DataAuto\_task

Variant 3

Evgenii Berdinskikh

2024-10-27

Table of Contents

[Чтение данных 2](#_Toc180952279)

[Выведите общее описание данных 2](#_Toc180952280)

[Очистка данных 4](#_Toc180952281)

[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: 4](#_Toc180952282)

[Переименуйте переменные в человекочитаемый вид (что делать с пробелами в названиях?). В соответствии с описанием данных приведите переменные к нужному типу (numeric или factor). Отсортируйте данные по возрасту по убыванию: 5](#_Toc180952283)

[Присвойте получившийся датасет переменной “cleaned\_data”: 6](#_Toc180952284)

[Сколько осталось переменных? Сколько осталось случаев? Есть ли в данных идентичные строки? 6](#_Toc180952285)

[Сколько всего переменных с пропущенными значениями в данных и сколько пропущенных точек в каждой такой переменной? 6](#_Toc180952286)

[Сохраните в файл outliers.csv субъектов, которые являются выбросами (например, по правилу трёх сигм) — это необязательное задание со звёздочкой; 7](#_Toc180952287)

[Описательные статистики 7](#_Toc180952288)

[Количественные переменные 7](#_Toc180952289)

[Категориальные переменные 13](#_Toc180952290)

[Визуализация 15](#_Toc180952291)

[Количественные переменные 15](#_Toc180952292)

[Категориальные переменные 17](#_Toc180952293)

[Статистические оценки 20](#_Toc180952294)

[Проверка на нормальность 20](#_Toc180952295)

[Сравнение групп 24](#_Toc180952296)

[Количественные переменные 24](#_Toc180952297)

[Категориальные переменные 25](#_Toc180952298)

[Далее идут **необязательные** дополнительные задания, которые могут принести вам дополнительные баллы в том числе в случае ошибок в предыдущих 26](#_Toc180952299)

[Корреляционный анализ 26](#_Toc180952300)

[Моделирование 26](#_Toc180952301)

# 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)

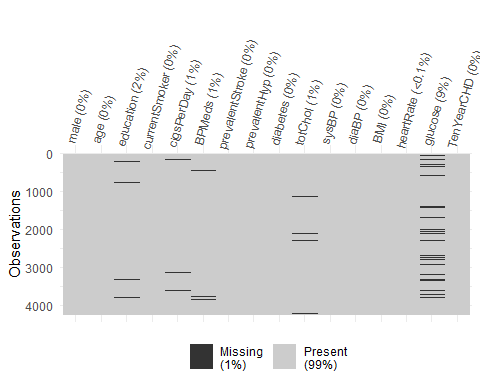
Data summary

|  |  |
| --- | --- |
| 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

## 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()) )) >= 2  
 )

## # 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**: наибольшее количество пропусков в переменной *glucose* - **9%**, поэтому было решено не удалять ни одну из переменных, а найти тех субьктов, для котороых есть пропуски по двум переменным и более.

## Переименуйте переменные в человекочитаемый вид (что делать с пробелами в названиях?). В соответствии с описанием данных приведите переменные к нужному типу (numeric или factor). Отсортируйте данные по возрасту по убыванию:

data %>%   
 transmute(  
 Пол = factor(male, levels = c(0, 1), labels = c("Ж", "М")), Возраст = age, Образование = factor(education),   
 Курильщик = factor(currentSmoker, levels = c(0, 1), labels = c("Нет", "Да")),   
 `Число сигарет в день` = cigsPerDay,   
 `Прием препаратов для АД` = factor(BPMeds, levels = c(0, 1), labels = c("Нет", "Да")),   
 `Перенесенный инсульт` = factor(prevalentStroke, levels = c(0, 1), labels = c("Нет", "Да")),   
 Гипертония = factor(prevalentHyp, levels = c(0, 1), labels = c("Нет", "Да")),   
 Диабет = factor(diabetes, levels = c(0, 1), labels = c("Нет", "Да")),   
 `Уровень общего холестерина` = totChol, `Систолическое АД` = sysBP, `Диастолическое АД` = diaBP,   
 `Индекс массы тела` = BMI, `Частота сердечных сокращений` = heartRate, `Уровень глюкозы` = glucose,   
 `Риск развития ИБС` = factor(TenYearCHD, levels = c(0, 1), labels = c("Нет", "Да"))  
 ) %>%   
 arrange(desc(Возраст))

## # A tibble: 4,238 × 16  
## Пол Возраст Образование Курильщик `Число сигарет в день`  
## <fct> <dbl> <fct> <fct> <dbl>  
## 1 Ж 70 1 Нет 0  
## 2 Ж 70 2 Нет 0  
## 3 Ж 69 2 Нет 0  
## 4 Ж 69 3 Нет 0  
## 5 М 69 1 Да 4  
## 6 Ж 69 3 Нет 0  
## 7 М 69 1 Да 1  
## 8 М 69 <NA> Да 23  
## 9 Ж 69 2 Нет 0  
## 10 М 68 1 Нет 0  
## # ℹ 4,228 more rows  
## # ℹ 11 more variables: `Прием препаратов для АД` <fct>,  
## # `Перенесенный инсульт` <fct>, Гипертония <fct>, Диабет <fct>,  
## # `Уровень общего холестерина` <dbl>, `Систолическое АД` <dbl>,  
## # `Диастолическое АД` <dbl>, `Индекс массы тела` <dbl>,  
## # `Частота сердечных сокращений` <dbl>, `Уровень глюкозы` <dbl>,  
## # `Риск развития ИБС` <fct>

## Присвойте получившийся датасет переменной “cleaned\_data”:

cleaned\_data <- data %>%   
 rowwise() %>%   
 filter(  
 sum(is.na( across(everything()) )) < 2  
 ) %>%   
 ungroup() %>%   
 transmute(  
 Пол = factor(male, levels = c(0, 1), labels = c("Ж", "М")), Возраст = age, Образование = factor(education),   
 Курильщик = factor(currentSmoker, levels = c(0, 1), labels = c("Нет", "Да")),   
 `Число сигарет в день` = cigsPerDay,   
 `Прием препаратов для АД` = factor(BPMeds, levels = c(0, 1), labels = c("Нет", "Да")),   
 `Перенесенный инсульт` = factor(prevalentStroke, levels = c(0, 1), labels = c("Нет", "Да")),   
 Гипертония = factor(prevalentHyp, levels = c(0, 1), labels = c("Нет", "Да")),   
 Диабет = factor(diabetes, levels = c(0, 1), labels = c("Нет", "Да")),   
 `Уровень общего холестерина` = totChol, `Систолическое АД` = sysBP, `Диастолическое АД` = diaBP,   
 `Индекс массы тела` = BMI, `Частота сердечных сокращений` = heartRate, `Уровень глюкозы` = glucose,   
 `Риск развития ИБС` = factor(TenYearCHD, levels = c(0, 1), labels = c("Нет", "Да"))  
 ) %>%   
 arrange(desc(Возраст))

## Сколько осталось переменных? Сколько осталось случаев? Есть ли в данных идентичные строки?

cat(paste("В очищенных данных осталось", ncol(cleaned\_data), "переменных и", nrow(cleaned\_data), "случаев, количество идентичных строк -", sum(duplicated(cleaned\_data))  
 ))

## В очищенных данных осталось 16 переменных и 4177 случаев, количество идентичных строк - 0

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

skim(cleaned\_data) %>%   
 group\_by(skim\_type) %>%  
 filter(n\_missing > 0) %>%  
 ungroup() %>%  
 transmute("Переменная" = skim\_variable, "Количество пропущенных точек" = n\_missing)

## # A tibble: 7 × 2  
## Переменная `Количество пропущенных точек`  
## <chr> <int>  
## 1 Образование 93  
## 2 Прием препаратов для АД 51  
## 3 Число сигарет в день 23  
## 4 Уровень общего холестерина 9  
## 5 Индекс массы тела 13  
## 6 Частота сердечных сокращений 1  
## 7 Уровень глюкозы 331

## Сохраните в файл outliers.csv субъектов, которые являются выбросами (например, по правилу трёх сигм) — это необязательное задание со звёздочкой;

#За выбросы принимаем значения ниже Q1 - 1.5 × IQR и выше Q3 + 1.5 × IQR.  
  
outliers <- cleaned\_data %>%   
 mutate(across(where(is.numeric),   
 function(x)   
 x < quantile(x, 0.25, na.rm = TRUE) - 1.5 \* IQR(x, na.rm = TRUE) |   
 x > quantile(x, 0.75, na.rm = TRUE) + 1.5 \* IQR(x, na.rm = TRUE),  
 .names = "outlier\_{.col}")) %>%   
 rowwise() %>%   
 mutate(outliers = sum(c\_across(starts\_with("outlier\_")), na.rm = TRUE)) %>%   
 ungroup() %>%   
 filter(outliers > 0) %>%   
 select(1:16)  
  
