МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

# Практикум №3

з курсу «Аналіз даних в інформаційних системах»

на тему: «Описова статистика»

Викладач:
Ліхоузова Т.А.

Виконав:
студент 2 курсу
групи ІП-14 ФІОТ
Шляхтун Денис

Київ-2023

**Тема:** Описова статистика.

**Мета роботи:** ознайомитись з методикою первинної обробки статистичних даних; проаналізувати вплив способу представлення даних на їх інформативність.

**Основне завдання**

Скачати дані із файлу Data2.csv

1. Записати дані у data frame
2. Дослідити структуру даних
3. Виправити помилки в даних
4. Побудувати діаграми розмаху та гістограми
5. Додати стовпчик із щільністю населення

**Додаткове завдання**

Відповісти на питання (файл Data2.csv):

1. Чи є пропущені значення? Якщо є, замінити середніми
2. Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу?
3. В якому регіоні середня площа країни найбільша?
4. Знайдіть країну з найбільшою щільністю населення у світі? У Європі та центральній Азії?
5. Чи співпадає в якомусь регіоні середнє та медіана ВВП?
6. Вивести топ 5 країн та 5 останніх країн по ВВП та кількості CO2 на душу населення.

**Виконання основного завдання.**

Виконання комп’ютерного практикуму здійснювалося засобами R та RStudio.

1. Записати дані у data frame

Завантаження даних у data frame здійснювалося із CSV файлу Data2.csv за допомогою функції read.csv. Параметри функції: роздільник – «;»,

перший рядок – рядок заголовків стовпчиків, десяткові дробові числа пишуться через «,».

2. Дослідження структури

Дослідження структури data frame здійснено за допомогою функції str().

```
> str(data)
'data.frame':   217 obs. of  6 variables:
 $ Country.Name  : chr  "Afghanistan" "Albania" "Algeria" "American Samoa"
...
 $ Region        : chr  "South Asia" "Europe & Central Asia" "Middle East &
North Africa" "East Asia & Pacific" ...
 $ GDP.per.capita: num  562 4125 3917 11835 36989 ...
 $ Populatiion   : int  34656032 2876101 40606052 55599 77281 28813463 10096
3 43847430 2924816 104822 ...
 $ CO2.emission  : num  9809 5717 145400 NA 462 ...
 $ Area          : num  652860 28750 2381740 200 470 ...
```

Рис. 1 – структура data frame

За допомогою цієї функції можна побачити назви колонок, типи даних і перші значення.

3. Виправити помилки в даних

По структурі даних видно помилку у назві колонки, тому назва Populatiion була змінена на Population.

Для аналізу помилок в числових даних відсортуємо їх по зростанню.

```
> print(head(data$GDP.per.capita[order(data$GDP.per.capita)]))
[1] -6722.2235   285.7274   300.3077   382.0693   382.2132   401.7423
> print(head(data$Population[order(data$Population)]))
[1] 11097 13049 21503 30661 31949 33203
> print(head(data$CO2.emission[order(data$CO2.emission)]))
[1]  11.001  44.004  47.671  62.339 102.676 113.677
> print(head(data$Area[order(data$Area)]))
[1] -676590.0       2.0      10.0      20.0      30.0      30.3
```

Рис. 2 – сортування даних

По рис. 2 видно, що у data frame присутні від’ємні значення, тому колонки GDP.per.capita і Area було взято по модулю.

