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:

```
- 1 -------------------------
-----------------------------
- 2 -------------------------
-----------------------------
- 3 -------------------------
Структура данных df2:
'data.frame':   7187 obs. of  7 variables:
 $ name            : Factor w/ 7185 levels "-The з©ґйЪьґґ–3D- еђ›ґґ®иЁ\u0098ж†¶ељ›хеџќе'„зґћзµћьг,'е•џґґ†! nSћMr.CUR
VEґґ‹г,%ґґ®жњ'ж€¦зљ¶ nSћ",..: 528 813 5004 253 5126 2085 4893 1156 1149 2636 ...
 $ size_bytes      : num  9.28e+07 2.23e+08 1.33e+08 1.76e+08 1.56e+08 ...
 $ price           : num  0 0 0 0 0 1.99 0 0 0 0 ...
 $ rating_count_tot: num  985920 961794 878563 824451 706110 ...
 $ user_rating     : num  4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 ...
 $ prime_genre     : Factor w/ 23 levels "Book","Business",..: 17 8 12 8 8 8 8 8 8 8 ...
 $ lang_num        : num  45 24 18 10 1 13 11 10 13 13 ...
NULL

-----------------------------
- 4 -------------------------

Суммарная статистика переменных:
     price          user_rating       lang_num         size_bytes
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :5.898e+05
 1st Qu.: 0.000   1st Qu.:3.500   1st Qu.: 1.000   1st Qu.:4.687e+07
 Median : 0.000   Median :4.000   Median : 1.000   Median :9.704e+07
 Mean   : 1.648   Mean   :3.526   Mean   : 5.424   Mean   :1.989e+08
 3rd Qu.: 1.990   3rd Qu.:4.500   3rd Qu.: 8.000   3rd Qu.:1.817e+08
 Max.   :99.990   Max.   :5.000   Max.   :75.000   Max.   :4.026e+09

-----------------------------
- 5 -------------------------
Приложение с максимальным числом языков: [1] Google Photos - unlimited photo and video storage
7185 Levels: -The з©ґйЪьґґ–3D- еђ›ґґ®иЁ\u0098ж†¶ељ›хеџќе'„зґћзµћьг,'е•џґґ†! nSћMr.CURVEґґ‹г,%ґґ®жњ'ж€¦зљ¶ nSћ ...

-----------------------------
- 6 -------------------------
Квантели цены:  0 0 0 1.99 99.99
Квантели рейтинга пользователей:  0 3.5 4 4.5 5
Квантели количества языков:  0 1 1 8 75
-----------------------------
- 7 -------------------------
Коэффициент эксцесса:
      size_bytes            price rating_count_tot      user_rating         lang_num
        35.03544        155.99892        115.56071          3.97504         14.76900

Коэффициент ассиметрии:
      size_bytes            price rating_count_tot      user_rating         lang_num
        4.932380         8.953334         9.211774        -1.523169         2.876981

Коэффициент вариации:

size_bytes:  180.5198
price:  218.9761
rating_count_tot:  442.2274
user_rating:  43.07252
lang_num:  145.7713
-----------------------------
```

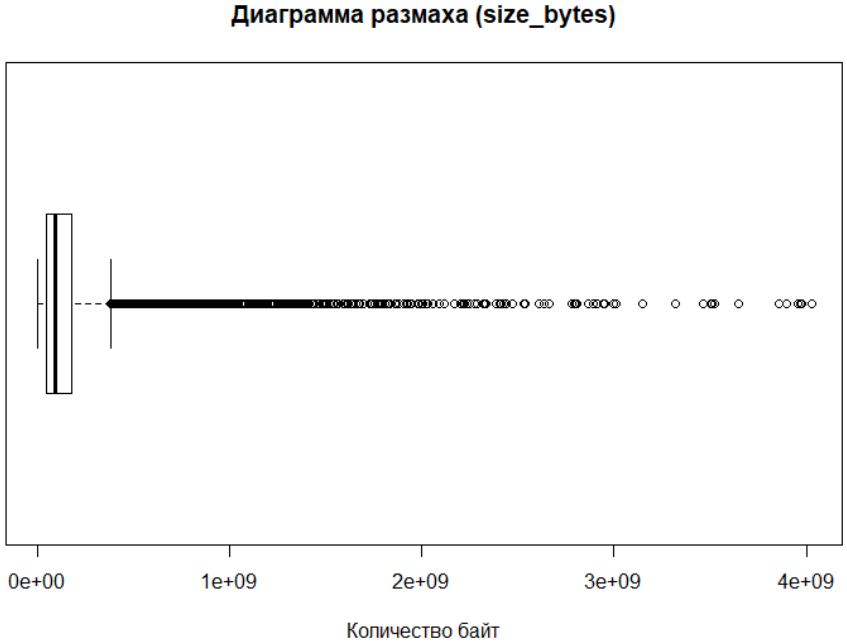


Диаграмма размаха (price)

