание, приведение переменных к нужному типу, сортировка по возрасту

cleaned\_data <- framingham\_shortened %>%   
 rename (Male = male, Age = age, Education = education,   
 'Current Smoker'= currentSmoker, 'Cigarettes per day' = cigsPerDay,  
 'Blood pressure medication' = BPMeds, 'Stroke in history'= prevalentStroke,  
 'Hypertension in history' = prevalentHyp, 'Diabets in history' = diabetes,   
 'Total cholesterol' = totChol, 'Systolic blood pressure' = sysBP,   
 'Diastolic blood pressure' = diaBP, 'Heart rate' = heartRate,  
 Glucose = glucose, '10 year risk of CHD' = TenYearCHD) %>%   
 mutate\_at (c('Male', 'Education', 'Current Smoker', 'Blood pressure medication',  
 'Stroke in history', 'Hypertension in history', 'Diabets in history',  
 '10 year risk of CHD'), as.factor) %>%   
 mutate\_at (c('Age','Cigarettes per day', 'Total cholesterol', 'Systolic blood pressure',  
 'Diastolic blood pressure', 'BMI', 'Heart rate', 'Glucose'), as.numeric) %>%   
 arrange (desc (Age))

## Создание файла outliers.csv с субъектами, которые являются выбросами

outliers <- cleaned\_data %>%   
 mutate\_at (c('Male', 'Education', 'Current Smoker', 'Blood pressure medication',  
 'Stroke in history', 'Hypertension in history', 'Diabets in history',  
 '10 year risk of CHD'), as.numeric) %>%  
 filter(if\_any(.cols = everything(), .fns = function(x) abs(x - mean(x, na.rm = TRUE)) > 3\*sd(x, na.rm = TRUE)))  
  
write\_csv (outliers, "outliers.csv")

# Сколько осталось переменных?

ncol(cleaned\_data)

## [1] 16

# Сколько осталось случаев?

nrow (cleaned\_data)

## [1] 4177

# Есть ли в данных идентичные строки?

ifelse (nrow(unique(cleaned\_data)) == nrow (cleaned\_data),   
 "Идентичных строк нет",  
 paste ("В данных", nrow(cleaned\_data) - nrow(unique(cleaned\_data)), "повторяющихся строк" ))

## [1] "Идентичных строк нет"

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

paste("В данных",   
 ncol(cleaned\_data %>% select(where(anyNA))),   
 "переменных с пропущенными значениями")

## [1] "В данных 7 переменных с пропущенными значениями"

Ниже представлено количество пропущенных точек в каждой из таких переменных

cleaned\_data %>%  
 select(where(anyNA)) %>%  
 sapply (function (x) sum(is.na(x)))

## Education Cigarettes per day Blood pressure medication   
## 93 23 51   
## Total cholesterol BMI Heart rate   
## 9 13 1   
## Glucose   
## 331

# Описательные статистики

## Количественные переменные

statistics <- list(  
 `Количество субъектов` = ~length(.x) %>% as.character(),  
 `Количество пропущенных значений` = ~sum(is.na(.x)) %>% as.character(),  
 `Среднее` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", mean(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `Станд. отклон.` = ~ifelse(sum(!is.na(.x)) < 3, "Н/П\*", sd(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `мин. - макс.` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", paste0(min(.x, na.rm = TRUE) %>% round(2), " - ", max(.x, na.rm = TRUE) %>% round(2))),  
 `Медиана` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", median(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `Q1 - Q3` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", paste0(quantile(.x, 0.25, na.rm = TRUE) %>% round(2), " - ", quantile(.x, 0.75, na.rm = TRUE) %>% round(2))),  
 `Межквартильный размах` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", IQR(.x, na.rm = TRUE) %>% round(2) %>% as.character())  
)  
  
