Задания

Задача 1:

1. Загрузите данные из файла **VILLA.xls**
2. Определите тип данных, с которыми Вы работаете.
3. Рассчитайте и проинтерпретируйте описательные статистики по каждой переменной, включая фиктивную переменную.
4. Проанализируйте исходную выборку на наличие статистических выбросов, используя анализ ящичковых диаграмм. Сделайте выводы.
5. Проверьте однородность всех переменных с помощью коэффициента вариации по каждой переменной. Сделайте выводы.
6. Проверьте нормальность распределения переменной **Price** с помощью:
    a. гистограммы
    b. коэффициентов асимметрии и эксцесса
    c. графика Q-Qplot
    d. проверки гипотезы о нормально распределении (на уровне значимости 0,05) с помощью критериев: Колмогорова-Смирнова, Шапиро-Уилка, Лиллифорса, Крамера-фон Мизеса и Андерсона-Дарлинга, Шапиро-Франсиа, хи-квадрат Пирсона. Сделайте выводы

Работа программы:

1

Filter

	N	Price	Dist	house	area	Eco	Стоимость.коттеджей.по.Киевскому.направлению.по.прайс.листу..Стройсервис.	NA.	NA..1	NA..2
1	1	300.0	20.0	400	22.0	1	NA	NA	NA	NA
2	2	60.0	18.0	170	6.0	0	N	номер по порядку	NA	NA
3	3	14.0	90.0	60	11.0	1	Price	цена в тыс. USD	NA	NA
4	4	38.0	18.0	65	6.0	1	Dist	расстояние от кольцевой автодороги в км.	NA	NA
5	5	85.0	25.0	320	20.0	0	House	площадь дома, кв.м.	NA	NA
6	6	85.0	19.0	210	20.0	0	Area	площадь участка, сотки	NA	NA
7	7	28.0	30.0	60	5.0	1	Eco	1, если рядом река, озеро	NA	NA
8	8	83.0	45.0	228	20.0	0	NA	NA	NA	NA
9	9	80.0	25.0	200	20.0	1	NA	NA	NA	NA
10	10	15.0	46.0	36	10.0	1	NA	NA	NA	NA
11	11	27.0	86.0	180	17.0	0	NA	NA	NA	NA
12	12	42.0	85.0	250	15.0	1	NA	NA	NA	NA
13	13	5.5	85.0	36	12.0	0	NA	NA	NA	NA
14	14	47.0	74.0	285	15.0	0	NA	NA	NA	NA
15	15	5.0	95.0	36	10.0	0	NA	NA	NA	NA
16	16	59.0	9.0	420	10.0	0	NA	NA	NA	NA
17	17	27.0	12.0	130	6.0	0	NA	NA	NA	NA