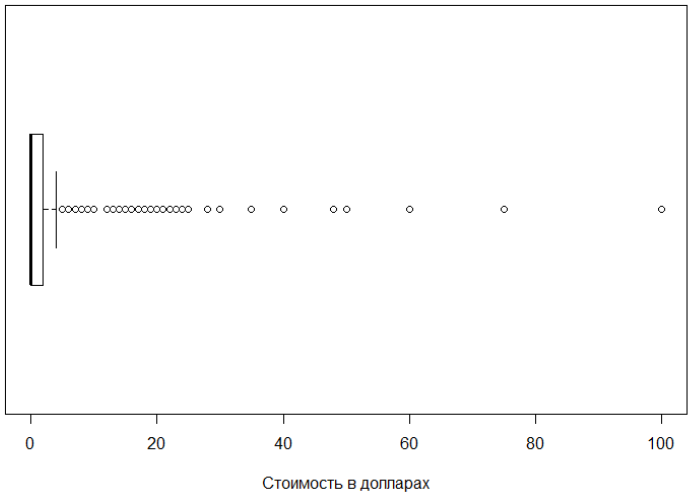


Диаграмма размаха (rating_count_tot)

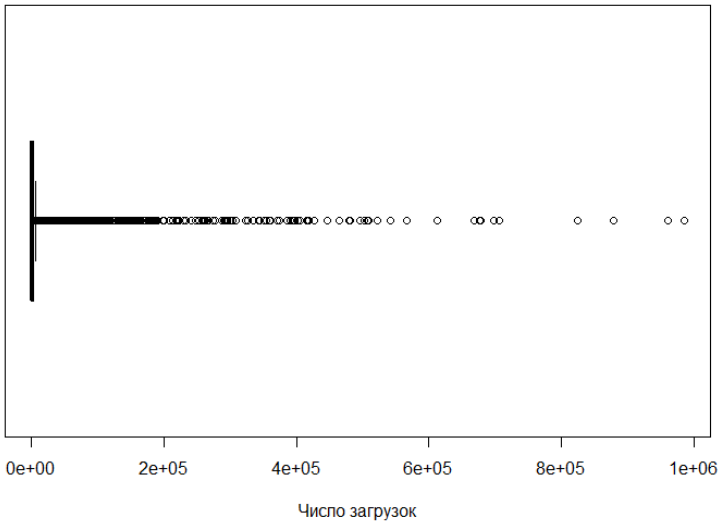


Диаграмма размаха (user_rating)

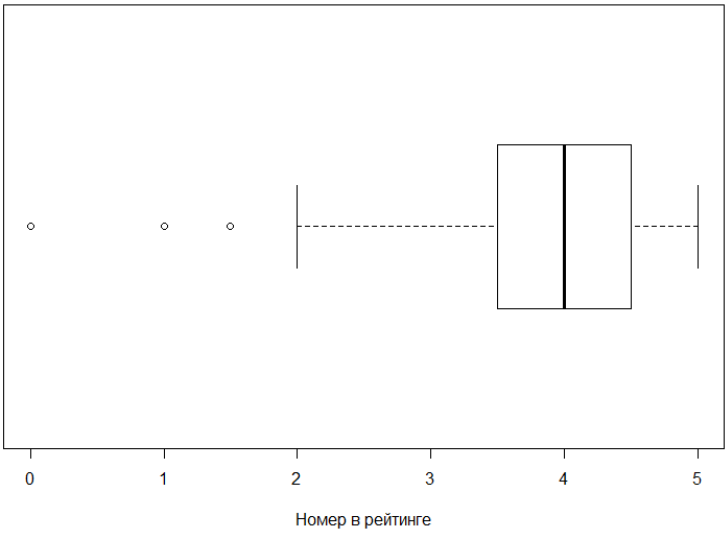
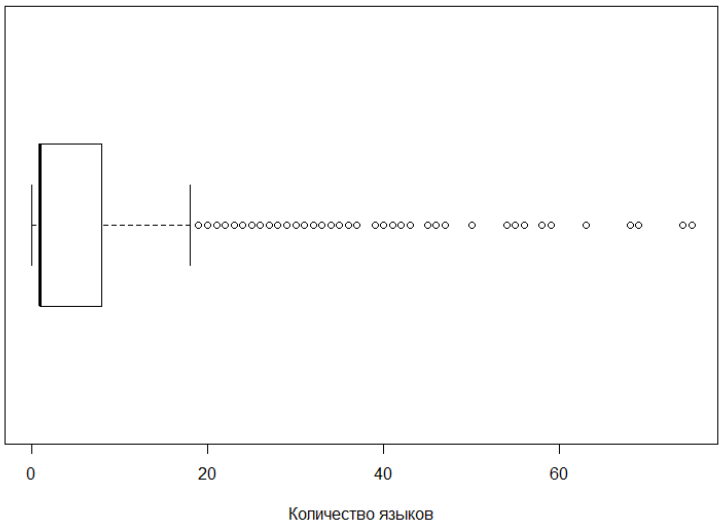
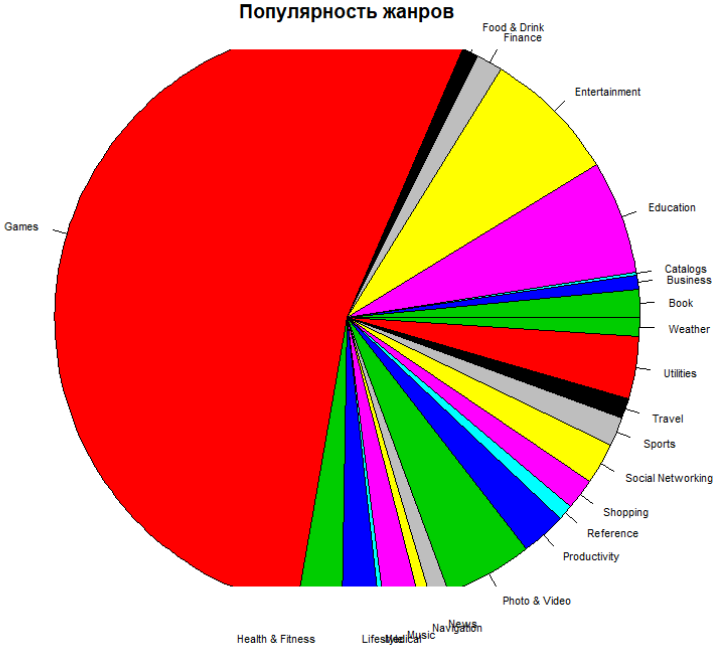


