# **Задания**

## **Задача 1:**

1. Импортируйте набор данных с именем «AppleStore» в R.

2. Создайте новый фрейм данных, который будет содержать все переменные, кроме «id» и «currency». Назовите этот новый фрейм данных как df2.

3. Изучите структуру нового набора данных (df2) и предоставьте анализ общей информации об этом наборе данных (что такое единица наблюдения, сколько переменных и наблюдений, какие переменные находятся в наборе данных и какие они типы) ,

4. Анализ суммарной статистики переменных «цена», «user\_rating» и «lang\_num», «size\_bytes».

5. Какое приложение имеет наибольшее количество языков?

6. Определите квантили переменных «цена», «user\_rating» и «lang\_num».

7. Для всех количественных переменных рассчитать коэффициенты эксцесса и асимметрии и коэффициент вариации. Сделать выводы.

8. Для всех количественных переменных построить Boxplot. Обязательно сделать подписи на графике. Сделать выводы о наличии выбросов.

9. Для всех качественных данных построить круговые диаграммы.

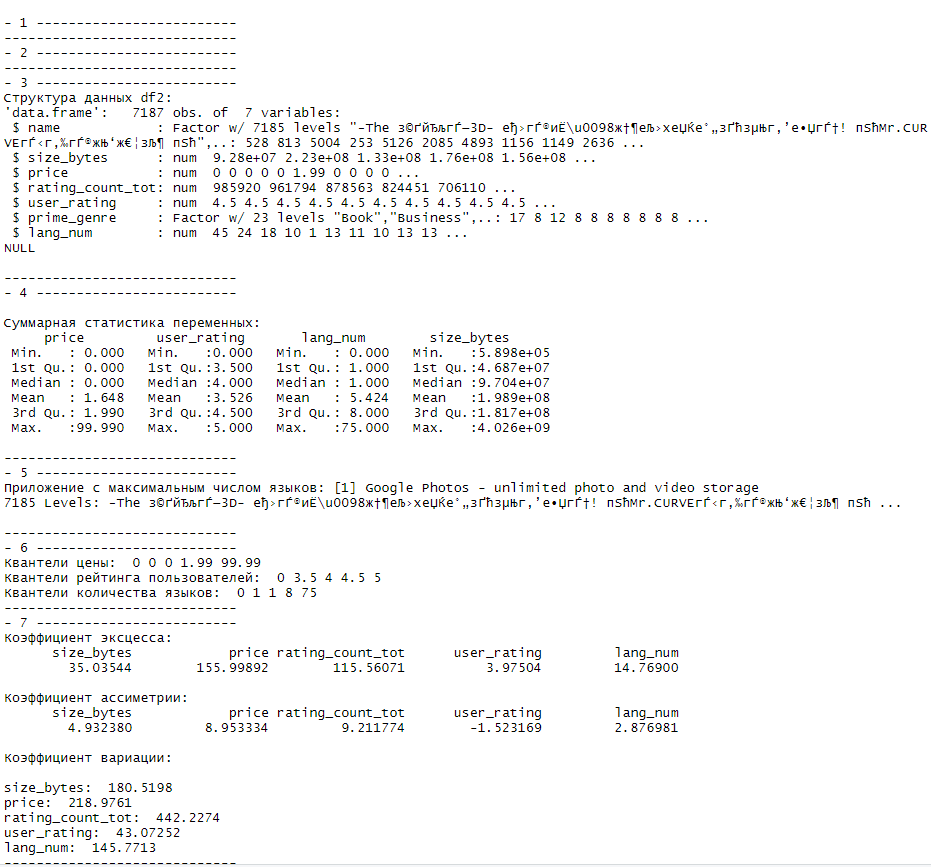
10. Для всех количественных переменных построить гистограммы с плотностью нормального распределения. Сделать выводы.

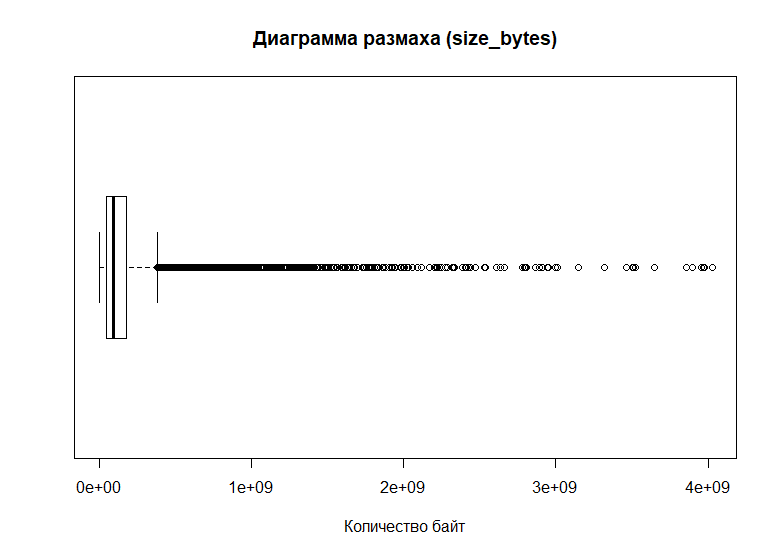
11. Какой жанр наиболее распространен? Подсказка: чтобы выяснить это, преобразуйте переменную prime\_genre в множитель и просмотрите ее сводную статистику.

