automatization\_notebook\_04

2024-01-07

# Чтение данных

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

path <- here('data', 'raw','healthcare-dataset-stroke-data.csv')  
df <- read\_csv(path)

## 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.

# Выведите общее описание данных

head(df)

## # A tibble: 6 × 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   
## # ℹ 5 more variables: Residence\_type <chr>, avg\_glucose\_level <dbl>, bmi <chr>,  
## # smoking\_status <chr>, stroke <dbl>

# 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”.

df <- df %>%  
 mutate(id = as.character(id), bmi = as.numeric(bmi)) %>%   
 mutate(across(c(gender, hypertension, heart\_disease, ever\_married, work\_type, Residence\_type, smoking\_status, stroke), as.factor))# Приведение к нужному типу

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `bmi = as.numeric(bmi)`.  
## Caused by warning:  
## ! NAs introduced by coercion

cleaned\_data <- df %>% select(where(~ (sum(is.na(.)) / nrow(df)) < 0.2)) %>% # Отбор и удаление столбцов, в которых пропущенных значений более 20%  
 filter(rowSums(is.na(.)) < length(df) \* 0.2) %>% # Отбор и удаление строк, в которых пропущенных значений более 20%. Т.к. количество пропущенных значений и в столюцах и строках было менее 20%, данные не изменились   
 rename\_with(function(x) x %>% stri\_replace\_all\_regex("\_", " ", vectorize\_all = F)) %>% # Переименование  
 arrange(desc(age)) # Сортировка данных по возрасту по убыванию

# Сколько осталось переменных?

cleaned\_data %>% glimpse()

## Rows: 5,110  
## Columns: 12  
## $ id <chr> "64778", "39373", "712", "61960", "38829", "37132"…  
## $ gender <fct> Male, Female, Female, Male, Female, Male, Male, Fe…  
## $ age <dbl> 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82…  
## $ hypertension <fct> 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ `heart disease` <fct> 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0,…  
## $ `ever married` <fct> Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, No, No…  
## $ `work type` <fct> Private, Self-employed, Private, Private, Private,…  
## $ `Residence type` <fct> Rural, Urban, Rural, Urban, Rural, Urban, Rural, U…  
## $ `avg glucose level` <dbl> 208.30, 196.92, 84.03, 144.90, 59.32, 200.59, 103.…  
## $ bmi <dbl> 32.5, 22.2, 26.5, 26.4, 33.2, 29.0, 25.0, 27.9, 29…  
## $ `smoking status` <fct> Unknown, never smoked, formerly smoked, smokes, ne…  
## $ stroke <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,…

print(paste('Осталось', length(cleaned\_data), 'переменных'))

## [1] "Осталось 12 переменных"

# Сколько осталось случаев?

cleaned\_data %>% glimpse()

## Rows: 5,110  
## Columns: 12  
## $ id <chr> "64778", "39373", "712", "61960", "38829", "37132"…  
## $ gender <fct> Male, Female, Female, Male, Female, Male, Male, Fe…  
## $ age <dbl> 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82…  
## $ hypertension <fct> 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ `heart disease` <fct> 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0,…  
## $ `ever married` <fct> Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, No, No…  
## $ `work type` <fct> Private, Self-employed, Private, Private, Private,…  
## $ `Residence type` <fct> Rural, Urban, Rural, Urban, Rural, Urban, Rural, U…  
## $ `avg glucose level` <dbl> 208.30, 196.92, 84.03, 144.90, 59.32, 200.59, 103.…  
## $ bmi <dbl> 32.5, 22.2, 26.5, 26.4, 33.2, 29.0, 25.0, 27.9, 29…  
## $ `smoking status` <fct> Unknown, never smoked, formerly smoked, smokes, ne…  
## $ stroke <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,…

print(paste('Осталось', nrow(cleaned\_data), 'случаев'))

## [1] "Осталось 5110 случаев"

# Есть ли в данных идентичные строки?

duplicates <- cleaned\_data %>% filter(duplicated(.))  
print(paste('Идентичных строк', nrow(duplicates)))