  
#Файл с выбросами сохраним в папку data/docs  
write\_excel\_csv(outliers, "data/docs/outliers.csv")

# Описательные статистики

## Количественные переменные

Рассчитайте для всех количественных переменных для каждой группы (TenYearCHD): - Количество значений; - Количество пропущенных значений; - Среднее; - Медиану; - Стандартное отклонение; - 25% квантиль и 75% квантиль; - Интерквартильный размах; - Минимум; - Максимум; - 95% ДИ для среднего - задание со звёздочкой.

statistics <- list(  
 'Количество значений' = ~ as.character( sum(!is.na(.x)) ),   
 'Количество пропущенных значений'= ~ as.character( sum(is.na(.x)) ),   
 'Среднее значение' = ~ as.character( round (mean(.x, na.rm = TRUE), 2) ),  
 '95% ДИ для среднего' = ~ paste0(round (t.test(.x)[["conf.int"]][1], 2), " - ",   
 round(t.test(.x)[["conf.int"]][2], 2)),  
 'Стандартное отклонение' = ~ as.character( round (sd(.x, na.rm = TRUE), 2) ),   
 'Медиана (Q1-Q3)' = ~ paste0(median(.x, na.rm = TRUE), " (", quantile(.x, 0.25, na.rm = TRUE), " - ",  
 quantile(.x, 0.75, na.rm = TRUE), ")"),  
 'IQR' = ~ as.character( IQR(.x, na.rm = TRUE) ),   
 'Минимум и Максимум' = ~ paste0(min(.x, na.rm = TRUE), " - ", max(.x, na.rm = TRUE))   
 )  
  
  
cleaned\_data %>%   
 group\_by(`Риск развития ИБС`) %>%   
 summarise(across(where(is.numeric), statistics)) %>%   
 pivot\_longer(!`Риск развития ИБС`) %>%   
 separate(name, into = c("Переменная", "Статистика"), sep = "\_") %>%   
 rename(`Значение` = value) %>%   
 flextable() %>%   
 theme\_box() %>%   
 merge\_v(c("Риск развития ИБС", "Переменная")) %>%   
 width(j = 1, width = 2) %>%   
 width(j = 2, width = 2) %>%   
 width(j = 3, width = 3.5) %>%   
 width(j = 4, width = 2)

| **Риск развития ИБС** | **Переменная** | **Статистика** | **Значение** |
| --- | --- | --- | --- |
| Нет | Возраст | Количество значений | 3545 |
| Количество пропущенных значений | 0 |
| Среднее значение | 48.72 |
| 95% ДИ для среднего | 48.44 - 49 |
| Стандартное отклонение | 8.39 |
| Медиана (Q1-Q3) | 48 (42 - 55) |
| IQR | 13 |
| Минимум и Максимум | 32 - 70 |
| Число сигарет в день | Количество значений | 3524 |
| Количество пропущенных значений | 21 |
| Среднее значение | 8.68 |
| 95% ДИ для среднего | 8.29 - 9.06 |
| Стандартное отклонение | 11.67 |
| Медиана (Q1-Q3) | 0 (0 - 20) |
| IQR | 20 |
| Минимум и Максимум | 0 - 70 |
| Уровень общего холестерина | Количество значений | 3537 |
| Количество пропущенных значений | 8 |
| Среднее значение | 235.17 |
| 95% ДИ для среднего | 233.73 - 236.62 |
| Стандартное отклонение | 43.83 |
| Медиана (Q1-Q3) | 232 (205 - 261) |
| IQR | 56 |
| Минимум и Максимум | 113 - 696 |
| Систолическое АД | Количество значений | 3545 |
| Количество пропущенных значений | 0 |
| Среднее значение | 130.3 |
| 95% ДИ для среднего | 129.63 - 130.97 |
| Стандартное отклонение | 20.38 |
| Медиана (Q1-Q3) | 127 (116 - 141) |
| IQR | 25 |
| Минимум и Максимум | 83.5 - 243 |
| Диастолическое АД | Количество значений | 3545 |
| Количество пропущенных значений | 0 |
| Среднее значение | 82.16 |
| 95% ДИ для среднего | 81.78 - 82.53 |
| Стандартное отклонение | 11.29 |
| Медиана (Q1-Q3) | 81 (74 - 88) |
| IQR | 14 |
| Минимум и Максимум | 50 - 142.5 |
| Индекс массы тела | Количество значений | 3539 |
| Количество пропущенных значений | 6 |
| Среднее значение | 25.67 |
| 95% ДИ для среднего | 25.54 - 25.8 |
| Стандартное отклонение | 3.98 |
| Медиана (Q1-Q3) | 25.23 (23 - 27.87) |
| IQR | 4.87 |
| Минимум и Максимум | 15.54 - 51.28 |
| Частота сердечных сокращений | Количество значений | 3545 |
| Количество пропущенных значений | 0 |
| Среднее значение | 75.76 |
| 95% ДИ для среднего | 75.36 - 76.15 |
| Стандартное отклонение | 11.99 |
| Медиана (Q1-Q3) | 75 (68 - 83) |
| IQR | 15 |
| Минимум и Максимум | 44 - 143 |
| Уровень глюкозы | Количество значений | 3252 |
| Количество пропущенных значений | 293 |
| Среднее значение | 80.68 |
| 95% ДИ для среднего | 80.03 - 81.33 |
| Стандартное отклонение | 18.97 |
| Медиана (Q1-Q3) | 78 (71 - 86) |
| IQR | 15 |
| Минимум и Максимум | 40 - 386 |
| Да | Возраст | Количество значений | 632 |
| Количество пропущенных значений | 0 |
| Среднее значение | 54.25 |
| 95% ДИ для среднего | 53.63 - 54.88 |
| Стандартное отклонение | 8.01 |
| Медиана (Q1-Q3) | 55 (48 - 61) |
| IQR | 13 |
| Минимум и Максимум | 35 - 70 |
| Число сигарет в день | Количество значений | 630 |
| Количество пропущенных значений | 2 |
| Среднее значение | 10.68 |
| 95% ДИ для среднего | 9.66 - 11.7 |
| Стандартное отклонение | 13.04 |
| Медиана (Q1-Q3) | 3 (0 - 20) |
| IQR | 20 |
| Минимум и Максимум | 0 - 60 |
| Уровень общего холестерина | Количество значений | 631 |
| Количество пропущенных значений | 1 |
| Среднее значение | 245.57 |
| 95% ДИ для среднего | 241.81 - 249.33 |
| Стандартное отклонение | 48.14 |
| Медиана (Q1-Q3) | 241 (214 - 272) |
| IQR | 58 |
| Минимум и Максимум | 107 - 600 |
| Систолическое АД | Количество значений | 632 |
| Количество пропущенных значений | 0 |
| Среднее значение | 143.62 |
| 95% ДИ для среднего | 141.53 - 145.72 |
| Стандартное отклонение | 26.8 |
| Медиана (Q1-Q3) | 139 (125 - 158) |
| IQR | 33 |
| Минимум и Максимум | 83.5 - 295 |
| Диастолическое АД | Количество значений | 632 |
| Количество пропущенных значений | 0 |
| Среднее значение | 86.96 |
| 95% ДИ для среднего | 85.86 - 88.06 |
| Стандартное отклонение | 14.1 |
| Медиана (Q1-Q3) | 85 (78 - 95) |
| IQR | 17 |
| Минимум и Максимум | 48 - 140 |
| Индекс массы тела | Количество значений | 625 |
| Количество пропущенных значений | 7 |
| Среднее значение | 26.55 |
| 95% ДИ для среднего | 26.19 - 26.9 |
| Стандартное отклонение | 4.53 |
| Медиана (Q1-Q3) | 26.13 (23.48 - 28.92) |
| IQR | 5.44 |
| Минимум и Максимум | 15.96 - 56.8 |
| Частота сердечных сокращений | Количество значений | 631 |
| Количество пропущенных значений | 1 |
| Среднее значение | 76.53 |
| 95% ДИ для среднего | 75.57 - 77.49 |
| Стандартное отклонение | 12.28 |
| Медиана (Q1-Q3) | 75 (68 - 85) |
| IQR | 17 |
| Минимум и Максимум | 50 - 120 |
| Уровень глюкозы | Количество значений | 594 |
| Количество пропущенных значений | 38 |
| Среднее значение | 89.01 |
| 95% ДИ для среднего | 85.69 - 92.32 |
| Стандартное отклонение | 41.14 |
| Медиана (Q1-Q3) | 79 (72 - 90) |
| IQR | 18 |
| Минимум и Максимум | 40 - 394 |

