Задания

Задача 1:

1. Загрузите данные из файла **VILLA.xls**
2. Определите тип данных, с которыми Вы работаете.
3. Проверьте нормальность распределения переменной **Price** с помощью критерия Лиллифорса.
4. Проверьте гипотезу о равенстве дисперсий Цены коттеджей (**Price**) в двух совокупностях (рядом с озером и нет).
5. В зависимости от предыдущего результата проверить гипотезу о равенстве Цены коттеджей (**Price**) в двух совокупностях (рядом с озером и нет). Сделайте выводы.

Работа программы:

Листинг:

Задача 2:

Будем работать с базой данных `psych_survey.csv`

Проверить гипотезу о равенстве среднего роста (height) у студентов с разным любимым предметом (subject)

Любимый предмет. Респонденту нужно было выбрать один ответ из 5 предложенных вариантов:

1. Математика;
2. Биология;
3. Русский язык;
4. Иностранный язык;

Ни один из вышеперечисленных предметов.

Работа программы:

1.1

	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

1.2

```
- 1.2 -------------------------
> class(df$Price)
[1] "numeric"

> class(df$Dist)
[1] "numeric"

> class(df$house)
[1] "numeric"

> class(df$area)
[1] "numeric"

> class(df$Eco)
[1] "numeric"
```

1.3

```
- 1.3 -------------------------
> lillie.test((df$Price))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  (df$Price)
D = 0.19257, p-value = 7.602e-05


> #отвергаем нулевую гипотезу о нормальности распределения (p-value < 0,05)
```

1.4

```
        F test to compare two variances

data:  df2$Price and df2$Eco
F = 6018.2, num df = 49, denom df = 49, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 4
95 percent confidence interval:
 13660.78 42420.93
sample estimates:
ratio of variances
          24072.87
```

1.5

```
        Two Sample t-test

data:  df2$Price and df2$Eco
t = 6.5165, df = 98, p-value = 3.124e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  54.05892 101.40108
sample estimates:
mean of x mean of y
    78.25      0.52
```