## [1] "Идентичных строк 0"

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

missing\_var <- cleaned\_data %>%  
 summarise\_all(~ sum(is.na(.))) %>%  
 gather('Переменные', 'Пропущенные значения')   
  
missing\_var

## # A tibble: 12 × 2  
## Переменные `Пропущенные значения`  
## <chr> <int>  
## 1 id 0  
## 2 gender 0  
## 3 age 0  
## 4 hypertension 0  
## 5 heart disease 0  
## 6 ever married 0  
## 7 work type 0  
## 8 Residence type 0  
## 9 avg glucose level 0  
## 10 bmi 201  
## 11 smoking status 0  
## 12 stroke 0

# 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.

statistics <- list(  
 `Количество субъектов` = ~length(.x) %>% as.character(),  
 `Количество (есть данные)` = ~sum(!is.na(.x)) %>% as.character(),  
 `Нет данных` = ~sum(is.na(.x)) %>% as.character(),  
 `Ср. знач.` = ~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()),  
 `95% ДИ для среднего` = ~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)))  
)  
  
cleaned\_data %>% select(stroke, where(is.numeric)) %>%   
 group\_by(stroke) %>%   
 summarise(across(where(is.numeric), statistics)) %>%   
 pivot\_longer(!stroke) %>%   
 separate(name, into = c('Переменная','Статистика'), sep = '\_') %>%   
 rename('Значения' = value) %>%   
 flextable()

| stroke | Переменная | Статистика | Значения |
| --- | --- | --- | --- |
| 0 | age | Количество субъектов | 4861 |
| 0 | age | Количество (есть данные) | 4861 |
| 0 | age | Нет данных | 0 |
| 0 | age | Ср. знач. | 41.97 |
| 0 | age | Станд. отклон. | 22.29 |
| 0 | age | 95% ДИ для среднего | 22.29 |
| 0 | age | мин. - макс. | 0.08 - 82 |
| 0 | age | Медиана | 43 |
| 0 | age | Q1 - Q3 | 24 - 59 |
| 0 | avg glucose level | Количество субъектов | 4861 |
| 0 | avg glucose level | Количество (есть данные) | 4861 |
| 0 | avg glucose level | Нет данных | 0 |
| 0 | avg glucose level | Ср. знач. | 104.8 |
| 0 | avg glucose level | Станд. отклон. | 43.85 |
| 0 | avg glucose level | 95% ДИ для среднего | 43.85 |
| 0 | avg glucose level | мин. - макс. | 55.12 - 267.76 |
| 0 | avg glucose level | Медиана | 91.47 |
| 0 | avg glucose level | Q1 - Q3 | 77.12 - 112.83 |
| 0 | bmi | Количество субъектов | 4861 |
| 0 | bmi | Количество (есть данные) | 4700 |
| 0 | bmi | Нет данных | 161 |
| 0 | bmi | Ср. знач. | 28.82 |
| 0 | bmi | Станд. отклон. | 7.91 |
| 0 | bmi | 95% ДИ для среднего | 7.91 |
| 0 | bmi | мин. - макс. | 10.3 - 97.6 |
| 0 | bmi | Медиана | 28 |
| 0 | bmi | Q1 - Q3 | 23.4 - 33.1 |
| 1 | age | Количество субъектов | 249 |
| 1 | age | Количество (есть данные) | 249 |
| 1 | age | Нет данных | 0 |
| 1 | age | Ср. знач. | 67.73 |
| 1 | age | Станд. отклон. | 12.73 |
| 1 | age | 95% ДИ для среднего | 12.73 |
| 1 | age | мин. - макс. | 1.32 - 82 |
| 1 | age | Медиана | 71 |
| 1 | age | Q1 - Q3 | 59 - 78 |
| 1 | avg glucose level | Количество субъектов | 249 |
| 1 | avg glucose level | Количество (есть данные) | 249 |
| 1 | avg glucose level | Нет данных | 0 |
| 1 | avg glucose level | Ср. знач. | 132.54 |
| 1 | avg glucose level | Станд. отклон. | 61.92 |
| 1 | avg glucose level | 95% ДИ для среднего | 61.92 |
| 1 | avg glucose level | мин. - макс. | 56.11 - 271.74 |
| 1 | avg glucose level | Медиана | 105.22 |
| 1 | avg glucose level | Q1 - Q3 | 79.79 - 196.71 |
| 1 | bmi | Количество субъектов | 249 |
| 1 | bmi | Количество (есть данные) | 209 |
| 1 | bmi | Нет данных | 40 |
| 1 | bmi | Ср. знач. | 30.47 |
| 1 | bmi | Станд. отклон. | 6.33 |
| 1 | bmi | 95% ДИ для среднего | 6.33 |
| 1 | bmi | мин. - макс. | 16.9 - 56.6 |
| 1 | bmi | Медиана | 29.7 |
| 1 | bmi | Q1 - Q3 | 26.4 - 33.7 |