## Категориальные переменные

Рассчитайте для всех категориальных переменных для каждой группы (TenYearCHD): - Абсолютное количество; - Относительное количество внутри группы; - 95% ДИ для доли внутри группы - задание со звёздочкой.

for(i in match(names(  
 cleaned\_data %>% select(where(is.factor) & !"Риск развития ИБС" )),  
 colnames(cleaned\_data))  
 ) {  
   
 print(  
 cleaned\_data %>%   
 count(`Риск развития ИБС`, cleaned\_data[,i], name = "Абсолютное количество") %>%   
 group\_by(`Риск развития ИБС`) %>%   
 mutate(`Относительное количество` = paste0 (round(  
 `Абсолютное количество` / sum(`Абсолютное количество`) \* 100, 2), " %"),   
 across(where(is.factor), ~ fct\_na\_value\_to\_level(.x, level = "Пропущенные значения"))) %>%   
 ungroup() %>%   
 flextable() %>%   
 theme\_box() %>%   
 merge\_v(c("Риск развития ИБС")) %>%   
 autofit()  
 ) }

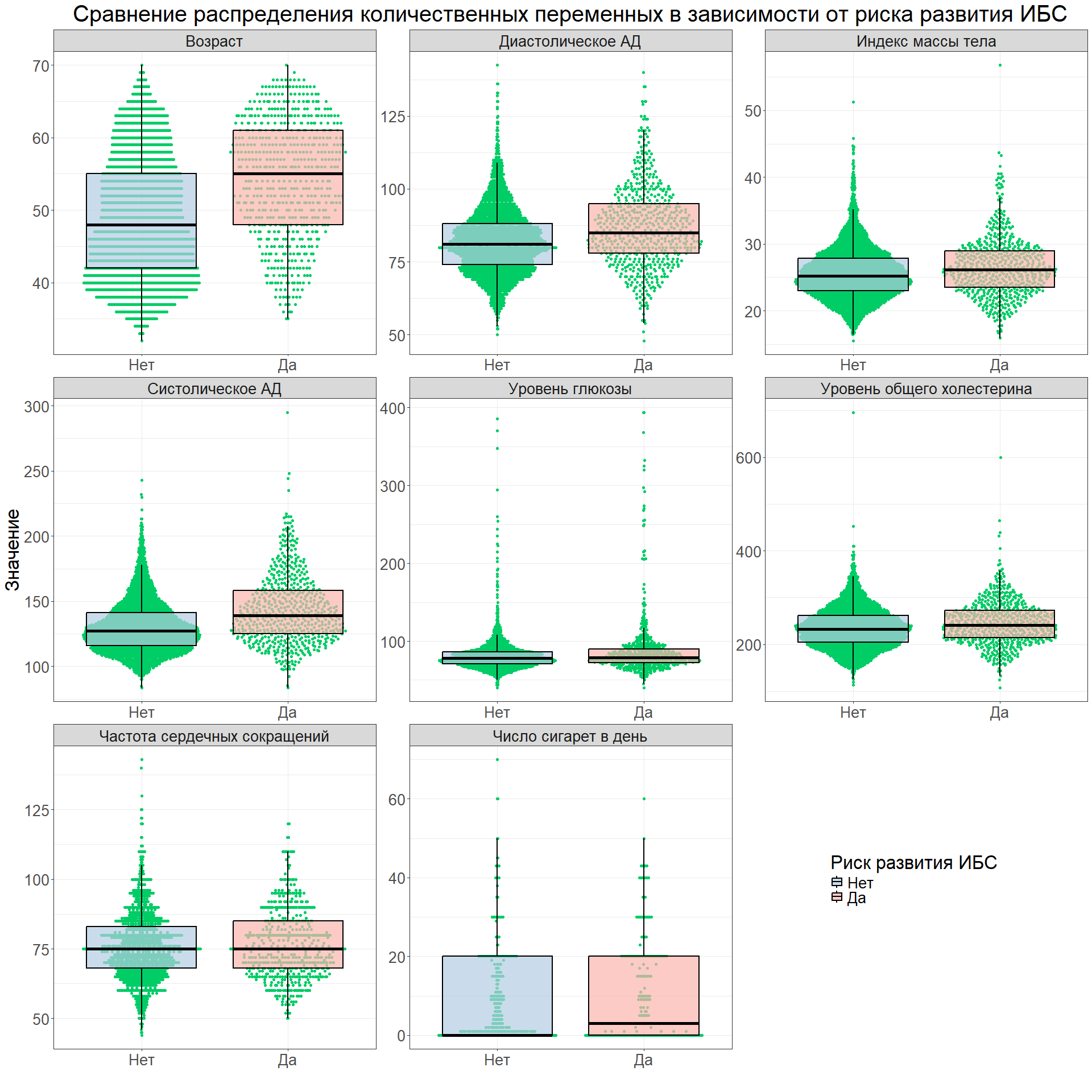
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Пол`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 4 row(s)   
## original dataset sample:   
## Риск развития ИБС Пол Абсолютное количество Относительное количество  
## 1 Нет Ж 2079 58.65 %  
## 2 Нет М 1466 41.35 %  
## 3 Да Ж 292 46.2 %  
## 4 Да М 340 53.8 %  
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Образование`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 10 row(s)   
## original dataset sample:   
## Риск развития ИБС Образование Абсолютное количество  
## 1 Нет 1 1381  
## 2 Нет 2 1095  
## 3 Нет 3 592  
## 4 Нет 4 399  
## 5 Нет Пропущенные значения 78  
## Относительное количество  
## 1 38.96 %  
## 2 30.89 %  
## 3 16.7 %  
## 4 11.26 %  
## 5 2.2 %  
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Курильщик`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 4 row(s)   
## original dataset sample:   
## Риск развития ИБС Курильщик Абсолютное количество Относительное количество  
## 1 Нет Нет 1816 51.23 %  
## 2 Нет Да 1729 48.77 %  
## 3 Да Нет 305 48.26 %  
## 4 Да Да 327 51.74 %  
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Прием препаратов для АД`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 6 row(s)   
## original dataset sample:   
## Риск развития ИБС Прием препаратов для АД Абсолютное количество  
## 1 Нет Нет 3424  
## 2 Нет Да 81  
## 3 Нет Пропущенные значения 40  
## 4 Да Нет 582  
## 5 Да Да 39  
## Относительное количество  
## 1 96.59 %  
## 2 2.28 %  
## 3 1.13 %  
## 4 92.09 %  
## 5 6.17 %  
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Перенесенный инсульт`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 4 row(s)   
## original dataset sample:   
## Риск развития ИБС Перенесенный инсульт Абсолютное количество  
## 1 Нет Нет 3531  
## 2 Нет Да 14  
## 3 Да Нет 621  
## 4 Да Да 11  
## Относительное количество  
## 1 99.61 %  
## 2 0.39 %  
## 3 98.26 %  
## 4 1.74 %  
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Гипертония`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 4 row(s)   
## original dataset sample:   
## Риск развития ИБС Гипертония Абсолютное количество Относительное количество  
## 1 Нет Нет 2565 72.36 %  
## 2 Нет Да 980 27.64 %  
## 3 Да Нет 313 49.53 %  
## 4 Да Да 319 50.47 %  
## a flextable object.  
## col\_keys: `Риск развития ИБС`, `Диабет`, `Абсолютное количество`, `Относительное количество`   
## header has 1 row(s)   
## body has 4 row(s)   
## original dataset sample:   
## Риск развития ИБС Диабет Абсолютное количество Относительное количество  
## 1 Нет Нет 3476 98.05 %  
## 2 Нет Да 69 1.95 %  
## 3 Да Нет 594 93.99 %  
## 4 Да Да 38 6.01 %

# Визуализация

## Количественные переменные

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

cleaned\_data %>%   
 select(where(is.numeric) | `Риск развития ИБС`) %>%   
 pivot\_longer(!`Риск развития ИБС`, names\_to = "Переменная", values\_to = "Значение") %>%   
 filter(!is.na(Значение)) %>%   
 ggplot(aes(`Риск развития ИБС`, Значение)) +  
 geom\_quasirandom(colour = "springgreen3")+  
 geom\_boxplot(aes(fill = `Риск развития ИБС`), outliers = FALSE, color = "black", linewidth = 1, alpha = 0.7)+  
 scale\_fill\_brewer(palette = "Pastel1", direction = -1)+  
 facet\_wrap( ~ Переменная, scales = "free")+  
 labs(x = "", title = "Сравнение распределения количественных переменных в зависимости от риска развития ИБС")+  
 theme\_custom+  
 theme(  
 legend.position = "inside",   
 legend.justification = c(0.90, 0.15)  
 )