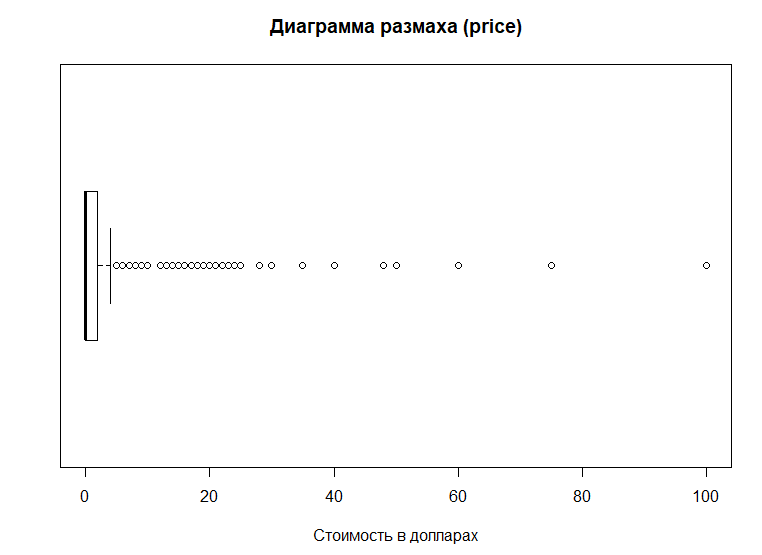
12. Создайте новый фрейм данных из существующего фрейма данных df2, чтобы новый фрейм данных содержал только приложения, соответствующие наиболее распространенному жанру. Рассчитайте сводную статистику переменных, которые вы проанализировали в (4) для нового фрейма данных, и сравните их с результатами в (4). Что вы можете сказать о цене, рейтинге пользователей и количестве языков приложений, относящихся к наиболее распространенному жанру, по сравнению со всей выборкой?

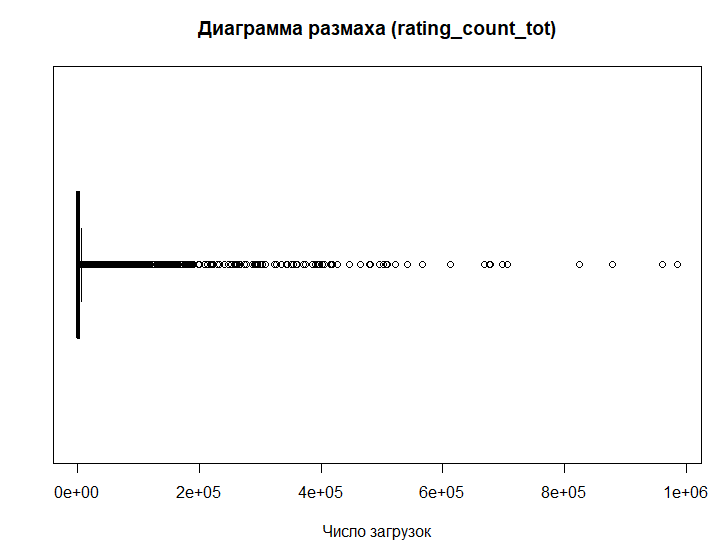
13. Проверьте, используя критерий Колмогорова-Смирнова, гипотезу о нормальности распределения показателя «цена»  по группам. Проделайте то же самое, используя критерий Шапиро-Уилка. Сделайте выводы.

