Colledge

2023-11-25

Предварительный анализ данных

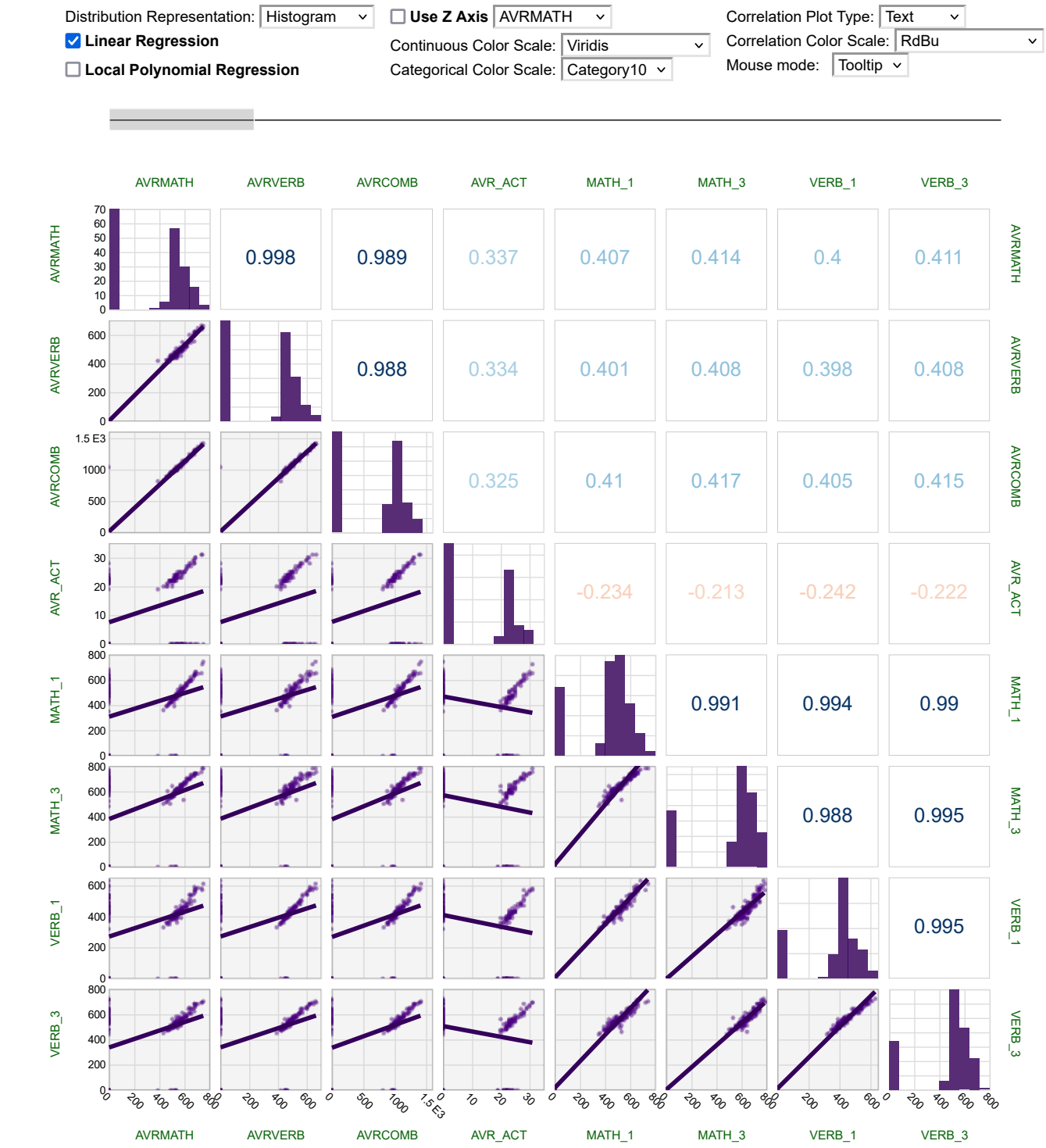
Загрузим данные

```
library(readxl)
library(tidyverse)
library(kableExtra)

data <- read_excel("C:/Users/yanas/Documents/7R/I_shortname.xls")


print_df <- function(df)
{
  df |>
    kable(format = "html") |>
    kable_styling() |>
    kableExtra::scroll_box(width = "100%", height = "100%")
}

head(data, 8) |>
 print_df()
```

...1	PPIND	FICE	STATE	TYPE	AVRMATH	AVRVERB	AVRCOMB	AVR_ACT	MATH_1	MATH_3	VERB_1	VERB_3
Auburn University-Ma	1	1009	AL	I	575	501	1076	24	520	638	453	559
University of Alabam	1	1051	AL	I	NA	NA	NA	23	NA	NA	NA	NA
University of Alabam	1	1052	AL	I	NA	NA	NA	21	NA	NA	NA	NA
University of Alaska	1	1063	AK	I	499	462	961	22	NA	NA	NA	NA
Arizona State Univer	1	1081	AZ	I	521	453	974	23	450	590	390	500
Northern Arizona Uni	1	1082	AZ	I	495	444	939	22	420	560	380	500
University of Arizon	1	1083	AZ	I	526	462	983	23	450	600	400	520
University of Arkans	1	1108	AR	I	NA	NA	NA	NA	NA	NA	NA	NA

Построим графики зависимостей признаков и выберем подходящие

```
library(scatterPlotMatrix)
library(dplyr)

data |>
  select(-PPIND, -FICE, -STATE, -TYPE, -...1) |>
  scatterPlotMatrix(regressionType = 1,
                    #categorical = categorical,
                    corrPlotType = "Text",
                    plotProperties = list(noCatColor = "Indigo"),
                    controlWidgets = TRUE,
                    height = 1050,
                    width = 1000)
```



Оставим в новом датасете только рассматриваемые

признаки

```
df <-
  data |>
  select(...1,
         PPIND,
         ADD_FEE,
         BOOK,
         NEW10,
         PH_D,
         SAL_ALL,
         SF_RATIO,
         GRADUAT,
         INSTRUCT)
```

Посмотрим на моды:

```
library(dplyr)
modes<-summarize(df, across(ADD_FEE:INSTRUCT, function(x) max(table(x))))
print(modes)
```

```
## # A tibble: 1 × 8
##   ADD_FEE  BOOK NEW10  PH_D SAL_ALL SF_RATIO GRADUAT INSTRUCT
##     <int> <int> <int> <int>   <int>    <int>   <int>    <int>
## 1       4    29    10    10       5        5       7        3
```

Рассматриваемые признаки(первые 4 есть в задании, остальные — те, которые влияют на переменную NEW10):

1. ***College name*** — название колледжа — **качественный признак**
2. ***PPIND*** — Гос/частное заведение (гос = 1, частный = 2) — **качественный признак**
3. ***ADD_FEE*** — дополнительные сборы — **количественный непрерывный признак**
4. ***BOOK*** — примерная стоимость учебников — **количественный дискретный признак**
5. ***NEW10(Процент студентов из лучших слоев школьных выпускников)*** — **количественный дискретный признак**
6. ***PH_D***— количество преподавателей со степенью Ph.D.— **количественный дискретный признак**
7. ***SAL_ALL*** — Средняя заработная плата — **количественный непрерывный признак**
8. ***SF_RATIO*** — соотношение студентов к преподавателям — **количественный непрерывный признак**
9. ***GRADUAT*** — процент выпускников — **количественный дискретный признак**
10. ***INSTRUCT*** — расходы на обучение в расчете на одного учащегося — **количественный непрерывный признак**

Посмотрим описательные статстистики

```
summary(df|>select(-PPIND, -...1))
```

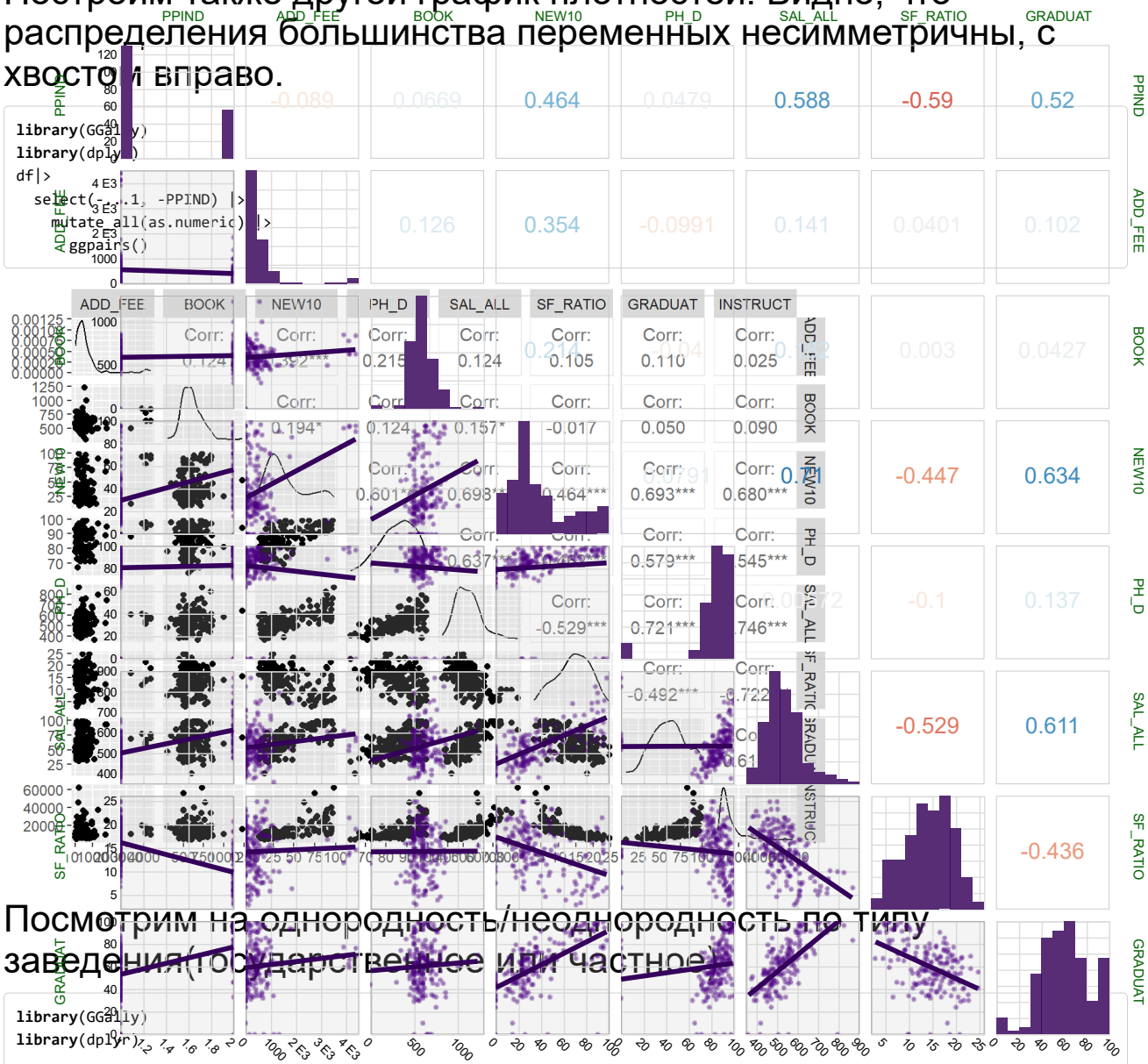
```
##     ADD_FEE            BOOK            NEW10            PH_D
##  Min.   :  20.0   Min.   : 300.0   Min.   : 8.00   Min.   :63.00
##  1st Qu.: 210.0   1st Qu.: 500.0   1st Qu.:24.00   1st Qu.:80.50
##  Median : 425.5   Median : 600.0   Median :32.00   Median :87.00
##  Mean   : 648.1   Mean   : 603.1   Mean   :41.48   Mean   :85.72
##  3rd Qu.: 694.0   3rd Qu.: 673.8   3rd Qu.:57.00   3rd Qu.:92.00
##  Max.   :4374.0   Max.   :1230.0   Max.   :98.00   Max.   :99.00
##  NA's   :40       NA's   :2        NA's   :16      NA's   :9
##     SAL_ALL         SF_RATIO        GRADUAT         INSTRUCT
##  Min.   :362.0   Min.   : 2.90   Min.   :10.00   Min.   : 3605
##  1st Qu.:472.2   1st Qu.:10.88   1st Qu.:47.50   1st Qu.: 7604
##  Median :522.5   Median :14.50   Median :62.00   Median : 9840
##  Mean   :534.0   Mean   :14.23   Mean   :62.02   Mean   :12832
##  3rd Qu.:578.2   3rd Qu.:18.02   3rd Qu.:74.50   3rd Qu.:14340
##  Max.   :866.0   Max.   :24.70   Max.   :99.00   Max.   :62469
##                                  NA's   :5
```

Построим MatrixPlot для рассматриваемых признаков:

```
library(scatterPlotMatrix)
library(dplyr)


categories <- list( c(1,2), NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL)

df |>
  select(-...1)|>
  scatterPlotMatrix(regressionType = 1,
                        corrPlotType = "Text",
                        categorical = categories,
                        plotProperties = list(noCatColor =  "Indigo"),
                        controlWidgets = TRUE,
                        height = 1050, width = 1000)
```

Distribution Representation: Histogram
Linear Regression
Local Polynomial Regression
Use Z Axis PPIND
Continuous Color Scale: Viridis
Categorical Color Scale: Category10
Correlation Plot Type: Text
Correlation Color Scale: RdBu
Mouse mode: Tooltip

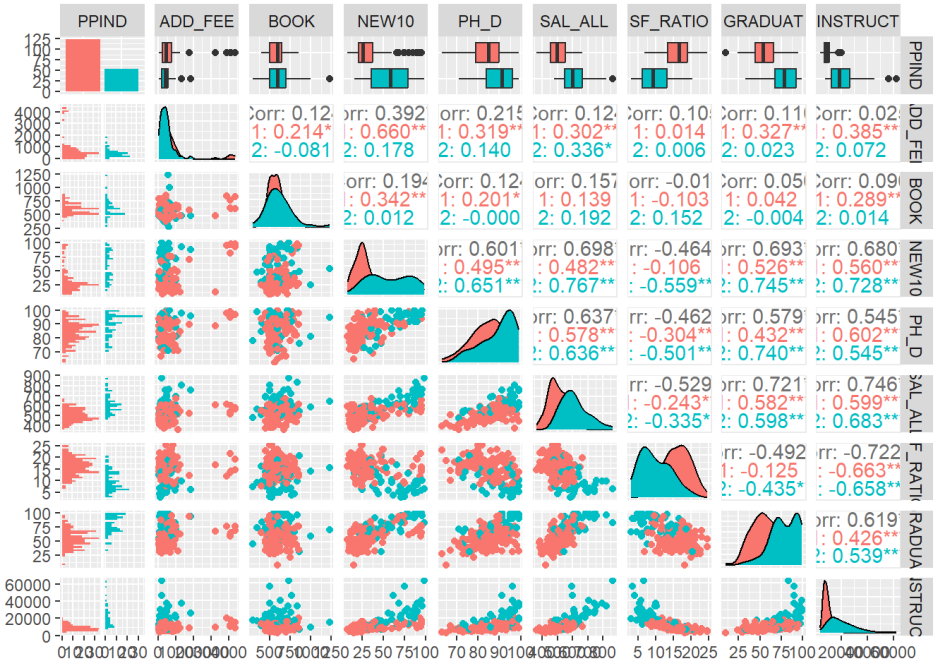
Построим также другой график плотностей. Видно, что распределения большинства переменных несимметричны, с хвостом вправо.