cleaned\_data %>%  
 select('10 year risk of CHD', where(is.numeric)) %>%  
 group\_by (`10 year risk of CHD`) %>%   
 summarise(across(where(is.numeric), statistics)) %>%  
 mutate\_at (1:ncol(cleaned\_data), as.character) %>%   
 pivot\_longer(!'10 year risk of CHD') %>%   
 separate(name, into = c('Переменная', 'Статистика'), sep = '\_') %>%   
 rename ('Значение' = value) %>%   
 flextable() %>%   
 theme\_box() %>%   
 merge\_v(c("10 year risk of CHD", "Переменная")) %>%   
 width(j= "Статистика", width=3) %>%   
 align (j = c("10 year risk of CHD", "Переменная"), align = "center")

| **10 year risk of CHD** | **Переменная** | **Статистика** | **Значение** |
| --- | --- | --- | --- |
| 0 | Age | Количество субъектов | 3545 |
| Количество пропущенных значений | 0 |
| Среднее | 48.72 |
| Станд. отклон. | 8.39 |
| мин. - макс. | 32 - 70 |
| Медиана | 48 |
| Q1 - Q3 | 42 - 55 |
| Межквартильный размах | 13 |
| Cigarettes per day | Количество субъектов | 3545 |
| Количество пропущенных значений | 21 |
| Среднее | 8.68 |
| Станд. отклон. | 11.67 |
| мин. - макс. | 0 - 70 |
| Медиана | 0 |
| Q1 - Q3 | 0 - 20 |
| Межквартильный размах | 20 |
| Total cholesterol | Количество субъектов | 3545 |
| Количество пропущенных значений | 8 |
| Среднее | 235.17 |
| Станд. отклон. | 43.83 |
| мин. - макс. | 113 - 696 |
| Медиана | 232 |
| Q1 - Q3 | 205 - 261 |
| Межквартильный размах | 56 |
| Systolic blood pressure | Количество субъектов | 3545 |
| Количество пропущенных значений | 0 |
| Среднее | 130.3 |
| Станд. отклон. | 20.38 |
| мин. - макс. | 83.5 - 243 |
| Медиана | 127 |
| Q1 - Q3 | 116 - 141 |
| Межквартильный размах | 25 |
| Diastolic blood pressure | Количество субъектов | 3545 |
| Количество пропущенных значений | 0 |
| Среднее | 82.16 |
| Станд. отклон. | 11.29 |
| мин. - макс. | 50 - 142.5 |
| Медиана | 81 |
| Q1 - Q3 | 74 - 88 |
| Межквартильный размах | 14 |
| BMI | Количество субъектов | 3545 |
| Количество пропущенных значений | 6 |
| Среднее | 25.67 |
| Станд. отклон. | 3.98 |
| мин. - макс. | 15.54 - 51.28 |
| Медиана | 25.23 |
| Q1 - Q3 | 23 - 27.87 |
| Межквартильный размах | 4.87 |
| Heart rate | Количество субъектов | 3545 |
| Количество пропущенных значений | 0 |
| Среднее | 75.76 |
| Станд. отклон. | 11.99 |
| мин. - макс. | 44 - 143 |
| Медиана | 75 |
| Q1 - Q3 | 68 - 83 |
| Межквартильный размах | 15 |
| Glucose | Количество субъектов | 3545 |
| Количество пропущенных значений | 293 |
| Среднее | 80.68 |
| Станд. отклон. | 18.97 |
| мин. - макс. | 40 - 386 |
| Медиана | 78 |
| Q1 - Q3 | 71 - 86 |
| Межквартильный размах | 15 |
| 1 | Age | Количество субъектов | 632 |
| Количество пропущенных значений | 0 |
| Среднее | 54.25 |
| Станд. отклон. | 8.01 |
| мин. - макс. | 35 - 70 |
| Медиана | 55 |
| Q1 - Q3 | 48 - 61 |
| Межквартильный размах | 13 |
| Cigarettes per day | Количество субъектов | 632 |
| Количество пропущенных значений | 2 |
| Среднее | 10.68 |
| Станд. отклон. | 13.04 |
| мин. - макс. | 0 - 60 |
| Медиана | 3 |
| Q1 - Q3 | 0 - 20 |
| Межквартильный размах | 20 |
| Total cholesterol | Количество субъектов | 632 |
| Количество пропущенных значений | 1 |
| Среднее | 245.57 |
| Станд. отклон. | 48.14 |
| мин. - макс. | 107 - 600 |
| Медиана | 241 |
| Q1 - Q3 | 214 - 272 |
| Межквартильный размах | 58 |
| Systolic blood pressure | Количество субъектов | 632 |
| Количество пропущенных значений | 0 |
| Среднее | 143.62 |
| Станд. отклон. | 26.8 |
| мин. - макс. | 83.5 - 295 |
| Медиана | 139 |
| Q1 - Q3 | 125 - 158 |
| Межквартильный размах | 33 |
| Diastolic blood pressure | Количество субъектов | 632 |
| Количество пропущенных значений | 0 |
| Среднее | 86.96 |
| Станд. отклон. | 14.1 |
| мин. - макс. | 48 - 140 |
| Медиана | 85 |
| Q1 - Q3 | 78 - 95 |
| Межквартильный размах | 17 |
| BMI | Количество субъектов | 632 |
| Количество пропущенных значений | 7 |
| Среднее | 26.55 |
| Станд. отклон. | 4.53 |
| мин. - макс. | 15.96 - 56.8 |
| Медиана | 26.13 |
| Q1 - Q3 | 23.48 - 28.92 |
| Межквартильный размах | 5.44 |
| Heart rate | Количество субъектов | 632 |
| Количество пропущенных значений | 1 |
| Среднее | 76.53 |
| Станд. отклон. | 12.28 |
| мин. - макс. | 50 - 120 |
| Медиана | 75 |
| Q1 - Q3 | 68 - 85 |
| Межквартильный размах | 17 |
| Glucose | Количество субъектов | 632 |
| Количество пропущенных значений | 38 |
| Среднее | 89.01 |
| Станд. отклон. | 41.14 |
| мин. - макс. | 40 - 394 |
| Медиана | 79 |
| Q1 - Q3 | 72 - 90 |
| Межквартильный размах | 18 |