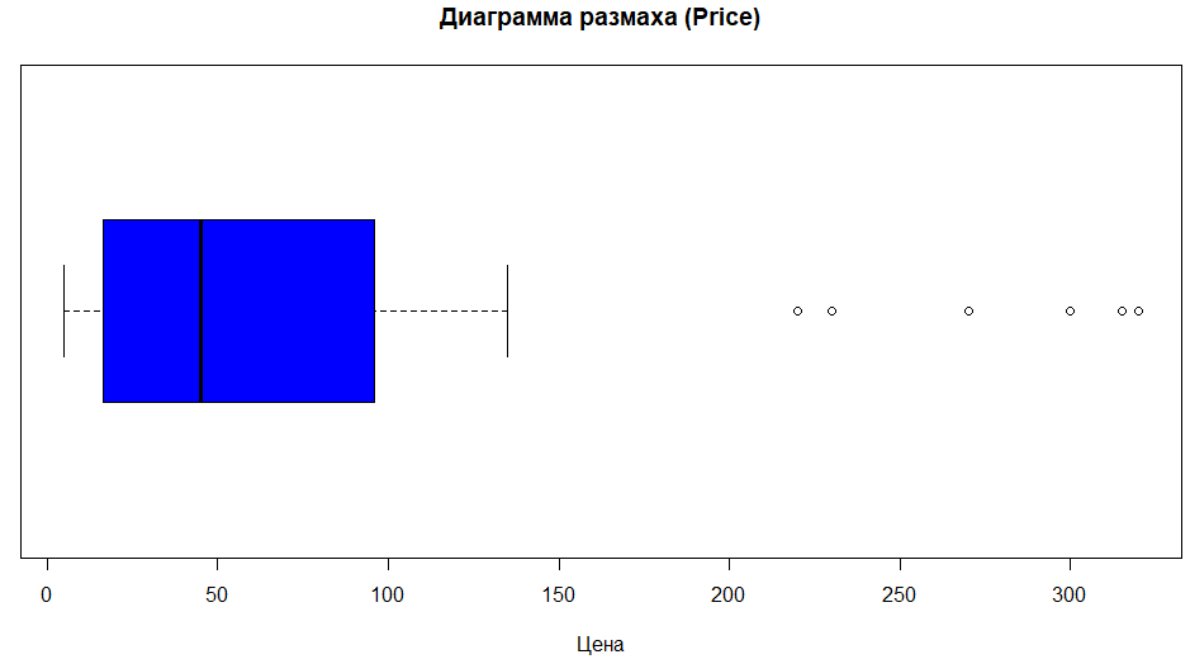
2

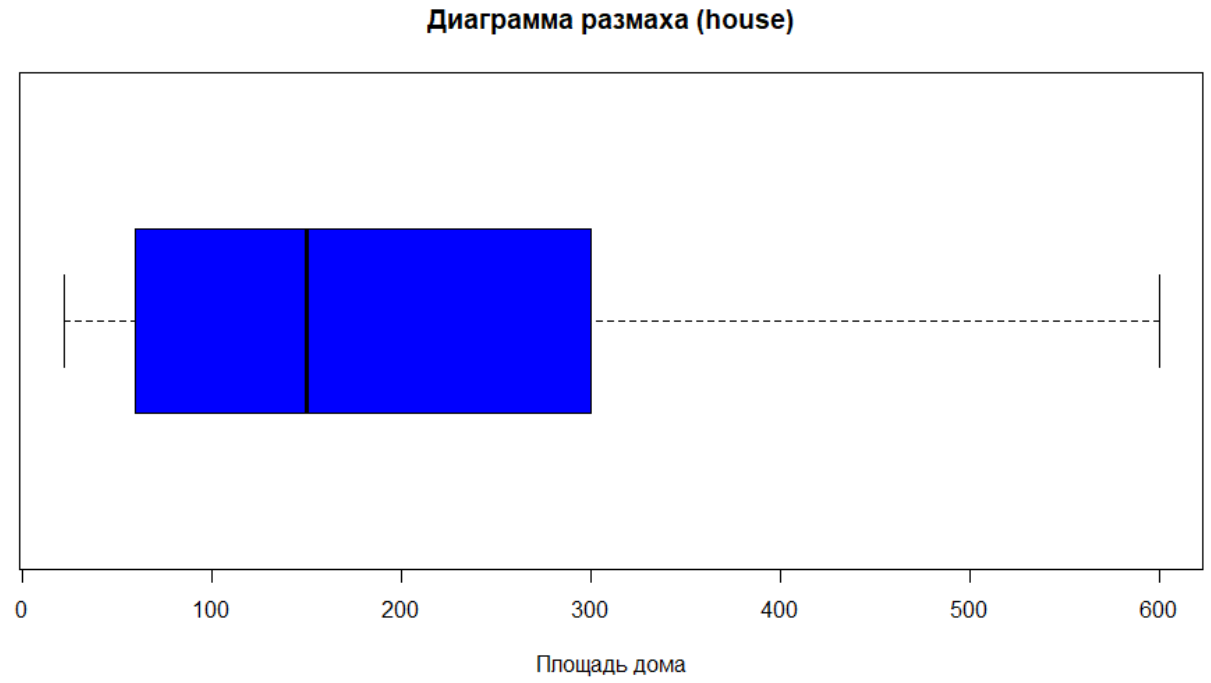
```
- 2 ------------------------'data.frame':      50 obs. of  10 variables:
 $ N                                                                 : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Price                                                             : num  300 60 14 38 85 85 28 83 80 15 ...
 $ Dist                                                              : num  20 18 90 18 25 19 30 45 25 46 ...
 $ house                                                             : num  400 170 60 65 320 210 60 228 200 36 ...
 $ area                                                              : num  22 6 11 6 20 20 5 20 20 10 ...
 $ Eco                                                               : num  1 0 1 1 0 0 1 0 1 1 ...
 $ Стоимость.коттеджей.по.Киевскому.направлению.по.прайс.листу..Стройсервис.: Factor w/ 6 levels "Area","Dist",..: NA 5 6 2 4 1 3 NA NA NA ...
 $ NA.                                                               : Factor w/ 6 levels "1, если рядом река, озеро",..: NA 2 6 5 3 4 1 NA NA NA ...
 $ NA..1                                                             : logi  NA NA NA NA NA NA ...
 $ NA..2                                                             : logi  NA NA NA NA NA NA ...
```