## 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.

cleaned\_data %>%   
 select(stroke, where(is.factor)) %>%   
 tbl\_summary(by = stroke) %>%   
 add\_ci()

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **0**, N = 4,8611 | **95% CI**2 | **1**, N = 2491 | **95% CI**2 |
| --- | --- | --- | --- | --- |
| gender |  |  |  |  |
| Female | 2,853 (59%) | 57%, 60% | 141 (57%) | 50%, 63% |
| Male | 2,007 (41%) | 40%, 43% | 108 (43%) | 37%, 50% |
| Other | 1 (<0.1%) | 0.00%, 0.13% | 0 (0%) | 0.00%, 1.9% |
| hypertension |  |  |  |  |
| 0 | 4,429 (91%) | 90%, 92% | 183 (73%) | 67%, 79% |
| 1 | 432 (8.9%) | 8.1%, 9.7% | 66 (27%) | 21%, 33% |
| heart disease |  |  |  |  |
| 0 | 4,632 (95%) | 95%, 96% | 202 (81%) | 76%, 86% |
| 1 | 229 (4.7%) | 4.1%, 5.4% | 47 (19%) | 14%, 24% |
| ever married | 3,133 (64%) | 63%, 66% | 220 (88%) | 84%, 92% |
| work type |  |  |  |  |
| children | 685 (14%) | 13%, 15% | 2 (0.8%) | 0.14%, 3.2% |
| Govt\_job | 624 (13%) | 12%, 14% | 33 (13%) | 9.4%, 18% |
| Never\_worked | 22 (0.5%) | 0.29%, 0.70% | 0 (0%) | 0.00%, 1.9% |
| Private | 2,776 (57%) | 56%, 59% | 149 (60%) | 53%, 66% |
| Self-employed | 754 (16%) | 15%, 17% | 65 (26%) | 21%, 32% |
| Residence type |  |  |  |  |
| Rural | 2,400 (49%) | 48%, 51% | 114 (46%) | 40%, 52% |
| Urban | 2,461 (51%) | 49%, 52% | 135 (54%) | 48%, 60% |
| smoking status |  |  |  |  |
| formerly smoked | 815 (17%) | 16%, 18% | 70 (28%) | 23%, 34% |
| never smoked | 1,802 (37%) | 36%, 38% | 90 (36%) | 30%, 42% |
| smokes | 747 (15%) | 14%, 16% | 42 (17%) | 13%, 22% |
| Unknown | 1,497 (31%) | 30%, 32% | 47 (19%) | 14%, 24% |
| 1n (%) | | | | |
| 2CI = Confidence Interval | | | | |

# 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.

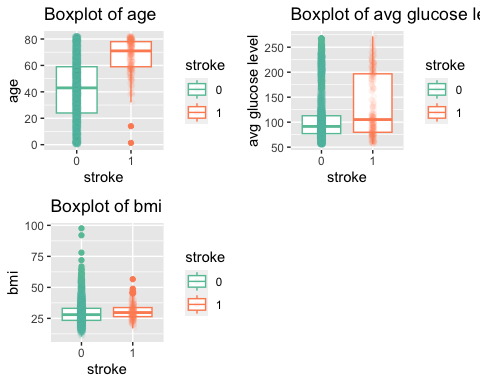
numeric\_columns <- cleaned\_data %>%  
 select(where(is.numeric))  
  
plots <- map(names(numeric\_columns), function(var) {  
 ggplot(cleaned\_data, aes\_string(x = "stroke", y = paste0("`", var, "`"), group = "stroke")) +  
 geom\_boxplot(aes(color = stroke)) +   
 geom\_point(aes(color = stroke), position = position\_jitter(width = 0.03), alpha = 0.05) +  
 scale\_color\_brewer(type = 'div', palette = 'Set2') +  
 labs(x = "stroke", y = var) +  
 ggtitle(paste("Boxplot of", var))  
})

## Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
## ℹ Please use tidy evaluation idioms with `aes()`.  
## ℹ See also `vignette("ggplot2-in-packages")` for more information.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

grid.arrange(grobs = plots, ncol = 2)