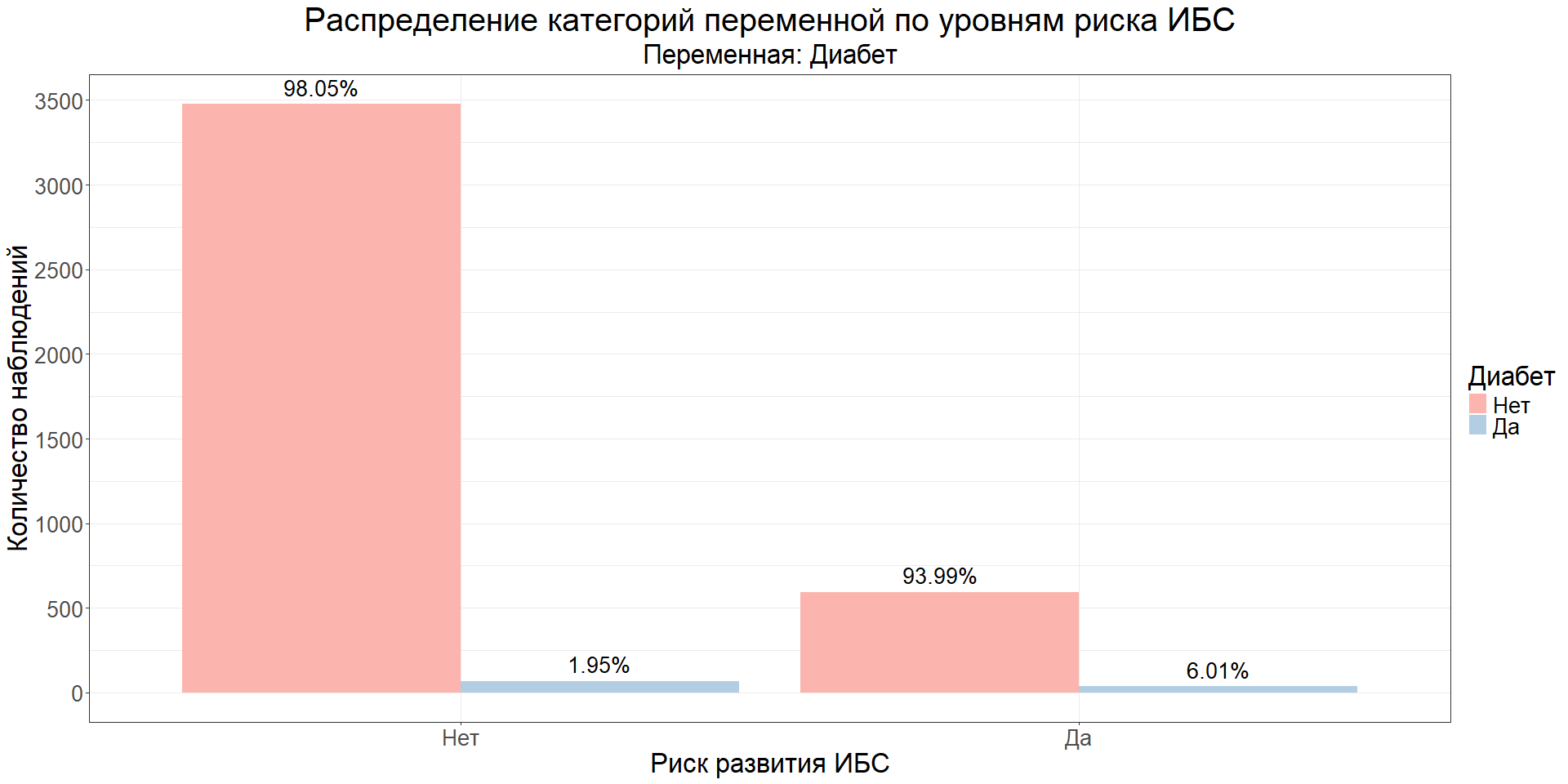
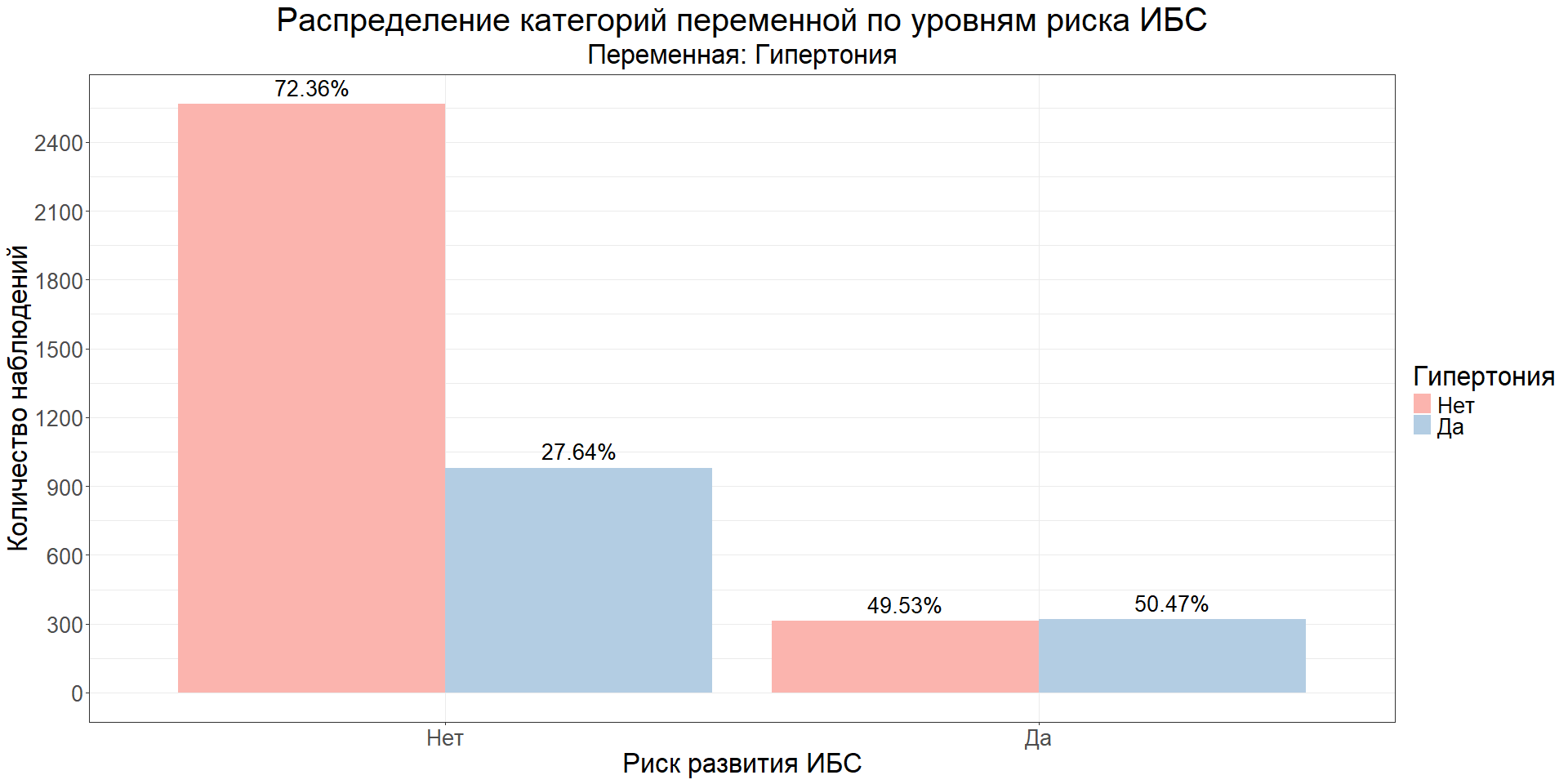
