R\_course\_tasks

Dmitrii Belousov

2024-09-01

Table of Contents

plot(pressure)



# Data Import

url <- "https://stepik.org/media/attachments/lesson/790859/numeric\_data.rds"  
destination\_directory <- "./"  
destination\_file <- paste0(destination\_directory, "numeric\_data.rds")  
download.file(url, destfile = destination\_file, method = "auto")

numeric\_data <- read\_rds("./numeric\_data.rds")

url <- "https://stepik.org/media/attachments/lesson/790859/factor\_data.rds"  
destination\_directory <- "./"  
destination\_file <- paste0(destination\_directory, "factor\_data.rds")  
download.file(url, destfile = destination\_file, method = "auto")

factor\_data <- read\_rds("./factor\_data.rds")

url <- "https://stepik.org/media/attachments/lesson/790863/data\_tsv.tsv"  
destination\_directory <- "./raw/"  
destination\_file <- paste0(destination\_directory, "data.tsv")  
download.file(url, destfile = destination\_file, method = "auto")

data <- read.csv("./raw/data.tsv", sep='\t', na.strings = "")

# Statistics

## Summary

summary(numeric\_data)

## Возраст Рост Базофилы\_E1 Эозинофилы\_E1   
## Min. :21.00 Min. :155.0 Min. :-0.2188 Min. :-1.227   
## 1st Qu.:28.00 1st Qu.:164.0 1st Qu.: 0.4020 1st Qu.: 2.325   
## Median :30.50 Median :168.0 Median : 0.6509 Median : 3.728   
## Mean :30.25 Mean :167.7 Mean : 0.6509 Mean : 3.707   
## 3rd Qu.:33.00 3rd Qu.:171.2 3rd Qu.: 0.8644 3rd Qu.: 5.083   
## Max. :42.00 Max. :181.0 Max. : 1.7186 Max. : 8.434   
## Гемоглобин\_E1 Эритроциты\_E1 Базофилы\_E2 Эозинофилы\_E2   
## Min. : 5.352 Min. :2.821 Min. :0.1854 Min. :-0.2124   
## 1st Qu.:10.681 1st Qu.:3.605 1st Qu.:0.8062 1st Qu.: 3.3403   
## Median :11.711 Median :4.082 Median :1.0551 Median : 4.7428   
## Mean :11.860 Mean :4.104 Mean :1.0551 Mean : 4.7216   
## 3rd Qu.:13.175 3rd Qu.:4.599 3rd Qu.:1.2686 3rd Qu.: 6.0976   
## Max. :16.232 Max. :5.728 Max. :2.1228 Max. : 9.4492   
## Гемоглобин\_E2 Эритроциты\_E2   
## Min. : 6.073 Min. :5.136   
## 1st Qu.:11.402 1st Qu.:5.920   
## Median :12.432 Median :6.398   
## Mean :12.581 Mean :6.420   
## 3rd Qu.:13.896 3rd Qu.:6.914   
## Max. :16.952 Max. :8.044

psych::describe(numeric\_data)

## vars n mean sd median trimmed mad min max range  
## Возраст 1 100 30.25 3.98 30.50 30.31 3.71 21.00 42.00 21.00  
## Рост 2 100 167.70 5.77 168.00 167.80 5.93 155.00 181.00 26.00  
## Базофилы\_E1 3 100 0.65 0.38 0.65 0.64 0.35 -0.22 1.72 1.94  
## Эозинофилы\_E1 4 100 3.71 2.15 3.73 3.71 2.06 -1.23 8.43 9.66  
## Гемоглобин\_E1 5 100 11.86 1.78 11.71 11.80 1.75 5.35 16.23 10.88  
## Эритроциты\_E1 6 100 4.10 0.67 4.08 4.10 0.73 2.82 5.73 2.91  
## Базофилы\_E2 7 100 1.06 0.38 1.06 1.04 0.35 0.19 2.12 1.94  
## Эозинофилы\_E2 8 100 4.72 2.15 4.74 4.73 2.06 -0.21 9.45 9.66  
## Гемоглобин\_E2 9 100 12.58 1.78 12.43 12.52 1.75 6.07 16.95 10.88  
## Эритроциты\_E2 10 100 6.42 0.67 6.40 6.41 0.73 5.14 8.04 2.91  
## skew kurtosis se  
## Возраст -0.06 -0.10 0.40  
## Рост -0.11 -0.66 0.58  
## Базофилы\_E1 0.31 0.16 0.04  
## Эозинофилы\_E1 -0.02 -0.50 0.21  
## Гемоглобин\_E1 -0.01 0.88 0.18  
## Эритроциты\_E1 0.08 -0.74 0.07  
## Базофилы\_E2 0.31 0.16 0.04  
## Эозинофилы\_E2 -0.02 -0.50 0.21  
## Гемоглобин\_E2 -0.01 0.88 0.18  
## Эритроциты\_E2 0.08 -0.74 0.07

factor\_data$Группа <- iconv(factor\_data$Группа, from = "Windows-1251", to = "UTF-8")  
  
table(factor\_data$Группа, factor\_data$`Группа крови`)

