Lab3

Завантаження пакетів

```
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(pacman, ggplot2,
   plotly, rio, rmarkdown, moments, agricolae, corrplot, tidyverse, corrr, model4you,
   Metrics, readxl, MASS, stats)
```

```
library(pacman)
library(moments)
library(agricolae)
library(corrplot)
library(tidyverse)
library(corrr)
library(model4you)
library(Metrics)
library(readxl)
library(MASS)
library(stats)
```

Завантаження даних

```
# CSV
data_csv <- import("/Users/victoria/Downloads/kc_house_data.csv")
```

```
## Warning in (function (input = "", file = NULL, text = NULL, cmd =
## NULL, : Stopped early on line 21431. Expected 21 fields but found 20.
## Consider fill=TRUE and comment.char=. First discarded non-empty line:
## <<20140617000000,1.33E+06,4,4,4420,16526,2,0,0,3,11,4420,0,2013,0,98075,47.5914,-122.027,3510,50447>>
```

Опис даних

Цей набір даних містить ціни продажу будинків для округу Кінг, який включає Сіетл. Сюди входять будинки, продані з травня 2014 року по травень 2015 року. Джерело: https://www.kaggle.com/harlfoxem/housesalesprediction

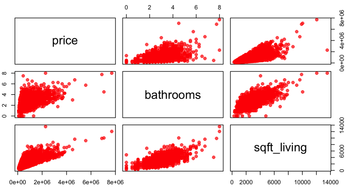
Розглянемо перші рядки датасету:

```
head(data_csv)
```

```
##           id             date  price bedrooms bathrooms sqft_living sqft_lot
## 1 7129300520 20141013T000000 221900        3      1.00        1180     5650
## 2 6414100192 20141209T000000 538000        3      2.25        2570     7242
## 3 5631500400 20150225T000000 180000        2      1.00         770    10000
## 4 2487200875 20141209T000000 604000        4      3.00        1960     5000
## 5 1954400510 20150218T000000 510000        3      2.00        1680     8080
## 6 1321400060 20140627T000000 257500        3      2.25        1715     6819
##   floors waterfront view condition grade sqft_above sqft_basement yr_built
## 1      1          0    0         3     7       1180            0      1955
## 2      2          0    0         3     7       2170          400      1951
## 3      1          0    0         3     6        770            0      1933
## 4      1          0    0         5     7       1050          910      1965
## 5      1          0    0         3     8       1680            0      1987
## 6      2          0    0         3     7       1715            0      1995
##   yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1            0   98178 47.5112 -122.257          1340       5650
## 2         1991   98125 47.7210 -122.319          1690       7639
## 3            0   98028 47.7379 -122.233          2720       8062
## 4            0   98136 47.5208 -122.393          1360       5000
## 5            0   98074 47.6168 -122.045          1800       7503
## 6            0   98003 47.3097 -122.327          2238       6819
```

Для кількісних даних(EU_Sales, JP_Sales, Other_Sales, Global_Sales, Year) зобразимо матричну діаграму розсіювання.

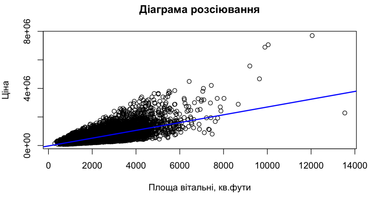
```
pairs(~price+bathrooms+sqft_living, data=data_csv, col="red", pch=10)
```



За графіками бачимо кореляційний зв'язок між величинами, наприклад, bathrooms і sqft_living, і price і sqft_living і price.

Побудуємо відповідну регресійну модель між sqft_living і price та зобразимо на одному графіку з діаграмою розсіювання.

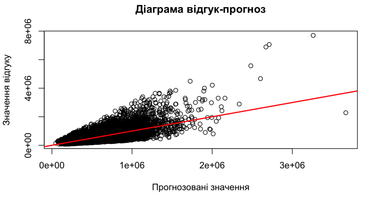
```
x <- data_csv$sqft_living
y <- data_csv$price
plot(x, y, main="Діаграма розсіювання", xlab="Площа вітальні, кв.фути", ylab="Ціна")
#b1<-cov(x,y)/var(x)
#b0<-mean(y)-b1*mean(x)
model<-lm(y~x)
#rse(y,y-resid(model))
#1-mean((resid(model))^2)/mean((y-mean(y))^2)
abline(model, col="blue", lwd=2)
```




```
x <- data_csv$sqft_living
y <- data_csv$price
b1<-cov(x,y)/var(x) # slope
b0<-mean(y)-b1*mean(x) # intercept
model<-lm(y~x)
#b0
#b1
#rse(y,y-resid(model)) # residual standart error
#1-mean((resid(model))^2)/mean((y-mean(y))^2) # determination
```

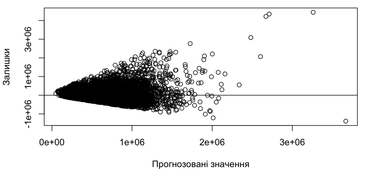
Побудуємо діаграму "Відгук-прогноз"

```
# Response vs prediction plot
resid<-resid(model)
pred<-y-resid
plot(pred,y,xlab="Прогнозовані значення",ylab="Значення відгуку",main="Діаграма відгук-прогноз")
abline(0,1, col="red", lwd=2)
```



Побудуємо діаграму "відгук-залишки"

```
# Residuals vs prediction plot
plot(pred,resid,xlab="Прогнозовані значення",ylab="Залишки",main="Діаграма прогноз(відгук)-залишки")
abline(0,0,lwd=1)
```



Побудуємо Q-Q діаграму для залишків:

```
qqnorm(model$residuals, xlab="Теоретичні квантилі",  ylab="Вибіркові квантилі")
qqline(model$residuals, col="red")
```