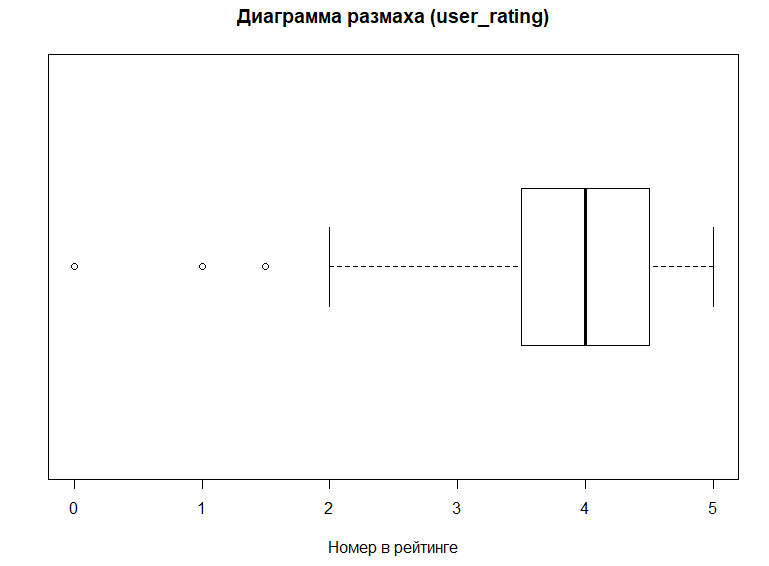
### **Работа программы:**

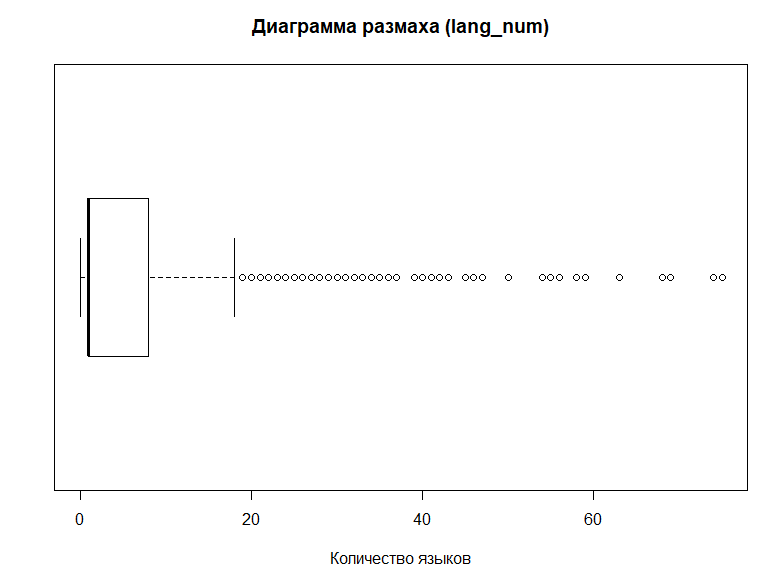


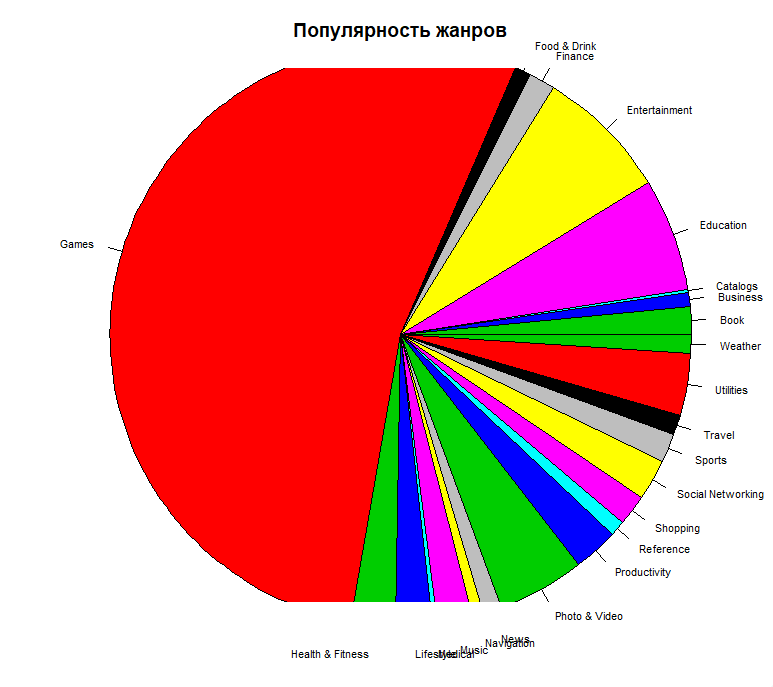


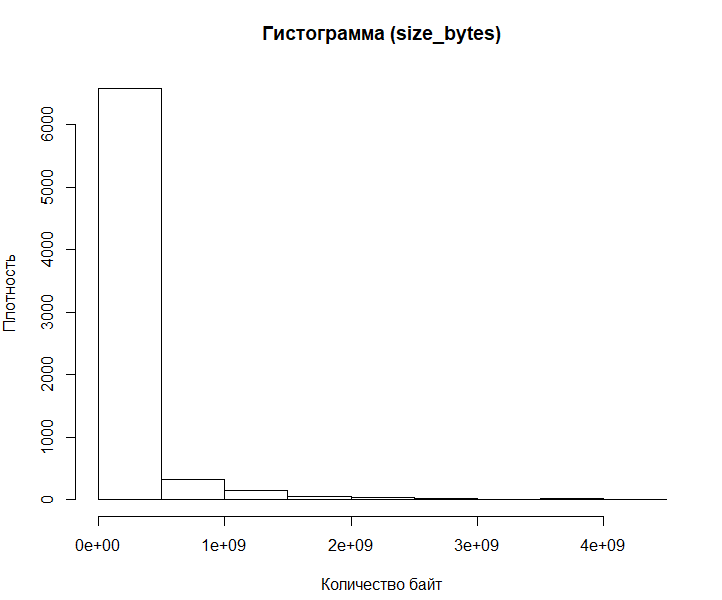


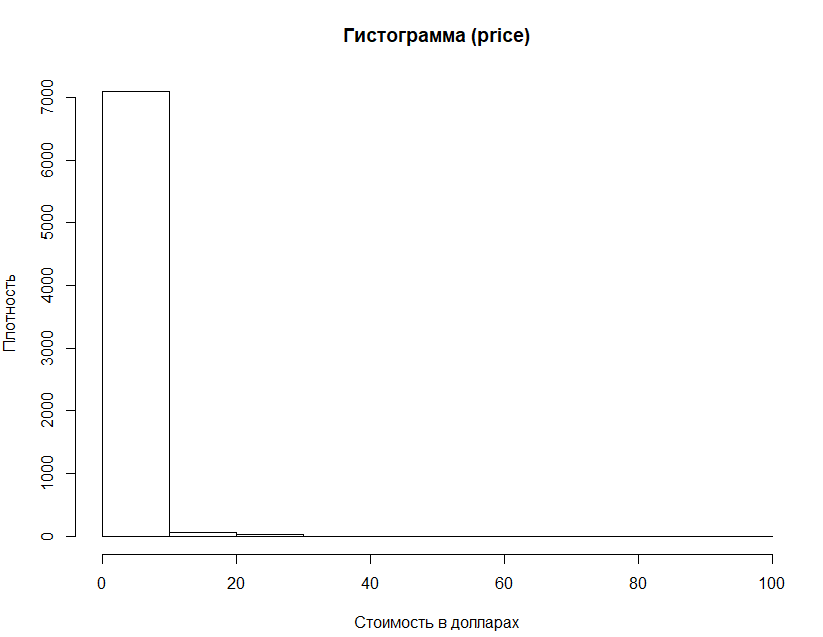


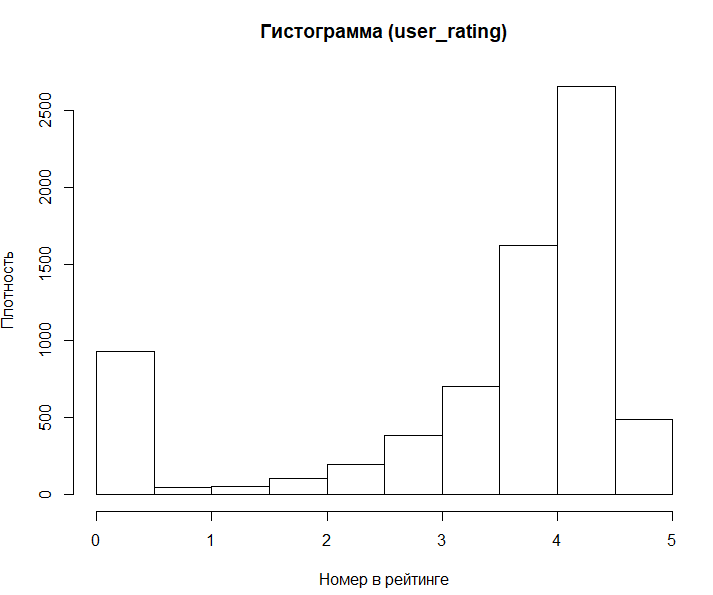


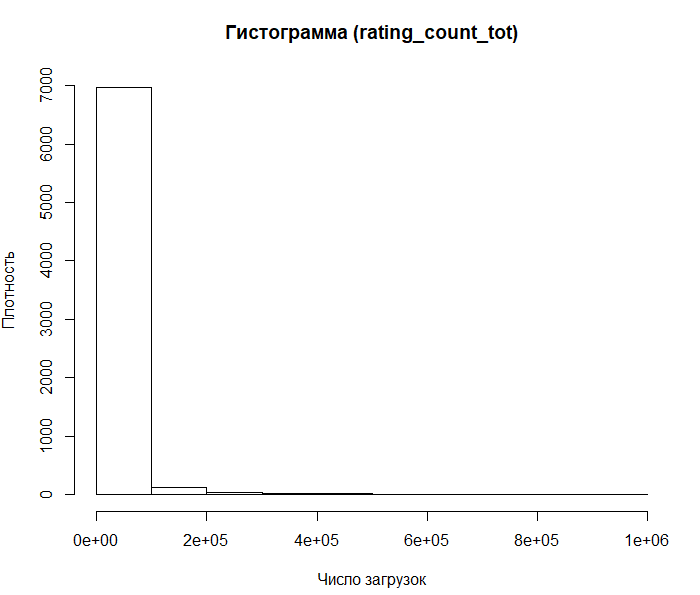


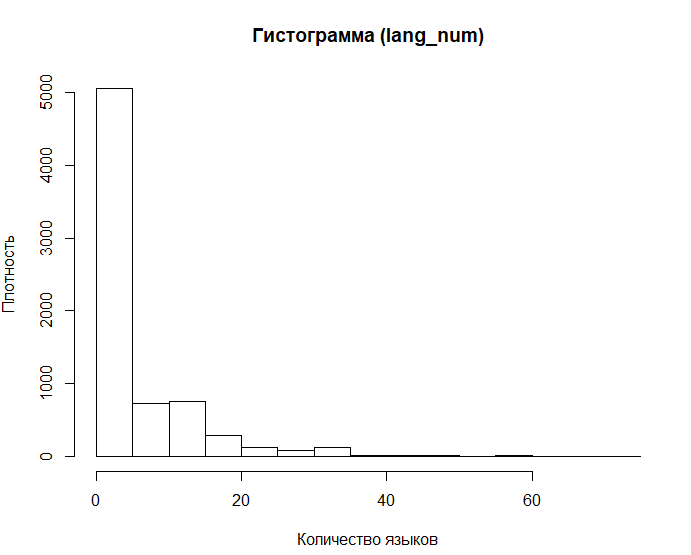


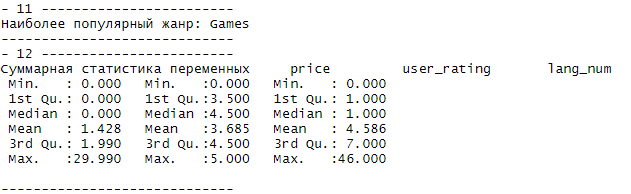


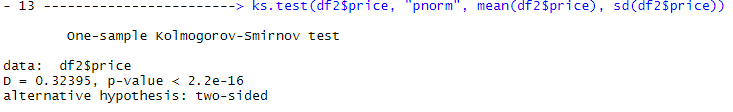


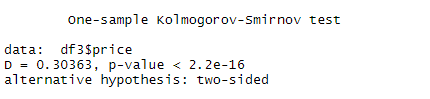












### **Листинг:**

#install.packages(c("xlsx", "dplyr", "moments", "psych", "ggplot2", "DescTools"))

library(ggplot2)

library(DescTools)

library(xlsx)

library(dplyr)

library(moments)

library(psych)

Task <- function()

{

cat("\n- 1 -------------------------")

df <- read.xlsx("AppleStore.xlsx", 1,encoding = "UTF-8")

cat("\n-----------------------------")

cat("\n- 2 -------------------------")

df2 <- subset(df[which(colnames(df) != "id" & colnames(df) != "currency")])

cat("\n-----------------------------")

cat("\n- 3 -------------------------")

cat("\nСтруктура данных df2:\n")

print(str(df2))

cat("\n-----------------------------")

cat("\n- 4 -------------------------\n")

cat("\nСуммарная статистика переменных:\n")

print(summary(subset(df, select = c(price, user\_rating, lang\_num, size\_bytes))))

cat("\n-----------------------------")

cat("\n- 5 -------------------------")

cat("\nПриложение с максимальным числом языков: ")

print(df$name[df$lang\_num==max(df$lang\_num)])

cat("\n-----------------------------")

cat("\n- 6 -------------------------")

cat("\nКвантели цены: ", quantile(df$price))

cat("\nКвантели рейтинга пользователей: ", quantile(df$user\_rating))

cat("\nКвантели количества языков: ", quantile(df$lang\_num))

cat("\n-----------------------------")

cat("\n- 7 -------------------------")

cat("\nКоэффициент эксцесса: \n")

print(kurtosis(df2[, sapply(df2, is.numeric)], na.rm = TRUE))

cat("\nКоэффициент ассиметрии: \n")