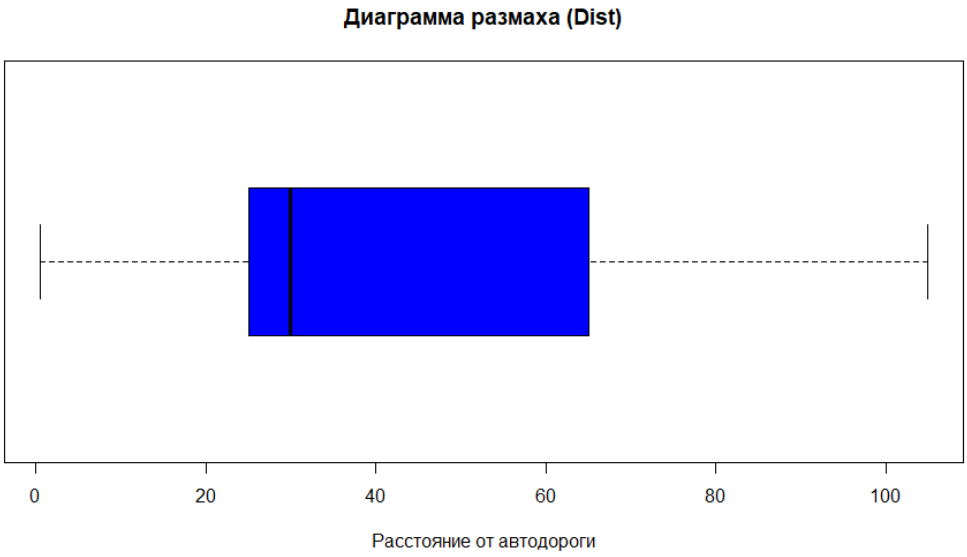
3

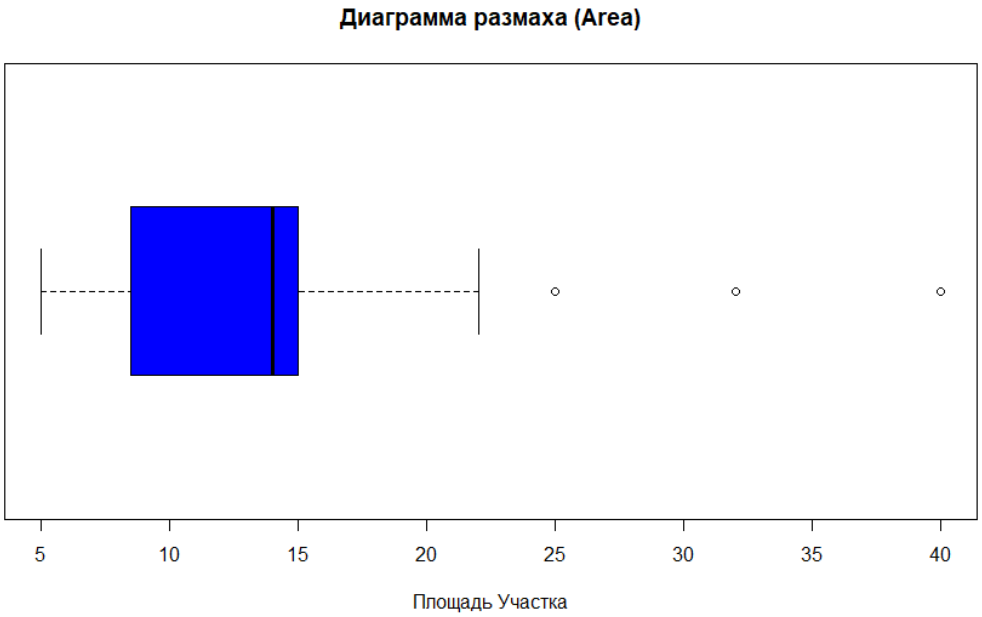
```
- 3 -------------------------
> summary(df)
       N             Price             Dist            house            area            Eco
 Min.   : 1.00   Min.   :  5.00   Min.   :  0.50   Min.   : 22.00   Min.   : 5.000   Min.   :0.00
 1st Qu.:13.25   1st Qu.: 16.62   1st Qu.: 25.00   1st Qu.: 61.25   1st Qu.: 8.875   1st Qu.:0.00
 Median :25.50   Median : 46.00   Median : 30.00   Median :160.00   Median :14.000   Median :0.50
 Mean   :25.50   Mean   : 78.25   Mean   : 44.05   Mean   :192.24   Mean   :13.750   Mean   :0.52
 3rd Qu.:37.75   3rd Qu.: 99.00   3rd Qu.: 63.75   3rd Qu.:300.00   3rd Qu.:15.000   3rd Qu.:1.00
 Max.   :50.00   Max.   :320.00   Max.   :105.00   Max.   :600.00   Max.   :40.000   Max.   :2.00

 Стоимость.коттеджей.по.Киевскому.направлению.по.прайс.листу..Стройсервис.                                  NA.      NA..1          NA..2
 Area : 1                                                                    1, если рядом река, озеро          : 1   Mode:logical   Mode:logical
 Dist : 1                                                                    номер по порядку                   : 1   NA's:50        NA's:50
 Eco  : 1                                                                    площадь дома, кв.м.                : 1
 House: 1                                                                    площадь участка, сотки             : 1
 N    : 1                                                                    расстояние от кольцевой автодороги в км.: 1
 Price: 1                                                                    цена в тыс. USD                    : 1
 NA's :44                                                                    NA's                               :44

> df <- subset(df, Eco >= 0 & Eco <= 1);
```