Посмотрим на однородность/неоднородность по типу заведения(государственное или частное)

```
library(GGally)
library(dplyr)
df|>
  select(-...1, -PPIND) |>
  mutate_all(as.numeric) |>
  ggpairs()
```

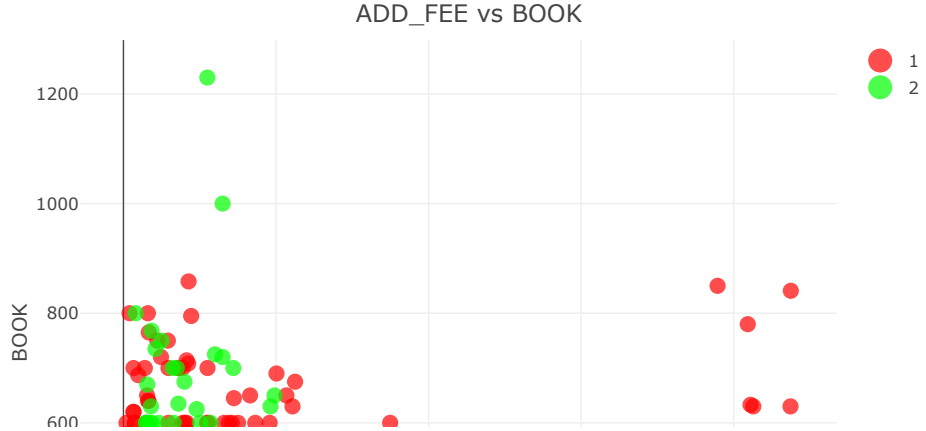
```
library(GGally)
library(dplyr)

df$PPIND <- factor(df$PPIND)

# Построим графики с разделением по 'PPIND'
ggpairs(df, columns = 2:ncol(df), aes(colour = PPIND))
```

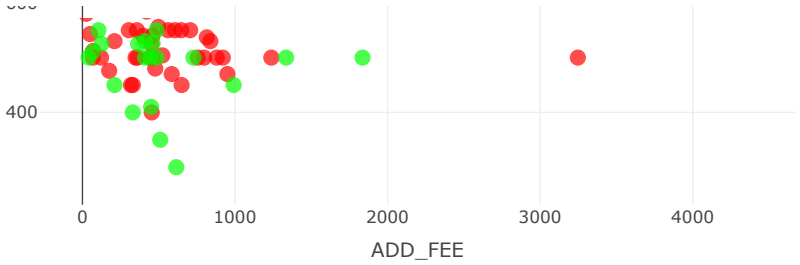


Неоднородности в данных видны по переменным INSTRUCT, NEW10, SF_RATIO, GRADUATE,

Видно неодородность по ADD_FEE(выбросы по государственным учрежденрям). По графику можно увидеть, что выбросы среди гос университетов — это Калифорнийский университет(разные филиалы) и Массачусетский.

```
library(plotly)

df$PPIND <- factor(df$PPIND)

plot_ly(df |> mutate(row_num = row_number()),
        x = ~ADD_FEE,
        y = ~BOOK, color = ~PPIND,
        colors = c("red","green"),
        text = ~paste("PPIND: ", PPIND,
                      "<br>...1: ", ...1,
                      "<br>Row Number: ", row_num)) %>%
  add_markers(size = 6) %>%
  layout(title = "ADD_FEE vs BOOK",
         xaxis = list(title = "ADD_FEE"),
         yaxis = list(title = "BOOK"),
         showlegend = TRUE,
         hoverlabel = list(bgcolor = "white",
                           font = list(family = "Arial",
                                       size = 12,
                                       color = "black")))
```





Найдем несимметричные с хвостом вправо(коэффициент асимметрии > 0) распределения.Выведем коэффициенты асимметрии.

```
library(e1071)

skewness(na.omit(df$ADD_FEE))
```

```
## [1] 3.203984
```

```
skewness(na.omit(df$BOOK))
```

```
## [1] 1.152447
```

```
skewness(na.omit(df$NEW10))
```

```
## [1] 0.8547832
```

```
skewness(na.omit(df$PH_D))
```

```
## [1] -0.4084976
```

```
skewness(na.omit(df$SAL_ALL))
```

```
## [1] 0.9074192
```

```
skewness(na.omit(df$SF_RATIO))
```

```
## [1] -0.2398104
```

```
skewness(na.omit(df$GRADUAT))
```

```
## [1] 0.108287
```

```
skewness(na.omit(df$INSTRUCT))
```

```
## [1] 2.645673
```

Прологорифмируем несимметричные с хвостом вправо распределения(“SAL_ALL”, “NEW10”, “BOOK”, “ADD_FEE”, “INSTRUCT”, “GRADUAT”) и снова построим графики:

```
library(scatterPlotMatrix)

data_log <- df |>
  mutate_at(vars("SAL_ALL",
                 "NEW10",
                 "BOOK",
                 "GRADUAT",
                 "ADD_FEE",
                 "INSTRUCT"),
            ~log(.))

categories <- list( c(1,2), NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL)


data_log|>
  select(-...1) |>
  scatterPlotMatrix(regressionType = 1,
                        corrPlotType = "Text",
                        categorical = categories,
                        plotProperties = list(noCatColor =  "Indigo"),
                        controlWidgets = TRUE,
                        height = 1050, width = 1000)
```

Distribution Representation: Histogram
☑ **Linear Regression**
☐ **Local Polynomial Regression**
☐ **Use Z Axis** PPIND
Continuous Color Scale: Viridis
Categorical Color Scale: Category10
Correlation Plot Type: Text
Correlation Color Scale: RdBu
Mouse mode: Tooltip

Выбросы, видимые на графиках, как значения равные нулю — это значения NA. Не будем их удалять. Удалим выброс Texas Southern Uni – выделяется по низкому значению GRADUATE относительно остальных наблюдений:

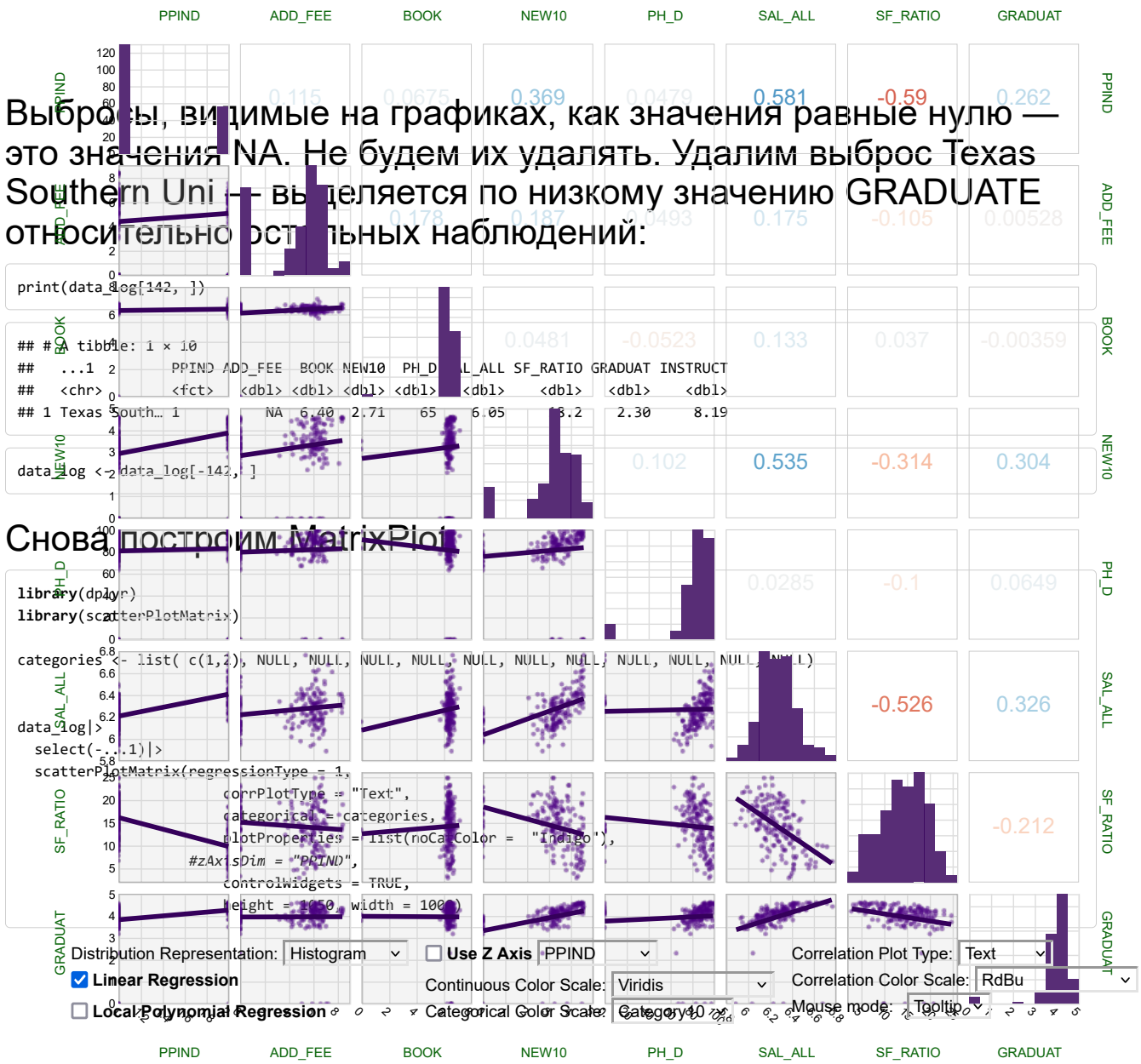
```
print(data_log[142, ])
```

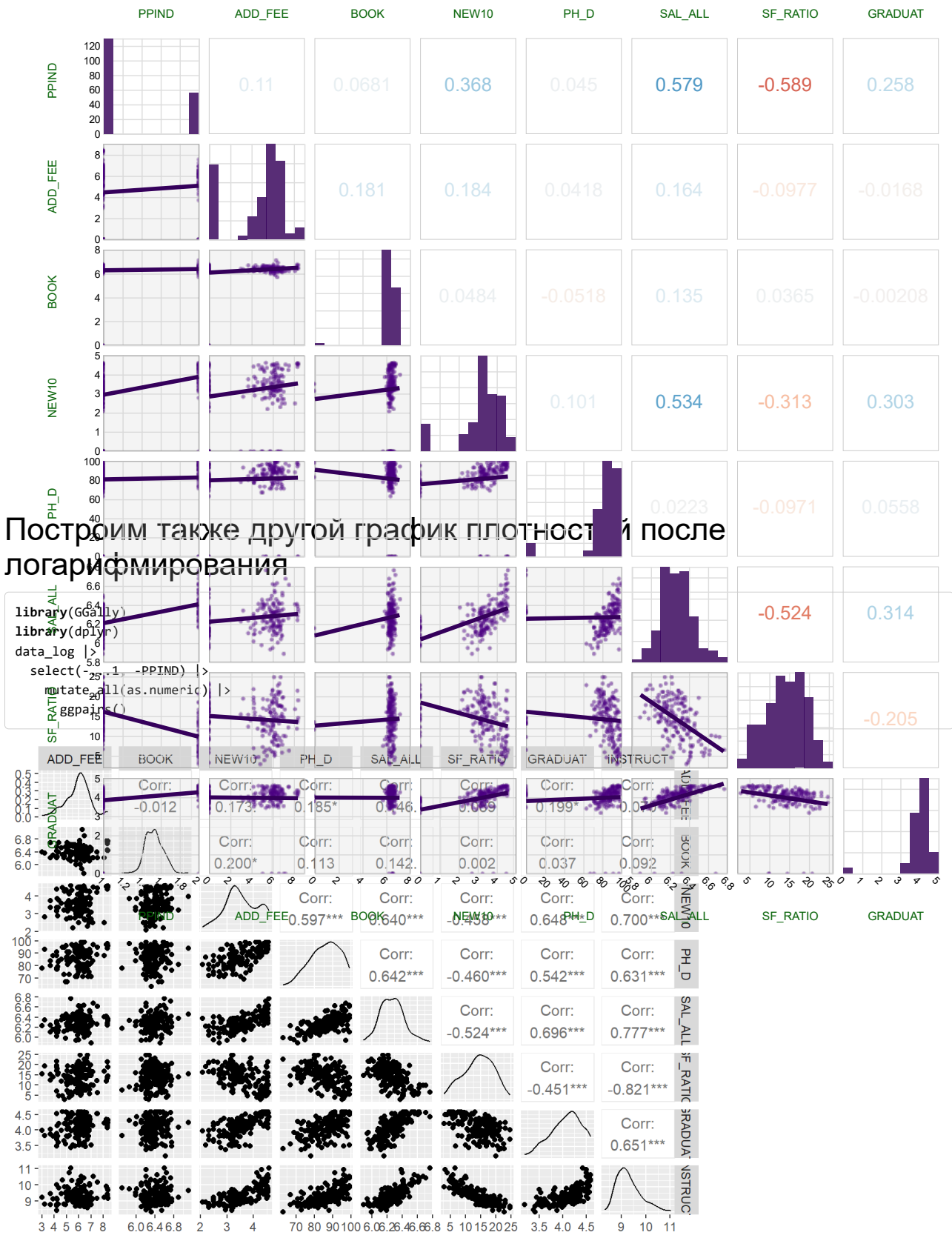
```
## # A tibble: 1 × 10
##   ...1  PPIND ADD_FEE  BOOK NEW10  PH_D SAL_ALL SF_RATIO GRADUAT INSTRUCT
##   <chr>  <fct>   <dbl> <dbl> <dbl> <dbl>   <dbl>    <dbl>   <dbl>    <dbl>
## 1 Texas… 1          NA  6.40  2.71    65    6.05     18.2    2.30     8.19
```

```
data_log <- data_log[-142, ]
```

Снова построим MatrixPlot

```
library(dplyr)
library(scatterPlotMatrix)

categories <- list( c(1,2), NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL)

data_log|>
  select(-...1)|>
  scatterPlotMatrix(regressionType = 1,
                    corrPlotType = "Text",
                    categorical = categories,
                    plotProperties = list(noCatColor = "indigo"),
                    #zAxisDim = "PPIND",
                    controlWidgets = TRUE,
                    height = 1050, width = 1000)
```





Построим также другой график плотностей после логарифмирования

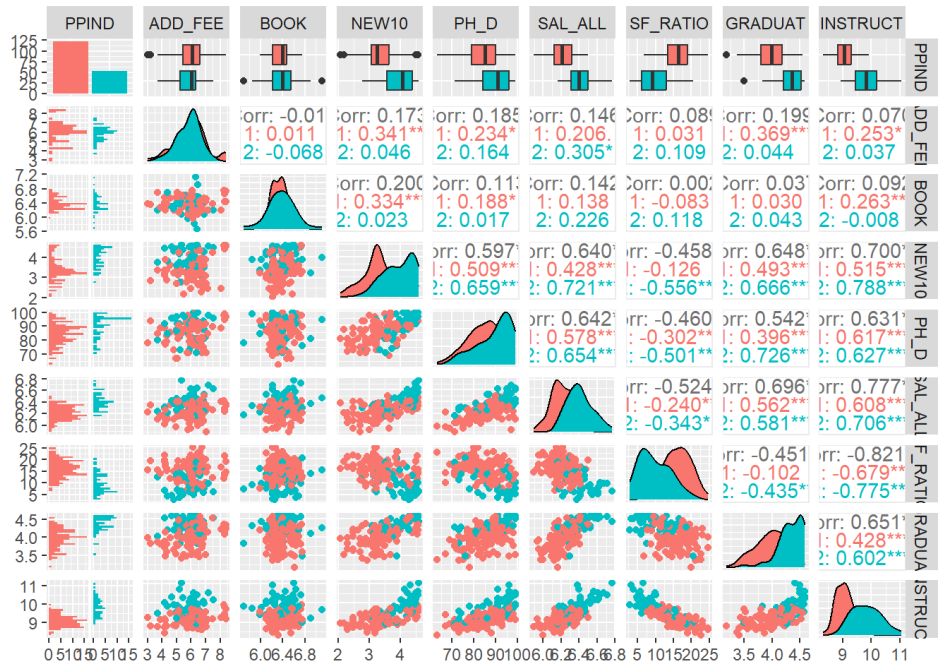
```
library(GGally)
library(dplyr)
data_log |>
  select(-...1, -PPIND) |>
    mutate_all(as.numeric) |>
      ggpairs()
```


Видно, что после логарифмирования распределения стали более симметричными. Это также можно увидеть по значениям skew(асимметрия) до и после логарифмирования. Положительные значения (хвост вправо) стали

меньше(ближе к нулю).