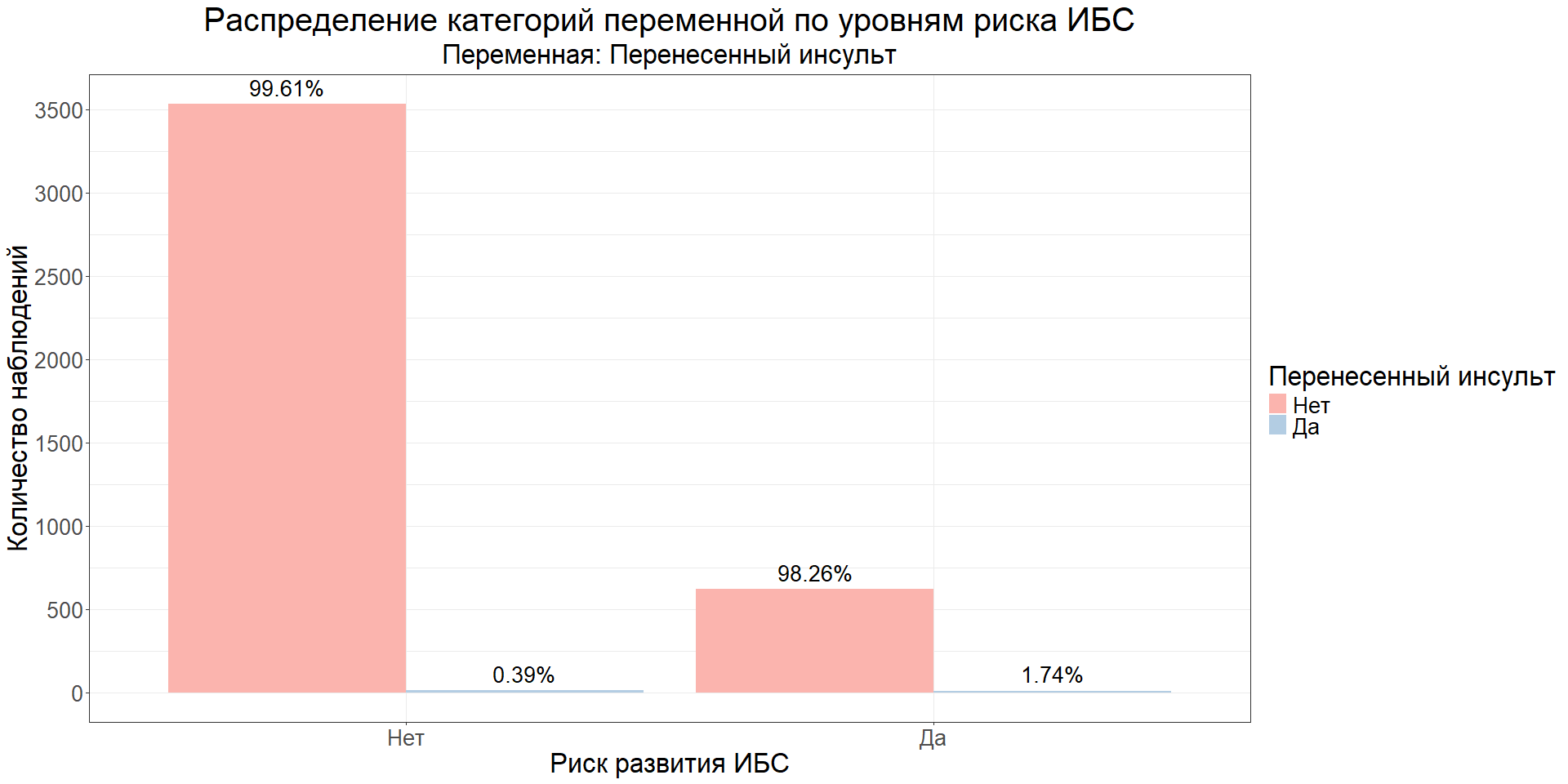
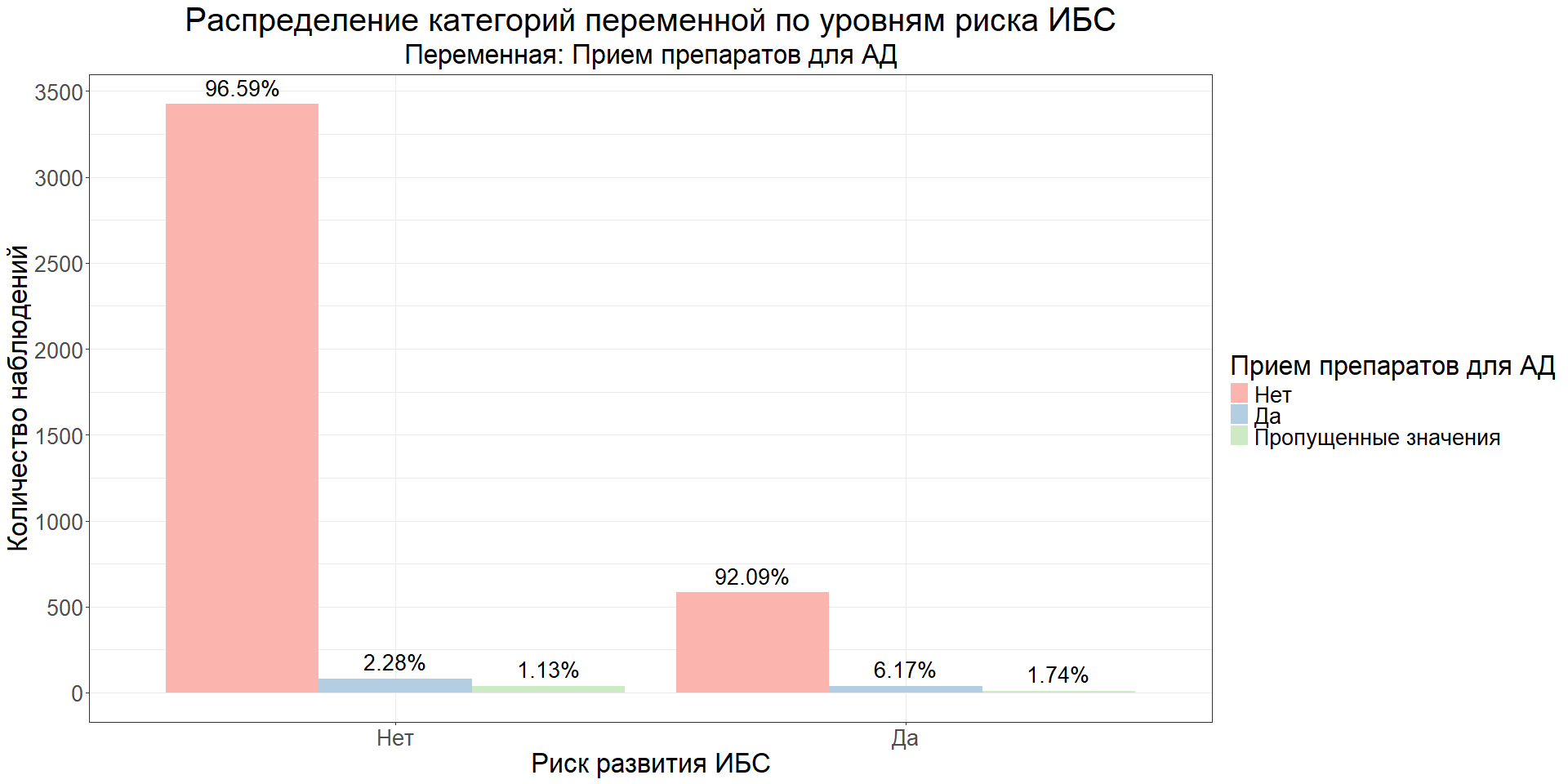
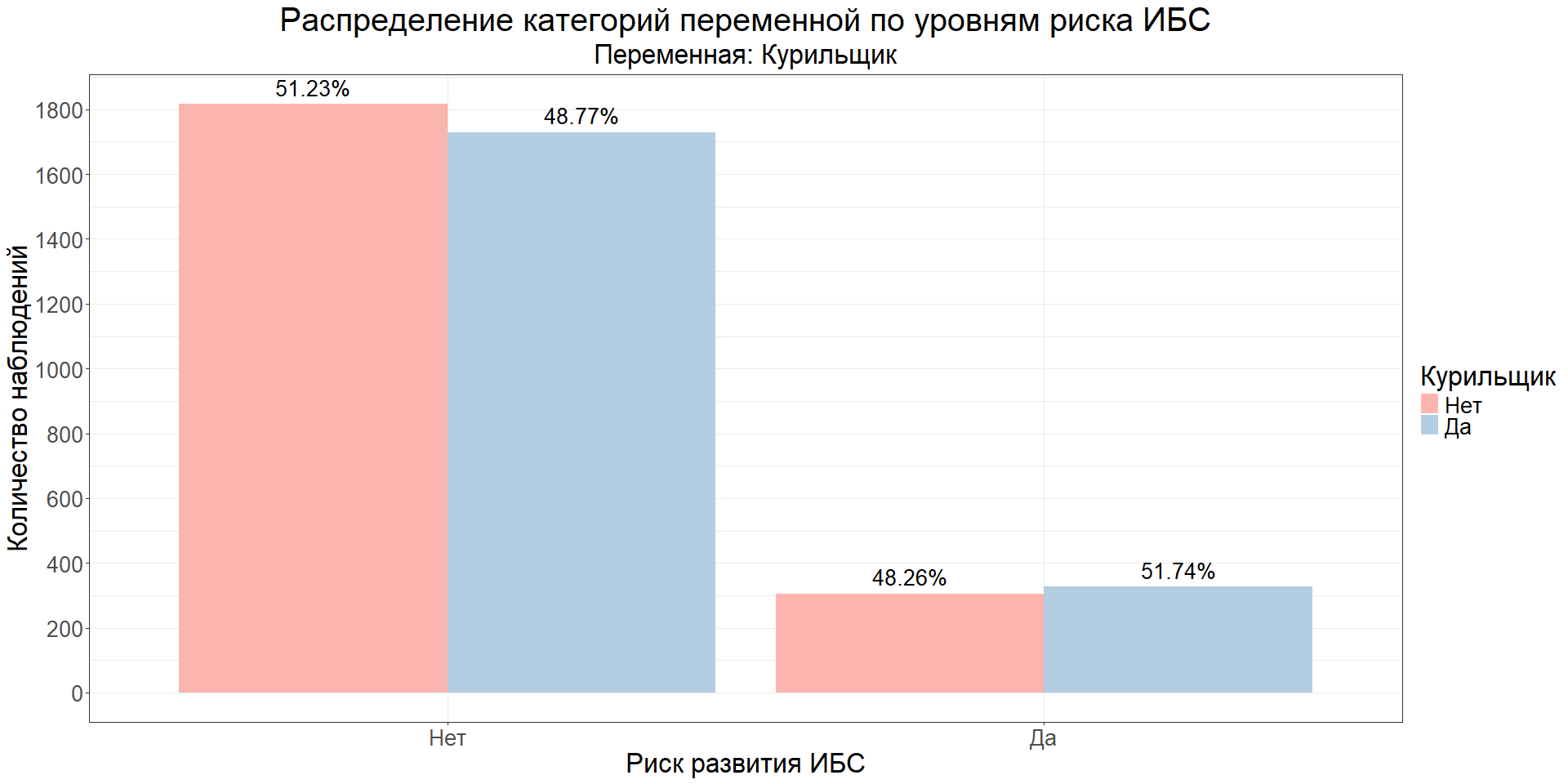