## Категориальные переменные

#### 10-летний риск ИБС и пол

cleaned\_data %>%  
 select('10 year risk of CHD', Male) %>%  
 count (`10 year risk of CHD`, Male) %>%  
 group\_by (`10 year risk of CHD`) %>%  
 mutate ('Relative quantity' = (n/sum(n)) %>% round (3) %>% `\*` (100) %>% str\_c ("%")) %>%   
 rename ('Absolute quantity' = n) %>%   
 flextable() %>% theme\_box() %>%   
 merge\_v("10 year risk of CHD") %>%   
 align (align = "center") %>%   
 set\_caption ("Numbers of people depending on '10 year risk of CHD' and 'Male gender'")

Numbers of people depending on '10 year risk of CHD' and 'Male gender'

| **10 year risk of CHD** | **Male** | **Absolute quantity** | **Relative quantity** |
| --- | --- | --- | --- |
| 0 | 0 | 2,079 | 58.6% |
| 1 | 1,466 | 41.4% |
| 1 | 0 | 292 | 46.2% |
| 1 | 340 | 53.8% |

#### 10-летний риск ИБС и образование

cleaned\_data %>%  
 select('10 year risk of CHD', Education) %>%  
 mutate (`Education` = `Education` %>% as.character() %>% replace\_na ("Нет данных") %>% as.factor()) %>%   
 count (`10 year risk of CHD`, Education) %>%  
 group\_by (`10 year risk of CHD`) %>%  
 mutate ('Relative quantity' = (n/sum(n)) %>% round (3) %>% `\*` (100) %>% str\_c ("%")) %>%   
 rename ('Absolute quantity' = n) %>%   
 flextable() %>% theme\_box() %>%   
 merge\_v("10 year risk of CHD") %>%   
 align (align = "center") %>%   
 set\_caption ("Numbers of people depending on '10 year risk of CHD' and 'Education'")

