Домашнее задание

Загрузим все пакеты для работы.

```
library("pander")
library("knitr")
library("lmtest")
library("psych")
library("memisc")
library("psych")
library("dplyr")
library("knitr")
library("rlms")
library("sandwich")
library("ggplot2")
library("scales")
```

Зададим директорий и подгрузим данные.

```
df <- rlms_read("r22i_os_32.sav")
```

Подготовим данные к анализу и выберем переменные.

```
data <- mutate(df, sex = as.numeric(rh5 == 2),
age = 2013 - rh6,
wage = rj13.2,
educ_l = as.numeric(as.numeric(r_diplom) < 4),
educ_m = as.numeric(as.numeric(r_diplom) == 4),
educ_ms = as.numeric(as.numeric(r_diplom) == 5),
educ_h = as.numeric(as.numeric(r_diplom) == 6),
city = as.numeric(as.numeric(status) < 3),
udovl = as.numeric(as.numeric(rj1.1.1) < 3)) %>%
dplyr::select(sex, age, wage, educ_l, educ_m, educ_ms, educ_h,
city, udovl)
```

Так как задание сделано в учебных целях, облегчим себе жизнь и очистим данные от пропусков.

```
data <- na.omit(data)
```

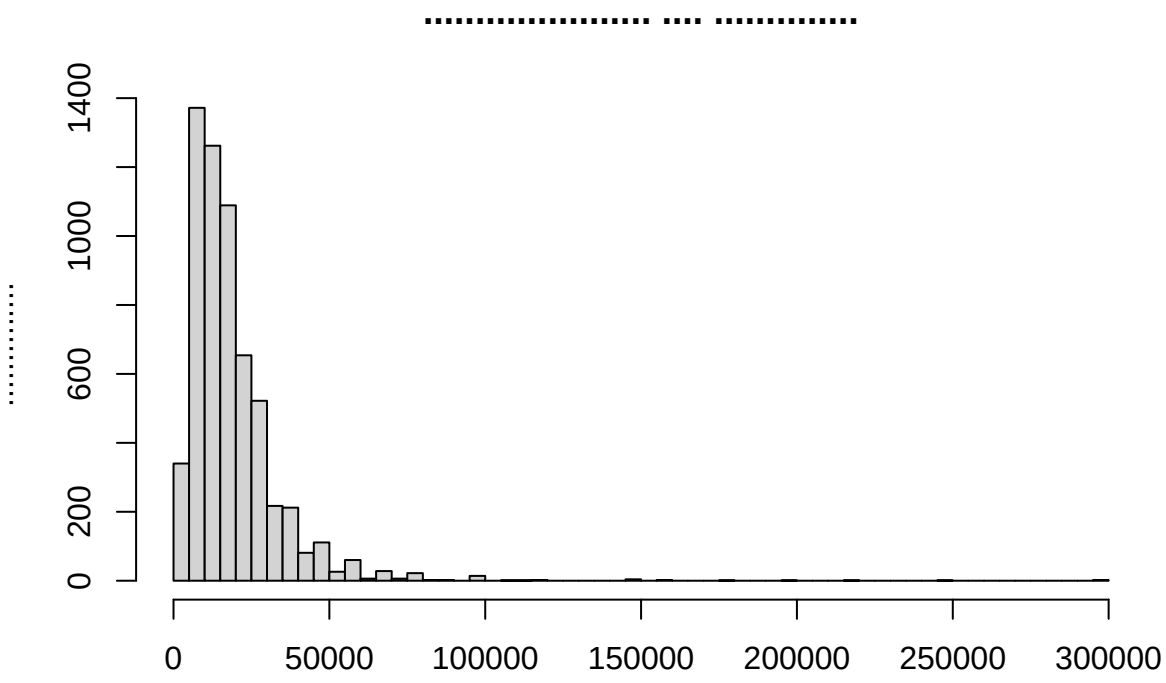
Посмотрим быстро на описание массива. Отберём некоторые характеристики.

```
desc <- describe(data) # from psych package
desc_selected <- as.data.frame(desc[, c(1, 2, 3, 4, 5)]) # not all statistics
pander(desc_selected)
```