Посмотрим на однородность/неоднородность по типу заведения(государственное или частное) после логарифмирования

```
library(GGally)
library(dplyr)

data_log$PPIND <- factor(data_log$PPIND)

# Построим графики с разделением по 'PPIND'
ggpairs(data_log, columns = 2:ncol(data_log), aes(colour = PPIND))
```



Можно сказать, что по всем, кроме первых двух перменных видны неоднородности(сдвиг), которые порождены отличиями в государственным и частных учебных заведениях.

Описательные статистики до и после логарифмирования:

До:

```
library(psych)
describe(df|>select(-...1,  -PPIND))
```

```
##          vars   n     mean      sd median  trimmed     mad    min     max
## ADD_FEE     1 136   648.12  859.97  425.5   461.92  351.38   20.0  4374.0
## BOOK        2 174   603.07  120.97  600.0   593.63  133.43  300.0  1230.0
## NEW10       3 160    41.48   25.33   32.0    38.81   18.53    8.0    98.0
## PH_D        4 167    85.72    8.19   87.0    86.15    8.90   63.0    99.0
## SAL_ALL     5 176   533.99   87.85  522.5   526.54   80.06  362.0   866.0
## SF_RATIO    6 176    14.23    4.92   14.5    14.37    5.34    2.9    24.7
## GRADUAT     7 171    62.02   19.11   62.0    61.70   20.76   10.0    99.0
## INSTRUCT    8 176 12831.88 8883.95 9840.5 11065.37 4401.84 3605.0 62469.0
##             range  skew kurtosis     se
## ADD_FEE   4354.0  3.20    10.17  73.74
## BOOK       930.0  1.15     3.52   9.17
## NEW10       90.0  0.85    -0.56   2.00
## PH_D        36.0 -0.41    -0.53   0.63
## SAL_ALL    504.0  0.91     1.24   6.62
## SF_RATIO    21.8 -0.24    -0.68   0.37
## GRADUAT     89.0  0.11    -0.65   1.46
## INSTRUCT 58864.0  2.65     8.88 669.65
```

После:

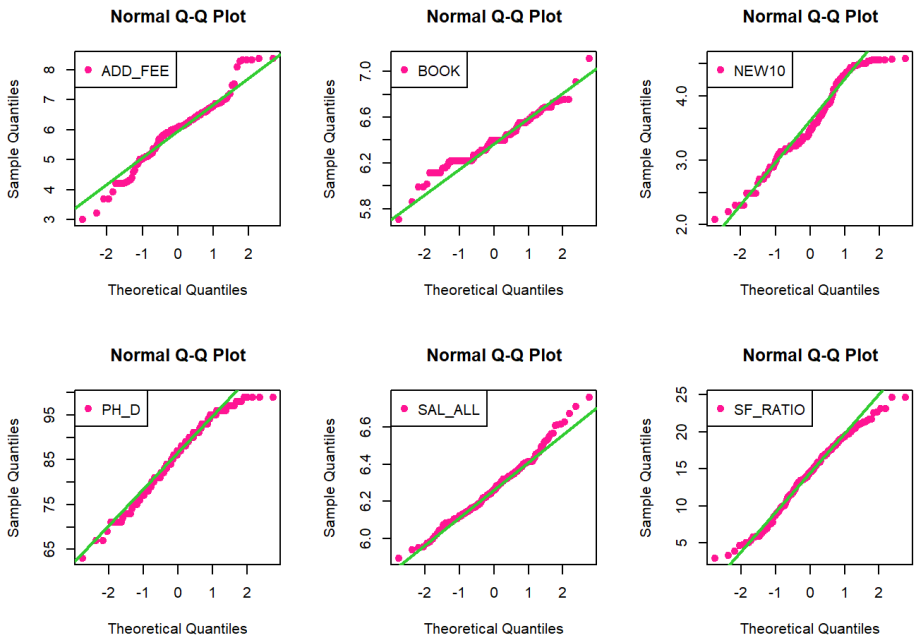
```
library(psych)
describe(data_log|>select(-...1, -PPIND))
```

```
##          vars   n  mean   sd median trimmed  mad   min   max range  skew
## ADD_FEE     1 136  5.95 1.03   6.05    5.97 0.80  3.00  8.38  5.39 -0.17
## BOOK        2 173  6.38 0.19   6.40    6.38 0.23  5.70  7.11  1.41  0.20
## NEW10       3 159  3.55 0.61   3.47    3.56 0.57  2.08  4.58  2.51 -0.01
## PH_D        4 166 85.84 8.05  87.00   86.24 8.90 63.00 99.00 36.00 -0.38
## SAL_ALL     5 175  6.27 0.16   6.26    6.26 0.15  5.89  6.76  0.87  0.38
## SF_RATIO    6 175 14.20 4.93  14.50   14.34 5.34  2.90 24.70 21.80 -0.23
## GRADUAT     7 170  4.08 0.32   4.14    4.10 0.32  3.18  4.60  1.42 -0.40
## INSTRUCT    8 175  9.31 0.51   9.21    9.26 0.47  8.32 11.04  2.72  0.93
##          kurtosis   se
## ADD_FEE      0.55 0.09
## BOOK         1.12 0.01
## NEW10       -0.75 0.05
## PH_D        -0.59 0.62
## SAL_ALL      0.24 0.01
## SF_RATIO    -0.68 0.37
## GRADUAT     -0.52 0.02
## INSTRUCT     0.64 0.04
```

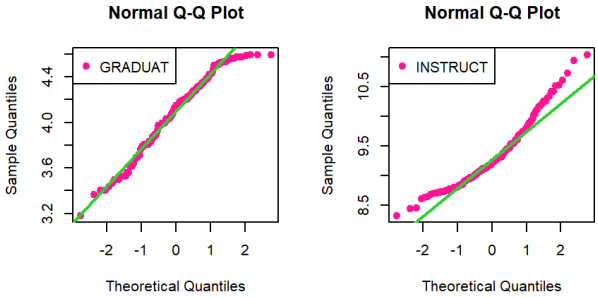
О виде распределений и о сравнении распределений

Проверим распределения на нормальность с помощью QQ-plot (не по категоризующей перменной, а в общем):

```
plot_qq_graph <- function(data, column_name)
{
  data<-data
  expected_quantiles <- qnorm(ppoints(length(data)))
  qqnorm(data,
         pch = 19,
         col = "deeppink")
  qqline(data,
         distribution = qnorm,
         lwd = 2,
         col = "limegreen")
  legend("topleft",
         legend = column_name,
         col = "deeppink",
         pch = 19)
}

par(mfrow = c(2, 3))

plot_qq_graph(data_log$ADD_FEE, "ADD_FEE")
plot_qq_graph(data_log$BOOK, "BOOK")
plot_qq_graph(data_log$NEW10, "NEW10")
plot_qq_graph(data_log$PH_D, "PH_D")
plot_qq_graph(data_log$SAL_ALL, "SAL_ALL")
plot_qq_graph(data_log$SF_RATIO, "SF_RATIO")
```




```
plot_qq_graph(data_log$GRADUAT, "GRADUAT")
plot_qq_graph(data_log$INSTRUCT, "INSTRUCT")
```



Распределения PH_D, SAL_ALL и SF_RATIO похожи на нормальное.

Применим критерии Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка для проверки на нормальность. Нулевая гипотеза: распредление нормально.

```
library(nortest) # для критериев Лиллиефорса и Anderson-Darling
library(ggplot2) # для критерия

feature_names <- c("ADD_FEE", "BOOK", "NEW10", "PH_D", "SAL_ALL", "SF_RATIO", "GRADUAT", "INSTRUCT")

results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)

check_normality <- function(data, feature_name)
  {
  lillie_p_value <- lillie.test(data)$p.value
  ad_p_value <- ad.test(data)$p.value
  shapiro_p_value <- shapiro.test(data)$p.value

  new_row <- data.frame(
    Feature = feature_name,
    Lillie = lillie_p_value,
    Anderson_Darling = ad_p_value,
    Shapiro_Wilk = shapiro_p_value
  )

  results_table <<- rbind(results_table, new_row)
}

for (feature in feature_names)
{
  check_normality(data_log[[feature]], feature)
}

results_table$Normality <- ifelse(results_table$Lillie > 0.05 & results_table$Anderson_Darling > 0.05 & results_table$Shapir
o_Wilk > 0.05, "+", "-")


results_table
```

```
##    Feature       Lillie Anderson_Darling Shapiro_Wilk Normality
## 1  ADD_FEE 0.0014165662     3.870942e-04 3.177860e-03         -
## 2     BOOK 0.0005213534     1.421674e-03 2.653837e-03         -
## 3    NEW10 0.0047618561     3.130766e-04 4.471203e-04         -
## 4     PH_D 0.0419057504     8.696945e-03 2.260817e-03         -
## 5  SAL_ALL 0.5727571564     1.815191e-01 1.237259e-01         +
## 6 SF_RATIO 0.2869955943     4.914540e-02 3.142859e-02         -
## 7  GRADUAT 0.0266871075     9.202347e-03 1.431225e-03         -
## 8 INSTRUCT 0.0000379426     1.420356e-07 1.236022e-06         -
```

По тестам почти для всех признаков критерии нашли отклонение от нормального распределения.

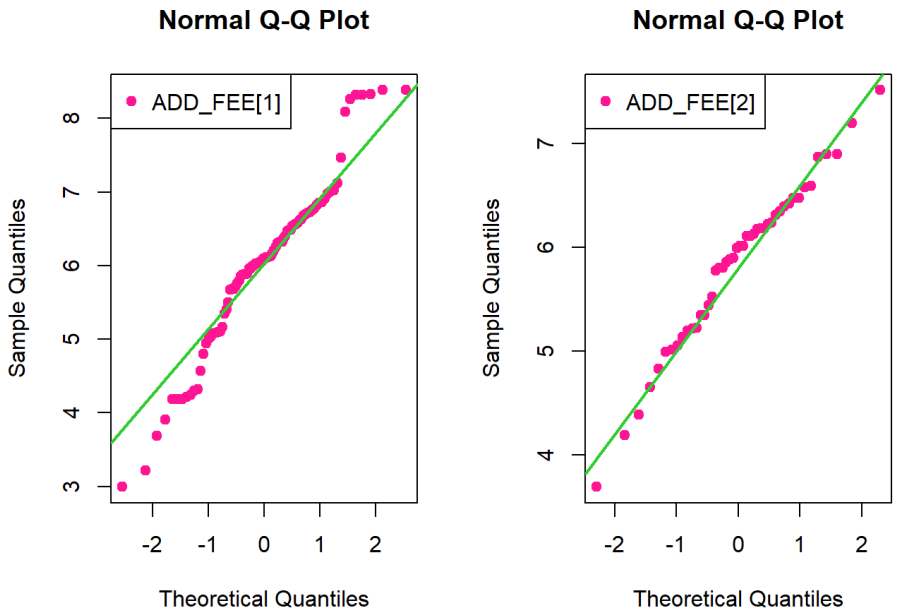
Однако, так как в данных есть неоднородности по государственным и частным учебным заведениям, то проверим на нормальность каждый признак по категоризующей переменной.

Так как проверяем различия между числовыми признаками в зависимости от типа учебного заведения, то это будут независимые выборки. Далее все выводы будут делаться на уровне значимости 0,05.

Признак ADD_FEE.

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$ADD_FEE[data_log$PPIND == "1"], "ADD_FEE[1]")
plot_qq_graph(data_log$ADD_FEE[data_log$PPIND == "2"], "ADD_FEE[2]")
```



По qq-plot нельзя сказать, что этот признак по группам нормально распределен

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

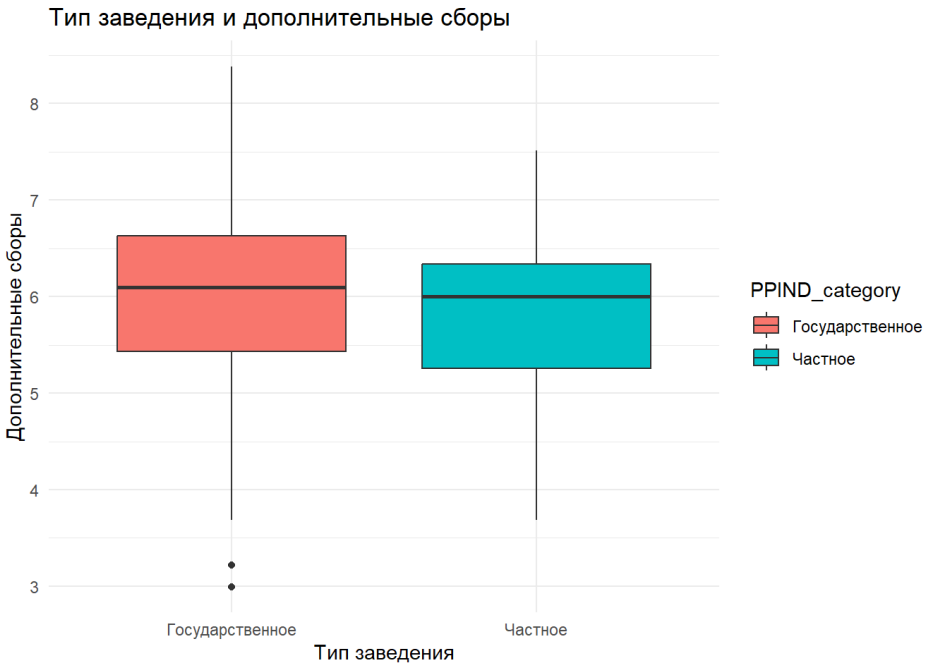
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)

  check_normality(data_log$ADD_FEE[data_log$PPIND == "1"], "ADD_FEE[1]")
  check_normality(data_log$ADD_FEE[data_log$PPIND == "2"], "ADD_FEE[2]")


results_table
```

```
##      Feature      Lillie Anderson_Darling Shapiro_Wilk
## 1 ADD_FEE[1] 0.005229149      0.001958166  0.009932777
## 2 ADD_FEE[2] 0.086477279      0.244846706  0.502103434
```

Результаты показывают, что для признака ADD_FEE по гос завдедениям критерии нашли отклонения от нормального распределения, а по частным — нет.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = ADD_FEE, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "Дополнительные сборы",
       title = "Тип заведения и дополнительные сборы") +
  theme_minimal()
```



По ящикам с усами можно сказать, что медианы по дополнительным взносам в разных завдениях похожи. То есть сумма дополнительных взносов мало отличается в государственных и частных заведениях.

t-test (можем использовать для данных не распределенным нормально, так как он асимптотический)