4. Побудувати діаграми розмаху та гістограми

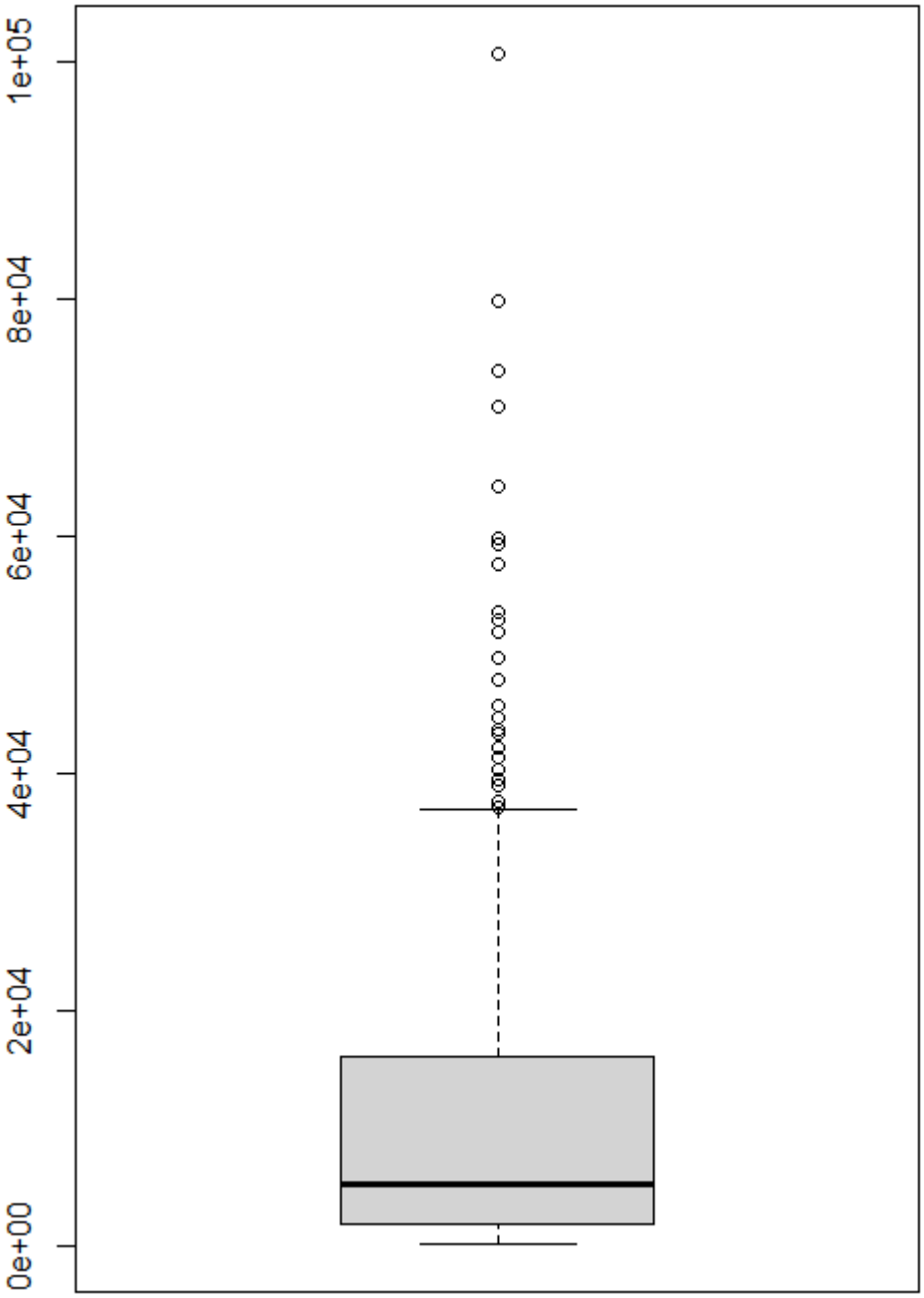


Рис. 3 – діаграма розмаху по колонці GDP.per.capita

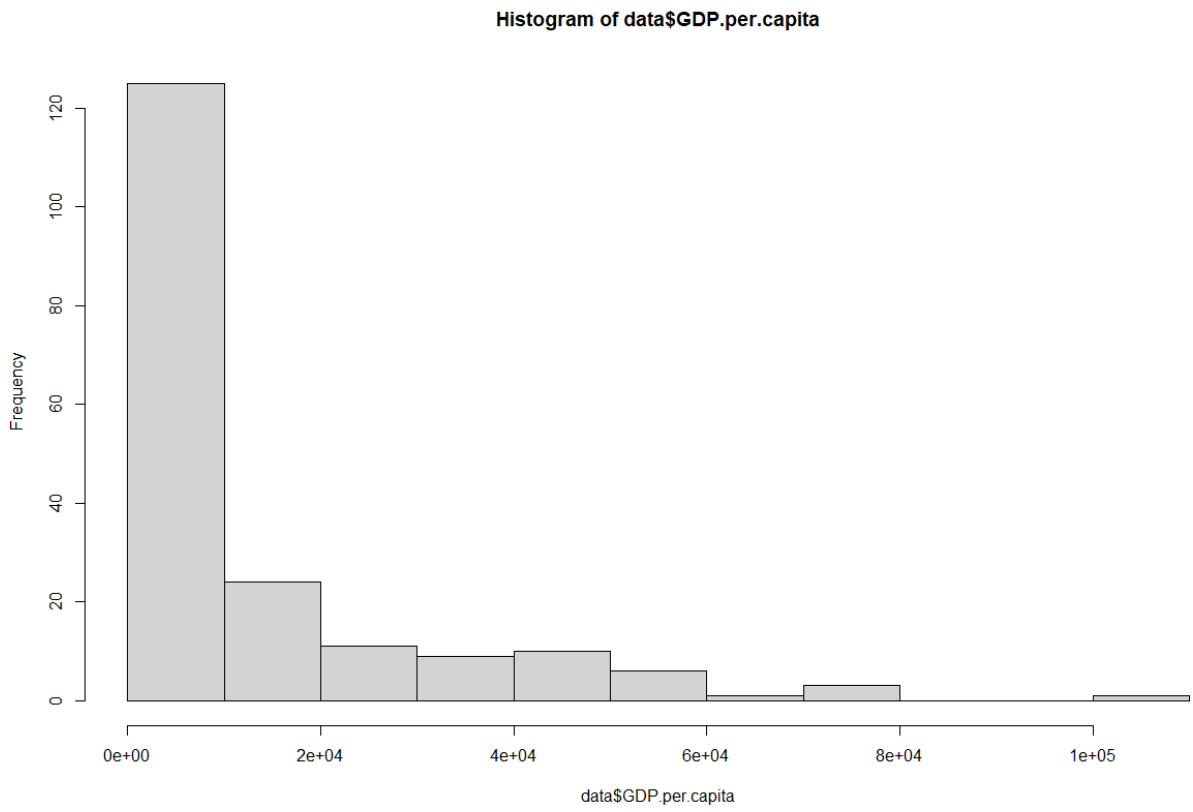


Рис. 4 – гістограма по колонці GDP.per.capita

5. Додати стовпчик із щільністю населення

У data frame було додано нову колонку, яка є часткою населення на площу країни.

| Population | CO2.emission | Area | Population.per.area |
|---|---|---|---|
| 34656032 | 9809.225 | 652860.0 | 5.308341e+01 |
| 2876101 | 5716.853 | 28750.0 | 1.000383e+02 |
| 40606052 | 145400.217 | 2381740.0 | 1.704890e+01 |
| 55599 | 165114.116 | 200.0 | 2.779950e+02 |
| 77281 | 462.042 | 470.0 | 1.644277e+02 |
| 28813463 | 34763.160 | 1246700.0 | 2.311179e+01 |
| 100963 | 531.715 | 440.0 | 2.294614e+02 |
| 43847430 | 204024.546 | 2780400.0 | 1.577019e+01 |
| 2924816 | 5529.836 | 29740.0 | 9.834620e+01 |
| 104822 | 872.746 | 180.0 | 5.823444e+02 |
| 24127159 | 361261.839 | 7741220.0 | 3.116713e+00 |
| 8747358 | 58712.337 | 83879.0 | 1.042854e+02 |

Рис. 5 – значення нової колонки густоти населення

**Виконання додаткового завдання.**

1. Чи є пропущені значення? Якщо є, замінити середніми

Дослідимо колонки на пропущені значення:

```
> length(data$Country.Name[is.na(data$Country.Name)])
[1] 0
> length(data$Region[is.na(data$Region)])
[1] 0
> length(data$GDP.per.capita[is.na(data$GDP.per.capita)]) #пропущені значення
[1] 27