## Warning: Removed 201 rows containing non-finite values (`stat\_boxplot()`).

## Warning: Removed 201 rows containing missing values (`geom\_point()`).

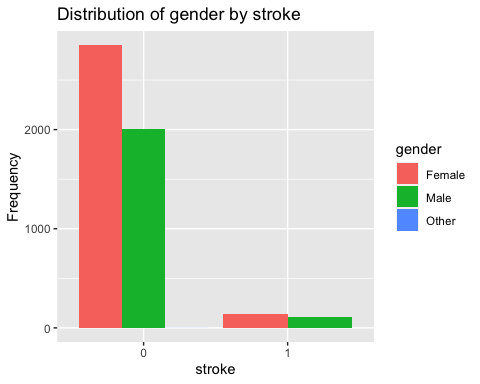


## Категориальные переменные

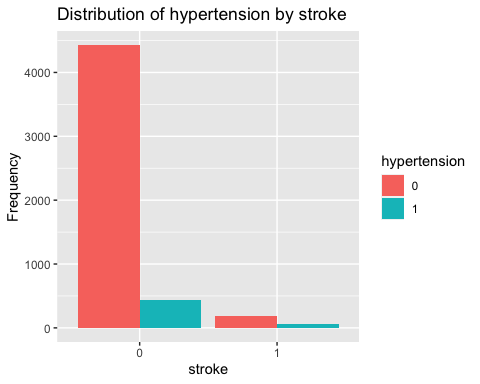
1. Сделайте подходящие визуализации категориальных переменных. Обоснуйте, почему выбрали именно этот тип.

categorical\_columns <- cleaned\_data %>%  
 select(where(is.factor))  
  
bars <- map(names(categorical\_columns), function(var) {  
 ggplot(cleaned\_data, aes\_string(x = "stroke", fill = sprintf("`%s`", var))) +  
 geom\_bar(position = "dodge") +  
 labs(x = "stroke", y = "Frequency", fill = var) +  
 ggtitle(paste("Distribution of", var, "by stroke"))  
})  
  
bars

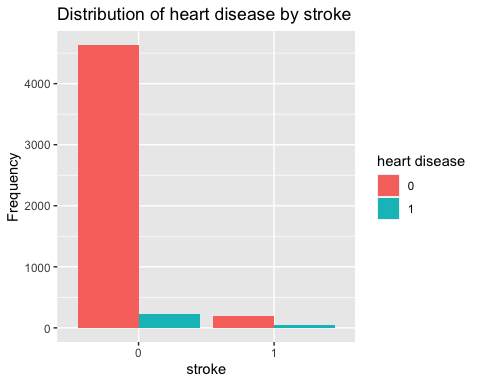
## [[1]]



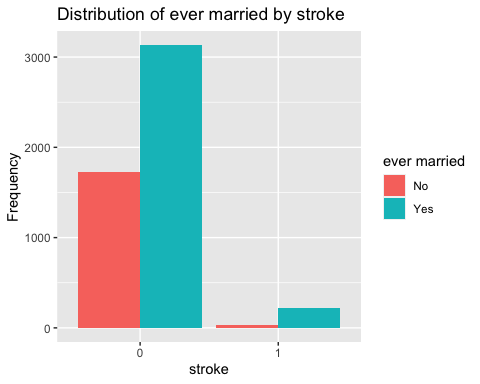
##   
## [[2]]



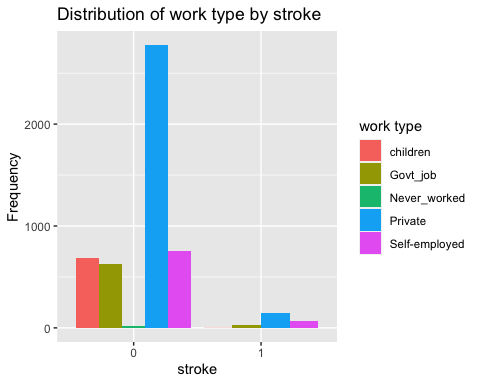
##   
## [[3]]