```
library(car)

perform_t_test <- function(feature, categorical_var, data) {
  # Провекра равенства дисперсий
  levene_test <- car::leveneTest(as.formula(paste(feature, "~", categorical_var)), data = data)

  # Выбор метода для сравнения средних
  if (levene_test$'Pr(>F)'[1] > 0.05) {
    # Равенство дисперсий, используем t-тест с равными дисперсиями
    t_test_result <- t.test(as.formula(paste(feature, "~", categorical_var)), data = data, var.equal = TRUE)
  } else {
    # Неравенство дисперсий, используем t-тест с неравными дисперсиями
    t_test_result <- t.test(as.formula(paste(feature, "~", categorical_var)), data = data, var.equal = FALSE)
  }

  return(t_test_result)
}
result <- perform_t_test("ADD_FEE", "PPIND", data_log)
print(result)
```

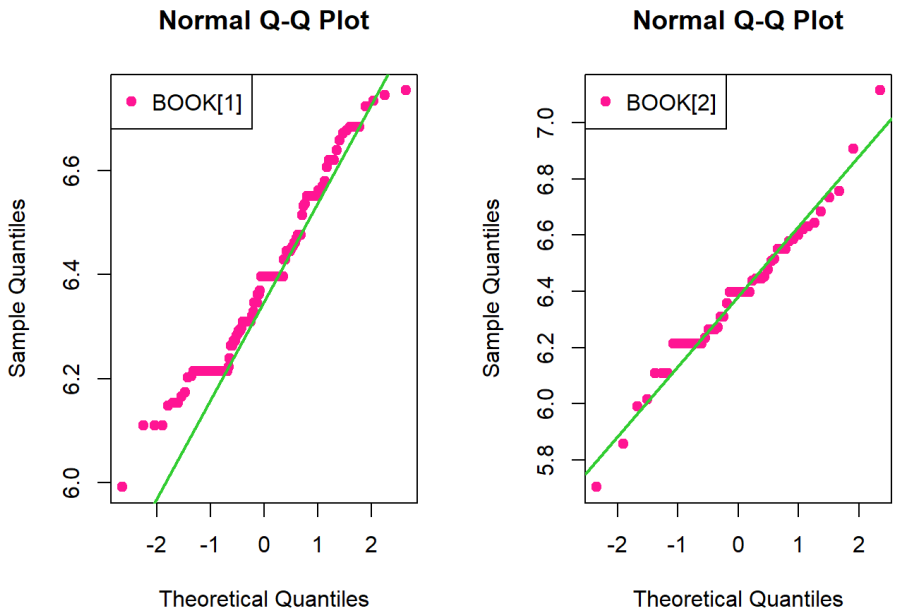
```
##
##  Two Sample t-test
##
## data:  ADD_FEE by PPIND
## t = 0.94085, df = 134, p-value = 0.3485
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.1934170  0.5443931
## sample estimates:
## mean in group 1 mean in group 2
##        6.014273        5.838785
```

По результатам теста не отвергаем нулевую гипотезу о равенстве средних

Признак BOOK

Провреим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$BOOK[data_log$PPIND == "1"], "BOOK[1]")
plot_qq_graph(data_log$BOOK[data_log$PPIND == "2"], "BOOK[2]")
```



По qq-plot нельзя сказать, что этот признак по группам нормально распределен

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

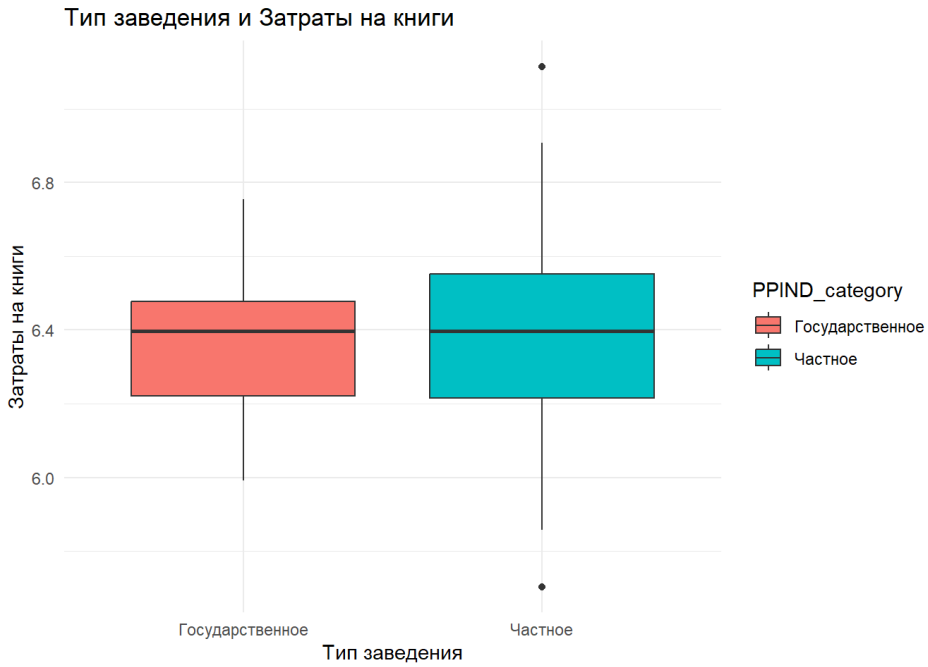
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)
  check_normality(data_log$BOOK[data_log$PPIND == "1"], "BOOK[1]")
  check_normality(data_log$BOOK[data_log$PPIND == "2"], "BOOK[2]")

results_table
```

```
##   Feature      Lillie Anderson_Darling Shapiro_Wilk
## 1 BOOK[1] 0.001400773      0.001304235  0.005342335
## 2 BOOK[2] 0.072034012      0.258330658  0.421168063
```

Результаты показывают, что для признака BOOK по гос завдедениям критерии нашли отклонения от нормального распределения, а по частным заведениям — нет.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = BOOK, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "Затраты на книги",
       title = "Тип заведения и Затраты на книги") +
  theme_minimal()
```



По ящикам с усами можно сказать, что медианы не отличаются. Затраты на книги в частных и государственных заведениях мало отличаются друг от друга.

t-test

```
library(car)
result <- perform_t_test("BOOK", "PPIND", data_log)
print(result)
```

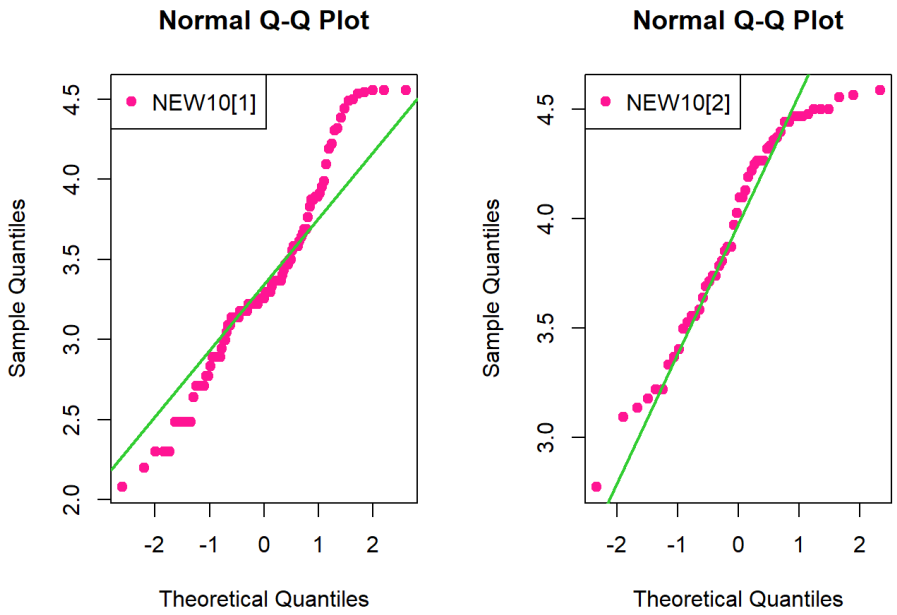
```
##
##  Welch Two Sample t-test
##
## data:  BOOK by PPIND
## t = 0.0051294, df = 73.053, p-value = 0.9959
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.07393097  0.07431251
## sample estimates:
## mean in group 1 mean in group 2
##        6.383287        6.383096
```

По результатам теста не отвергаем нулевую гипотезу о равенстве средних

Признак NEW10

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$NEW10[data_log$PPIND == "1"], "NEW10[1]")
plot_qq_graph(data_log$NEW10[data_log$PPIND == "2"], "NEW10[2]")
```



По qq-plot нельзя сказать, что этот признак по группам нормально распределен

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

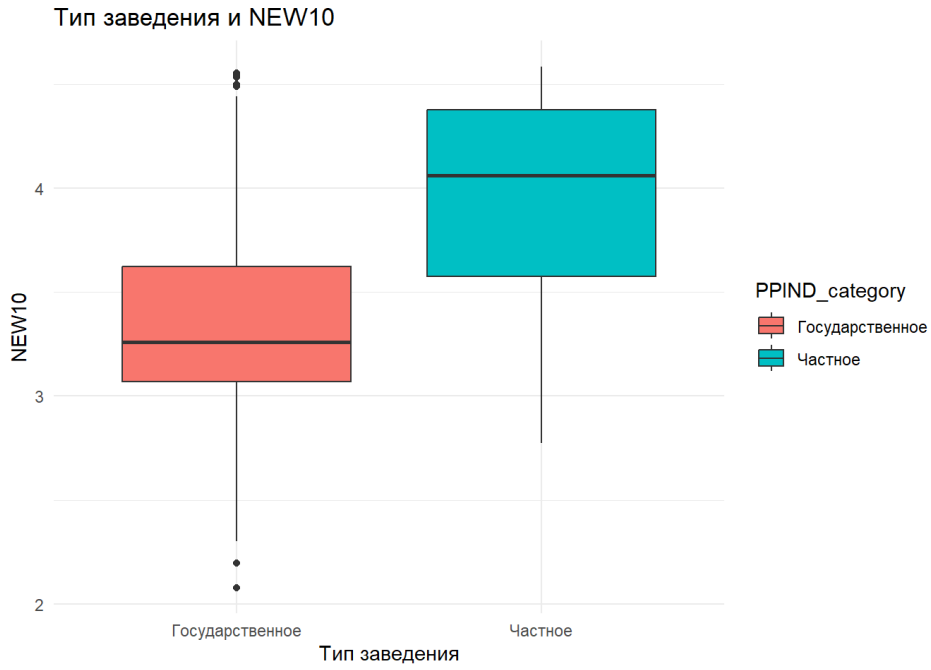
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)

  check_normality(data_log$NEW10[data_log$PPIND == "1"], "NEW10[1]")
  check_normality(data_log$NEW10[data_log$PPIND == "2"], "NEW10[2]")

results_table
```

```
##    Feature      Lillie Anderson_Darling Shapiro_Wilk
## 1 NEW10[1] 0.001606804      0.001726780  0.006230225
## 2 NEW10[2] 0.020209900      0.004572878  0.004791182
```

Результаты показывают, что для признака NEW10 по гос и по частным заведениям критерии нашли отклонения от нормального распределения.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = NEW10, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "NEW10",
       title = "Тип заведения и NEW10") +
  theme_minimal()
```



По ящикам с усами можно сказать, медианы сильно различаются. Есть большие отличия в проценте студентов, которые были отлчиниками в школе: в частных заведениях процент таких студентов выше.

t-test

```
library(car)
result <- perform_t_test("NEW10", "PPIND", data_log)
print(result)
```

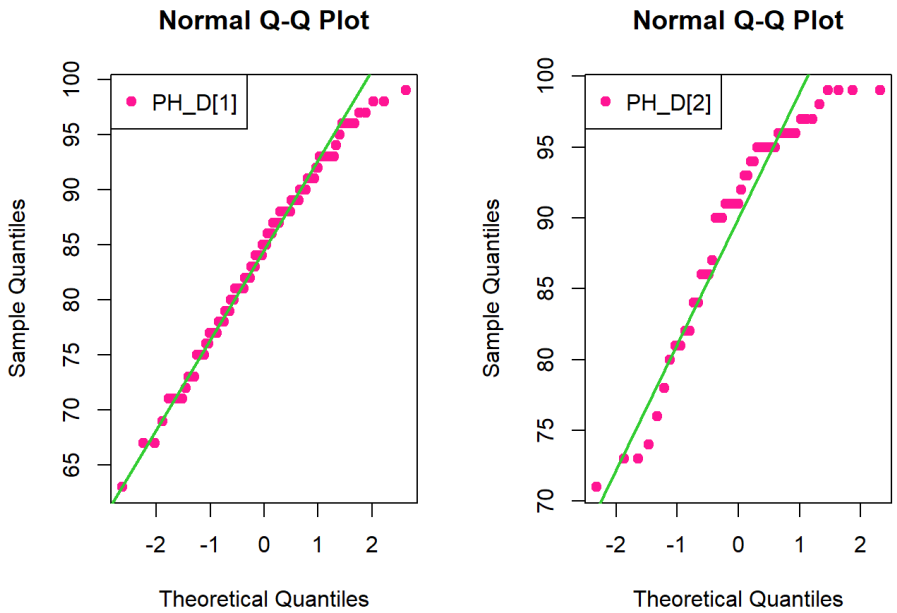
```
##
##  Two Sample t-test
##
## data:  NEW10 by PPIND
## t = -6.5716, df = 157, p-value = 6.957e-10
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.7861350 -0.4227798
## sample estimates:
## mean in group 1 mean in group 2
##        3.349782        3.954240
```

По результатам теста отвергаем нулевую гипотезу о равенстве средних

Признак PH_D

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$PH_D[data_log$PPIND == "1"], "PH_D[1]")
plot_qq_graph(data_log$PH_D[data_log$PPIND == "2"], "PH_D[2]")
```



По qq-plot видно, что распределние по гос заведениям похоже на нормальное, по частным — нет.