print(skewness(df2[, sapply(df2, is.numeric)], na.rm = TRUE))

cat("\nКоэффициент вариации: \n")

cat("\nsize\_bytes: ", sd(df2$size\_bytes) / mean(df2$size\_bytes) \* 100)

cat("\nprice: ", sd(df2$price) / mean(df2$price) \* 100)

cat("\nrating\_count\_tot: ", sd(df2$rating\_count\_tot) / mean(df2$rating\_count\_tot) \* 100)

cat("\nuser\_rating: ", sd(df2$user\_rating) / mean(df2$user\_rating) \* 100)

cat("\nlang\_num: ", sd(df2$lang\_num) / mean(df2$lang\_num) \* 100)

cat("\n-----------------------------")

cat("\n- 8 -------------------------")

boxplot((df2[, sapply(df2, is.numeric)])$size\_bytes, xlab = "Количество байт", main = "Диаграмма размаха (size\_bytes)", horizontal = TRUE)

boxplot((df2[, sapply(df2, is.numeric)])$price, xlab = "Стоимость в долларах", main = "Диаграмма размаха (price)", horizontal = TRUE)

boxplot((df2[, sapply(df2, is.numeric)])$rating\_count\_tot, xlab = "Число загрузок", main = "Диаграмма размаха (rating\_count\_tot)", horizontal = TRUE)

boxplot((df2[, sapply(df2, is.numeric)])$user\_rating, xlab = "Номер в рейтинге", main = "Диаграмма размаха (user\_rating)", horizontal = TRUE)

boxplot((df2[, sapply(df2, is.numeric)])$lang\_num, xlab = "Количество языков", main = "Диаграмма размаха (lang\_num)", horizontal = TRUE)

cat("\n-----------------------------")

cat("\n- 9 -------------------------")

pie(table(df2$prime\_genre), cex = 0.7, radius = 2, main = "Популярность жанров", col = c(3:16))

cat("\n-----------------------------")

cat("\n- 10 ------------------------")

hist((df2[, sapply(df2, is.numeric)])$size\_bytes, xlab = "Количество байт", ylab = "Плотность", main = "Гистограмма (size\_bytes)", horizontal = TRUE)

curve(dnorm(x, mean(df$size\_bytes), sd = sd(df$size\_bytes)), add = TRUE)

hist((df2[, sapply(df2, is.numeric)])$price, xlab = "Стоимость в долларах", ylab = "Плотность", main = "Гистограмма (price)", horizontal = TRUE)

curve(dnorm(x, mean(df$price), sd = sd(df$price)), add = TRUE)

hist((df2[, sapply(df2, is.numeric)])$rating\_count\_tot, xlab = "Число загрузок", ylab = "Плотность", main = "Гистограмма (rating\_count\_tot)", horizontal = TRUE)

curve(dnorm(x, mean(df$rating\_count\_tot), sd = sd(df$rating\_count\_tot)), add = TRUE)

hist((df2[, sapply(df2, is.numeric)])$user\_rating, xlab = "Номер в рейтинге", ylab = "Плотность", main = "Гистограмма (user\_rating)", horizontal = TRUE)

curve(dnorm(x, mean(df$user\_rating), sd = sd(df$user\_rating)), add = TRUE)

hist((df2[, sapply(df2, is.numeric)])$lang\_num, xlab = "Количество языков", ylab = "Плотность ", main = "Гистограмма (lang\_num)", horizontal = TRUE)

curve(dnorm(x, mean(df$lang\_num), sd = sd(df$lang\_num)), add = TRUE)

cat("\n-----------------------------")

cat("\n- 11 ------------------------")

df %>% group\_by(df$prime\_genre) %>% summarise(count = n()) %>% arrange(count) %>% tail()

cat("\nНаиболее популярный жанр: Games")

cat("\n-----------------------------")

cat("\n- 12 ------------------------")

df3 = subset(df2, prime\_genre == "Games")

cat("\nСуммарная статистика переменных")

print(summary(subset(df3, select = c(price, user\_rating, lang\_num))))

cat("\n-----------------------------")

cat("\n- 13 ------------------------")

ks.test(df2$price, "pnorm", mean(df2$price), sd(df2$price))

ks.test(df3$price, "pnorm", mean(df3$price), sd(df3$price))

shapiro.test(df[df$state == TRUE]$price)

cat("\n-----------------------------")

}

## **Задача 2:**

Все графики строятся с помощью библиотеки ggplot2.

1. Загрузите файл demography.csv. В нём содержатся данные по населению Белгородской и Калужской областей за 2016 год (источник — Росстат).

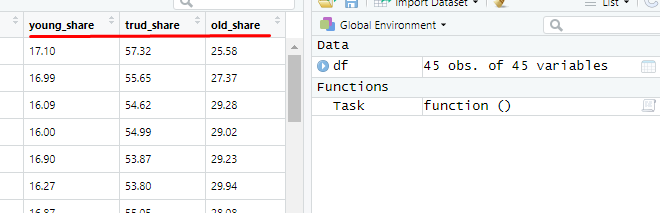
df <- read.csv("https://raw.githubusercontent.com/allatambov/R-programming-3/master/seminars/sem8-09-02/demography.csv", encoding = "UTF-8")