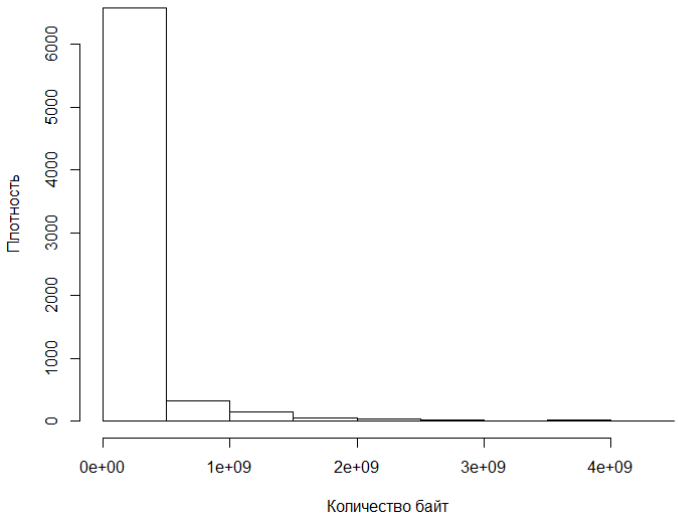
Диаграмма размаха (lang_num)

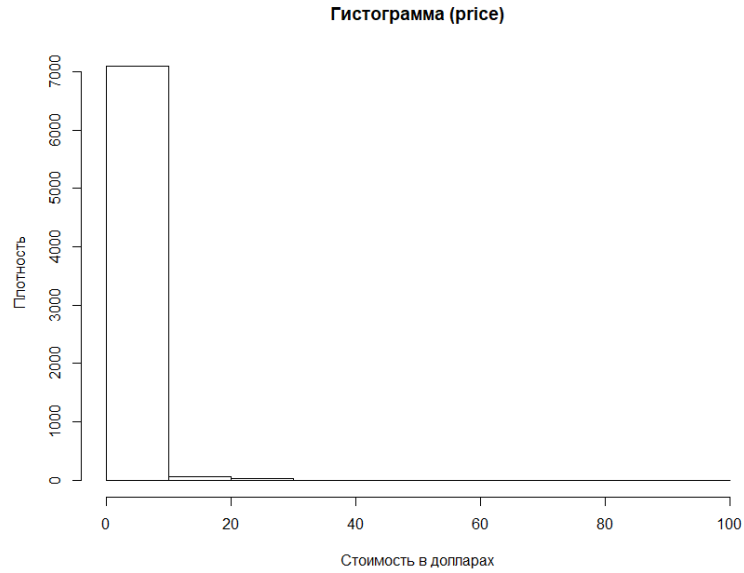


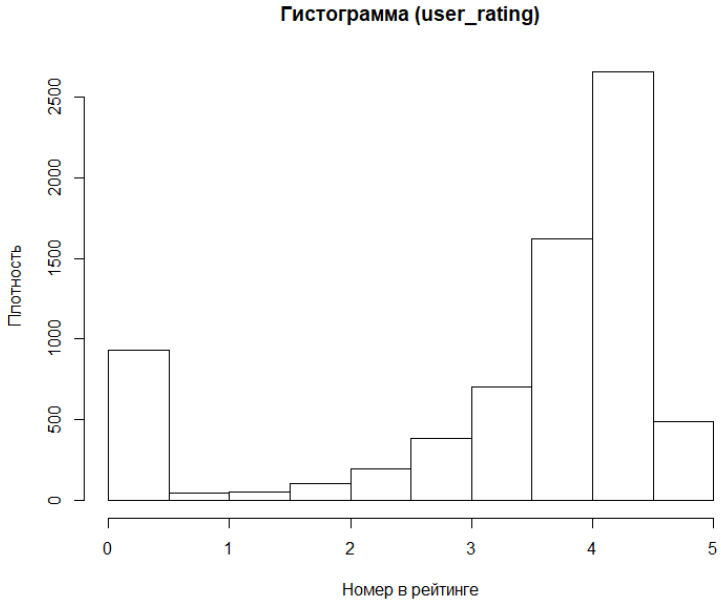
Популярность жанров

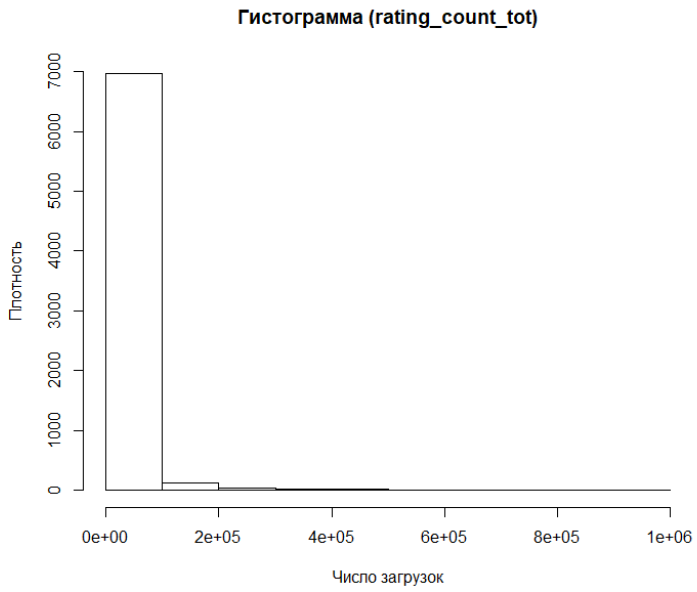


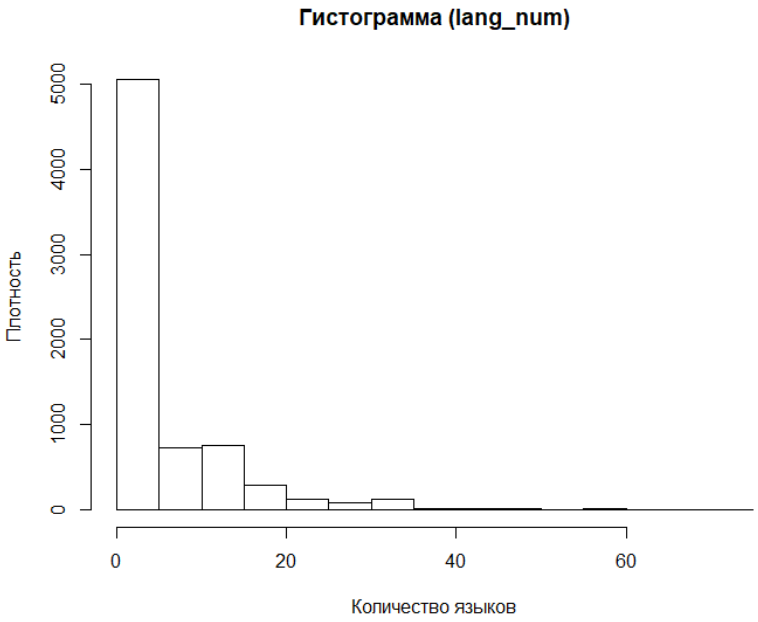
Гистограмма (size_bytes)




Гистограмма (price)
Плотность
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
Стоимость в доллрах


Гистограмма (user_rating)
Плотность
0
500
1000
1500
2000
2500
0
1
2
3
4
5
Номер в рейтинге


Гистограмма (rating_count_tot)
Плотность
0
1000
2000
3000
4000
5000
6000
7000
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
Число загрузок




```
- 11 -------------------------
Наиболее популярный жанр: Games
------------------------------
- 12 -------------------------
Суммарная статистика переменных     price          user_rating       lang_num
 Min.    : 0.000   Min.    :0.000   Min.    : 0.000
 1st Qu.: 0.000   1st Qu.:3.500   1st Qu.: 1.000
 Median : 0.000   Median :4.500   Median : 1.000
 Mean    : 1.428   Mean    :3.685   Mean    : 4.586
 3rd Qu.: 1.990   3rd Qu.:4.500   3rd Qu.: 7.000
 Max.    :29.990   Max.    :5.000   Max.    :46.000

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

        One-sample Kolmogorov-Smirnov test

data:  df2$price
D = 0.32395, p-value < 2.2e-16
alternative hypothesis: two-sided


        One-sample Kolmogorov-Smirnov test

data:  df3$price
D = 0.30363, p-value < 2.2e-16
alternative hypothesis: two-sided
```

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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. Постройте столбиковую диаграмму (*bar plot*), которая показывала бы, сколько в базе данных районов Белгородской области, а сколько — Калужской.

Работа программы:

young_share	trud_share	old_share
17.10	57.32	25.58
16.99	55.65	27.37
16.09	54.62	29.28
16.00	54.99	29.02
16.90	53.87	29.23
16.27	53.80	29.94