> length(data$Population[is.na(data$Population)]) #пропущені значення
[1] 1
> length(data$CO2.emission[is.na(data$CO2.emission)]) #пропущені значення
[1] 12
> length(data$Area[is.na(data$Area)])
[1] 0
```

Рис. 6 – кількість пустих записів по колонкам

У колонках GDP.per.capita, Population і CO2.emission є пропущені значення, тому вони будуть замінені на середні.

2. Яка країна має найбільший ВВП на людину (GDP per capita)? Яка має найменшу площу?

```
> head(data[order(data$GDP.per.capita, decreasing = TRUE),], 1)
    Country.Name               Region GDP.per.capita Population CO2.emission Area
116   Luxembourg Europe & Central Asia       100738.7     582972     9658.878 2590
> head(head(data[order(data$Area),]), 1)
    Country.Name               Region GDP.per.capita Population CO2.emission Area
131       Monaco Europe & Central Asia       13445.59      38499     165114.1    2
```

Рис. 7 – країна з найбільшим ВВП та країна з найменшою площею

Країна з найбільшим ВВП – Люксембург, з найменшою площею – Монако.

3. В якому регіоні середня площа країни найбільша?

Для дослідження середньої площі регіону data frame треба групувати за регіонами та знаходити середнє значення кожного регіону.

```
> group <- data%>%group_by(Region)%>%summarise(avArea = mean(Area))
> head(as.data.frame(group), 1)
              Region   avArea
1 East Asia & Pacific 669979.9
```

Рис. 8 – регіон за найбільшою середньою площею країн

Отже, це регіон East Asia & Pacific.