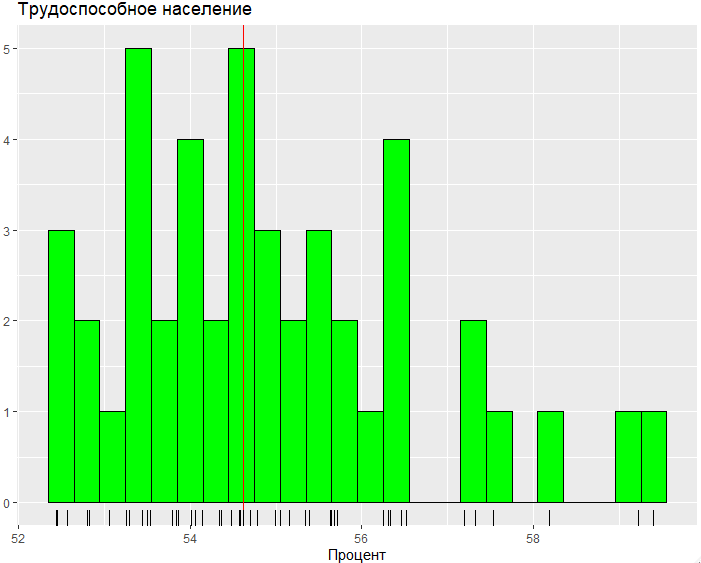
**Переменные:**

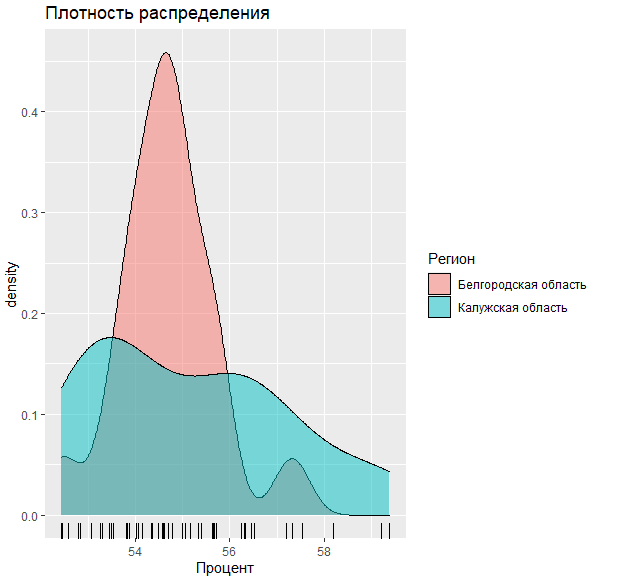
* + region: название региона;
  + district: название района;
  + empl\_total: численность занятого населения;
  + A-O: занятость по отраслям (как на сайте Росстата: сельское хозяйство, );
  + popul\_total: численность населения;
  + urban\_total: численность городского населения;
  + rural\_total: численность сельского населения;
  + wa\_total: численность трудоспособного населения;
  + wa\_female: численность трудоспособного населения (женский пол);
  + wa\_male: численность трудоспособного населения (мужской пол);
  + ret\_total: численность пенсионеров;
  + ret\_female: численность пенсионеров (женский пол);
  + ret\_male: численность пенсионеров (мужской пол);
  + young\_total: численность населения, моложе трудоспособного возраста;
  + young\_female: численность населения, моложе трудоспособного возраста (женский пол);
  + young\_male: численность населения, моложе трудоспособного возраста (мужской пол);
  + X18\_19 - X70\_plus: численность населения по возрастным группам.

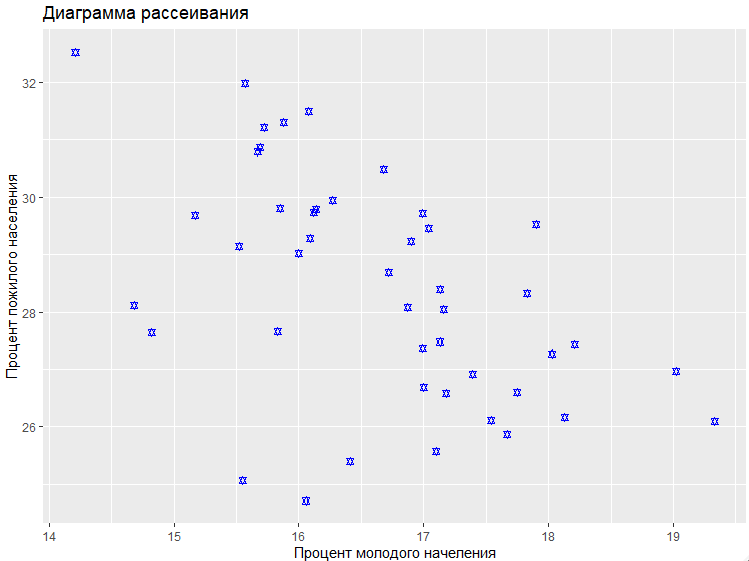
1. Создайте переменную young\_share — процент населения возраста, моложе трудоспособного. Создайте переменную trud\_share — процент населения трудоспособного возраста и old\_share — процент населения возраста, старше трудоспособного.
2. Постройте гистограмму для доли трудоспособного населения в процентах. Измените цвет гистограммы, добавьте *rugs*. Добавьте вертикальную линию, которая отчерчивает медианное значение доли трудоспособного населения в процентах.
3. Постройте сглаженные графики плотности распределения для доли трудоспособного населения в процентах по регионам (два графика в одной плоскости). Настройте цвета и прозрачность заливки. По графикам плотности определите, имеет ли смысл для визуализации распределения доли трудоспособного населения строить скрипичные диаграммы (*violin plot*). Если да, постройте их (так же по группам). Если нет, постройте ящики с усами.
4. Постройте диаграмму рассеяния для переменных young\_share и old\_share. Можно ли сказать, что чем больше процент молодого населения (моложе трудоспособного населения), тем меньше процент пожилых людей (старше трудоспособного возраста)? Поменяйте цвет и тип маркера для точек.
5. Создайте переменную male\_share — доля мужского населения в районе/городе (в процентах). Создайте переменную male, которая принимает значение 1, если доля мужчин в муниципальном районе/городе больше доли женщин, и значение 0 — во всех остальных случаях.
6. Постройте пузырьковую диаграмму (*bubble plot*) для переменных young\_share и old\_share, учитывая информацию о доле мужчин в районе и о том, преобладают ли мужчины в районе или нет.
7. Постройте столбиковую диаграмму (*bar plot*), которая показывала бы, сколько в базе данных районов Белгородской области, а сколько — Калужской.