Import Dataset | List

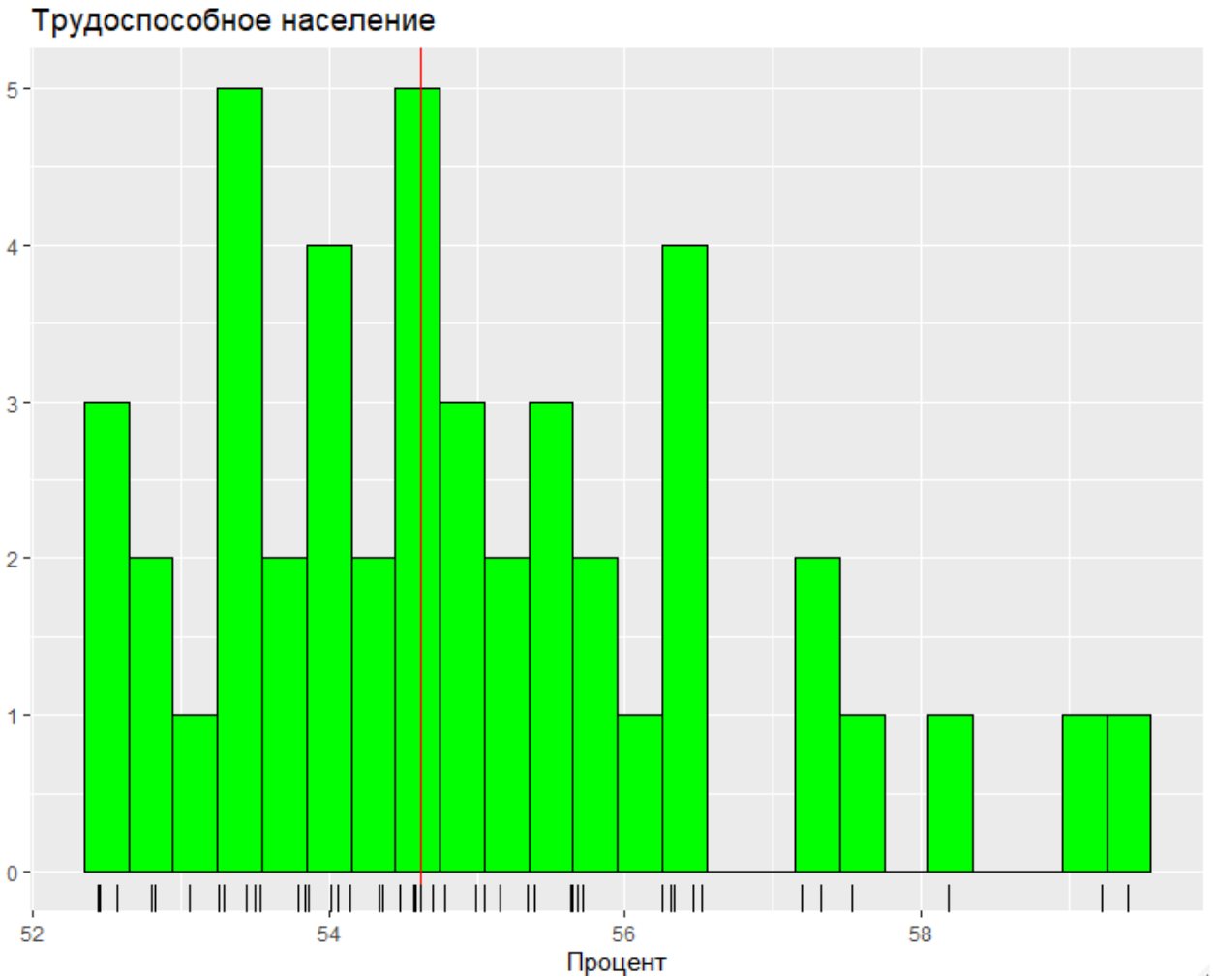
Global Environment

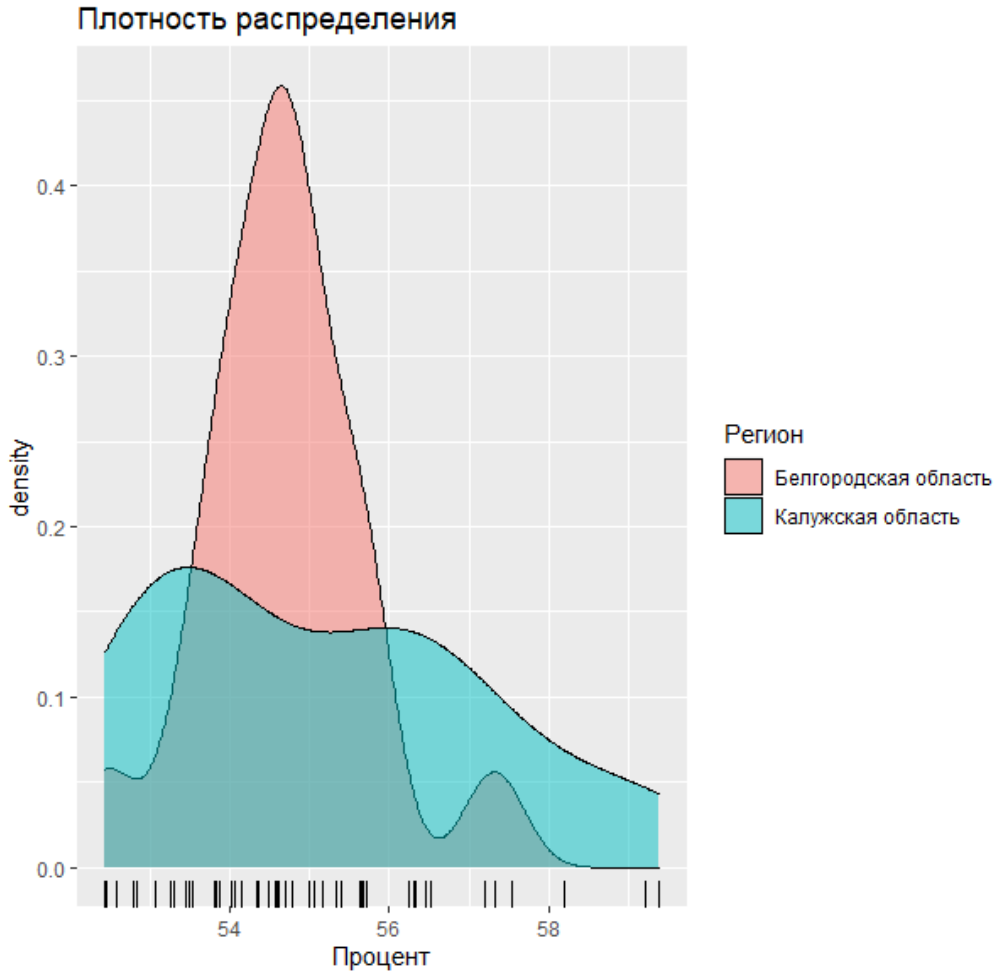
Data

df 45 obs. of 45 variables

Functions

Task function ()