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

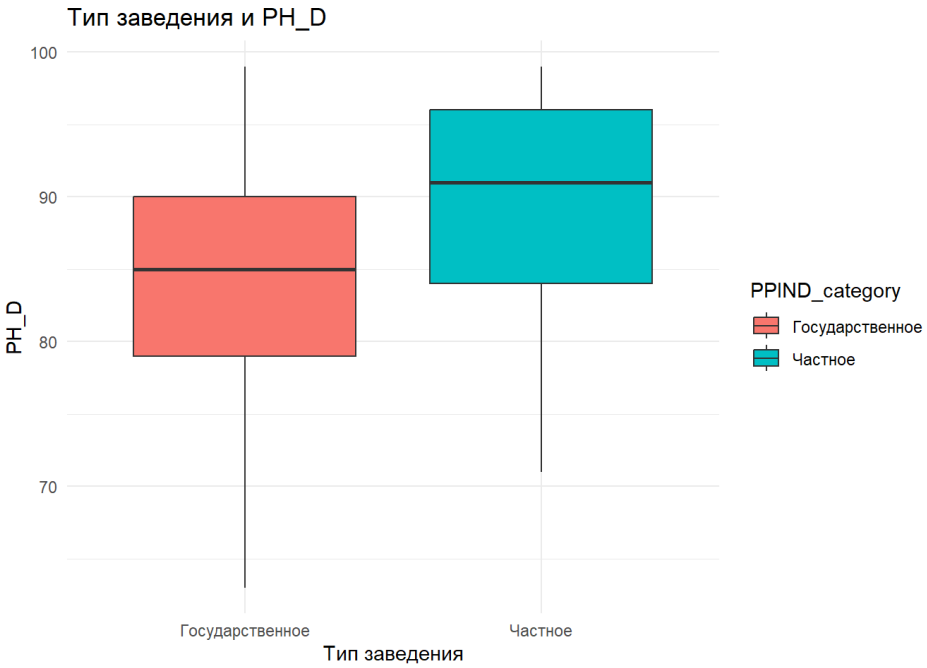
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)

  check_normality(data_log$PH_D[data_log$PPIND == "1"], "PH_D[1]")
  check_normality(data_log$PH_D[data_log$PPIND == "2"], "PH_D[2]")
results_table
```

```
##   Feature       Lillie Anderson_Darling Shapiro_Wilk
## 1 PH_D[1] 0.0976181318     0.2716997346 0.1934208406
## 2 PH_D[2] 0.0007729636     0.0001544774 0.0004137112
```

Результаты показывают, что для признака PH_D по гос заведениям критерии не нашли отклонений от нормального распределения, а по частным — нашли.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = PH_D, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "PH_D",
       title = "Тип заведения и PH_D") +
  theme_minimal()
```



По ящикам с усами можно сказать, что медианы заметно отличаются. Преподавателей с PH_D больше в частных учебных заведениях.

t-test

```
library(car)
result <- perform_t_test("PH_D", "PPIND", data_log)
print(result)
```

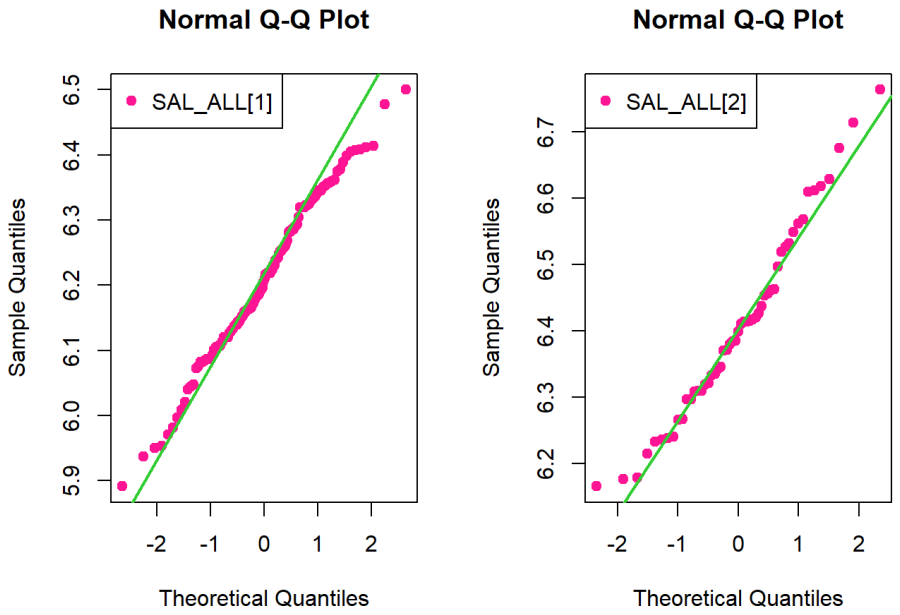
```
##
##  Two Sample t-test
##
## data:  PH_D by PPIND
## t = -4.0604, df = 164, p-value = 7.572e-05
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -7.904603 -2.732062
## sample estimates:
## mean in group 1 mean in group 2
##        84.27350        89.59184
```

По результатам теста отвергаем нулевую гипотезу о равенстве средних

Признак SAL_ALL

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$SAL_ALL[data_log$PPIND == "1"], "SAL_ALL[1]")
plot_qq_graph(data_log$SAL_ALL[data_log$PPIND == "2"], "SAL_ALL[2]")
```



По qq-plot можно сказать, что

этот признак по группам распределен нормально

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

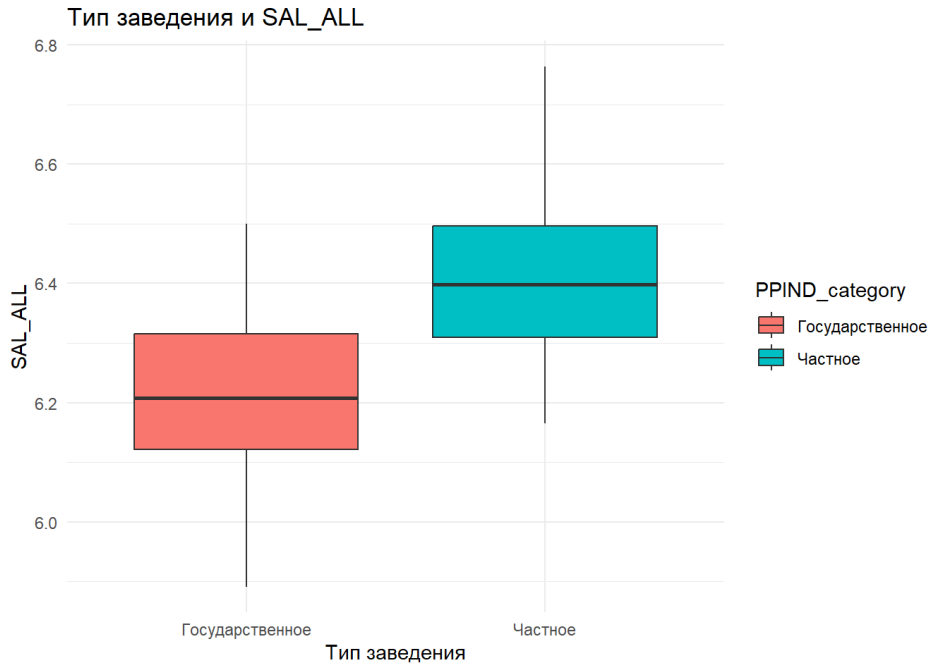
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)

  check_normality(data_log$SAL_ALL[data_log$PPIND == "1"], "SAL_ALL[1]")
  check_normality(data_log$SAL_ALL[data_log$PPIND == "2"], "SAL_ALL[2]")


results_table
```

```
##      Feature    Lillie Anderson_Darling Shapiro_Wilk
## 1 SAL_ALL[1] 0.1669333        0.3989354    0.5571695
## 2 SAL_ALL[2] 0.3893919        0.3382660    0.2679110
```

Результаты показывают, что для признака SAL_ALL по гос и по частным заведениям распределен критерии не нашли отклонения от нулевой гипотезы о нормальном распределении.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = SAL_ALL, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "SAL_ALL",
       title = "Тип заведения и SAL_ALL") +
  theme_minimal()
```



По ящикам с усами можно сказать, что медианы отличаются. Зарплата преподавателдей в частных учебных заведениях значительно выше.

t-test

```
library(car)
result <- perform_t_test("SAL_ALL", "PPIND", data_log)
print(result)
```

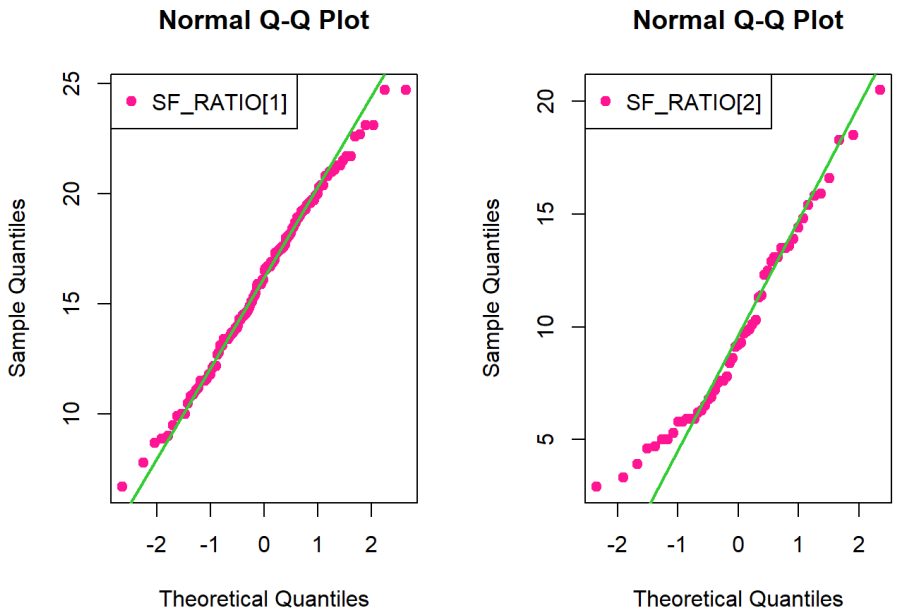
```
##
##  Two Sample t-test
##
## data:  SAL_ALL by PPIND
## t = -9.3463, df = 173, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.2404504 -0.1566004
## sample estimates:
## mean in group 1 mean in group 2
##        6.208777        6.407302
```

По результатам теста отвергаем нулевую гипотезу о равенстве средних

Признак SF_RATIO

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$SF_RATIO[data_log$PPIND == "1"], "SF_RATIO[1]")
plot_qq_graph(data_log$SF_RATIO[data_log$PPIND == "2"], "SF_RATIO[2]")
```



По qq-plot видно, что распределние по гос заведениям похоже на нормальное, по частным — нет.

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

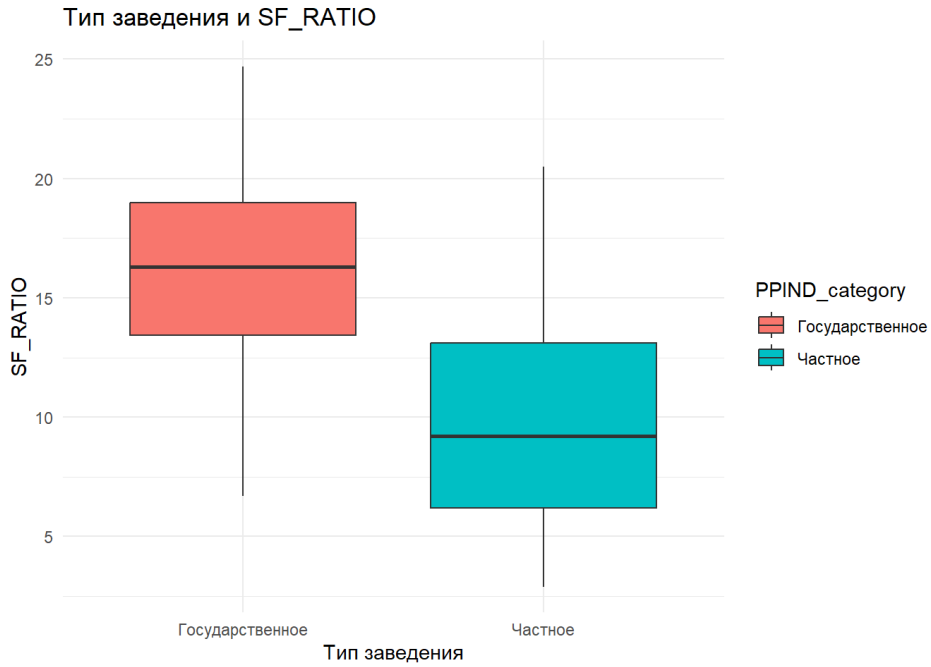
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)
  check_normality(data_log$SF_RATIO[data_log$PPIND == "1"], "SF_RATIO[1]")
  check_normality(data_log$SF_RATIO[data_log$PPIND == "2"], "SF_RATIO[2]")
results_table
```

```
##       Feature     Lillie Anderson_Darling Shapiro_Wilk
## 1 SF_RATIO[1] 0.85501479       0.80501728   0.75749892
## 2 SF_RATIO[2] 0.08809977       0.04256109   0.05277412
```

Результаты показывают, что для признака SF_RATIO по гос и по частным заведениям критерии не нашли отклонения от нормального распределения.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = SF_RATIO, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "SF_RATIO",
       title = "Тип заведения и SF_RATIO") +
   theme_minimal()
```



По ящикам с усами можно сказать, что медианы сильно отличаются. Соотношение студентов к преподавателям выше в государственных учебных заведениях.

t-test

```
library(car)
result <- perform_t_test("SF_RATIO", "PPIND", data_log)
print(result)
```

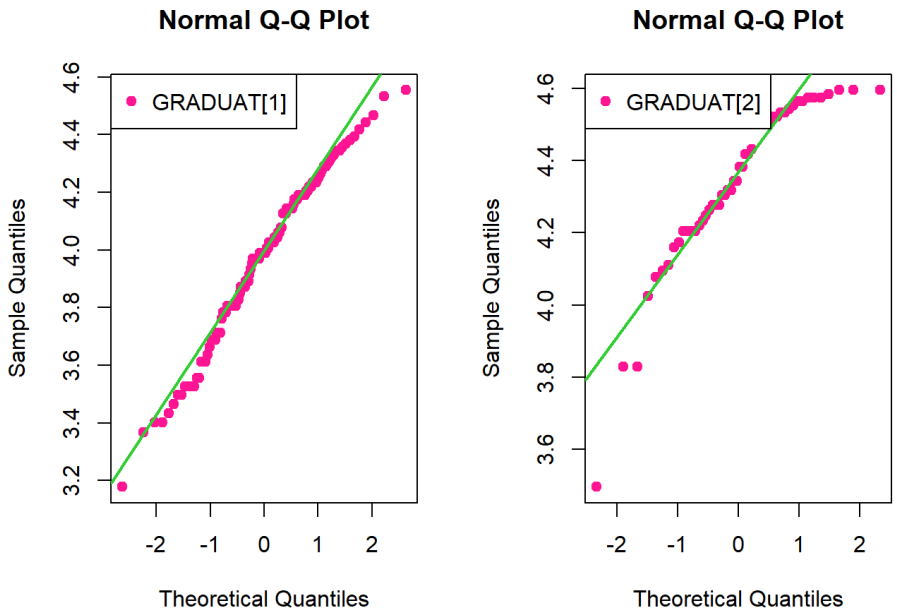
```
##
##  Two Sample t-test
##
## data:  SF_RATIO by PPIND
## t = 9.5879, df = 173, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  5.002351 7.595824
## sample estimates:
## mean in group 1 mean in group 2
##       16.112295        9.813208
```