4





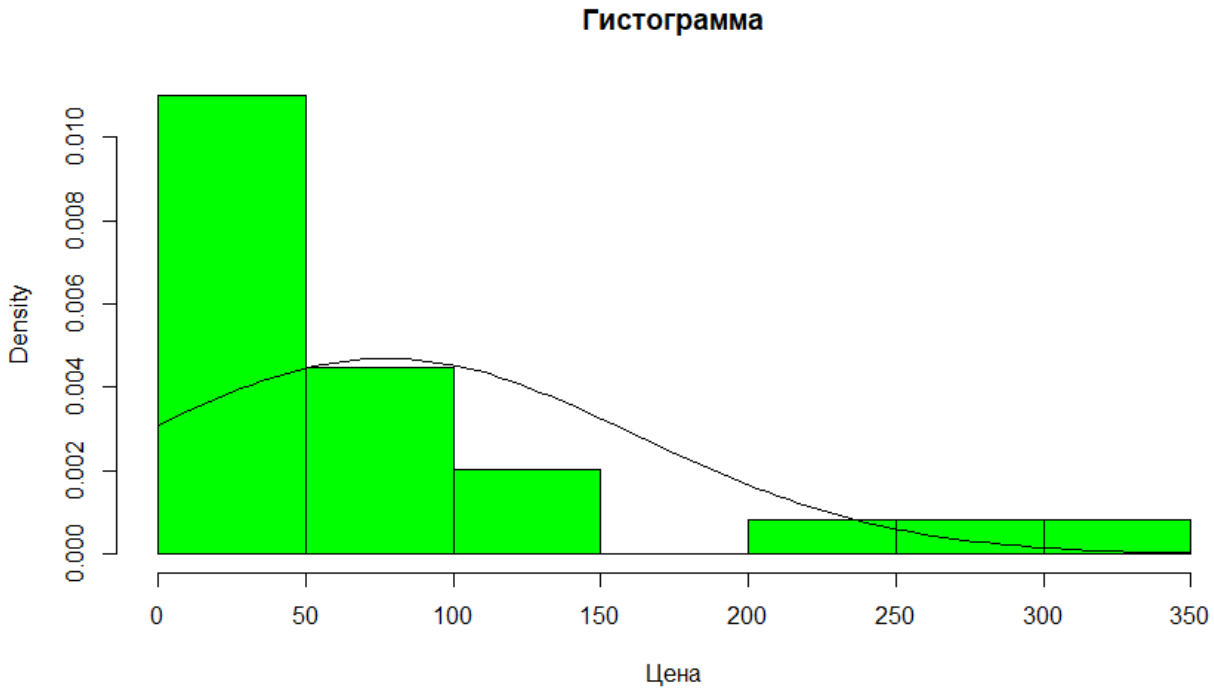




5

```
- 5 -------------------------
> cat("\nКоэффициент вариации (Price): ", sd(df$Price) / mean(df$Price) * 100);

Коэффициент вариации (Price):  109.8215
> #  >33%, совокупность неоднородная
>
> cat("\nКоэффициент вариации (Dist): ",sd(df$Dist) / mean(df$Dist) * 100);

Коэффициент вариации (Dist):  64.86475
> # >33%, совокупность неоднородная
>
> cat("\nКоэффициент вариации (house): ",sd(df$house) / mean(df$house) * 100);

Коэффициент вариации (house):  80.30613
> # >33%, совокупность неоднородная
>
> cat("\nКоэффициент вариации (area): ",sd(df$area) / mean(df$area) * 100);

Коэффициент вариации (area):  50.39244
> # >33%, совокупность неоднородная
>
> cat("\nКоэффициент вариации (Eco): ",sd(df$Eco) / mean(df$Eco) * 100);

Коэффициент вариации (Eco):  103.1197
> # >33%, совокупность неоднородная
```