4. Знайдіть країну з найбільшою щільністю населення у світі? У Європі та центральній Азії?

Країна з найбільшою щільністю населення у світі визначається простим сортуванням, для Європи і центральної Азії потрібно використовувати фільтрування.

```
> head(datanew[order(datanew$Population.per.area, decreasing = TRUE),], 1)
        Country.Name             Region GDP.per.capita Population CO2.emission Area Population.per.area
117 Macao SAR, China East Asia & Pacific       74017.18     612167      1283.45 30.3            20203.53
> head(filter(datanew[order(datanew$Population.per.area, decreasing = TRUE),], datanew$Region == "Europe & Central Asia"), 1)
  Country.Name               Region GDP.per.capita Population CO2.emission Area Population.per.area
1       Monaco Europe & Central Asia       13445.59      38499     165114.1    2             19249.5
```

Рис. 9 – країни з найбільшою щільністю населення

Отже, Макао має найбільшу щільність населення у світі, а Монако – у Європі та центральній Азії.

5. Чи співпадає в якомусь регіоні середнє та медіана ВВП?

Під час виконання цього завдання розглядається середнє і медіана окремих регіонів, а не всього data frame. Для вирішення задачі було згруповано дані по регіонам і знайдено медіану та середнє ВВП. Регіонів, у яких медіана та середнє співпадають, знайдено не було.

6. Вивести топ 5 країн та 5 останніх країн по ВВП та кількості CO2 на душу населення.

Для виконання завдання було створено нову колонку з кількістю CO2 на душу населення

```
> datanew <- cbind(data, Population.per.area = data$Population / data$Area, CO2.per.capita = data$CO2.emission / data$Population)
> head(datanew[order(datanew$GDP.per.capita,datanew$CO2.per.capita, decreasing = c(TRUE, TRUE)),], 5)
        Country.Name             Region GDP.per.capita Population CO2.emission     Area Population.per.area CO2.per.capita
116       Luxembourg Europe & Central Asia      100738.68     582972     9658.878   2590.0           225.08571     0.016568339
189      Switzerland Europe & Central Asia       79887.52    8372098    35305.876  41290.0           202.76333     0.004217088
117 Macao SAR, China   East Asia & Pacific       74017.18     612167     1283.450     30.3         20203.53135     0.002096568
147           Norway Europe & Central Asia       70868.12    5232929    47626.996 385178.0            13.58574     0.009101403
93           Ireland Europe & Central Asia       64175.44    4773095    34066.430  70280.0            67.91541     0.007137178
> tail(datanew[order(datanew$GDP.per.capita,datanew$CO2.per.capita, decreasing = c(TRUE, TRUE)),], 5)
               Country.Name             Region GDP.per.capita Population CO2.emission   Area Population.per.area CO2.per.capita
119              Madagascar Sub-Saharan Africa       401.7423   24894551     3076.613 587295            42.38849   1.235858e-04
38  Central African Republic Sub-Saharan Africa       382.2132    4594621      300.694 622980             7.37523   6.544479e-05
135              Mozambique Sub-Saharan Africa       382.0693   28829476     8426.766 799380            36.06480   2.922969e-04
120                  Malawi Sub-Saharan Africa       300.3077   18091575     1276.116 118480           152.69729   7.053648e-05
32                  Burundi Sub-Saharan Africa       285.7274   10524117      440.040  27830           378.15728   4.181253e-05
```

Рис. 10 – топ 5 країн та 5 останніх країн

Через те, що ВВП не співпадає у десятці країн, CO2 на душу населення не впливає на сортування.

**Висновок.**

При виконанні комп’ютерного практикуму було досліджено дані файлу Data2.csv. При виконанні роботи було записано дані у data frame; досліджено його структуру; визначено та виправлено помилки (від’ємні значення даних та назва колонки); побудовано діаграму розмаху та гістограму; додано стовпчик із щільністю населення, що є часткою кількості населення на площу країн. Для виконання додаткових завдань було

знайдено пропущенні значення та заповнено їх середніми; знайдено країни з найбільшим ВВП (Люксембург) та найменшою площею (Монако); згруповано регіони для знаходження одного з найбільшою середньою площею країн (East Asia & Pacific); знайдено країну з найбільшою щільністю населення у світі (Макао) та Європі й центральній Азії (Монако); досліджено, чи співпадає в якомусь регіоні середнє та медіана ВВП (внаслідок групування і фільтрування результат виявився негативним); виведено топ 5 країн та 5 останніх країн по ВВП та кількості CO2 на душу населення (рисунок 10).