Numbers of people depending on '10 year risk of CHD' and 'Education'

| **10 year risk of CHD** | **Education** | **Absolute quantity** | **Relative quantity** |
| --- | --- | --- | --- |
| 0 | 1 | 1,381 | 39% |
| 2 | 1,095 | 30.9% |
| 3 | 592 | 16.7% |
| 4 | 399 | 11.3% |
| Нет данных | 78 | 2.2% |
| 1 | 1 | 319 | 50.5% |
| 2 | 142 | 22.5% |
| 3 | 87 | 13.8% |
| 4 | 69 | 10.9% |
| Нет данных | 15 | 2.4% |

#### 10-летний риск ИБС и курение

cleaned\_data %>%  
 select('10 year risk of CHD', `Current Smoker`) %>%  
 count (`10 year risk of CHD`, `Current Smoker`) %>%  
 group\_by (`10 year risk of CHD`) %>%  
 mutate ('Relative quantity' = (n/sum(n)) %>% round (3) %>% `\*` (100) %>% str\_c ("%")) %>%   
 rename ('Absolute quantity' = n) %>%   
 flextable() %>% theme\_box() %>%   
 merge\_v("10 year risk of CHD") %>%   
 align (align = "center") %>%   
 set\_caption ("Numbers of people depending on '10 year risk of CHD' and Smoking status")

Numbers of people depending on '10 year risk of CHD' and Smoking status

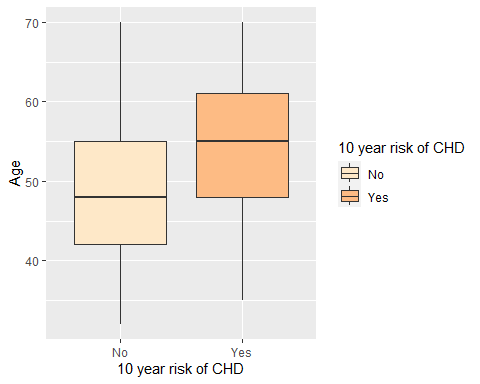
| **10 year risk of CHD** | **Current Smoker** | **Absolute quantity** | **Relative quantity** |
| --- | --- | --- | --- |
| 0 | 0 | 1,816 | 51.2% |
| 1 | 1,729 | 48.8% |
| 1 | 0 | 305 | 48.3% |
| 1 | 327 | 51.7% |

# Визуализация

## Количественные переменные

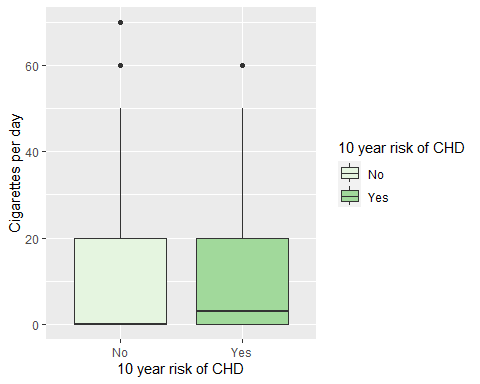
#### 10-летний риск ИБС и возраст

library (ggplot2)  
library(RColorBrewer)  
  
numeric\_data <- cleaned\_data %>%   
 select (`10 year risk of CHD`, where (is.numeric)) %>%  
 mutate (`10 year risk of CHD` = ifelse (`10 year risk of CHD` == 1, "Yes", "No"))  
  
ggplot (numeric\_data, aes(x = `10 year risk of CHD`, y = Age, fill =`10 year risk of CHD`)) +  
 geom\_boxplot() +  
 scale\_fill\_brewer(palette="OrRd")

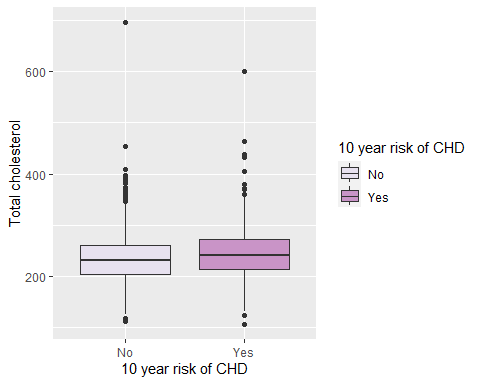


#### 10-летний риск ИБС и количество сигарет в день

library (ggplot2)  
library(RColorBrewer)  
  
ggplot (numeric\_data, aes(x = `10 year risk of CHD`, y = `Cigarettes per day`, fill =`10 year risk of CHD`)) +  
 geom\_boxplot(na.rm = T) +  
 scale\_fill\_brewer(palette="Greens")