##   
## O (I) A (II) B (III) AB (IV) <NA>  
## Группа 1 10 16 11 4 9  
## Группа 2 15 18 5 4 8

## Dplyr

data <- as\_tibble(data)  
head(data)

## # A tibble: 6 × 13  
## Группа Возраст Пол Рост Группа.крови Базофилы\_E1 Эозинофилы\_E1  
## <chr> <int> <chr> <int> <chr> <dbl> <dbl>  
## 1 Группа 1 31 Женский 174 A (II) 0.422 0.646  
## 2 Группа 1 28 Женский 157 A (II) 0.327 4.97   
## 3 Группа 1 33 Женский 166 NA 0.799 3.39   
## 4 Группа 1 26 Женский 168 O (I) 0.0237 4.54   
## 5 Группа 1 33 Женский 170 A (II) 0.664 3.32   
## 6 Группа 1 28 Мужской 172 B (III) 0.481 2.79   
## # ℹ 6 more variables: Гемоглобин\_E1 <dbl>, Эритроциты\_E1 <dbl>,  
## # Базофилы\_E2 <dbl>, Эозинофилы\_E2 <dbl>, Гемоглобин\_E2 <dbl>,  
## # Эритроциты\_E2 <dbl>

data %>%   
 select(`Возраст`, function(x) any(str\_detect(x, "Жен")))

## # A tibble: 100 × 2  
## Возраст Пол   
## <int> <chr>   
## 1 31 Женский  
## 2 28 Женский  
## 3 33 Женский  
## 4 26 Женский  
## 5 33 Женский  
## 6 28 Мужской  
## 7 27 Мужской  
## 8 31 Мужской  
## 9 23 Женский  
## 10 29 Женский  
## # ℹ 90 more rows

### Select

columns\_to\_print <- c("Батч", "Базофилы\_E2")  
  
data %>%   
 select(`Группа`, where(is.numeric)) %>%   
 select(`Батч` = `Группа`, starts\_with("Б")) %>%   
 select(all\_of(columns\_to\_print)) %>%   
 slice\_sample(prop = .1)

## # A tibble: 10 × 2  
## Батч Базофилы\_E2  
## <chr> <dbl>  
## 1 Группа 1 1.20   
## 2 Группа 1 1.30   
## 3 Группа 1 0.580  
## 4 Группа 2 1.51   
## 5 Группа 2 1.73   
## 6 Группа 1 0.842  
## 7 Группа 2 1.20   
## 8 Группа 1 1.30   
## 9 Группа 2 0.841  
## 10 Группа 1 2.00

### Filter

data %>%   
 filter(`Пол` == 'Женский'   
 & `Группа.крови` %in% c('A (II)', 'O (I)')   
 & between(`Возраст`, 20, 50)   
 & if\_any(.cols = contains("\_E2"), .fns = function(x) near(x, 1.5, tol = 0.3)))

## # A tibble: 9 × 13  
## Группа Возраст Пол Рост Группа.крови Базофилы\_E1 Эозинофилы\_E1  
## <chr> <int> <chr> <int> <chr> <dbl> <dbl>  
## 1 Группа 1 31 Женский 174 A (II) 0.422 0.646  
## 2 Группа 1 32 Женский 165 A (II) 0.803 2.46   
## 3 Группа 1 28 Женский 162 O (I) 0.834 2.59   
## 4 Группа 1 31 Женский 164 A (II) 0.894 4.39   
## 5 Группа 2 25 Женский 168 O (I) 1.33 6.90   
## 6 Группа 2 33 Женский 162 A (II) 1.10 6.34   
## 7 Группа 2 29 Женский 176 O (I) 1.11 1.42   
## 8 Группа 2 35 Женский 169 A (II) 1.36 5.92   
## 9 Группа 2 34 Женский 168 O (I) 0.950 0.803  
## # ℹ 6 more variables: Гемоглобин\_E1 <dbl>, Эритроциты\_E1 <dbl>,  
## # Базофилы\_E2 <dbl>, Эозинофилы\_E2 <dbl>, Гемоглобин\_E2 <dbl>,  
## # Эритроциты\_E2 <dbl>

### Mutate

data <- data %>%   
 mutate(Elderly = ifelse(`Возраст` >= median(`Возраст`, na.rm = TRUE), "Yes", "No")) %>%   
 mutate(Age\_Group = case\_when(`Возраст` < quantile(`Возраст`, .25) ~ "Young",   
 between(`Возраст`, quantile(`Возраст`, .25),   
 quantile(`Возраст`, .75)) ~ "Average",  
 `Возраст` > quantile(`Возраст`, .75) ~ "Old")) %>%   
 mutate(across(where(is.numeric), function(x) (x - mean(x, na.rm=T))/(sd(x, na.rm = T))))