Плотность распределения
density
0.4
0.3
0.2
0.1
0.0
54
56
58
Процент
Регион
Белгородская область
Калужская область

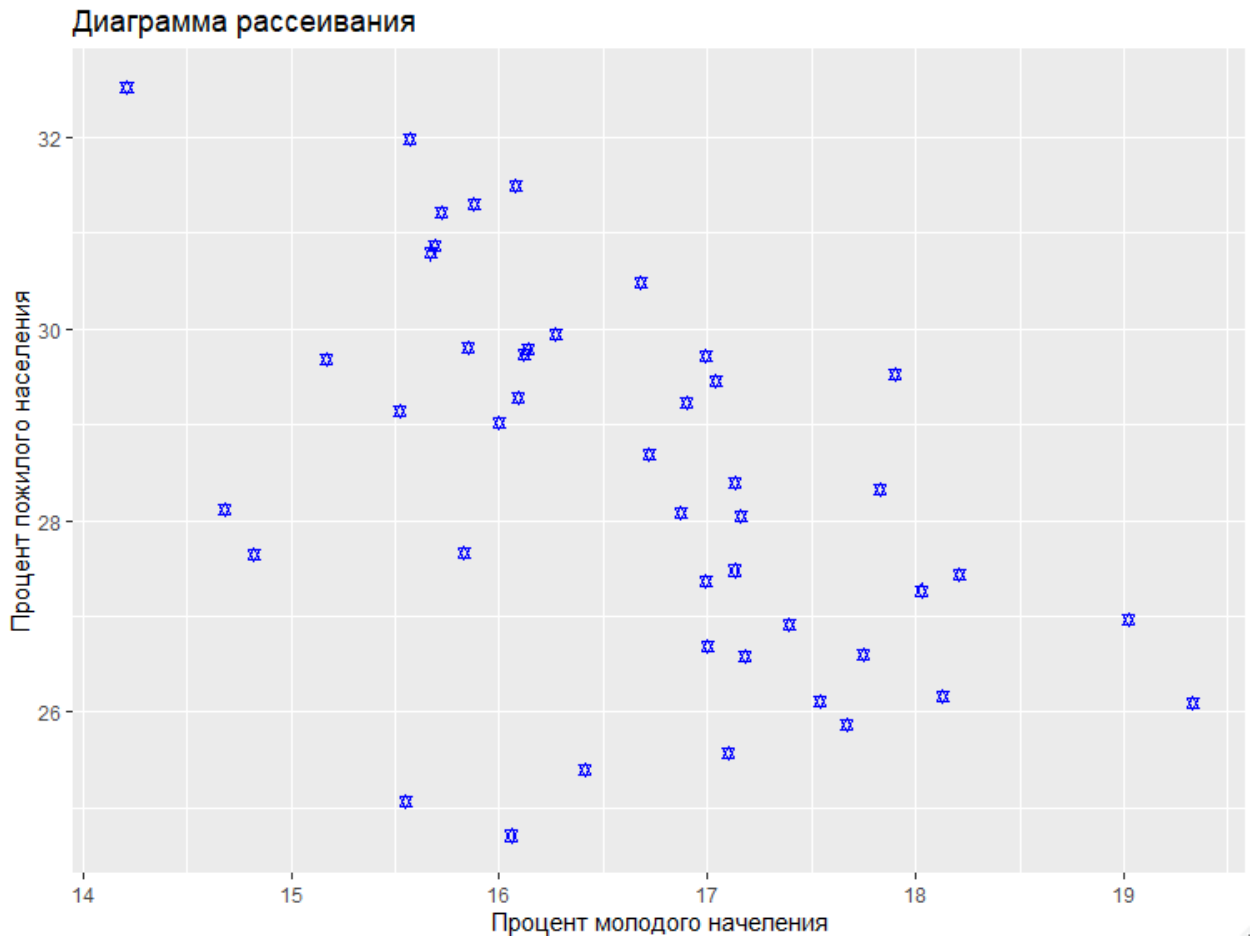
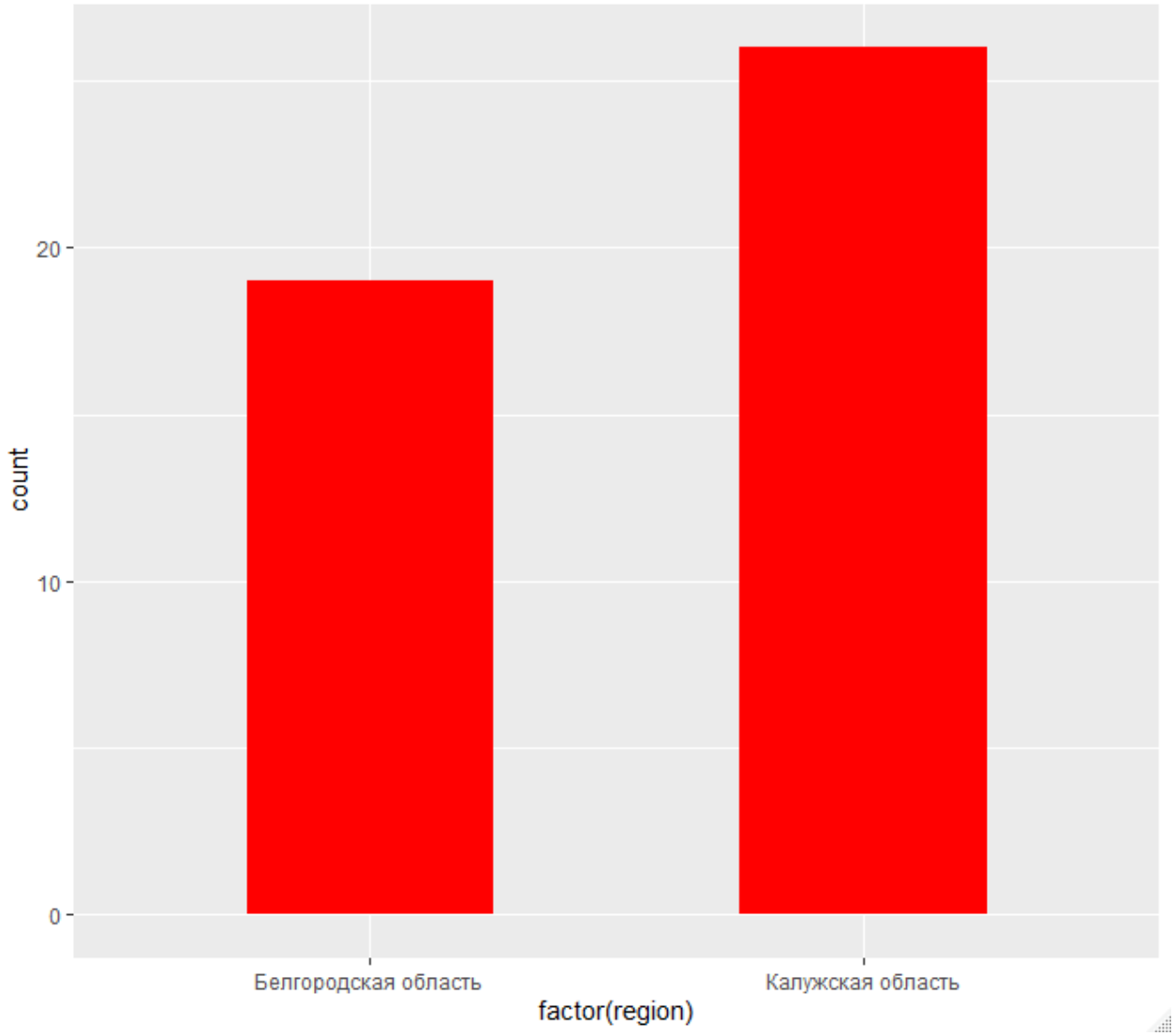