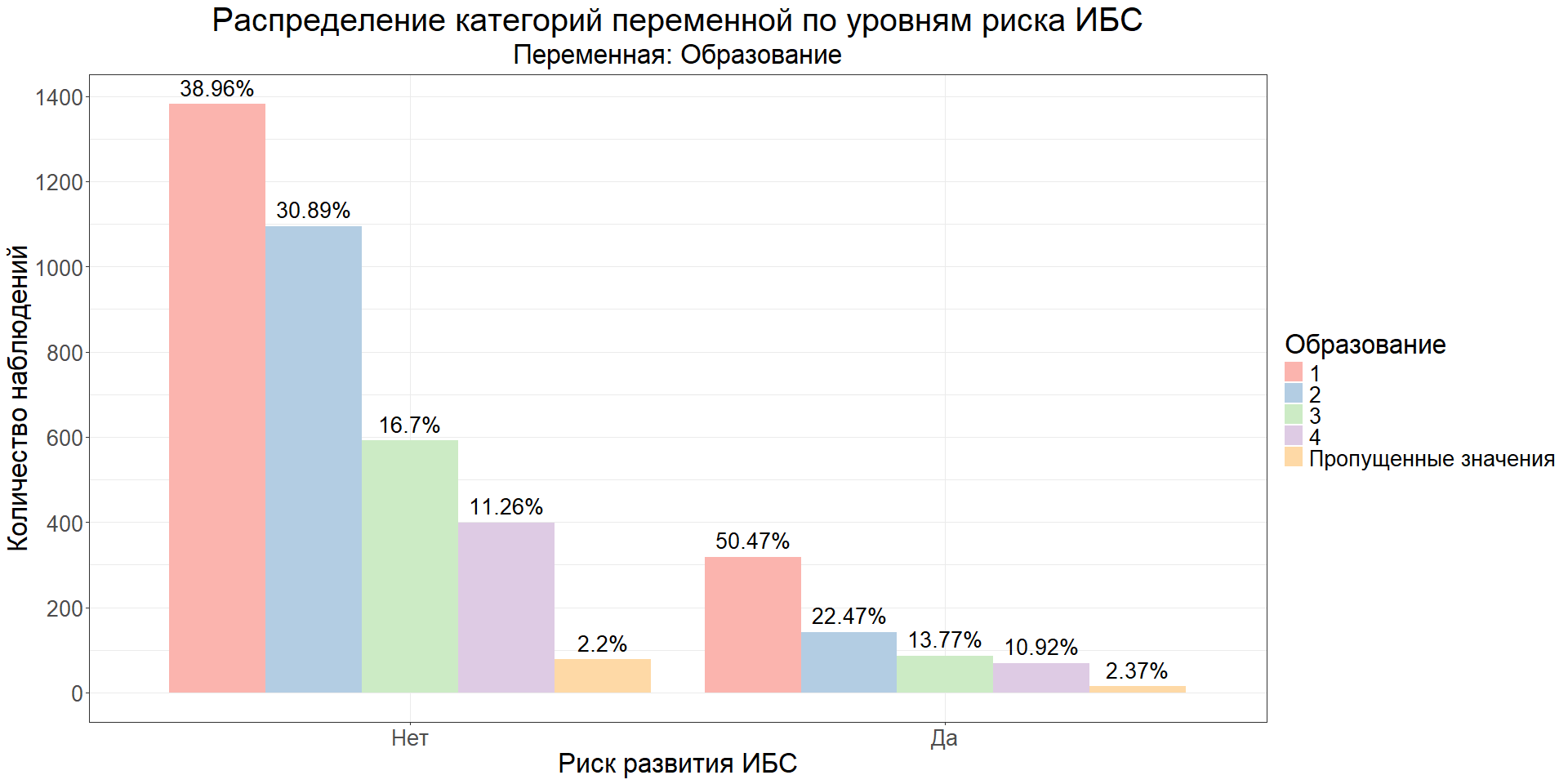
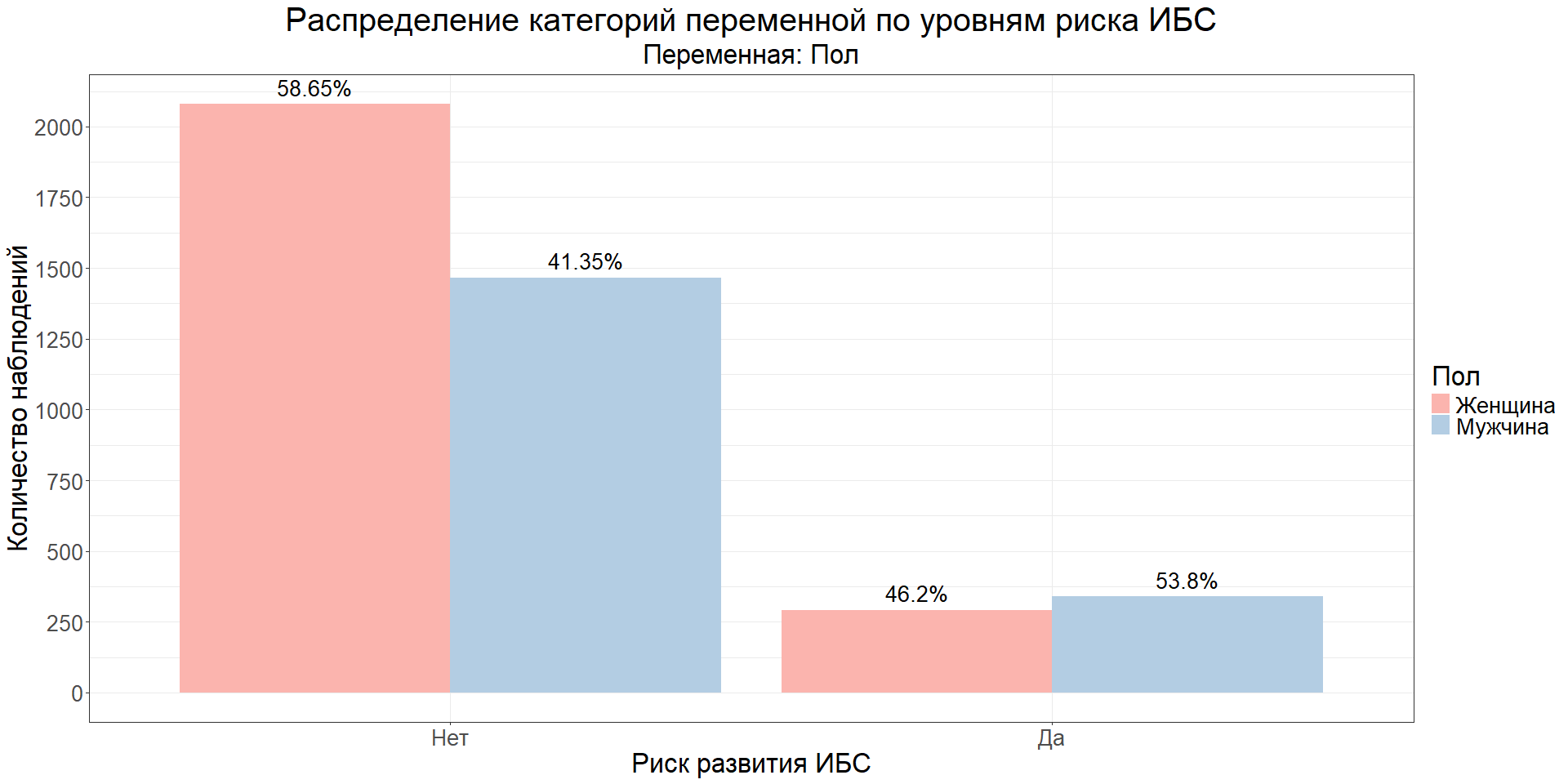