По результатам теста отвергаем нулевую гипотезу о равенстве средних

Признак GRADUAT

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$GRADUAT[data_log$PPIND == "1"], "GRADUAT[1]")
plot_qq_graph(data_log$GRADUAT[data_log$PPIND == "2"], "GRADUAT[2]")
```



По qq-plot нельзя сказать, что этот признак по группам нормально распределен

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

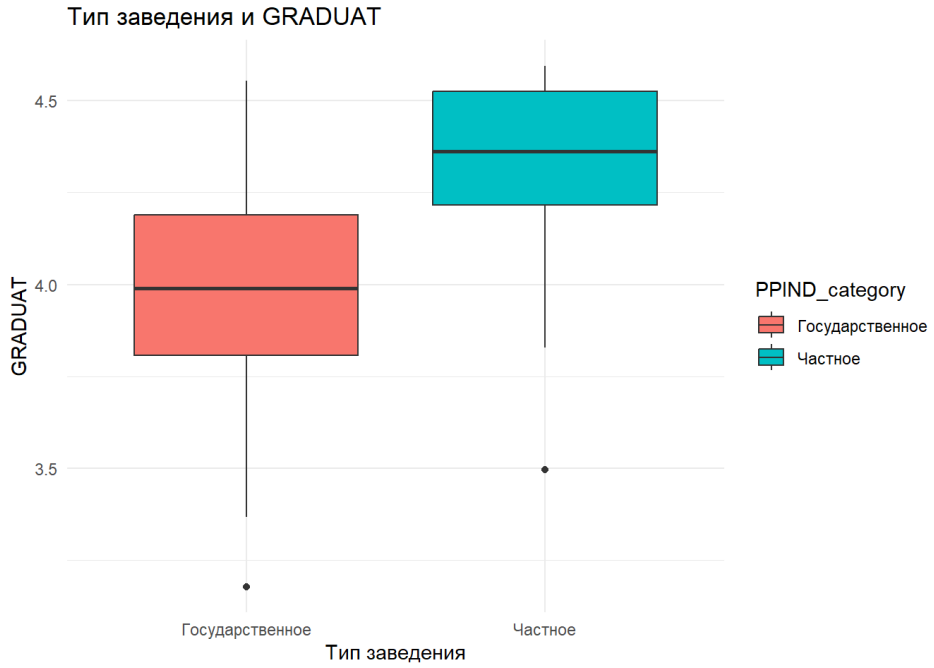
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)
  check_normality(data_log$GRADUAT[data_log$PPIND == "1"], "GRADUAT[1]")
  check_normality(data_log$GRADUAT[data_log$PPIND == "2"], "GRADUAT[2]")
results_table
```

```
##      Feature      Lillie Anderson_Darling Shapiro_Wilk
## 1 GRADUAT[1] 0.038301928     0.1174846887 1.650298e-01
## 2 GRADUAT[2] 0.002156807     0.0008638095 7.071745e-05
```

Результаты показывают, что для признака GRADUAT по гос и по частным заведениям критерии нашли отклонения от нормального распределения.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = GRADUAT, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "GRADUAT",
       title = "Тип заведения и GRADUAT") +
   theme_minimal()
```



По ящикам с усами можно сказать, что медианы сильно отличаются. В частных заведениях процент выпускающихсмя студентов выше, чем в государственных.

t-test

```
library(car)
result <- perform_t_test("GRADUAT", "PPIND", data_log)
print(result)
```

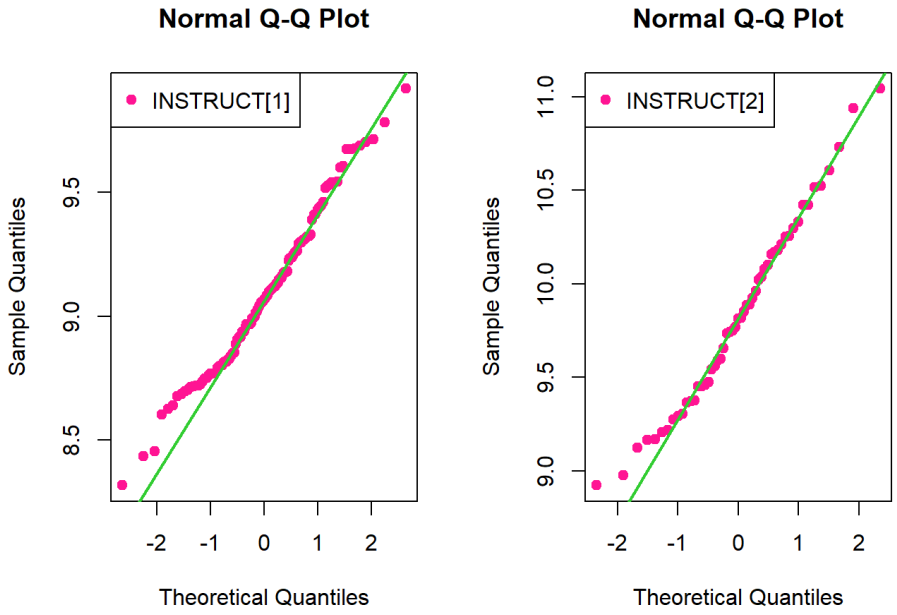
```
##
##  Two Sample t-test
##
## data:  GRADUAT by PPIND
## t = -8.328, df = 168, p-value = 2.769e-14
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.4603920 -0.2839438
## sample estimates:
## mean in group 1 mean in group 2
##        3.970564        4.342732
```

По результатам теста отвергаем нулевую гипотезу о равенстве средних

Признак INSTRUCT

Проверим нормальность визуально:

```
par(mfrow = c(1, 2))
plot_qq_graph(data_log$INSTRUCT[data_log$PPIND == "1"], "INSTRUCT[1]")
plot_qq_graph(data_log$INSTRUCT[data_log$PPIND == "2"], "INSTRUCT[2]")
```



По qq-plot нельзя сказать, что этот признак по группам нормально распределен

Проверим нормальность по тестам Лиллиефорса, Андерсена-Дарлинга и Шапиро-Уилка:

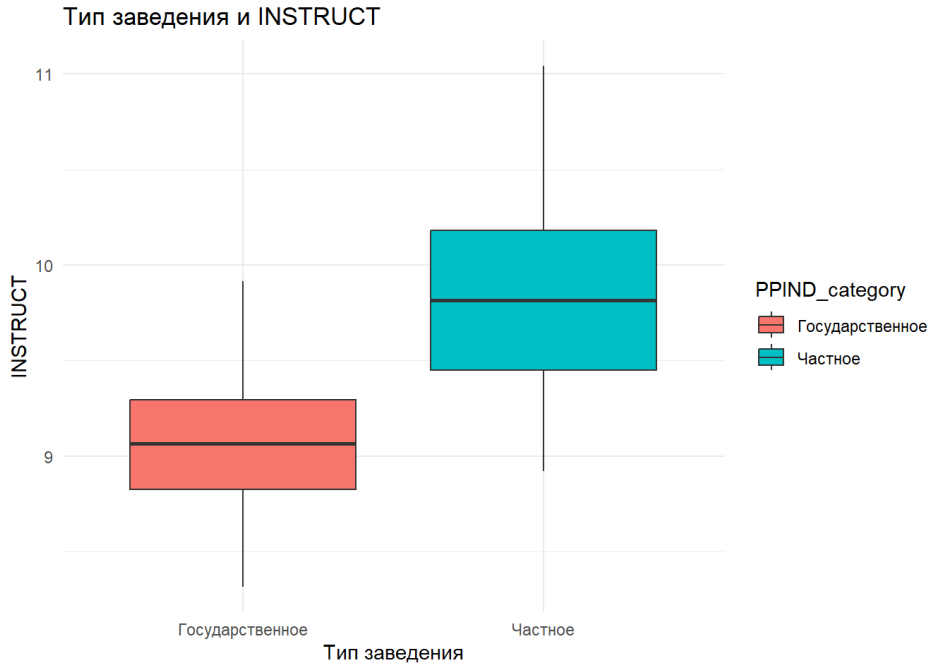
```
results_table <- data.frame(
  Feature = character(),
  Lillie = numeric(),
  Anderson_Darling = numeric(),
  Shapiro_Wilk = numeric(),
  stringsAsFactors = FALSE
)
  check_normality(data_log$INSTRUCT[data_log$PPIND == "1"], "INSTRUCT[1]")
  check_normality(data_log$INSTRUCT[data_log$PPIND == "2"], "INSTRUCT[2]")
results_table
```

```
##       Feature    Lillie Anderson_Darling Shapiro_Wilk
## 1 INSTRUCT[1] 0.2812896       0.07475714    0.1425812
## 2 INSTRUCT[2] 0.5493138       0.60569460    0.4738598
```

Результаты показывают, что для признака INSTRUCT по гос и по частным заведениям критерии не нашли отклонений от нормального распределения.

Box-plot

```
library(ggplot2)
library(dplyr)

data_log |>
  mutate(PPIND_category = ifelse(PPIND == 1, "Государственное", "Частное")) |>
  ggplot(aes(x = PPIND_category, y = INSTRUCT, fill = PPIND_category)) +
  geom_boxplot() +
  labs(x = "Тип заведения", y = "INSTRUCT",
       title = "Тип заведения и INSTRUCT") +
   theme_minimal()
```



По ящикам с усами можно сказать, что медианы заметно отличаются. В частных заведениях затраты на одного студента выше, чем в государственных.

t-test

```
library(car)
result <- perform_t_test("INSTRUCT", "PPIND", data_log)
print(result)
```

```
##
##  Welch Two Sample t-test
##
## data:  INSTRUCT by PPIND
## t = -9.8994, df = 70.134, p-value = 5.954e-15
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.8961327 -0.5956012
## sample estimates:
## mean in group 1 mean in group 2
##        9.084125        9.829992
```

По результатам теста отвергаем нулевую гипотезу о равенстве средних

Результаты t-test.

```
library(car)

features <- c("ADD_FEE", "BOOK", "NEW10", "PH_D", "SAL_ALL", "SF_RATIO", "GRADUAT", "INSTRUCT")

perform_t_tests <- function(data_frame, features, categorical_var)
  {
  results <- data.frame(Feature = character(),
                        Equal_Variance = character(),
                        P_Value = numeric(),
                        Hypothesis = character(),
                        stringsAsFactors = FALSE)

  for (feature in features) {
    # Проверка равенства дисперсий
    levene_test <- leveneTest(as.formula(paste(feature, "~", categorical_var)), data = data_frame)

    # Выбор метода для сравнения средних
    if (levene_test$'Pr(>F)'[1] > 0.05) {
      # Равенство дисперсий, используем t-тест с равными дисперсиями
      t_test_result <- t.test(as.formula(paste(feature, "~", categorical_var)), data = data_frame, var.equal = TRUE)
    } else {
      # Неравенство дисперсий, используем t-тест с неравными дисперсиями
      t_test_result <- t.test(as.formula(paste(feature, "~", categorical_var)), data = data_frame, var.equal = FALSE)
    }

    # Определение равенства дисперсий
    equal_var <- ifelse(levene_test$'Pr(>F)'[1] > 0.05, "+", "-")

    hypothesis <- ifelse(t_test_result$p.value < 0.05, "откл (!=)", "не откл (=)")
    result_row <- data.frame(
      Feature = feature,
      Equal_Variance = equal_var,
      P_Value = t_test_result$p.value,
      Hypothesis = hypothesis
    )
    results <- rbind(results, result_row)
  }
  return(results)
}
results <- perform_t_tests(data_log, features, "PPIND")
print(results)
```

```
##    Feature Equal_Variance      P_Value  Hypothesis
## 1  ADD_FEE              + 3.484740e-01 не откл (=)
## 2     BOOK              - 9.959214e-01 не откл (=)
## 3    NEW10              + 6.957361e-10   откл (!=)
## 4     PH_D              + 7.572281e-05   откл (!=)
## 5  SAL_ALL              + 4.564766e-17   откл (!=)
## 6 SF_RATIO              + 9.955843e-18   откл (!=)
## 7  GRADUAT              + 2.768791e-14   откл (!=)
## 8 INSTRUCT              - 5.953510e-15   откл (!=)
```

Так как данные не распределены нормально, то применим критерий Манна-Уитни. Распределения признаков ADD_FEE, BOOK, SF_RATIO, INSTRUCT довольно симметричные (некоторые из них сдвинуты по категоризующей переменной), поэтому для них проверяется нулевая гипотеза о равенстве медиан в исходном распределении (равенство мат ожиданий в логарифмированном). Для остальных признаков с несимметричными распределениями — нулевая гипотеза: две выборки полностью однородны, т. е. принадлежат одному распределению.

```
apply_mann_whitney <- function(features, group_var, data_frame) {
  results <- data.frame(
    Feature = character(),
    P_Value = numeric(),
    Hypothesis = character(),
    stringsAsFactors = FALSE
  )

  for (feature in features)
    {
    mw_test <- wilcox.test(data_frame[[feature]] ~ data_frame[[group_var]], data = data_frame)
    hypothesis <- ifelse(mw_test$p.value < 0.05, "откл", "не откл")

    result_row <- data.frame(
      Feature = feature,
      P_Value = mw_test$p.value,
      Hypothesis = hypothesis
    )
    results <- rbind(results, result_row)
  }
  return(results)
}
features_to_test <- c("ADD_FEE", "BOOK", "NEW10", "PH_D", "SAL_ALL",  "SF_RATIO", "GRADUAT", "INSTRUCT")
grouping_variable <- "PPIND"

results_table1 <- apply_mann_whitney(features_to_test, grouping_variable, data_log)

print(results_table1)
```

```
##    Feature      P_Value Hypothesis
## 1  ADD_FEE 2.900678e-01    не откл
## 2     BOOK 8.908343e-01    не откл
## 3    NEW10 5.956459e-09       откл
## 4     PH_D 4.533451e-05       откл
## 5  SAL_ALL 8.594768e-14       откл
## 6 SF_RATIO 9.710700e-14       откл
## 7  GRADUAT 1.058917e-13       откл
## 8 INSTRUCT 6.442331e-17       откл
```

Тест Манна-Уитни дал такие же результаты как и t-test.

Тест Колмогорова-Смирнова. Проверяем гипотезу о равенстве распределений:

```
perform_ks_group_test <- function(features, categorical_var, data_frame) {
  results <- data.frame(
    Feature = character(),
    P_Value = numeric(),
    Hypothesis = character(),
    stringsAsFactors = FALSE
  )

  for (feature in features) {
    for (group_val in unique(data_frame[[categorical_var]])) {
      ks_result <- ks.test(data_frame[[feature]][data_frame[[categorical_var]] == group_val],data_frame[[feature]][data_fram
e[[categorical_var]] != group_val])
    }
      hypothesis <- ifelse(ks_result$p.value < 0.05, "откл", "не откл")

      result_row <- data.frame(
        Feature = as.character(feature),
        P_Value = ks_result$p.value,
        Hypothesis = hypothesis
      )

      results <- rbind(results, result_row)

  }
  return(results)
}
features_to_test <- c("ADD_FEE", "BOOK", "NEW10", "PH_D", "SAL_ALL",  "SF_RATIO", "GRADUAT", "INSTRUCT")
results_table <- perform_ks_group_test(features_to_test, "PPIND", data_log)


results_table
```

```
##    Feature      P_Value Hypothesis
## 1  ADD_FEE 2.938388e-01    не откл
## 2     BOOK 7.103304e-01    не откл
## 3    NEW10 1.131360e-08       откл
## 4     PH_D 1.246575e-05       откл
## 5  SAL_ALL 7.012522e-11       откл
## 6 SF_RATIO 6.242531e-11       откл
## 7  GRADUAT 3.231907e-13       откл
## 8 INSTRUCT 4.951998e-14       откл
```

По результатам теста распределения по категоризующей переменной ADD_FEE и BOOK являются одинаковыми, про остальные признаки нельзя сделать такой вывод.

По результатам можно сделать заключения о государственных и частных учебных заведедениях:

1. Нельзя сказать, что ***дополнительные взносы*** для студентов государственных и частных учебных заведений значимо отличаются.
2. ***Затраты на книги*** тоже не имеют заметных отличий в государственных и частных заведениях.
3. ***Процент студентов, которые были отличниками в школе*** значительно больше в частных заведениях, чем в государственных
4. ***Количество преподавателей с PH_D*** заметно выше в частных заведениях, чем в государственных.
5. ***Средняя заработная плата преподавателей*** значительно выше в частных учебных заведениях.
6. ***Соотношение студентов к преподавателям*** выше в государственных завдедениях(больше поступающих).
7. ***Процент выпускающихся студентов*** выше в частных учебных заведениях.
8. ***Расходы на обучение в расчете на одного учащегося*** больше в частных учебных заведениях, чем в государственных.

Сравнение величину дополнительных взносов (ADD FEE) и затраты на книги (BOOK).

t-test для парных выборок. Нулевая гипотеза: равенство мат ожиданий(в логарифмированных данных или медиан в исходных).