Диаграмма рассеивания
Процент пожилого населения
32
30
28
26
14
15
16
17
18
19
Процент молодого начеления


Диаграмма рассеивания
Процент пожилого населения
32
30
28
26
14
15
16
17
18
19
Процент молодого населения
male
1.00
0.75
0.50
0.25
0.00
male_share
46
48
50
52



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:

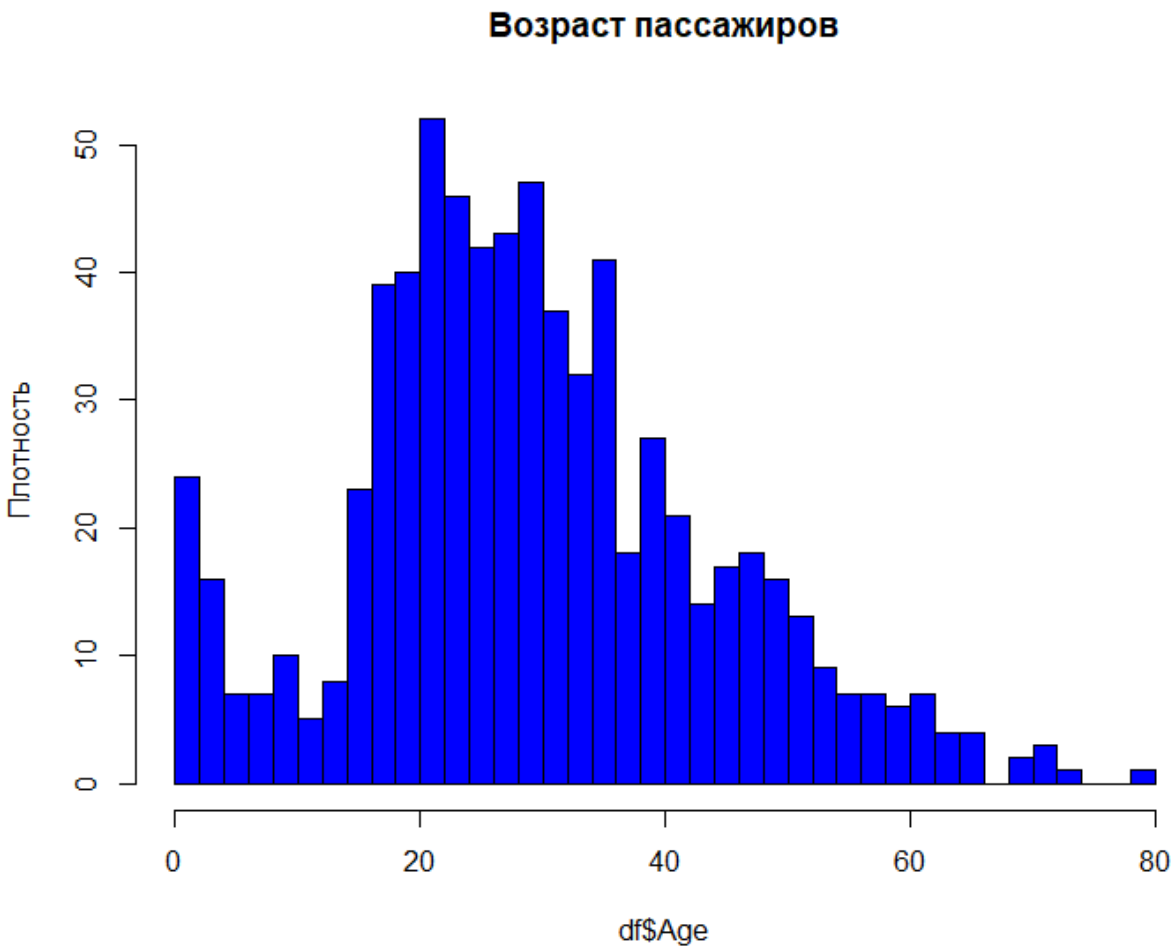
1. Загрузите данные из файла Titanic.csv, с которым вы уже работали.
2. 1. Выведите описательные статистики для всех переменных в таблице. Выберите два показателя (один количественный, один качественный) и проинтерпретируйте все выведенные по ним значения статистик.
3. 2. Постройте для показателя Age 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:

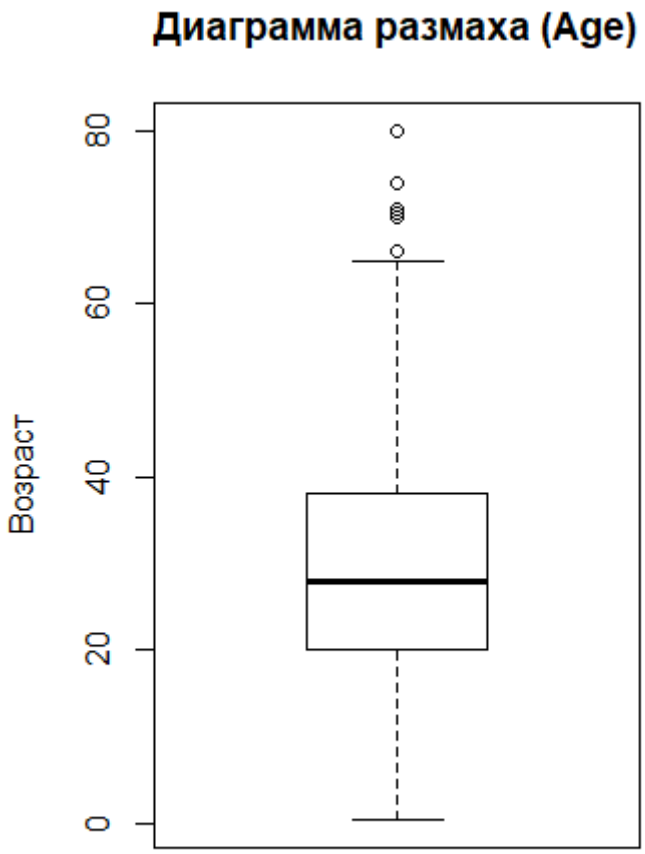
```
> Task()

- 1 -------------------------[1] "Описательные статистики"
  PassengerId       Survived          Pclass                                      Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                   :  1
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :  1
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :  1
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :  1
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :  1
                                                  (Other)                             :885
     Sex           Age            SibSp           Parch             Ticket         Fare
 female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000   1601     :  7   Min.   :  0.00
 male  :577   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   347082   :  7   1st Qu.:  7.91
              Median :28.00   Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
              Mean   :29.70   Mean   :0.523   Mean   :0.3816   3101295  :  6   Mean   : 32.20
              3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000   347088   :  6   3rd Qu.: 31.00
              Max.   :80.00   Max.   :8.000   Max.   :6.0000   CA 2144  :  6   Max.   :512.33
              NA's   :177                                      (Other)  :852
        Cabin     Embarked
           :687     :  2
 B96 B98   :  4   C:168
 C23 C25 C27:  4   Q: 77
 G6        :  4   S:644
 C22 C26   :  3
 D         :  3
 (Other)   :186

----------------------------
```


Возраст пассажиров
Плотность
0
10
20
30
40
50
0
20
40
60
80
df$Age

```
- 3 -------------------------
выбросов:  8
-------------------------------
```



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-----------------------------")
}
```