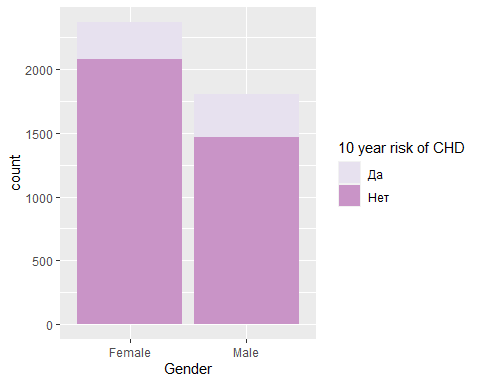


library (ggplot2)  
library(RColorBrewer)  
  
ggplot (numeric\_data, aes(x = `10 year risk of CHD`, y = `Total cholesterol`, fill =`10 year risk of CHD`)) +  
 geom\_boxplot(na.rm = T) +  
 scale\_fill\_brewer(palette="PuRd")



## Категориальные переменные

library (ggplot2)  
library(RColorBrewer)  
  
categorical\_data <- cleaned\_data %>%   
 select(`10 year risk of CHD`, where(is.factor)) %>%   
 mutate (`10 year risk of CHD` = ifelse (`10 year risk of CHD` == 1, "Да", "Нет")) %>%  
 rename (Gender = Male) %>%   
 mutate (`Gender` = ifelse (`Gender` == 1, "Male", "Female"))  
  
  
ggplot (categorical\_data, aes(x = `Gender`, fill =`10 year risk of CHD`)) +  
 geom\_bar (na.rm = T) +  
 scale\_fill\_brewer(palette="PuRd")