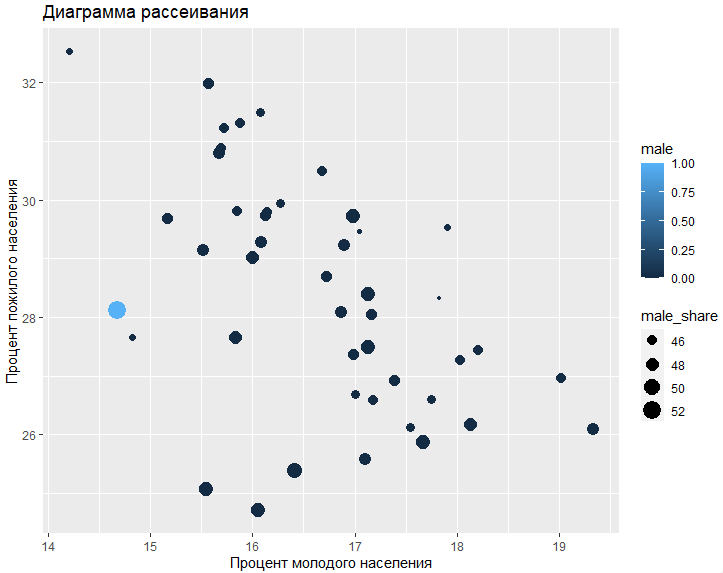
### **Работа программы:**

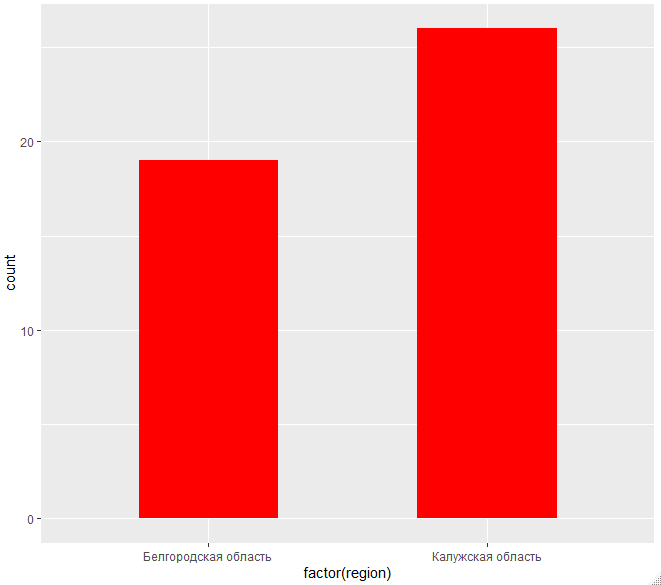












### **Листинг:**

#install.packages(c("xlsx", "dplyr", "moments", "psych", "ggplot2", "DescTools"))

library(ggplot2)

library(DescTools)

library(xlsx)

library(dplyr)

library(moments)

library(psych)

Task <- function()

{

cat("\n- 1 -------------------------")

df <- read.csv("https://raw.githubusercontent.com/allatambov/R-programming-3/master/seminars/sem8-09-02/demography.csv", encoding = "UTF-8")

cat("\n-----------------------------")

cat("\n- 2 -------------------------")

df$young\_share <- as.double(round(df$young\_total/df$popul\_total\*100, 2))

df$trud\_share <- as.double(round(df$wa\_total/df$popul\_total\*100, 2))

df$old\_share <- as.double(round(df$ret\_total/df$popul\_total\*100, 2))

cat("\n-----------------------------")

cat("\n- 3 -------------------------")

ggplot(data = df, aes(x = trud\_share)) +

geom\_histogram(binwidth = 0.3, fill = "green", color = "black") +

labs(x = "Процент", y = "Частота", title = "Трудоспособное население") +

geom\_vline(xintercept = median(df$trud\_share),color = "red") +

geom\_rug()

cat("\n-----------------------------")

cat("\n- 4 -------------------------")

print(ggplot(data = df, aes(x = trud\_share, group = region, fill = region)) +

geom\_density(alpha = 0.5) +

geom\_rug() +

labs(x = "Процент", title = "Плотность распределения") +

scale\_fill\_manual(values = c("black","blue")) +

scale\_fill\_discrete(name = "Регион"))

print(ggplot(data = df, aes(x = "", y = trud\_share, group = region, fill = region)) +

geom\_violin() +

geom\_rug() +

labs(x = "Плотность", y = "Процент") +

scale\_fill\_manual(values = c("red","green")) +

scale\_fill\_discrete(name = "Регион"))

cat("\n-----------------------------")

cat("\n- 5 -------------------------")

print(ggplot(data = df, aes(x = young\_share, y = old\_share)) +

geom\_point(color = "blue",pch = 11) +

labs(title = "Диаграмма рассеивания", x = "Процент молодого начеления", y = "Процент пожилого населения"))

cat("\n-----------------------------")

cat("\n- 6 -------------------------")

df$male\_share <- as.double(round((df$wa\_male + df$ret\_male + df$young\_male)/df$popul\_total \* 100, 2))

df$male <- as.integer(df$male\_share > 50)

cat("\n-----------------------------")

cat("\n- 7 -------------------------")

print(ggplot(data = df, aes(x = young\_share, y = old\_share)) +

geom\_point(aes(size = male\_share, color = male)) +

labs(title = "Диаграмма рассеивания", x = "Процент молодого населения", y = "Процент пожилого населения"))

cat("\n-----------------------------")

cat("\n- 8 -------------------------")

ggplot(df, aes(x = factor(region))) +

geom\_bar(stat = "count", width = 0.5, fill = "red")