2

```
- 2 -------------------------
> df <- read.csv("psych_survey.csv", sep = ";")

> View(df)

> str(df)
'data.frame':   123 obs. of  9 variables:
 $ height   : Factor w/ 39 levels "145","150","151",..: 19 17 13 8 36 18 30 27 33 25 ...
 $ math     : int  76 62 80 75 62 70 NA 79 50 70 ...
 $ bio      : int  96 79 94 77 74 65 68 69 63 79 ...
 $ subject  : int  1 2 2 4 2 1 4 1 2 4 ...
 $ gender   : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 2 1 2 1 ...
 $ residence: int  2 1 2 1 2 1 1 1 1 1 ...
 $ length   : Factor w/ 26 levels "0","11,7","12",..: 1 1 6 24 8 14 19 22 22 23 ...
 $ angle    : Factor w/ 26 levels "0","11","12",..: 1 1 2 3 4 4 4 4 4 4 ...
 $ soft     : Factor w/ 2 levels "R","SPSS": 1 1 1 1 1 1 1 2 1 1 ...

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

> df$subject <- as.factor((df$subject))

> summary(df$subject)
   1    2    3    4    5 NA's
  29   37   18   18   19    2

> df <- subset(df, subject != "NA")

> summary(df$subject)
 1  2  3  4  5
29 37 18 18 19

> #Проверка данных на нормальность
> math <- subset(df, subject == 1)

> bio <- subset(df, subject == 2)

> rus <- subset(df, subject == 3)

> inostr <- subset(df, subject == 4)

> no <- subset(df, subject == 5)

> ks.test(math$height, "pnorm",
+         mean = mean(math$height, na.rm = T),
+         sd = sd(math$height, na.rm = T))

        One-sample Kolmogorov-Smirnov test

data:  math$height
D = 0.12571, p-value = 0.7492
alternative hypothesis: two-sided
```

```
        One-sample Kolmogorov-Smirnov test

data:  bio$height
D = 0.073658, p-value = 0.988
alternative hypothesis: two-sided


> ks.test(rus$height, "pnorm",
+         mean = mean(rus$height, na.rm = T),
+         sd = sd(rus$height, na.rm = T))

        One-sample Kolmogorov-Smirnov test

data:  rus$height
D = 0.16273, p-value = 0.7271
alternative hypothesis: two-sided


> ks.test(inostr$height, "pnorm",
+         mean = mean(inostr$height, na.rm = T),
+         sd = sd(inostr$height, na.rm = T))

        One-sample Kolmogorov-Smirnov test

data:  inostr$height
D = 0.152, p-value = 0.7999
alternative hypothesis: two-sided


> ks.test(no$height, "pnorm",
+         mean = mean(no$height, na.rm = T),
+         sd = sd(no$height, na.rm = T))

        One-sample Kolmogorov-Smirnov test

data:  no$height
D = 0.087669, p-value = 0.9986
alternative hypothesis: two-sided


> #Согласно всем тестам, во всех выборках нулевая гипотеза о нормальном распределении не отвергается.
> #Следовательно, поэтому мы выбираем однофакто .... [TRUNCATED]

> summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
subject       4    654  163.53   2.145 0.0796 .
Residuals   116   8843   76.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Листинг:

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

library(nortest);
library(xlsx);

cat("\n- 1.1 -------------------------")
df <- read.xlsx("villa_new.xlsx", 1,encoding = "UTF-8");
View(df);
cat("\n------------------------------")

cat("\n- 1.2 -------------------------")
class(df$Price)
class(df$Dist)
class(df$house)
class(df$area)
class(df$Eco)

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

cat("\n- 1.3 -------------------------")
lillie.test((df$Price))
#отвергаем нулевую гипотезу о нормальности распределения (p-value < 0,05)
cat("\n------------------------------")

cat("\n- 1.4 -------------------------")
var.test(df$Price ~ df$Eco, data = df, alternative = "two.sided")
#гипотеза о равенстве дисперсий отвергается (p-value < 0,05)
cat("\n------------------------------")

cat("\n- 1.5 -------------------------")
```

```
#исходя из результатов предыдущего номера используем двухвыборочный критерий Стьюдента равенства средних
#(t-критерий в модификации Уэлча (Welch) с неравными дисперсиями)

t.test(Price ~ Eco, df, var.equal = FALSE)
#вывод: нулевая гипотеза о равенстве средних отвергается, поскольку p-value меньше уровня значимости 0,05
cat("\n-----------------------------")

cat("\n- 2 -------------------------")
df <- read.csv("psych_survey.csv", sep = ";")
View(df)
str(df)

df$height <- as.numeric((df$height))
df$subject <- as.factor((df$subject))

summary(df$subject)
df <- subset(df, subject != "NA")
summary(df$subject)

#Проверка данных на нормальность
math <- subset(df, subject == 1)
bio <- subset(df, subject == 2)
rus <- subset(df, subject == 3)
inostr <- subset(df, subject == 4)
no <- subset(df, subject == 5)

ks.test(math$height, "pnorm",
        mean = mean(math$height, na.rm = T),
        sd = sd(math$height, na.rm = T))

ks.test(bio$height, "pnorm",
        mean = mean(bio$height, na.rm = T),
        sd = sd(bio$height, na.rm = T))

ks.test(rus$height, "pnorm",
        mean = mean(rus$height, na.rm = T),
        sd = sd(rus$height, na.rm = T))

ks.test(inostr$height, "pnorm",
        mean = mean(inostr$height, na.rm = T),
        sd = sd(inostr$height, na.rm = T))

ks.test(no$height, "pnorm",
        mean = mean(no$height, na.rm = T),
        sd = sd(no$height, na.rm = T))

#Согласно всем тестам, во всех выборках нулевая гипотеза о нормальном распределении не отвергается.
#Следовательно, поэтому мы выбираем однофакторный дисперсионный анализ для сравнения средних в нескольких группах.

anova <- aov(height ~ subject, data = df)
summary(anova)
#В данном случае мы не отвергаем нулевую гипотезу об отсутствии различий между всеми средними против альтернативы о том, что хотя бы одно среднее отличается.
cat("\n-----------------------------")
```