# Статистические оценки

## Проверка на нормальность

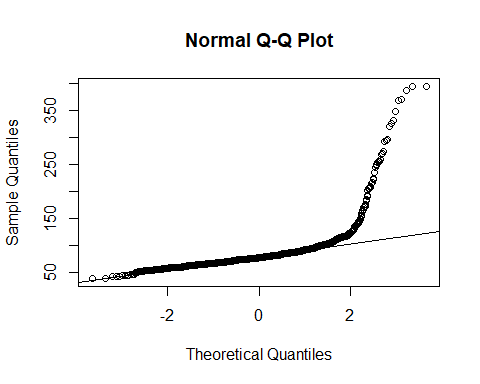
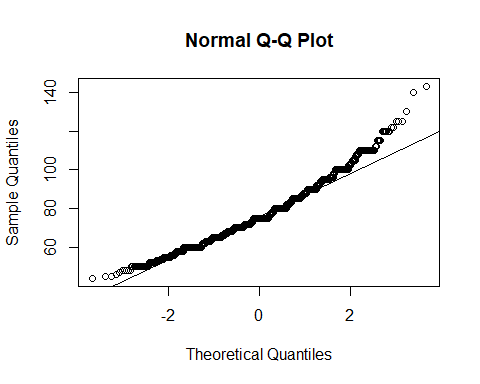
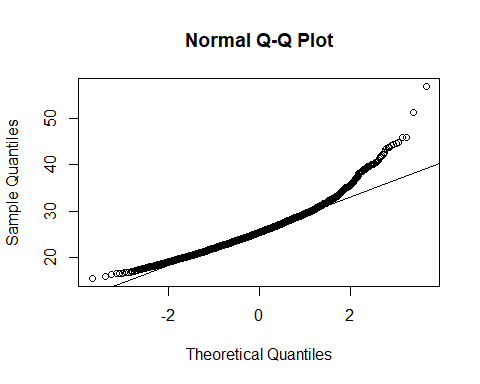
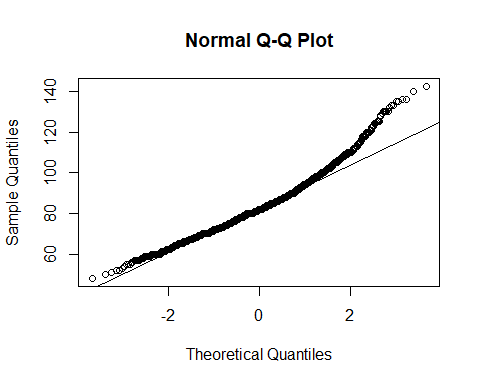
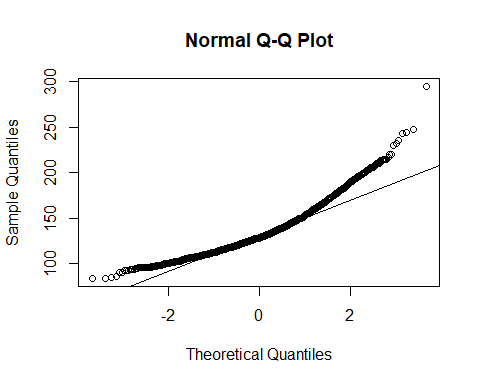
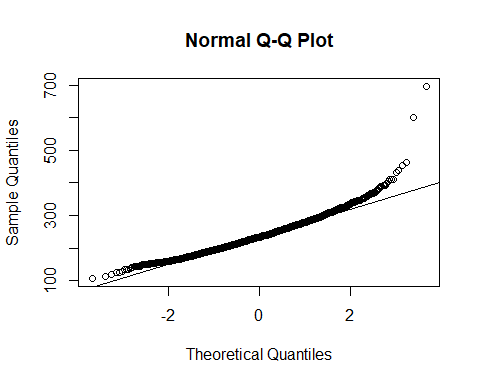
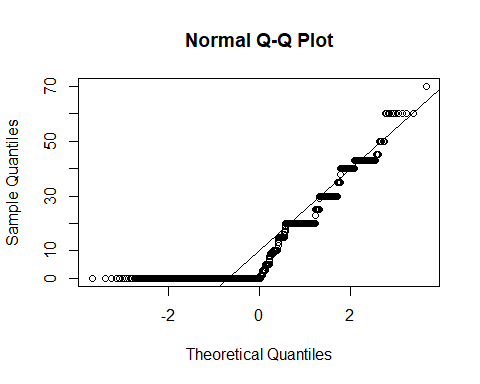
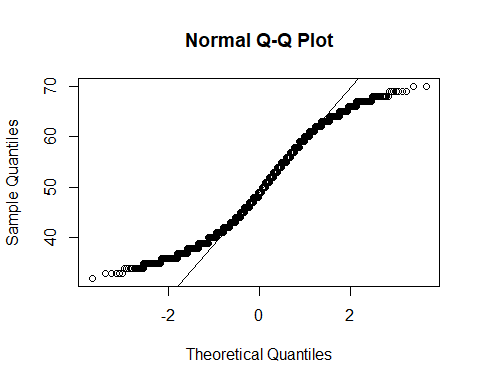
1. Оценка переменных на соответствие нормальному распределению с помощью теста Шапиро-Уилка

shapiro\_p <- cleaned\_data %>%   
 mutate\_at (c('Male', 'Education', 'Current Smoker', 'Blood pressure medication', 'Stroke in history', 'Hypertension in history', 'Diabets in history', '10 year risk of CHD'), as.numeric) %>%   
 summarise(across (where (is.numeric),   
 function (x) shapiro.test(x)$p)) %>%   
 select(where (function(x) sum (x) > 0.05))  
  
ifelse (ncol (shapiro\_p) == 0, "В датасете данные всех переменнных распределены нормально, т.к. отсутствуют переменные, для которых p.value, полученное с помощью теста Шапиро–Уилка, менее 0.05", "В данных имеются переменные, распределение которых значимо отличается от нормального")