##   
## [[4]]



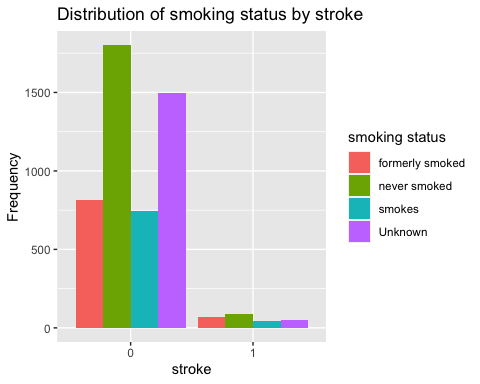
##   
## [[5]]



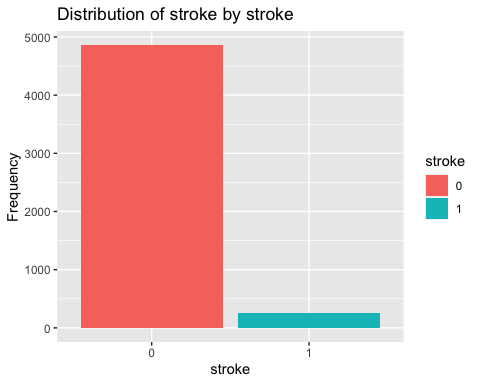
##   
## [[6]]



##   
## [[7]]



##   
## [[8]]



# Выбор оптимального способа визуализации категориальных переменных зависит от задачи, которую мы решаем. Выбор в пользу столбчатой диаграммы был сделан в силу хорошей демонстации распределения переменных в группах.

# Статистические оценки

## Проверка на нормальность

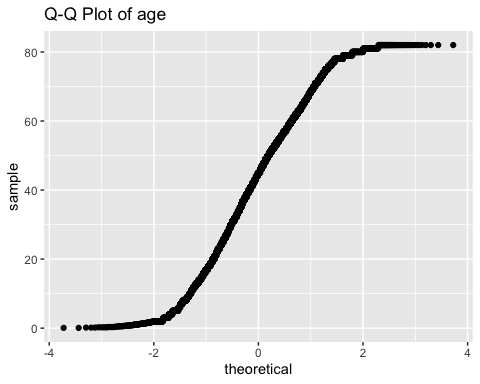
1. Оцените каждую переменную на соответствие нормальному распределению с помощью теста Шапиро-Уилка. Какие из переменных являются нормальными и как как вы это поняли?

# Тест Шапиро-Уилка имеет ограничение при применении на выборке <3 и > 5000, поскольку даже незначительные отклонения от нормальности будут квалифицироваться как значимые на обычных уровнях, следовательно тест Шапиро-Уилка в данном случае для всех количественных переменных будет <0.05.

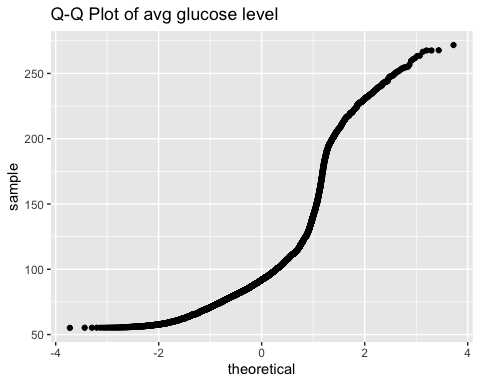
1. Постройте для каждой количественной переменной QQ-плот. Отличаются ли выводы от теста Шапиро-Уилка? Какой метод вы бы предпочли и почему?

numeric\_columns <- cleaned\_data %>%  
 select(where(is.numeric))  
  
qq\_plots <- lapply(names(numeric\_columns), function(var) {  
 ggplot(cleaned\_data, aes(sample = .data[[var]])) +  
 geom\_qq() +  
 labs(title = paste("Q-Q Plot of", var))  
})  
  
print(qq\_plots)

## [[1]]

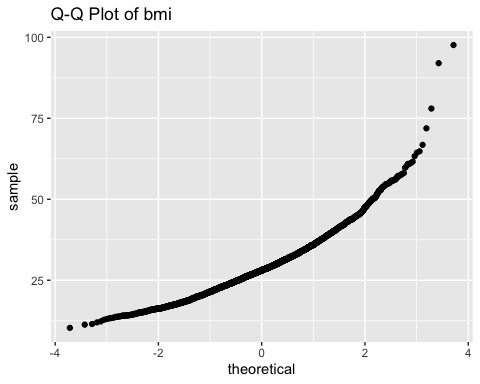


##   
## [[2]]



##   
## [[3]]

## Warning: Removed 201 rows containing non-finite values (`stat\_qq()`).



# Ни один метод не подтверждает нормальность окончательно, и у разных методов есть свои преимущества и ограничения. Визуальные методы нагляднее для быстрой оценки характера распределения, а также не имеют ограничений по размеру выборкм.  
#Распеределение возраста является равномерным и напоминает распределение Стьюдента. Ср. уровень глю имеет двугорбое распределение, а ИМТ напоминает экспоненциальное распределение.