### Rename

data %>%   
 rename(`Возрастная\_Группа` = Age\_Group, `Старшие` = Elderly)

## # A tibble: 100 × 15  
## Группа Возраст Пол Рост Группа.крови Базофилы\_E1 Эозинофилы\_E1  
## <chr> <dbl> <chr> <dbl> <chr> <dbl> <dbl>  
## 1 Группа 1 0.188 Женский 1.09 A (II) -0.608 -1.43   
## 2 Группа 1 -0.565 Женский -1.85 A (II) -0.862 0.591  
## 3 Группа 1 0.690 Женский -0.294 NA 0.395 -0.149  
## 4 Группа 1 -1.07 Женский 0.0520 O (I) -1.67 0.388  
## 5 Группа 1 0.690 Женский 0.398 A (II) 0.0338 -0.182  
## 6 Группа 1 -0.565 Мужской 0.745 B (III) -0.452 -0.429  
## 7 Группа 1 -0.816 Мужской -1.85 A (II) 0.636 -0.635  
## 8 Группа 1 0.188 Мужской 1.09 NA 0.550 0.127  
## 9 Группа 1 -1.82 Женский 1.26 A (II) -0.712 -0.148  
## 10 Группа 1 -0.314 Женский 0.745 A (II) -0.983 1.97   
## # ℹ 90 more rows  
## # ℹ 8 more variables: Гемоглобин\_E1 <dbl>, Эритроциты\_E1 <dbl>,  
## # Базофилы\_E2 <dbl>, Эозинофилы\_E2 <dbl>, Гемоглобин\_E2 <dbl>,  
## # Эритроциты\_E2 <dbl>, Старшие <chr>, Возрастная\_Группа <chr>

### Sorting

data %>%   
 select(`Группа`, `Пол`, where(is.numeric)) %>%   
 group\_by(`Пол`) %>%   
 arrange(desc(`Возраст`), .by\_group = T) %>%   
 summarize(across(where(is.numeric), function(x) mean(x, na.rm = T)))

## # A tibble: 2 × 11  
## Пол Возраст Рост Базофилы\_E1 Эозинофилы\_E1 Гемоглобин\_E1 Эритроциты\_E1  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Женский 0.0272 0.0847 -0.0607 -0.0387 -0.0639 -0.0377  
## 2 Мужской -0.0307 -0.0955 0.0685 0.0436 0.0720 0.0425  
## # ℹ 4 more variables: Базофилы\_E2 <dbl>, Эозинофилы\_E2 <dbl>,  
## # Гемоглобин\_E2 <dbl>, Эритроциты\_E2 <dbl>

stats <- list(  
 `\_N\_subjects` = ~length(.x) %>% as.character(),  
 `\_N\_entries` = ~sum(!is.na(.x)) %>% as.character(),  
 `\_No\_data` = ~sum(is.na(.x)) %>% as.character(),  
 `\_Mean` = ~ifelse(sum(!is.na(.x)) == 0, "NA\*", mean(.x, na.rm = T)   
 %>% round(2)   
 %>% as.character()),  
 `\_Standard\_Deviation` = ~ifelse(sum(!is.na(.x)) < 3, "NA\*", sd(.x, na.rm = T)   
 %>% round(2)   
 %>% as.character()),  
 `\_CI\_for\_Mean` = ~sd(.x, na.rm = T)  
 %>% round(2)   
 %>% as.character(),  
 `\_Min - Max` = ~ifelse(sum(!is.na(.x)) == 0, "NA\*", paste0(min(.x, na.rm = T) %>% round(2),   
 " - ", max(.x, na.rm = T) %>% round(2))),  
 `\_Q1 - Q3` = ~ifelse(sum(!is.na(.x)) == 0, "NA\*", paste0(quantile(.x, .25, na.rm = T) %>% round(2),   
 " - ", quantile(.x, .75, na.rm = T) %>% round(2)))  
  