## Категориальные переменные

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

Для графического представления двух категориальных переменных отлично подойдет **Столбчатая диаграмма**. Построенные графики четко визуализирует статистические данные:

for(i in match(names(  
 cleaned\_data %>% select(where(is.factor) & !"Риск развития ИБС" )),  
 colnames(cleaned\_data))  
 ) {  
   
 print(  
 cleaned\_data %>%   
 rename (fill = names(cleaned\_data[,i])) %>%   
 count(`Риск развития ИБС`, fill, name = "Абсолютное количество") %>%   
 group\_by(`Риск развития ИБС`) %>%   
 mutate(`Относительное количество` =   
 `Абсолютное количество` / sum(`Абсолютное количество`),   
 across(where(is.factor), ~ fct\_na\_value\_to\_level(.x, level = "Пропущенные значения"))) %>%   
 ungroup() %>%   
 ggplot(aes(x = `Риск развития ИБС`, y = `Абсолютное количество`, fill = fill)) +   
 geom\_col(position = position\_dodge(width = 0.9)) +  
 geom\_text(aes(label = paste0(round(`Относительное количество`\*100, 2), "%")),   
 position = position\_dodge(width = 0.9),   
 size = 7, vjust = -0.5) +   
 scale\_fill\_brewer(name = names(cleaned\_data[,i]),  
 palette = "Pastel1",  
 labels = c("Ж" = "Женщина", "М" = "Мужчина"))+  
 scale\_y\_continuous(n.breaks = 10) +  
 labs(y = "Количество наблюдений",   
 title = paste0("Распределение категорий переменной по уровням риска ИБС"),   
 subtitle = paste0("Переменная: ", names(cleaned\_data[,i])))+  
 theme\_custom  
 )  
 }



# 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?

Все количественные переменные в различных группах Риска развития ИБС соответствуют нормальному распределению, что подтверждено результатами теста Шапиро-Уилка. Если p-value, полученное в результате данного теста, меньше заранее установленного уровня значимости (0.05), мы отклоняем нулевую гипотезу о том, что данные ненормальны. Из-за множественных сравнений решение о нормальности распределения принималось на основании скорректированных значений p-value с использованием поправки Холма.

cleaned\_data %>%   
 group\_by(`Риск развития ИБС`) %>%  
 summarise(across(where(is.numeric),   
 ~ shapiro.test(.x)$p.value )) %>%   
 ungroup() %>%   
 pivot\_longer(!`Риск развития ИБС`, names\_to = "Переменная", values\_to = "p\_value") %>%   
 mutate(p\_value\_adjusted = p.adjust(p\_value, method = "holm"),   
 Результат = ifelse(p\_value\_adjusted < 0.05,  
 "Нормальное распределение",  
 "Ненормальное распределение"))

## # A tibble: 16 × 5  
## `Риск развития ИБС` Переменная p\_value p\_value\_adjusted Результат  
## <fct> <chr> <dbl> <dbl> <chr>   
## 1 Нет Возраст 6.77e-29 7.45e-28 Нормальн…  
## 2 Нет Число сигарет в день 3.57e-58 5.35e-57 Нормальн…  
## 3 Нет Уровень общего холес… 1.42e-25 1.28e-24 Нормальн…  
## 4 Нет Систолическое АД 4.13e-35 5.37e-34 Нормальн…  
## 5 Нет Диастолическое АД 1.27e-23 8.91e-23 Нормальн…  
## 6 Нет Индекс массы тела 3.98e-30 4.78e-29 Нормальн…  
## 7 Нет Частота сердечных со… 1.43e-24 1.15e-23 Нормальн…  
## 8 Нет Уровень глюкозы 1.93e-62 3.08e-61 Нормальн…  
## 9 Да Возраст 2.75e- 9 5.49e- 9 Нормальн…  
## 10 Да Число сигарет в день 1.41e-27 1.41e-26 Нормальн…  
## 11 Да Уровень общего холес… 1.12e-13 4.61e-13 Нормальн…  
## 12 Да Систолическое АД 9.22e-14 4.61e-13 Нормальн…  
## 13 Да Диастолическое АД 1.04e- 7 1.04e- 7 Нормальн…  
## 14 Да Индекс массы тела 3.51e-14 2.11e-13 Нормальн…  
## 15 Да Частота сердечных со… 2.92e-10 8.75e-10 Нормальн…  
## 16 Да Уровень глюкозы 2.40e-37 3.36e-36 Нормальн…

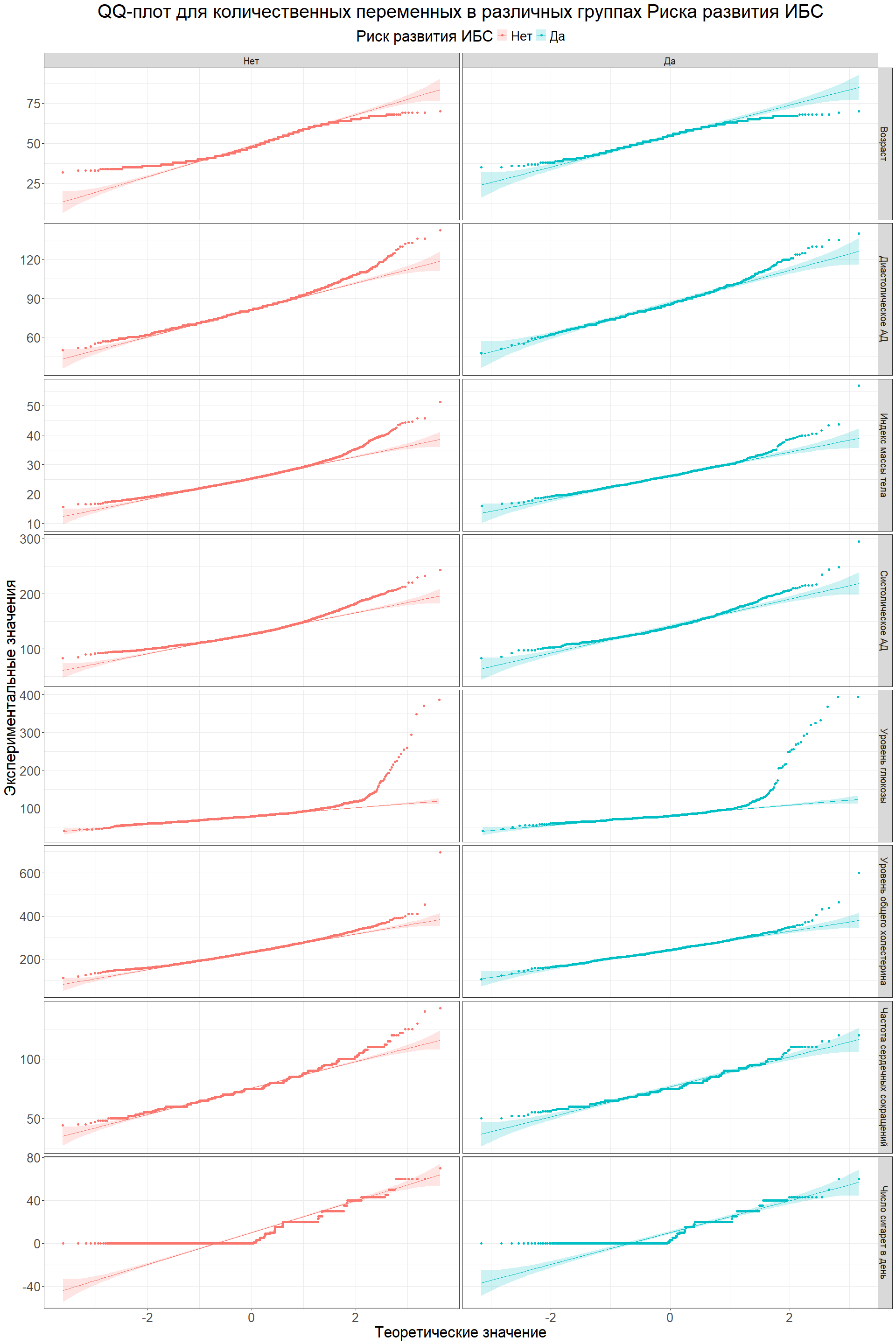
1. Постройте для каждой количественной переменной QQ-плот. Отличаются ли выводы от теста Шапиро-Уилка? Какой метод вы бы предпочли и почему?