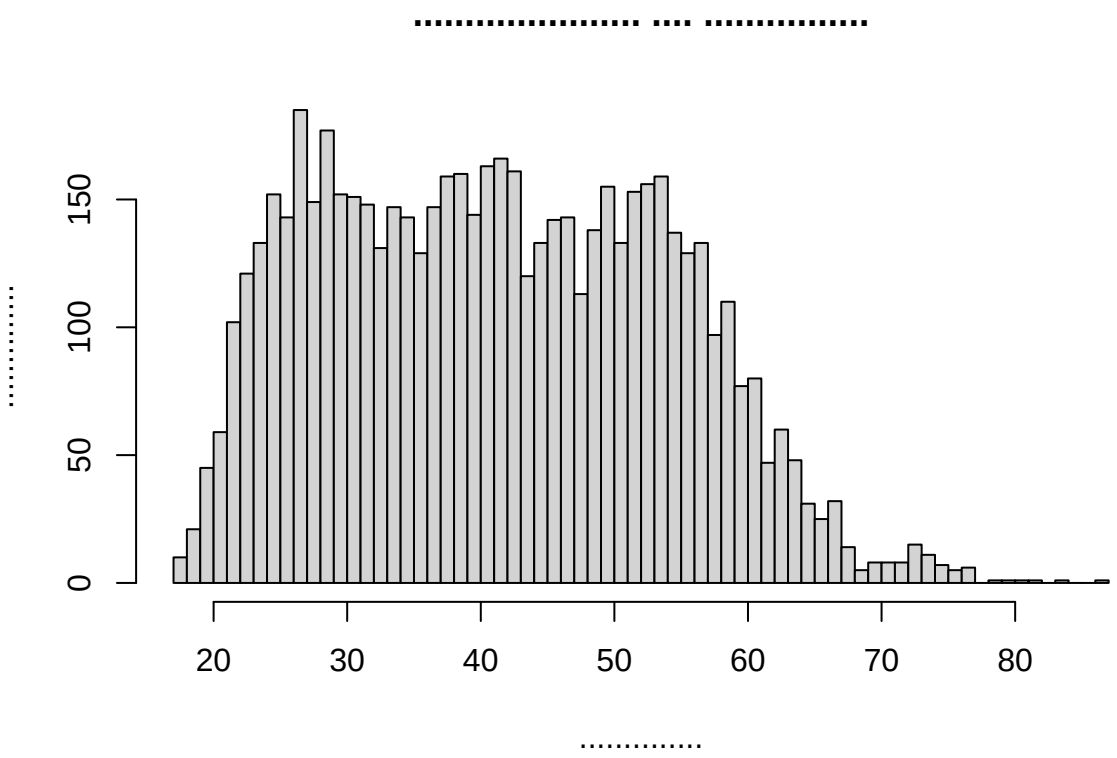
	vars	n	mean	sd	median
sex	1	6042	0.547	0.4978	1
age	2	6042	41.73	12.62	41
wage	3	6042	19806	15966	16000
educ_l	4	6042	0.08755	0.2827	0
educ_m	5	6042	0.3168	0.4653	0
educ_ms	6	6042	0.2719	0.445	0
educ_h	7	6042	0.3237	0.4679	0
city	8	6042	0.7143	0.4518	1
udovl	9	6042	0.6792	0.4668	1

Изобразим пару графиков (вариант 1).

```
hist(data$wage, breaks = 50, main = "Гистограмма по доходам", xlab = "Заработная плата", ylab = "Частота")
```