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

**Напишите текст здесь** 1. Тест Колмогорова-Смирнова. Ограничения: он может быть чувствителен к размеру выборки, критерий требует, чтобы выборка была достаточно большой и лучше подходит для непрерывных данных. 2. Тест Лиллиефорса. Ограничения: он тоже чувствителен к размеру выборки и требует, чтобы выборка была достаточно большой. 3. Оценка распределения с помощью гистограмм. Ограничения: нет колическтвенной оценки, субъективная интерпретация,внешний вид может меняться в зависимости от выбранного интервала. 4. Коэффициенты ассиметрии и эксцесса. Ограничения: не учитывают плотность распределения, мультиможальность.

## 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.

cleaned\_data %>%   
 select(!id) %>%   
 tbl\_summary(by = stroke) %>%   
 add\_p(list(all\_continuous() ~ "t.test", all\_categorical() ~ "fisher.test"))

## There was an error in 'add\_p()/add\_difference()' for variable 'work type', p-value omitted:  
## Error in stats::fisher.test(structure(c(4L, 5L, 4L, 4L, 4L, 2L, 2L, 2L, : FEXACT error 7(location). LDSTP=17700 is too small for this problem,  
## (pastp=261.195, ipn\_0:=ipoin[itp=96]=526, stp[ipn\_0]=253.832).  
## Increase workspace or consider using 'simulate.p.value=TRUE'

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **0**, N = 4,8611 | **1**, N = 2491 | **p-value**2 |
| --- | --- | --- | --- |
| gender |  |  | 0.6 |
| Female | 2,853 (59%) | 141 (57%) |  |
| Male | 2,007 (41%) | 108 (43%) |  |
| Other | 1 (<0.1%) | 0 (0%) |  |
| age | 43 (24, 59) | 71 (59, 78) | <0.001 |
| hypertension |  |  | <0.001 |
| 0 | 4,429 (91%) | 183 (73%) |  |
| 1 | 432 (8.9%) | 66 (27%) |  |
| heart disease |  |  | <0.001 |
| 0 | 4,632 (95%) | 202 (81%) |  |
| 1 | 229 (4.7%) | 47 (19%) |  |
| ever married | 3,133 (64%) | 220 (88%) | <0.001 |
| work type |  |  |  |
| children | 685 (14%) | 2 (0.8%) |  |
| Govt\_job | 624 (13%) | 33 (13%) |  |
| Never\_worked | 22 (0.5%) | 0 (0%) |  |
| Private | 2,776 (57%) | 149 (60%) |  |
| Self-employed | 754 (16%) | 65 (26%) |  |
| Residence type |  |  | 0.3 |
| Rural | 2,400 (49%) | 114 (46%) |  |
| Urban | 2,461 (51%) | 135 (54%) |  |
| avg glucose level | 91 (77, 113) | 105 (80, 197) | <0.001 |
| bmi | 28 (23, 33) | 30 (26, 34) | <0.001 |
| Unknown | 161 | 40 |  |
| smoking status |  |  | <0.001 |
| formerly smoked | 815 (17%) | 70 (28%) |  |
| never smoked | 1,802 (37%) | 90 (36%) |  |
| smokes | 747 (15%) | 42 (17%) |  |
| Unknown | 1,497 (31%) | 47 (19%) |  |
| 1n (%); Median (IQR) | | | |
| 2Fisher's exact test; Welch Two Sample t-test | | | |

# При проведении тестов Хи-квадрат, если ожидаемое количество наблюдений в одном из ячеек меньше 5, точный тест Фишера предпочтительнее.  
# Для сравнения средних в группах используется Welch Two Sample t-test, который позволяет сравнивать группы с неравными дисперсиями. Учитывая большой обхем выборки тест должен быть устойчив от отклонений от нормальности.

# 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