График квантиль-квантиль (Q-Q) используется для оценки соответствия набора данных определенному теоретическому распределению. Он сравнивает квантили наблюдаемых данных с квантилями теоретического распределения.

В большинстве случаев количественные переменные в представленных данных располагаются вдоль диагональной линии, что свидетельствует о нормальности распределения. Отклонения от диагонали в верхней и нижней части могут указывать на тяжелые хвосты распределения и наличие выбросов.

Я предпочел бы использовать оба метода — тест Шапиро-Уилка и график Q-Q. Тест дает количественную оценку, а визуализация позволяет быстро выявить возможные отклонения, что дает более полное представление о распределении данных.

cleaned\_data %>%   
select(where(is.numeric) | `Риск развития ИБС`) %>%   
 pivot\_longer(!`Риск развития ИБС`, names\_to = "Переменная", values\_to = "Значение") %>%   
 filter(!is.na(Значение)) %>%   
 ggqqplot(., x = "Значение", color = "Риск развития ИБС") +  
 facet\_grid( Переменная ~ `Риск развития ИБС`, scales = "free")+  
 labs(title = "QQ-плот для количественных переменных в различных группах Риска развития ИБС",   
 x = "Теоретические значение", y = "Экспериментальные значения")+  
 theme\_custom+  
 theme(legend.position = "top",   
 strip.text = element\_text(size = 15))



1. Ниже напишите, какие ещё методы проверки на нормальность вы знаете и какие у них есть ограничения.

Я использовал ПО Minitab, где помимо теста Шапиро-Уилка были следующие методы: - Критерий Андерсона-Дарлинга, который более эффективен для обнаружения ненормальности в хвостах распределения. - Критерий Колмогорова-Смирнова, который менее чувствителен к небольшим отклонениям от нормального распределения.

Кроме того, в качестве альтернативы можно использовать график плотности. Этот метод позволяет визуально оценить форму распределения и выявить возможные отклонения от нормальности.

## Сравнение групп

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.

### Количественные переменные

Для сравнения количественных переменных между группами по переменной **Риск развития ИБС** использовался двухвыборочный t-тест. Тест позволяет оценить различия средних значений между двумя независимыми группами. В данном случае, значения p-value оказались ниже уровня значимости (0.05) для всех переменных, кроме **Частоты сердечных сокращений**, что свидетельствует о наличии значимых различий между средними значениями в этих группах для большинства переменных. Учитывая множественные сравнения, окончательное решение о значимости принималось на основании скорректированных значений p-value с использованием поправки Холма.

test\_tibble <- tibble()  
n <- 0  
  
for(i in match(  
 names(cleaned\_data %>% select(where(is.numeric))),  
 colnames(cleaned\_data)  
 )) {  
  
n <- n+1  
   
test\_tibble[n,1] <- names(cleaned\_data[i])  
test\_tibble[n,2] <- t.test(cleaned\_data[[i]] ~ `Риск развития ИБС`, data = cleaned\_data)$p.value   
test\_tibble[n,3] <- round(diff(t.test(cleaned\_data[[i]] ~ `Риск развития ИБС`, data = cleaned\_data)$estimate),2)  
test\_tibble[n,4] <- paste0(round (t.test(cleaned\_data[[i]] ~ `Риск развития ИБС`, data = cleaned\_data)$conf.int[1], 2),  
 " ... ",  
 round (t.test(cleaned\_data[[i]] ~ `Риск развития ИБС`, data = cleaned\_data)$conf.int[2], 2))  
  
}  
  
  
test\_tibble %>%   
 transmute("Переменная" = ...1,   
 "p\_value" = ...2,   
 "p\_value\_adjusted" = p.adjust(p\_value, method = "holm"),   
 "Разница средних" = ...3,   
 "95% ДИ" = ...4)

## # A tibble: 8 × 5  
## Переменная p\_value p\_value\_adjusted `Разница средних` `95% ДИ`  
## <chr> <dbl> <dbl> <dbl> <chr>   
## 1 Возраст 2.94e-50 2.35e-49 5.54 -6.22 .…  
## 2 Число сигарет в день 3.25e- 4 6.50e- 4 2.01 -3.1 ..…  
## 3 Уровень общего холестери… 5.09e- 7 2.55e- 6 10.4 -14.42 …  
## 4 Систолическое АД 4.43e-30 3.10e-29 13.3 -15.52 …  
## 5 Диастолическое АД 1.84e-15 1.10e-14 4.81 -5.97 .…  
## 6 Индекс массы тела 6.04e- 6 1.81e- 5 0.88 -1.26 .…  
## 7 Частота сердечных сокращ… 1.44e- 1 1.44e- 1 0.77 -1.81 .…  
## 8 Уровень глюкозы 1.62e- 6 6.48e- 6 8.33 -11.71 …

### Категориальные переменные

Для сравнения категориальных переменных между группами по переменной **Риск развития ИБС** применялась мера ассоциации — V Крамера. Эта мера подходит для оценки силы связи между номинальными переменными, так как стандартизирована и принимает значения от 0 до 1. Для оценки значимости связи использовались значения p-value, полученные при расчете критерия хи-квадрат. Окончательное решение о значимости принималось на основе скорректированных p-value с использованием поправки Холма. Для всех категориальных переменных, кроме **Частоты сердечных сокращений**, значения p-value оказались ниже уровня значимости (0.05), что свидетельствует о значимых различиях между группами.

cramer\_tibble <- tibble()  
  
n <- 0  
  
for(i in match(  
 names(cleaned\_data %>% select(where(is.factor) & !"Риск развития ИБС")),  
 colnames(cleaned\_data)  
)) {  
   
n <- n + 1  
   
cramer\_tibble[n,1] <- names(cleaned\_data[i])  
cramer\_tibble[n,2] <- Cramers\_V(cleaned\_data$`Риск развития ИБС`, cleaned\_data[[i]])   
cramer\_tibble[n,3] <- chisq.test(cleaned\_data$`Риск развития ИБС`, cleaned\_data[[i]])$p.value  
  
   
}  
  
cramer\_tibble %>%   
 transmute("Переменная" = ...1,   
 "V\_Крамера" = ...2,   
 "p\_value" = ...3,   
 "p\_value\_adjusted" = p.adjust(p\_value, method = "holm"))

## # A tibble: 7 × 4  
## Переменная V\_Крамера p\_value p\_value\_adjusted  
## <chr> <dbl> <dbl> <dbl>  
## 1 Пол 0.0893 7.76e- 9 3.88e- 8  
## 2 Образование 0.0902 2.84e- 7 8.51e- 7  
## 3 Курильщик 0.0206 1.83e- 1 1.83e- 1  
## 4 Прием препаратов для АД 0.0824 1.19e- 7 4.74e- 7  
## 5 Перенесенный инсульт 0.0582 1.70e- 4 3.39e- 4  
## 6 Гипертония 0.176 5.52e-30 3.87e-29  
## 7 Диабет 0.0901 5.74e- 9 3.44e- 8

# 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