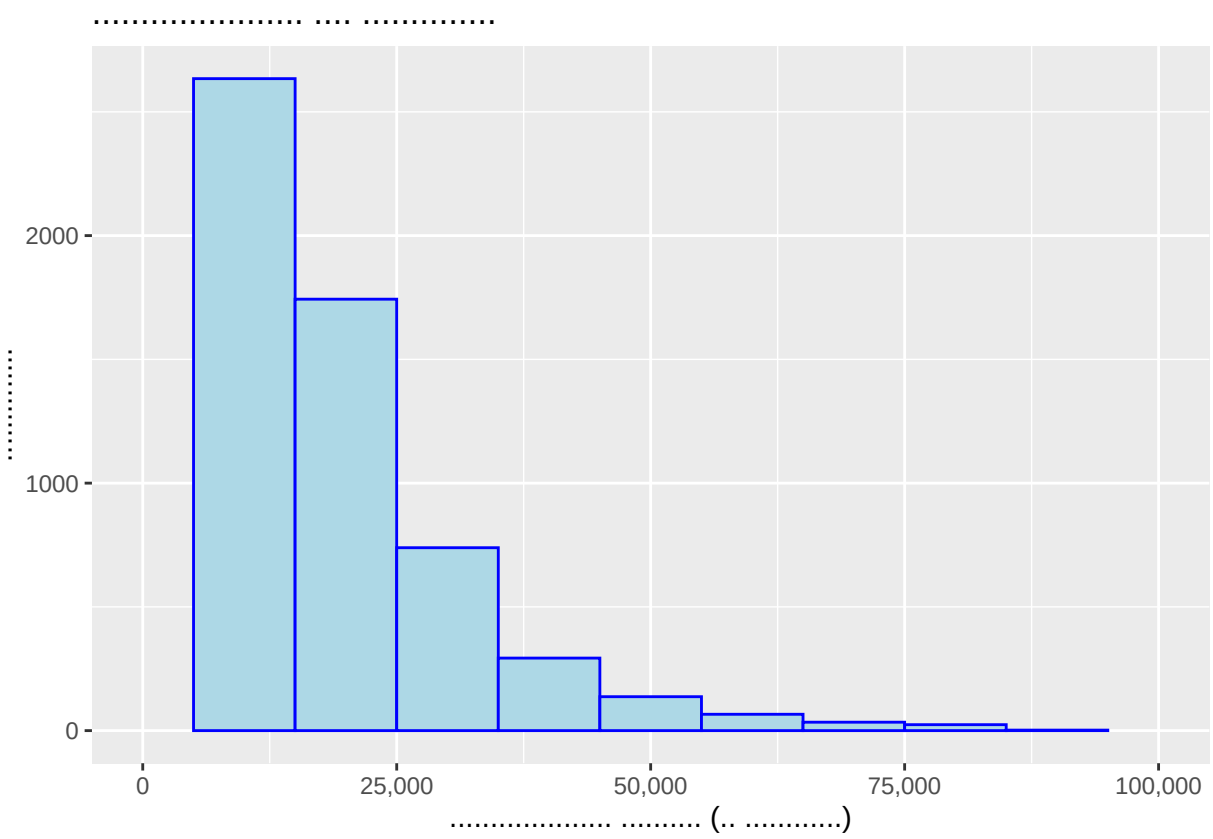



```
hist(data$age, breaks = 50, main = "Гистограмма по возрасту", xlab = "Возраст", ylab = "Частота")
```