## [1] "В датасете данные всех переменнных распределены нормально, т.к. отсутствуют переменные, для которых p.value, полученное с помощью теста Шапиро–Уилка, менее 0.05"

1. QQ-плоты для каждой количественной переменной

cleaned\_data %>%   
 select(where (is.numeric)) %>%   
 sapply (function(x) c (qqnorm(x), qqline (x)))



## Age Cigarettes per day Total cholesterol Systolic blood pressure  
## x numeric,4177 numeric,4177 numeric,4177 numeric,4177   
## y numeric,4177 numeric,4177 numeric,4177 numeric,4177   
## Diastolic blood pressure BMI Heart rate Glucose   
## x numeric,4177 numeric,4177 numeric,4177 numeric,4177  
## y numeric,4177 numeric,4177 numeric,4177 numeric,4177

*QQ-плот для переменной Glucose демонстрирует явное отклонение распределения от нормального. На мой взгляд, использование теста Шапиро–Уилка боллее предпочтительно, так как он лишен субъективности.*

1. Другие методы проверки на нормальность и их ограничения.

**1. Построение гистограммы с помощью *hist (x)***

* **метод субъективен**

**2. Тест Колмогорова Смирнова с помощью *ks.test (x, ‘pnorm’)***

* **необходима выборка размером более 50**

## Сравнение групп по 10 year risk of CHD

options (scipen = 999)  
cleaned\_data %>%  
 select('10 year risk of CHD', where(is.numeric)) %>%  
 summarise(across(where(is.numeric), function(x)   
 t.test (x ~ `10 year risk of CHD`)$p.value)) %>%   
 round(3) %>%   
 summarise(across(where(is.numeric), function(x)   
 ifelse (x < 0.001, "< 0.001", x))) %>%  
 flextable() %>% theme\_box() %>%   
 bold (i = 1, j = c(1,2,3,4,5,6,8), part = "body") %>%   
 set\_caption ("Results (p values) of comparing numerical variables by t.test")

Results (p values) of comparing numerical variables by t.test

| **Age** | **Cigarettes per day** | **Total cholesterol** | **Systolic blood pressure** | **Diastolic blood pressure** | **BMI** | **Heart rate** | **Glucose** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **< 0.001** | **< 0.001** | **< 0.001** | **< 0.001** | **< 0.001** | **< 0.001** | 0.144 | **< 0.001** |

*Для сравнения количественных переменных использовался t.test, так как все выбранные переменные имеют нормальное распределения по результатам теста Шапиро-Уилка*

options (scipen = 999)  
cleaned\_data %>%  
 select('10 year risk of CHD', where(is.factor)) %>%  
 mutate('10 year risk of CHD' = as.character('10 year risk of CHD')) %>%   
 summarise(across(where(is.factor),   
 function(x) chisq.test (table (x, `10 year risk of CHD`))$p.value)) %>%  
 round(3) %>%   
 summarise(across(where(is.numeric), function(x)  
 ifelse (x < 0.001, "< 0.001", x))) %>%  
 flextable() %>% theme\_box() %>%   
 bold (i = 1, j = c(1,2,4,5,6,7), part = "body") %>%   
 set\_caption ("Results (p values) of comparing categorical variables by chisq.test")

Results (p values) of comparing categorical variables by chisq.test

| **Male** | **Education** | **Current Smoker** | **Blood pressure medication** | **Stroke in history** | **Hypertension in history** | **Diabets in history** |
| --- | --- | --- | --- | --- | --- | --- |
| **< 0.001** | **< 0.001** | 0.315 | **< 0.001** | **< 0.001** | **< 0.001** | **< 0.001** |

## Корреляционный анализ

### Корреляционная матрица [[1]](#footnote-81)

library (psych)

##   
## Присоединяю пакет: 'psych'

## Следующие объекты скрыты от 'package:ggplot2':  
##   
## %+%, alpha

options (scipen = 999)  
  
as.data.frame(  
 (cleaned\_data %>%  
 select(where(is.numeric)) %>%   
 corr.test())$r %>% round (2)  
 )