```
t_test_result <- t.test(data_log$ADD_FEE, data_log$BOOK, paired = TRUE)

print(t_test_result)
```

```
##
##  Paired t-test
##
## data:  data_log$ADD_FEE and data_log$BOOK
## t = -4.7615, df = 135, p-value = 4.887e-06
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.6064077 -0.2504922
## sample estimates:
## mean difference
##        -0.42845
```

Таким образом, на основании данного теста можно сделать вывод о наличии статистически значимой разницы между средними значениями переменных ADD_FEE и BOOK. То есть затраты на книги в среднем больше, чем на дополнительные взносы.

Тест Вилкоксона для парных выборок. Нулевая гипотеза: равенство мат ожиданий(в логарифмированных данных или медиан в исходных).

```
wilcox.test <- wilcox.test(data_log$ADD_FEE, data_log$BOOK, paired = TRUE)

print(wilcox.test)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data_log$ADD_FEE and data_log$BOOK
## V = 2654, p-value = 1.349e-05
## alternative hypothesis: true location shift is not equal to 0
```

Такой же результат как и в t-test’е.

Тест Колмогорова-Смирнова. Нулевая гипотеза: распределения ADD_FEE и BOOK равны

```
ks_test_result <- ks.test(data_log$ADD_FEE, data_log$BOOK, paired = TRUE)

print(ks_test_result)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  data_log$ADD_FEE and data_log$BOOK
## D = 0.51046, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Отвергаем гипотезу о равенстве распределений

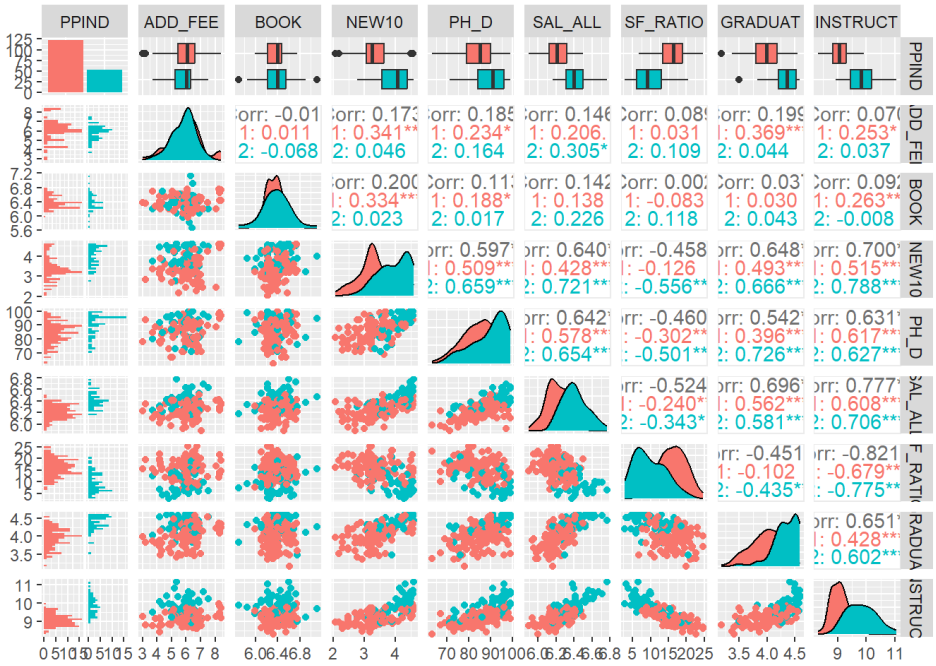
В результате можно заключить, что

1. Признаки ADD_FEE и BOOK имеют разные распределения
2. Затраты на книги в среднем больше, чем на дополнительные взносы

Анализ зависимостей

Еще раз посмотрим на pairs plot.

```
library(GGally)
library(dplyr)

data_log$PPIND <- factor(data_log$PPIND)
# Построим графики с разделением по 'PPIND'
ggpairs(data_log, columns = 2:ncol(data_log), aes(colour = PPIND))
```



Так как данные неоднородны по всем признакам, кроме ADD_FEE и BOOK, рассматриваем корреляции отдельно внутри групп.

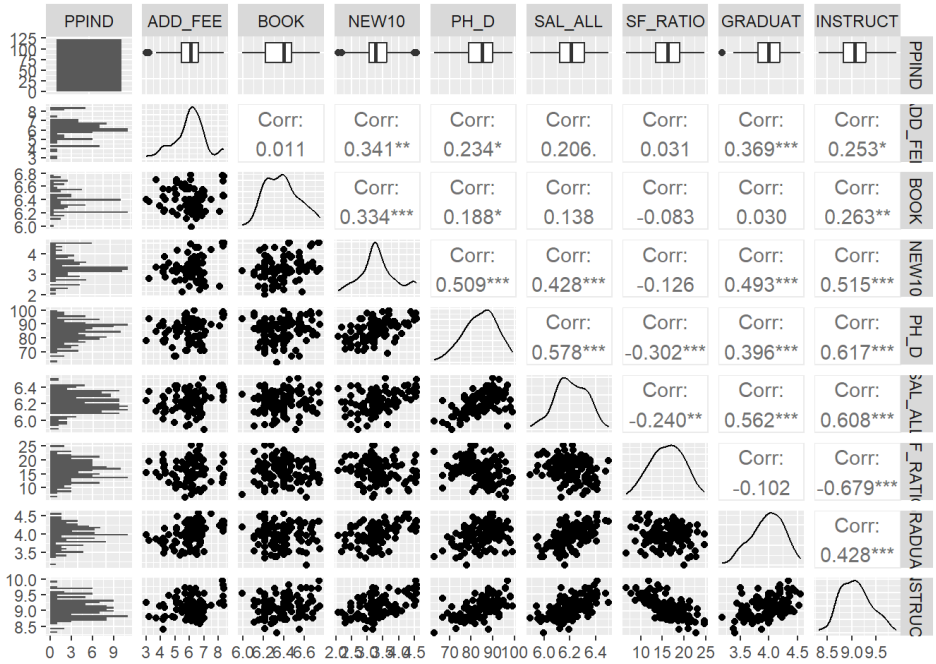
Коэффициенты Пирсона. Использую Pairwise Deletion: метод использует доступные данные для каждой пары переменных, игнорируя пропуски в других переменных, так как в некоторых признаках много пропусков (например, в одном – почти четверть).

Незначимые корреляции на графике не отображаются.

Государственные заведения

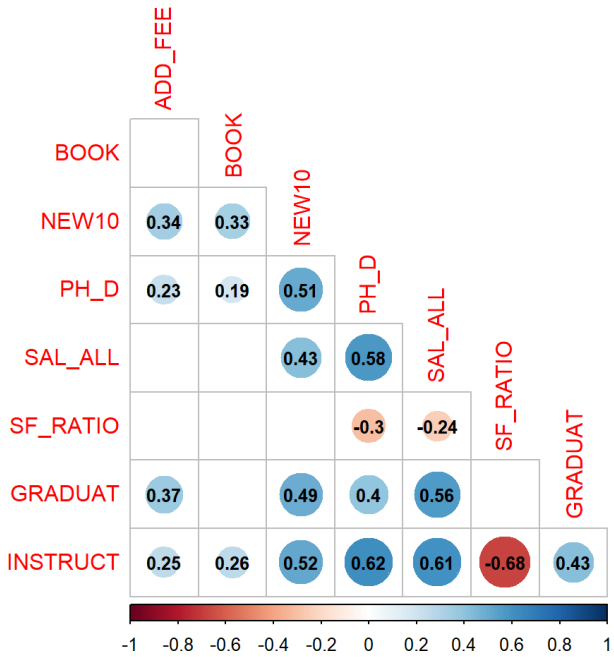
pairs plot для гос заведений

```
library(GGally)
library(dplyr)

data_log$PPIND <- factor(data_log$PPIND)
subset_data <- data_log[data_log$PPIND == 1, ]
ggpairs(subset_data, columns = 2:ncol(subset_data))
```



Коэффициенты корреляции Пирсона для гос заведений(незначимые не выводятся)

```
library(dplyr)
library(corrplot)


dfff <- data_log |>
  filter(PPIND == "1") |>
  select(ADD_FEE,
         BOOK,
         NEW10,
         PH_D,
         SAL_ALL,
         SF_RATIO,
         GRADUAT,
         INSTRUCT)

testRes = cor.mtest(dfff, conf.level = 0.95)
cor <- cor(dfff,
           method = "pearson",
           use = "pairwise.complete.obs")

  corrplot(cor,
           method = 'circle',
           type = 'lower',
           p.mat = testRes$p,
           insig='blank',
           sig.level = 0.05,
           addCoef.col ='black',
           number.cex = 0.8,
           diag=FALSE)
```



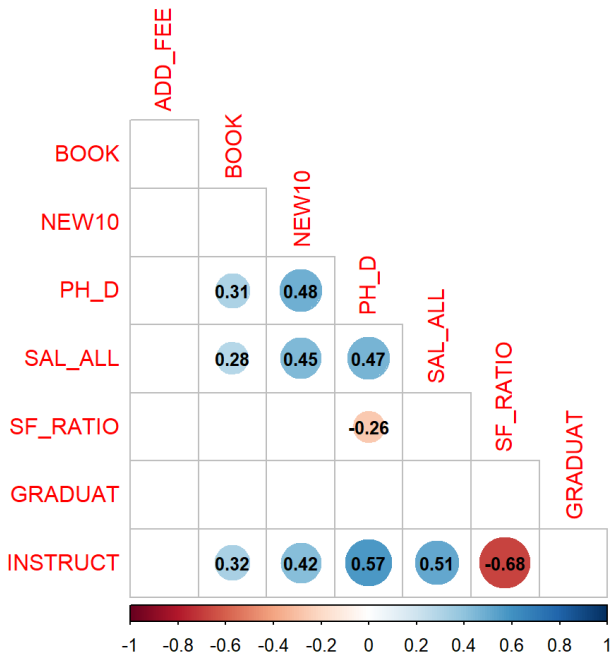
Некоторые корреляции сложно объяснить, поэтому посмотрим на частные корреляции за вычетом цены за обучение, которая может как-то влиять.

Добавляю столбец, показывающий стоимость обучения, чтобы с его помощью посмотреть на частные корреляции

```
library(ppcor)

data$log_column <- log(data$OUT_STAT)
data <- data[-142, ]

# Добавление логарифмированного столбца из data в data_log
data_log$OUT_STAT <- data$log_column
```

Частные корреляции за вычетом стоимости обучения, выведем только значимые

```
library(ppcor)

filtered_data <- data_log |>
  filter(PPIND == "1" &
         !is.na(ADD_FEE) &
         !is.na(BOOK) &
         !is.na(NEW10) &
         !is.na(PH_D)&
         !is.na(SAL_ALL)&
         !is.na(SF_RATIO)&
         !is.na(GRADUAT)&
         !is.na(INSTRUCT)&
         !is.na(OUT_STAT))

# Выбор всех столбцов кроме последнего (OUT_STAT)
features <- names(filtered_data)[3:(ncol(filtered_data) - 1)]

# Создание пустой матрицы результатов
results <- matrix(NA, nrow = length(features), ncol = length(features))

# Вычисление частных корреляций
for (i in 1:(length(features) - 1)) {
  for (j in (i + 1):length(features)) {
    result <- pcor.test(filtered_data[, features[i]], filtered_data[, features[j]], filtered_data[, ncol(filtered_data)], me
thod = "pearson")
    results[i, j] <- result$estimate
    results[j, i] <- result$estimate
  }
}
# Назначение имен строкам и столбцам
rownames(results) <- features
colnames(results) <- features

testRes = cor.mtest(results, conf.level = 0.95)

corrplot(results,
         method = 'circle',
         type = 'lower',
         p.mat = testRes$p,
         insig='blank',
         sig.level = 0.05,
         addCoef.col ='black',
         number.cex = 0.8,
         diag=FALSE)
```



Можно сделать вывод о том, что стоимость обучения в гос заведениях влияла на многие корреляции — они стали незначимы. ОСтальные изменились незначительно.

Выведем для каждой пары значение корреляции Пирсона и значение корреляции Спирмена(Государственные заведения):

```
library(dplyr)

filtered_data <- data_log |>
  filter(PPIND == "1") |>
   dplyr::select(ADD_FEE,
         BOOK,
         NEW10,
         PH_D,
         SAL_ALL,
         SF_RATIO,
         GRADUAT,
         INSTRUCT)

pairs <- combn(names(filtered_data),
               2,
               simplify = TRUE)
cor_table <- data.frame(Variables = apply(pairs,
                                          2,
                                          paste,
                                          collapse = " vs "))

cor_table$Pearson <- apply(pairs, 2, function(x)
  {
  cor(filtered_data[, x],
      use = "pairwise.complete.obs")[1, 2]
})

cor_table$Spearman <- apply(pairs, 2, function(x)
  {
  cor(filtered_data[, x], method = "spearman",
      use = "pairwise.complete.obs")[1, 2]
})

cor_table
```

```
##              Variables     Pearson    Spearman
## 1       ADD_FEE vs BOOK  0.01127714 -0.07326224
## 2      ADD_FEE vs NEW10  0.34065478  0.21007551
## 3       ADD_FEE vs PH_D  0.23382577  0.17931026
## 4    ADD_FEE vs SAL_ALL  0.20596537  0.18826740
## 5   ADD_FEE vs SF_RATIO  0.03125936  0.08481623
## 6    ADD_FEE vs GRADUAT  0.36894339  0.39387251
## 7   ADD_FEE vs INSTRUCT  0.25288963  0.12057902
## 8         BOOK vs NEW10  0.33401044  0.33476898
## 9          BOOK vs PH_D  0.18832139  0.20282765
## 10      BOOK vs SAL_ALL  0.13765933  0.15574579
## 11     BOOK vs SF_RATIO -0.08251567 -0.08056881
## 12      BOOK vs GRADUAT  0.03015766  0.06825304
## 13     BOOK vs INSTRUCT  0.26269901  0.26064187
## 14        NEW10 vs PH_D  0.50877971  0.49719953
## 15     NEW10 vs SAL_ALL  0.42806520  0.41741688
## 16    NEW10 vs SF_RATIO -0.12635324 -0.11714643
## 17     NEW10 vs GRADUAT  0.49319759  0.48048922
## 18    NEW10 vs INSTRUCT  0.51506578  0.46965813
## 19      PH_D vs SAL_ALL  0.57756907  0.59024226
## 20     PH_D vs SF_RATIO -0.30195686 -0.31157404
## 21      PH_D vs GRADUAT  0.39588162  0.40663269
## 22     PH_D vs INSTRUCT  0.61723922  0.64799146
## 23  SAL_ALL vs SF_RATIO -0.24032141 -0.23752630
## 24   SAL_ALL vs GRADUAT  0.56170103  0.57721790
## 25  SAL_ALL vs INSTRUCT  0.60787059  0.62757397
## 26  SF_RATIO vs GRADUAT -0.10220263 -0.05655679
## 27 SF_RATIO vs INSTRUCT -0.67881978 -0.66015728
## 28  GRADUAT vs INSTRUCT  0.42816448  0.42216515
```

Значения корреляций по Государственным заведениям в основном совпадают, что говорит о линейных или монотонных зависимостях.