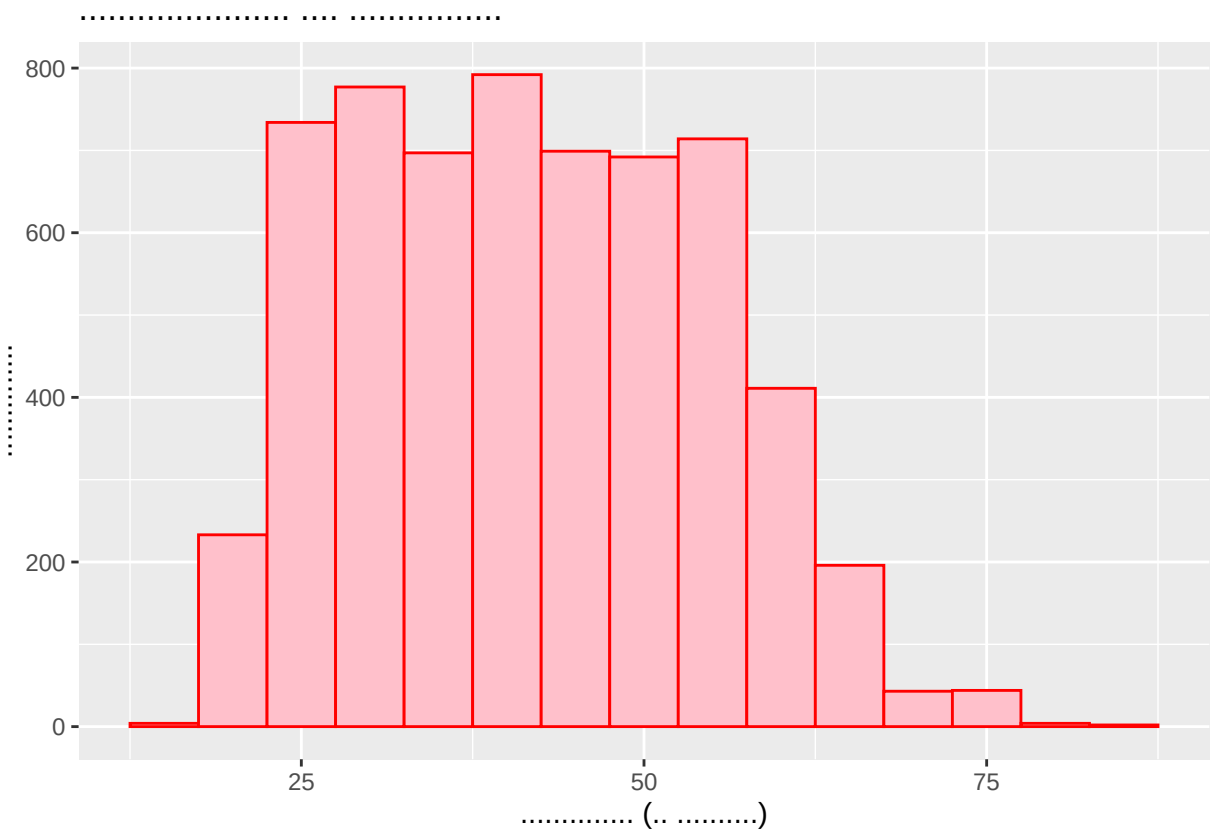


А теперь изобразим пару более красивых графиков (вариант 2).

```
ggplot(data = data, aes(wage)) +
  geom_histogram(binwidth = 10000, fill = "lightblue", color = "blue") +
  labs(title = "Гистограмма по доходам",
       x = "Заработная плата (в рублях)",
       y = "Частота") +
  scale_x_continuous(limits = c(0, 100000), labels = comma)
```




```
ggplot(data = data, aes(age)) +
  geom_histogram(fill = "pink", binwidth = 5, color = "red") +
  labs(title = "Гистограмма по возрасту",
       x = "Возраст (в годах)", y = "Частота")
```



Оценим модель и посмотрим на коэффициенты.

```
model <- lm(wage ~ sex + age + educ_m + educ_ms + educ_h + city + udovl,
            data = data)
```

Но для красоты можно сделать так:

```
coefs <- coeftest(model)
pander(coefs[, 1:4])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16414	947.9	17.32	1.321e-65
sex	-8271	388.2	-21.31	3.433e-97
age	-83.22	15.06	-5.524	3.453e-08
educ_m	2277	724.2	3.144	0.001672
educ_ms	4256	743.4	5.726	1.08e-08
educ_h	10403	734.5	14.16	8.012e-45
city	5024	422.9	11.88	3.453e-32
udovl	3757	408.1	9.207	4.523e-20

Основной вывод: все коэффициенты значимы. Все p-значения очень маленькие, то есть нулевая гипотеза ($H_0 : \beta_i = 0$) при проверке гипотезы о значимости для каждого отдельного коэффициента отвергается. Вспомним два важных момента для интерпретации:

1. Если переменная числовая и коэффициент при ней значим и, скажем, положителен, то увеличение этого показателя на одну единицу будет вызывать увеличение заработной платы на $\hat{\beta}_i$ при этой переменной.

2. Если переменная факторная и коэффициент при ней значим, то $\hat{\beta}_i$ при этой переменной будет означать, что заработная плата будет изменяться на $\hat{\beta}_i$ при переходе от базовой категории этой переменной к той категории, которая у нас отображена перед $\hat{\beta}_i$.

А теперь с робастными ошибками.

```
coefs2 <- coeftest(model, vcov. = vcovHC(model))
pander(coefs2[, 1:4])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16414	725	22.64	5.457e-109
sex	-8271	423.7	-19.52	2.363e-82
age	-83.22	13.71	-6.07	1.36e-09
educ_m	2277	466.5	4.882	1.079e-06
educ_ms	4256	508.5	8.37	7.073e-17
educ_h	10403	575	18.09	2.763e-71
city	5024	349.6	14.37	4.597e-46
udovl	3757	357.3	10.52	1.198e-25

Выводы не меняются. Все коэффициенты значимы, но стадартные отклонения теперь изменились.