## Age Cigarettes per day Total cholesterol  
## Age 1.00 -0.19 0.26  
## Cigarettes per day -0.19 1.00 -0.03  
## Total cholesterol 0.26 -0.03 1.00  
## Systolic blood pressure 0.39 -0.09 0.21  
## Diastolic blood pressure 0.20 -0.05 0.17  
## BMI 0.13 -0.09 0.12  
## Heart rate -0.01 0.07 0.09  
## Glucose 0.12 -0.06 0.05  
## Systolic blood pressure Diastolic blood pressure BMI  
## Age 0.39 0.20 0.13  
## Cigarettes per day -0.09 -0.05 -0.09  
## Total cholesterol 0.21 0.17 0.12  
## Systolic blood pressure 1.00 0.78 0.33  
## Diastolic blood pressure 0.78 1.00 0.38  
## BMI 0.33 0.38 1.00  
## Heart rate 0.18 0.18 0.07  
## Glucose 0.14 0.06 0.09  
## Heart rate Glucose  
## Age -0.01 0.12  
## Cigarettes per day 0.07 -0.06  
## Total cholesterol 0.09 0.05  
## Systolic blood pressure 0.18 0.14  
## Diastolic blood pressure 0.18 0.06  
## BMI 0.07 0.09  
## Heart rate 1.00 0.09  
## Glucose 0.09 1.00

library (psych)  
options (scipen = 999)  
  
as.data.frame ((cleaned\_data %>%  
 select(where(is.numeric)) %>%   
 corr.test())$p %>% round (4))

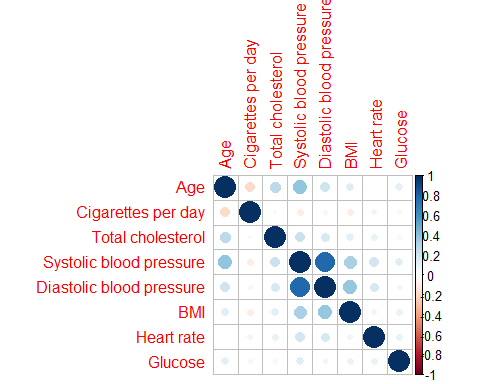
## Age Cigarettes per day Total cholesterol  
## Age 0.0000 0.0000 0.0000  
## Cigarettes per day 0.0000 0.0000 0.1667  
## Total cholesterol 0.0000 0.0834 0.0000  
## Systolic blood pressure 0.0000 0.0000 0.0000  
## Diastolic blood pressure 0.0000 0.0006 0.0000  
## BMI 0.0000 0.0000 0.0000  
## Heart rate 0.4752 0.0000 0.0000  
## Glucose 0.0000 0.0003 0.0039  
## Systolic blood pressure Diastolic blood pressure BMI  
## Age 0 0.0000 0  
## Cigarettes per day 0 0.0024 0  
## Total cholesterol 0 0.0000 0  
## Systolic blood pressure 0 0.0000 0  
## Diastolic blood pressure 0 0.0000 0  
## BMI 0 0.0000 0  
## Heart rate 0 0.0000 0  
## Glucose 0 0.0001 0  
## Heart rate Glucose  
## Age 0.4752 0.0000  
## Cigarettes per day 0.0000 0.0014  
## Total cholesterol 0.0000 0.0116  
## Systolic blood pressure 0.0000 0.0000  
## Diastolic blood pressure 0.0000 0.0008  
## BMI 0.0000 0.0000  
## Heart rate 0.0000 0.0000  
## Glucose 0.0000 0.0000

### Визуализация

library (corrplot)

## corrplot 0.92 loaded

cleaned\_data %>%  
 select(where(is.numeric)) %>%   
 cor (use = "complete.obs") %>%   
 corrplot()



*Корреляционные матрицы удобно использовать, чтобы быстро выявить сильно корреллирующие переменные. Недостаток корреляционных исследований в том, что они не могут ответить на вопрос о причинно-следственной связи и для расчёта коэффициента корреляции можно использовать только две переменные.*

## Моделирование

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

1. поправка на множественные сравнения пока не сделана. [↑](#footnote-ref-81)