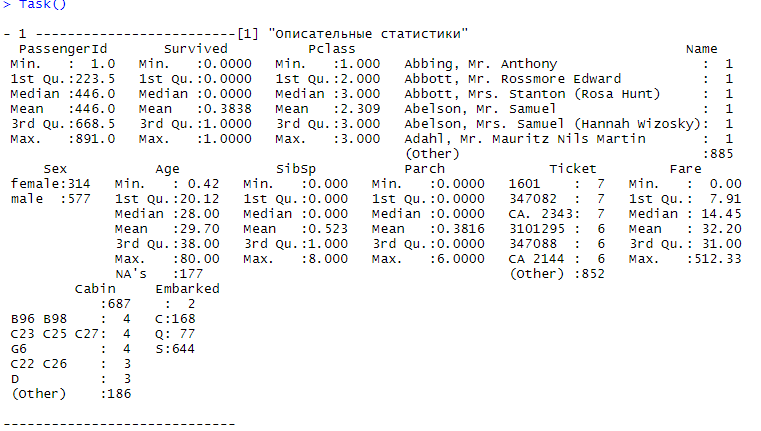
cat("\n-----------------------------")

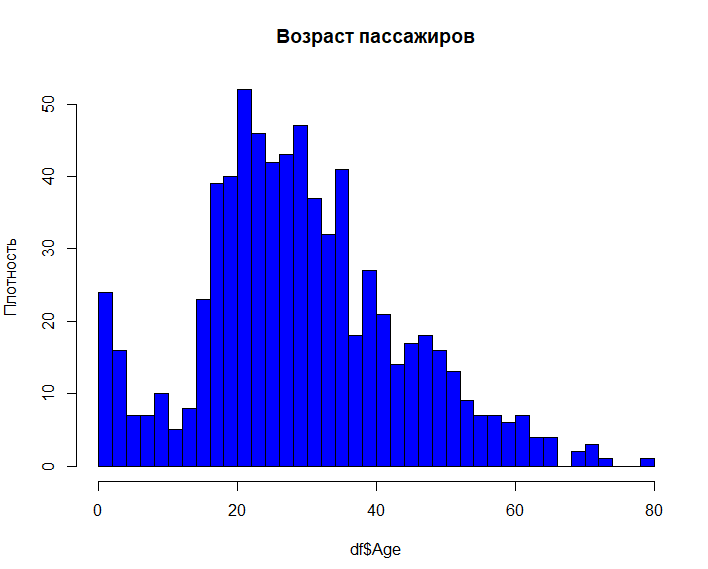
}

## **Задача 3:**

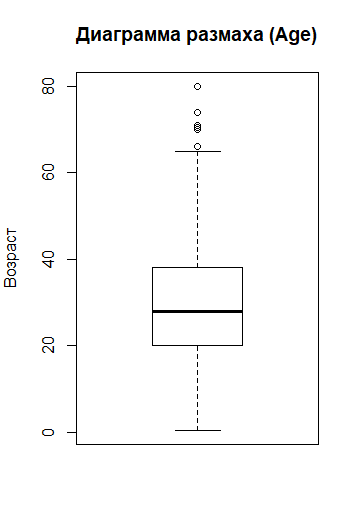
1. Загрузите данные из файла Titanic.csv, с которым вы уже работали.
2. 1. Выведите описательные статистики для всех переменных в таблице. Выберите два показателя (один количественный, один качественный) и проинтерпретируйте все выведенные по ним значения статистик.
3. 2. Постройте для показателя Age гистограмму, поменяйте ее цвет, добавьте название (заголовок) графика. Напишите, людей какого возраста в базе больше и меньше всего.
4. 3. Постройте для показателя Age ящик с усами. Напишите, есть ли в выборке нетипичные значения (выбросы), и если есть, то сколько.
5. 4. Постройте 95%-ный доверительный интервал для доли женщин среди выживших. Постройте 95%-ный доверительный интервал для доли мужчин среди выживших. Проинтерпретируйте полученные интервалы. Какой из доверительных интервалов длинее? Пересекаются ли доверительные интервалы?

### **Работа программы:**









### **Листинг:**

#install.packages(c("xlsx", "dplyr", "moments", "psych", "ggplot2", "DescTools"))

#library(ggplot2)

library(DescTools)

#library(xlsx)

#library(dplyr)

#library(moments)

#library(psych)

Task <- function()

{

cat("\n- 1 -------------------------")

df <- read.csv("https://raw.githubusercontent.com/agconti/kaggle-titanic/master/data/train.csv")

print("Описательные статистики")

print(summary(df))

#всего на Титанике было 314 женщин и 577 мужчин

#возраст 177 пассажиров неизвестен

#мин. возраст = 0,42, макс - 80.

#средний возраст = 29,70

#медиана = 28

cat("\n-----------------------------")

cat("\n- 2 -------------------------")

hist(df$Age, col = 'blue', breaks = 40, main = 'Возраст пассажиров', ylab = "Плотность")

#больше людей с возрастом 23-24

#меньше людей с возрастом 78-80

cat("\n-----------------------------")

cat("\n- 3 -------------------------")

cat("\nВыбросов: ",length(boxplot(df$Age, ylab = "Возраст", main = "Диаграмма размаха (Age)")$out))

cat("\n-----------------------------")

cat("\n- 4 -------------------------")

BinomCI(sum(df$Sex == "female" & df$Survived == 1), sum(df$Survived == 1), conf.level = 0.95)

# С 95%-ной уверенностью, доля женщин среди выживших в интервале от 0.63 до 0.73

BinomCI(sum(df$Sex == "male" & df$Survived == 1), sum(df$Survived == 1), conf.level = 0.95)

# С 95%-ной уверенностью, доля мужчин среди выживших в интервале от 0.27 до 0.37

cat("\n-----------------------------")

}