6a

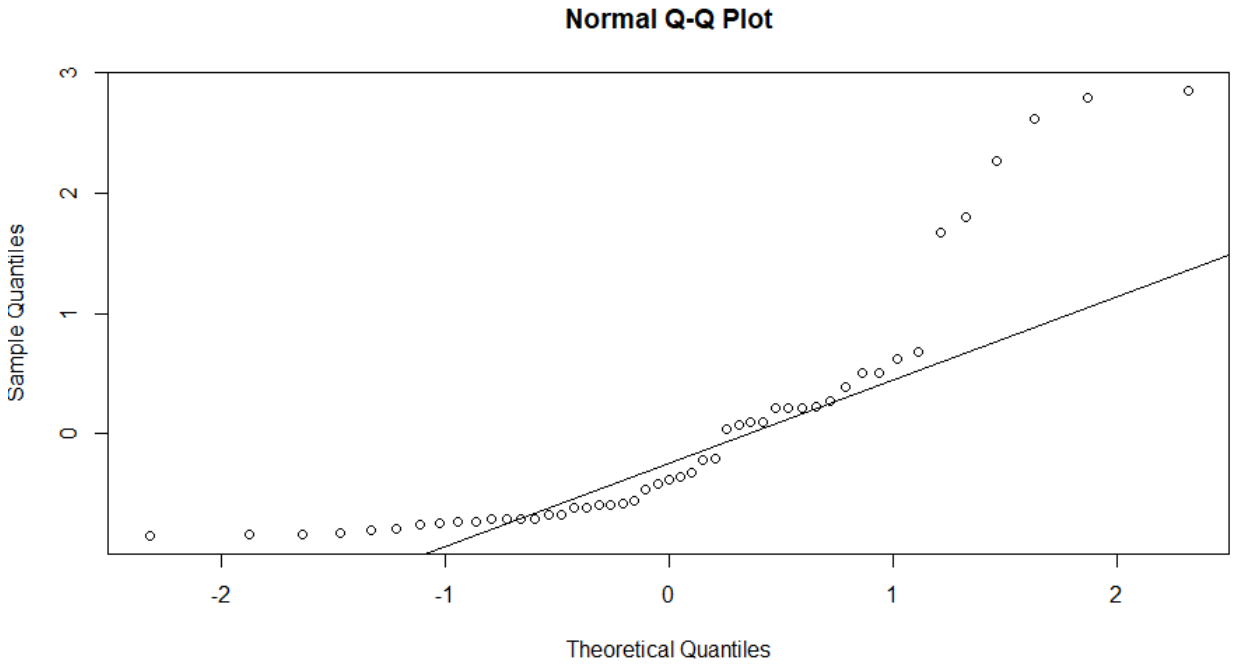


6b

```
- 6b ------------------------
> cat("\nКоэффициент вариации (Price): ", kurtosis(df$Price, na.rm = TRUE))  #Эксцесс > 0, распределение будет являться более высоким (островершинным)

Коэффициент вариации (Price):  4.807202
> cat("\nКоэффициент вариации (Price): ", skewness(df$Price, na.rm = TRUE))  #Коэффициент асимметрии > 0, правый хвост распределения длиннее левого

Коэффициент вариации (Price):  1.629149
```

6c

```
- 6c -------------------------
> df$Price_new <- scale(df$Price)

> df$Price_new <- as.numeric(df$Price_new)

> qnorm(0.1, mean = 0, sd = 1)
[1] -1.281552

> quantile(df$Price_new, 0.1)
       10%
-0.7964506

> qqnorm(df$Price_new)

> qqline(df$Price_new)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
0
1
2
3

6d

```
- 6d ------------------------
> #Критерий Колмогорова-Смирнова
> ks.test(df$Price, "pnorm",
+         mean = mean(df$Price, na.rm = T),
+         sd = sd(df$Price, na.rm = T))#не .... [TRUNCATED]

        One-sample Kolmogorov-Smirnov test

data:  df$Price
D = 0.19718, p-value = 0.04429
alternative hypothesis: two-sided


> #Критерий Шапиро-Уилка
> shapiro.test(df$Price)

        Shapiro-Wilk normality test

data:  df$Price
W = 0.76656, p-value = 1.992e-07


> #Критерий Лиллифорса
> lillie.test(df$Price) #не оттвергаем нулевую гипотезу о нормальности распределения

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  df$Price
D = 0.19718, p-value = 5.461e-05


> #Критерии Крамера-фон Мизеса и Андерсона-Дарлинга
> cvm.test(df$Price) #не оттвергаем нулевую гипотезу о нормальности распределения

        Cramer-von Mises normality test

data:  df$Price
W = 0.62971, p-value = 1.895e-07


> ad.test(df$Price)  #не оттвергаем нулевую гипотезу о нормальности распределения

        Anderson-Darling normality test

data:  df$Price
A = 3.8825, p-value = 8.035e-10


> #Критерий Шапиро-Франсиа
> sf.test(df$Price)  #не оттвергаем нулевую гипотезу о нормальности распределения

        Shapiro-Francia normality test

data:  df$Price
W = 0.76979, p-value = 1.577e-06


> #Критерий хи-квадрат Пирсона
> pearson.test(df$Price) #не оттвергаем нулевую гипотезу о нормальности распределения

        Pearson chi-square normality test

data:  df$Price
P = 63.857, p-value = 2.551e-11
```

Листинг:

```
#install.packages(c("xlsx", "moments", "nortest"));

library(nortest);
```

```
library(moments);
library(xlsx);

cat("\n- 1 -------------------------")
df <- read.xlsx("villa_new.xlsx", 1,encoding = "UTF-8");
View(df);
cat("\n-----------------------------")

cat("\n- 2 -------------------------")
str(df);
cat("\n-----------------------------")

cat("\n- 3 -------------------------")
summary(df)
df <- subset(df, Eco >= 0 & Eco <= 1);
cat("\n-----------------------------")

cat("\n- 4 -------------------------")
boxplot(df$Price, data = df, xlab = "Цена", main = "Диаграмма размаха (Price)", col =
"blue", horizontal = TRUE);
# 6 выбросов
boxplot(df$Dist, data = df,  xlab = "Расстояние от автодороги", main = "Диаграмма размаха
(Dist)", col = "blue", horizontal = TRUE);
# нет выбросов
boxplot(df$house, data = df, xlab = "Площадь дома", main = "Диаграмма размаха (House)",
col = "blue", horizontal = TRUE);
# нет выбросов
boxplot(df$area, data = df, xlab = "Площадь Участка", main = "Диаграмма размаха (Area)",
col = "blue", horizontal = TRUE);
# нет выбросов
cat("\n-----------------------------")

cat("\n- 5 -------------------------")
cat("\nКоэффициент вариации (Price): ", sd(df$Price) / mean(df$Price) * 100);
#  >33%, совокупность неоднородная

cat("\nКоэффициент вариации (Dist): ",sd(df$Dist) / mean(df$Dist) * 100);
# >33%, совокупность неоднородная

cat("\nКоэффициент вариации (house): ",sd(df$house) / mean(df$house) * 100);
# >33%, совокупность неоднородная

cat("\nКоэффициент вариации (area): ",sd(df$area) / mean(df$area) * 100);
# >33%, совокупность неоднородная

cat("\nКоэффициент вариации (Eco): ",sd(df$Eco) / mean(df$Eco) * 100);
# >33%, совокупность неоднородная
cat("\n-----------------------------")

cat("\n- 6a ------------------------")
hist(df$Price)
K <- round(1 + 3.32 * log(nrow(df),10),0)
hist(df$Price, breaks = K, freq = FALSE, col = "green",
     xlab = "Цена",
     main = "Гистограмма")
curve(dnorm(x, mean(df$Price), sd = sd(df$Price)), add = TRUE)
cat("\n-----------------------------")

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

```
cat("\nКоэффициент вариации (Price): ", kurtosis(df$Price, na.rm = TRUE))  #Эксцесс > 0,
распределение будет являться более высоким (островершинным)
cat("\nКоэффициент вариации (Price): ", skewness(df$Price, na.rm = TRUE))  #Коэффициент
асимметрии > 0, правый хвост распределения длиннее левого
cat("\n-----------------------------")

cat("\n- 6c ------------------------")
df$Price_new <- scale(df$Price)
df$Price_new <- as.numeric(df$Price_new)
qnorm(0.1, mean = 0, sd = 1)
quantile(df$Price_new, 0.1)
qqnorm(df$Price_new)
qqline(df$Price_new)
cat("\n-----------------------------")

cat("\n- 6d ------------------------")
#Критерий Колмогорова-Смирнова
ks.test(df$Price, "pnorm",
        mean = mean(df$Price, na.rm = T),
        sd = sd(df$Price, na.rm = T))#не оттвергаем нулевую гипотезу о нормальности
распределения

#Критерий Шапиро-Уилка
shapiro.test(df$Price)

#Критерий Лиллифорса
lillie.test(df$Price) #не оттвергаем нулевую гипотезу о нормальности распределения

#Критерии Крамера-фон Мизеса и Андерсона-Дарлинга
cvm.test(df$Price) #не оттвергаем нулевую гипотезу о нормальности распределения
ad.test(df$Price)  #не оттвергаем нулевую гипотезу о нормальности распределения

#Критерий Шапиро-Франсиа
sf.test(df$Price)  #не оттвергаем нулевую гипотезу о нормальности распределения

#Критерий хи-квадрат Пирсона
pearson.test(df$Price) #не оттвергаем нулевую гипотезу о нормальности распределения
cat("\n-----------------------------")
```