Выводы

Следующие пары признаков имеют значимые корреляции для государственных заведений:

1. ***NEW_10 и PH_D*** — Положительная корреляция между процентом студентов, которые были отличниками в школе и количеством преподавателей с докторской степенью может указывать на то, что более успешные школьники выбирают более престижные заведения, где больше преподавателей с PH_D (***NEW_10*** — **следствие**, ***PH_D*** — **причина**).
2. ***NEW_10 и SAL_ALL*** — Высокая положительная корреляция между процентом студентов, получивших отличные оценки в школе, и средней заработной платой преподавателей объясняется тем, что более успешные школьники выбирают более престижные заведения, где преодавателям платят больше. (***NEW_10*** — **следствие**, ***SAL_ALL*** — **причина**).
3. ***NEW_10 и INSTRUCT*** — Высокая положительная корреляция между процентом отличников и расходами на обучение на одного студента может указывать на то, что более высокие расходы могут предоставлять дополнительные ресурсы для успешного обучения и бывшие отличники в шокле выбирают именно такие заведения (***NEW_10*** — **следствие**, ***INSTRUCT*** — **причина**).
4. ***PH_D и SAL_ALL*** — Положительная корреляция между количеством преподавателей с докторской степенью и средней заработной платой преподавателей может указывать на привлечение более квалифицированных специалистов, что влияет на уровень их оплаты труда (***PH_D*** — **причина**, ***SAL_ALL*** — **следствие**)
5. ***PH_D и SF_RATIO*** — низкая отрицательная корреляция между количеством преподавателей с докторской степенью и соотношением студентов к преподавателям. (причина и следствие не ясны, но стоимость обучения в вузе никак не влияет на корреляцию)
6. ***PH_D и INSTRUCT*** — Положительная корреляция между количеством преподавателей с докторской степенью и расходами на обучение на одного студента (причина и следствие не ясны, но стоимость обучения в вузе никак не влияет на корреляцию)
7. ***SAL_ALL и INSTRUCT*** — Высокая положительная корреляция между средней заработной платой преподавателей и расходами на обучение на одного студента может свидетельствовать о том, что учебные заведения, где преподаватели получают более высокую заработную плату, могут тратить и большие суммы на обучение в рассчете на 1 человека (причина и следствие не ясны, но стоимость обучения в вузе никак не влияет на корреляцию)
8. ***INSTRUCT и SF_RATIO*** — Отрицательная корреляция между расходами на обучение на одного учащегося и соотношением студентов к преподавателям указывает на то, что чем больше человек, тем меньше средств тратят на каждого их них. (***SF_RATIO*** — **причина**, ***INSTRUCT*** — **следствие**).
9. ***BOOK и SAL_ALL*** — Положительная корреляция между расходами на книги и средней зарплатой преподавателей указывает на то,

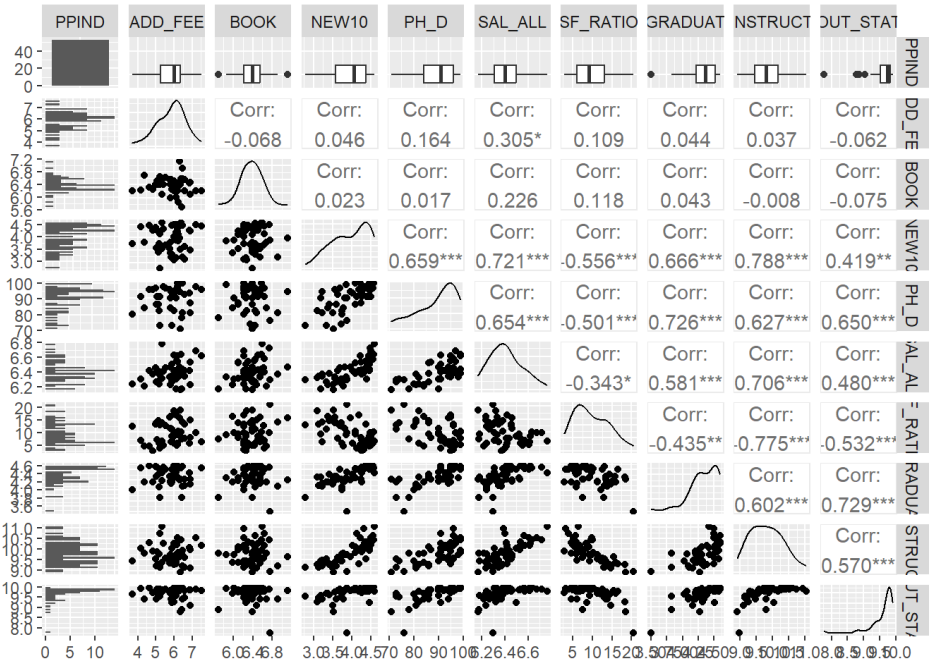
что учебные заведения, где преподаватели получают более высокую заработную плату, могут тратить и большие суммы на книги (причина и следствие не ясны, но стоимость обучения в вузе никак не влияет на корреляцию)

10. ***BOOK u INSTRUCT*** — как в прошлом пункте (причина и следствие не ясны, но стоимость обучения в вузе никак не влияет на корреляцию).

Также можно отметить, что государственнеы учреждения довольно сильно отличаются друг от друга по разным параметрам, в то время как частные более похожи.

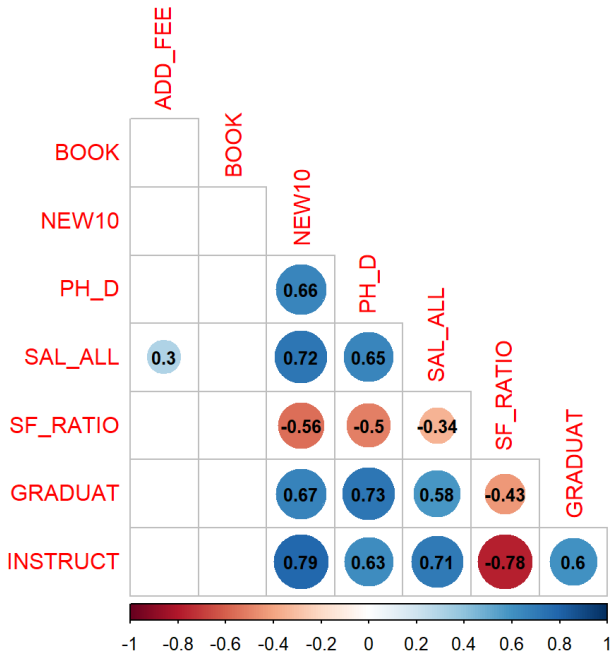
Частные заведения

```
library(GGally)
library(dplyr)

data_log$PPIND <- factor(data_log$PPIND)
subset_data <- data_log[data_log$PPIND == 2, ]
ggpairs(subset_data, columns = 2:ncol(subset_data))
```



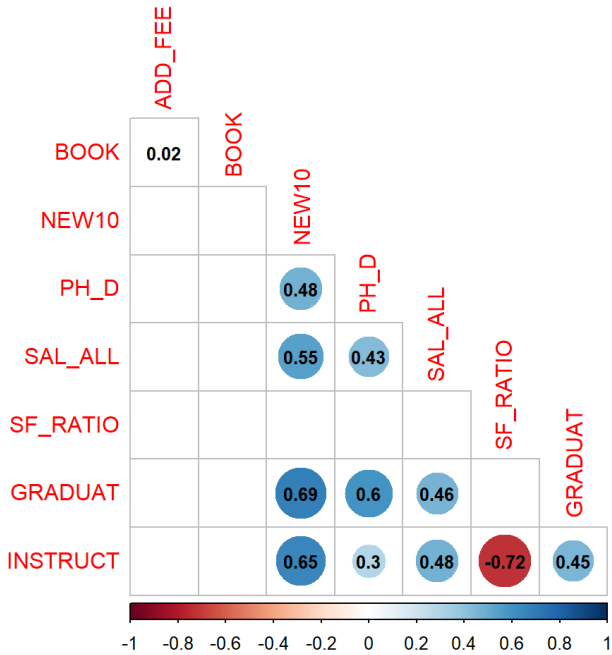
Коэффициенты корреляции Пирсона для частных заведений(незначимые не выводятся)

```
library(dplyr)

  dfff <- data_log |>
  filter(PPIND == "2") |>
  dplyr::select(ADD_FEE,
         BOOK,
         NEW10,
         PH_D,
         SAL_ALL,
         SF_RATIO,
         GRADUAT,
         INSTRUCT)

testRes = cor.mtest(dfff, conf.level = 0.95)
cor <- cor(dfff,
           method = "pearson",
           use = "pairwise.complete.obs")

  corrplot(cor,
           method = 'circle',
           type = 'lower',
           p.mat = testRes$p,
           insig='blank',
           sig.level = 0.05,
           addCoef.col ='black',
           number.cex = 0.8,
           diag=FALSE)
```



Частные корреляции за вычетом стоимости обучения, выведем только значимые

```
library(ppcor)
library(dplyr)

filtered_data <- data_log |>
  filter(PPIND == "2" &
         !is.na(ADD_FEE) &
         !is.na(BOOK) &
         !is.na(NEW10) &
         !is.na(PH_D)&
         !is.na(SAL_ALL)&
         !is.na(SF_RATIO)&
         !is.na(GRADUAT)&
         !is.na(INSTRUCT)&
         !is.na(OUT_STAT))

# Выбор всех столбцов кроме последнего (OUT_STAT)
features <- names(filtered_data)[3:(ncol(filtered_data) - 1)]

# Создание пустой матрицы результатов
results <- matrix(NA, nrow = length(features), ncol = length(features))

# Вычисление частных корреляций
for (i in 1:(length(features) - 1)) {
  for (j in (i + 1):length(features)) {
    result <- pcor.test(filtered_data[, features[i]], filtered_data[, features[j]], filtered_data[, ncol(filtered_data)], me
thod = "pearson")
    results[i, j] <- result$estimate
    results[j, i] <- result$estimate
  }
}
# Назначение имен строкам и столбцам
rownames(results) <- features
colnames(results) <- features

testRes = cor.mtest(results, conf.level = 0.95)

corrplot(results,
         method = 'circle',
         type = 'lower',
         p.mat = testRes$p,
         insig='blank',
         sig.level = 0.05,
         addCoef.col ='black',
         number.cex = 0.8,
         diag=FALSE)
```



Можно сделать вывод о том, что стоимость обучения в частных заведениях влияла на многие корреляции — они стали незначимы. ОСтальные изменились незначительно, кроме корреляции между INSTRUCT(затраты на одного студента) и PH_D — она почти полностью объяснялась скрытым фактором (стоимостью обучения).

Выведем для каждой пары значение корреляции Пирсона и значение корреляции Спирмена(Частные заведения):

```
library(dplyr)

filtered_data <- data_log |>
  filter(PPIND == "2") |>
   dplyr::select(ADD_FEE,
         BOOK,
         NEW10,
         PH_D,
         SAL_ALL,
         SF_RATIO,
         GRADUAT,
         INSTRUCT)

pairs <- combn(names(filtered_data),
               2,
               simplify = TRUE)
cor_table <- data.frame(Variables = apply(pairs,
                                          2,
                                          paste,
                                          collapse = " vs "))

cor_table$Pearson <- apply(pairs, 2, function(x)
  {
  cor(filtered_data[, x],
      use = "pairwise.complete.obs")[1, 2]
})

cor_table$Spearman <- apply(pairs, 2, function(x)
  {
  cor(filtered_data[, x], method = "spearman",
      use = "pairwise.complete.obs")[1, 2]
})

cor_table
```

```
##              Variables      Pearson     Spearman
## 1       ADD_FEE vs BOOK -0.068162484 -0.103697148
## 2      ADD_FEE vs NEW10  0.045991841  0.042715889
## 3       ADD_FEE vs PH_D  0.163542331  0.176117099
## 4    ADD_FEE vs SAL_ALL  0.304836024  0.260979523
## 5   ADD_FEE vs SF_RATIO  0.109321117  0.137855513
## 6    ADD_FEE vs GRADUAT  0.043950852 -0.031620953
## 7   ADD_FEE vs INSTRUCT  0.036706250  0.042374711
## 8         BOOK vs NEW10  0.023139972  0.073330643
## 9          BOOK vs PH_D  0.016600702  0.003337958
## 10      BOOK vs SAL_ALL  0.225631093  0.171621626
## 11     BOOK vs SF_RATIO  0.117528624  0.014893751
## 12      BOOK vs GRADUAT  0.043477345  0.097243936
## 13     BOOK vs INSTRUCT -0.008033684  0.014486401
## 14        NEW10 vs PH_D  0.658563027  0.652827971
## 15     NEW10 vs SAL_ALL  0.720521089  0.768251838
## 16    NEW10 vs SF_RATIO -0.556140843 -0.542915119
## 17     NEW10 vs GRADUAT  0.665653005  0.768960808
## 18    NEW10 vs INSTRUCT  0.788229954  0.832856533
## 19      PH_D vs SAL_ALL  0.654026120  0.672330434
## 20     PH_D vs SF_RATIO -0.500831354 -0.419786857
## 21      PH_D vs GRADUAT  0.725605157  0.717065358
## 22     PH_D vs INSTRUCT  0.627418213  0.628132215
## 23  SAL_ALL vs SF_RATIO -0.342950192 -0.321437215
## 24   SAL_ALL vs GRADUAT  0.581084216  0.670775972
## 25  SAL_ALL vs INSTRUCT  0.705893400  0.700413266
## 26  SF_RATIO vs GRADUAT -0.434850223 -0.375868130
## 27 SF_RATIO vs INSTRUCT -0.775453187 -0.783402904
## 28  GRADUAT vs INSTRUCT  0.602248723  0.652083151
```

Значения корреляций по частным заведениям в основном совпадают, что говорит о линейных или монотонных зависимостях, где-то есть незначительные отличия, вызванные выбросами.

Выводы

Следующие пары признаков имеют значимые корреляции для частных заведений:

1. ***NEW_10 и PH_D*** (***NEW_10*** — **следствие**, ***PH_D*** — **причина**).
2. ***NEW_10 и SAL_ALL*** (***NEW_10*** — **следствие**, ***SAL_ALL*** — **причина**).
3. ***NEW_10 и GRADUAT*** (***NEW_10*** — **следствие**, ***GRADUAT*** — **причина**).
4. ***NEW_10 и INSTRUCT*** (***NEW_10*** — **следствие**, ***INSTRUCT*** — **причина**).
5. ***PH_D и SAL_ALL*** (***PH_D*** — **причина**, ***SAL_ALL*** — **следствие**)
6. ***PH_D и GRADUAT*** (***PH_D*** — **причина**, ***GRADUAT*** — **следствие**).
7. ***PH_D и INSTRUCT*** — как выяснилось, без влияния стоимости обучения они почти не коррелируют между собой.
8. ***SAL_ALL и GRADUAT*** (***SAL_ALL*** — **причина**, ***GRADUAT*** — **следствие**).
9. ***SAL_ALL и INSTRUCT*** (***SAL_ALL*** — **причина**, ***INSTRUCT*** — **следствие**).
10. ***INSTRUCT и GRADUAT*** (***INSTRUCT*** — **причина**, ***GRADUAT*** — **следствие**).
11. ***INSTRUCT и SF_RATIO*** (***SF_RATIO*** — **причина**, ***INSTRUCT*** — **следствие**).

Можно сказать, что в частных заведениях корреляции интерпретируются более просто. Понятно, что там является причиной, а что — следствием. Это скорее особенность государственных заведений, которые более непохожи дргу на друга в отличие от частных.