)  
  
data %>%   
 select(`Группа`, where(is.numeric)) %>%   
 group\_by(`Группа`) %>%   
 summarise(across(where(is.numeric), stats)) %>%   
 pivot\_longer(!`Группа`) %>%   
 separate(name, into = c('Parameter', "Statistic"), sep="\_\_") %>%   
 rename(Value = value, Group = `Группа`) %>%   
 flextable() %>%   
 theme\_box() %>%   
 merge\_v(c("Group", "Parameter"))

| **Group** | **Parameter** | **Statistic** | **Value** |
| --- | --- | --- | --- |
| Группа 1 | Возраст | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.25 |
| Standard\_Deviation | 0.95 |
| CI\_for\_Mean | 0.95 |
| Min - Max | -2.32 - 1.95 |
| Q1 - Q3 | -0.82 - 0.44 |
| Рост | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.06 |
| Standard\_Deviation | 1.07 |
| CI\_for\_Mean | 1.07 |
| Min - Max | -2.2 - 2.3 |
| Q1 - Q3 | -0.64 - 0.74 |
| Базофилы\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.23 |
| Standard\_Deviation | 1 |
| CI\_for\_Mean | 1 |
| Min - Max | -2.31 - 2.5 |
| Q1 - Q3 | -0.86 - 0.49 |
| Эозинофилы\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.25 |
| Standard\_Deviation | 0.82 |
| CI\_for\_Mean | 0.82 |
| Min - Max | -1.87 - 2.06 |
| Q1 - Q3 | -0.71 - 0.22 |
| Гемоглобин\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.05 |
| Standard\_Deviation | 1.18 |
| CI\_for\_Mean | 1.18 |
| Min - Max | -3.66 - 2.46 |
| Q1 - Q3 | -0.68 - 0.79 |
| Эритроциты\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.1 |
| Standard\_Deviation | 1.01 |
| CI\_for\_Mean | 1.01 |
| Min - Max | -1.92 - 2.43 |
| Q1 - Q3 | -0.54 - 0.79 |
| Базофилы\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.23 |
| Standard\_Deviation | 1 |
| CI\_for\_Mean | 1 |
| Min - Max | -2.31 - 2.5 |
| Q1 - Q3 | -0.86 - 0.49 |
| Эозинофилы\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.25 |
| Standard\_Deviation | 0.82 |
| CI\_for\_Mean | 0.82 |
| Min - Max | -1.87 - 2.06 |
| Q1 - Q3 | -0.71 - 0.22 |
| Гемоглобин\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.05 |
| Standard\_Deviation | 1.18 |
| CI\_for\_Mean | 1.18 |
| Min - Max | -3.66 - 2.46 |
| Q1 - Q3 | -0.68 - 0.79 |
| Эритроциты\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.1 |
| Standard\_Deviation | 1.01 |
| CI\_for\_Mean | 1.01 |
| Min - Max | -1.92 - 2.43 |
| Q1 - Q3 | -0.54 - 0.79 |
| Группа 2 | Возраст | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.25 |
| Standard\_Deviation | 0.99 |
| CI\_for\_Mean | 0.99 |
| Min - Max | -2.32 - 2.95 |
| Q1 - Q3 | -0.31 - 0.94 |
| Рост | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.06 |
| Standard\_Deviation | 0.94 |
| CI\_for\_Mean | 0.94 |
| Min - Max | -1.51 - 1.61 |
| Q1 - Q3 | -0.64 - 0.57 |
| Базофилы\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.23 |
| Standard\_Deviation | 0.95 |
| CI\_for\_Mean | 0.95 |
| Min - Max | -1.81 - 2.84 |
| Q1 - Q3 | -0.44 - 0.78 |
| Эозинофилы\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.25 |
| Standard\_Deviation | 1.11 |
| CI\_for\_Mean | 1.11 |
| Min - Max | -2.3 - 2.2 |
| Q1 - Q3 | -0.51 - 1.08 |
| Гемоглобин\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.05 |
| Standard\_Deviation | 0.79 |
| CI\_for\_Mean | 0.79 |
| Min - Max | -1.69 - 1.84 |
| Q1 - Q3 | -0.61 - 0.65 |
| Эритроциты\_E1 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.1 |
| Standard\_Deviation | 0.99 |
| CI\_for\_Mean | 0.99 |
| Min - Max | -1.89 - 1.74 |
| Q1 - Q3 | -0.87 - 0.64 |
| Базофилы\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.23 |
| Standard\_Deviation | 0.95 |
| CI\_for\_Mean | 0.95 |
| Min - Max | -1.81 - 2.84 |
| Q1 - Q3 | -0.44 - 0.78 |
| Эозинофилы\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | 0.25 |
| Standard\_Deviation | 1.11 |
| CI\_for\_Mean | 1.11 |
| Min - Max | -2.3 - 2.2 |
| Q1 - Q3 | -0.51 - 1.08 |
| Гемоглобин\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.05 |
| Standard\_Deviation | 0.79 |
| CI\_for\_Mean | 0.79 |
| Min - Max | -1.69 - 1.84 |
| Q1 - Q3 | -0.61 - 0.65 |
| Эритроциты\_E2 | N\_subjects | 50 |
| N\_entries | 50 |
| No\_data | 0 |
| Mean | -0.1 |
| Standard\_Deviation | 0.99 |
| CI\_for\_Mean | 0.99 |
| Min - Max | -1.89 - 1.74 |
| Q1 - Q3 | -0.87 - 0.64 |