Київський національний університет
імені Т.Шевченка

# Звіт

до лабораторної роботи №1
на тему:

# *«Попередня обробка та кореляційний аналіз даних»*

***Студента третього курсу***
***Групи ДО-3***
***Факультету комп'ютерних наук***
***та кібернетики***

***Київ***
*2019*

# Вхідні данні

Для виконання лабораторної роботи було використано набір даних з ресурсу https://www.kaggle.com/tanuprabhu/population-by-country-2020. Даний набір містить наступні статистичні дані про країни світу (всього 235 країн): населення, його зміна в процентовому та абсолютному предаставлення, щільність, площа, кількість мігранітв, процент народжуваності, середній вік життя, процент міського населення та процентова частка площі окремо взятої країни.

# Хід роботи

## Попередня обробка

Постановка задачі: провести попередню обробку обраного набору даних. Для цього для кожної скалярної змінної, що обробляються:

- дати її класифікацію,
- графічно представити (емпіричну функцію щільності)/(полігон частот),
- побудувати зображення "скринька з вусами",
- підрахувати вибіркові значення: мінімального та максимального спостережень вибірки, медіани, квартилів, децилів,
- підрахувати вибіркові значення усіх характеристик положення центру значень, з якими Ви були ознайомлені і не тільки (тобто також тих характеристик, які на Вашу думку будуть корисні при подальшому аналізі цих даних),
- підрахувати вибіркові значення усіх характеристик розсіювання значень, з якими Ви були ознайомлені і не тільки (тобто також тих характеристик, які на Вашу думку будуть корисні при подальшому аналізі цих даних),
- провести аналіз скошеності та гостроверхості розподілу,
- провести інші процедури попереднього аналізу, які Ви вважаєте доцільними і будуть корисні при подальшому аналізі цих даних,
- сформулювати висновки по передньому аналізу.

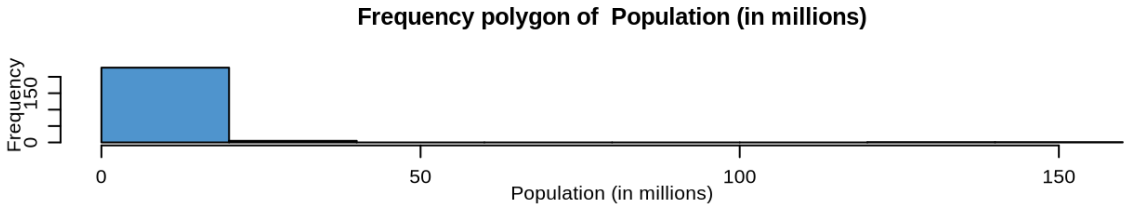
P.S. Усі графічні зображення повинні мати усі відповідні надписи.

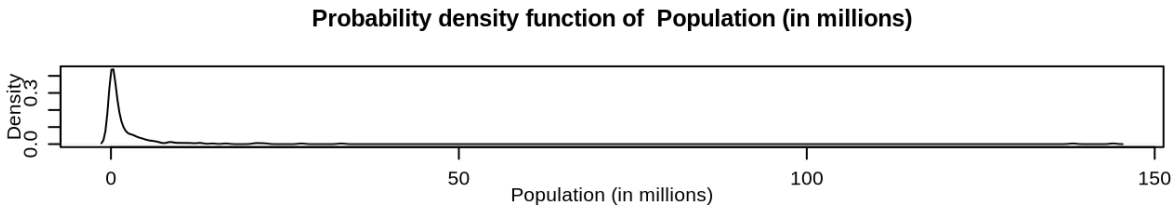
Виконання:

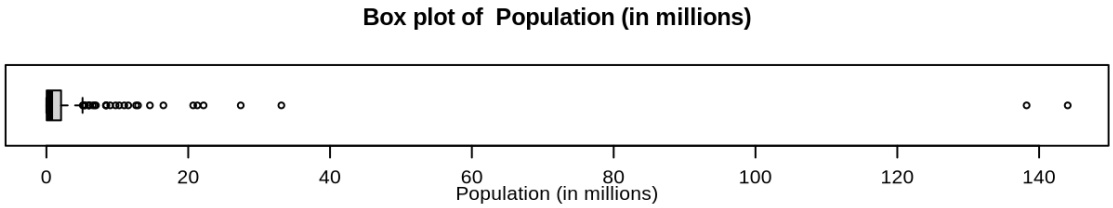
Для дослідження було обрано 4 скалярних змінних: площа (Area), населення (Population), кількіть мігрантів (Migrants) та середній вік (Age median). Кожна з цих змінних є кількісною. Для графічного представлення даних було побудовано було обрано полігон частот та графік щільності. Також для кожної із змінних були обраховані усі вибіркові значення та характеристики, котрі можуть знадобитися в подальшому дослідженні, а саме: мінімальне та максимальне спостереження вибірки, медіану, квартиль, децль, математичне сподівання, геометричне середнє, гармонічне середнє, мода, дисперсію, стандартне відхилення, коефіцієнт варіації, розмах вибірки

та інтервал концентрації розподілу. Також було проведено аналіз скошеності та гостроверхності розподілу.

Результати попередньої обробки даних про населення:

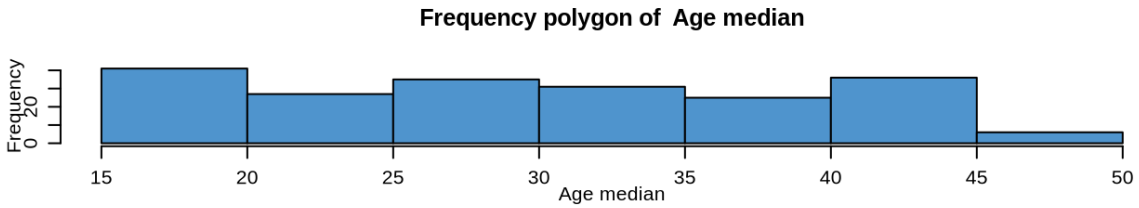


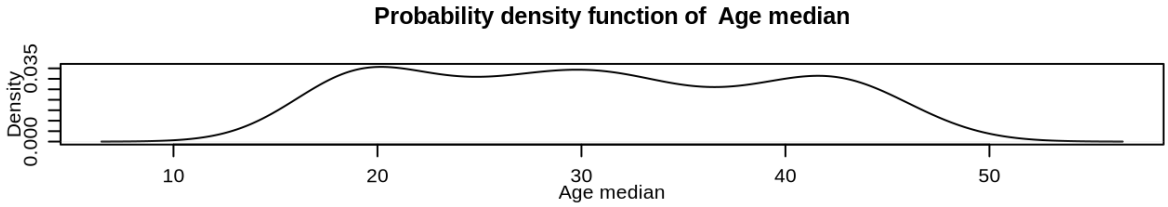



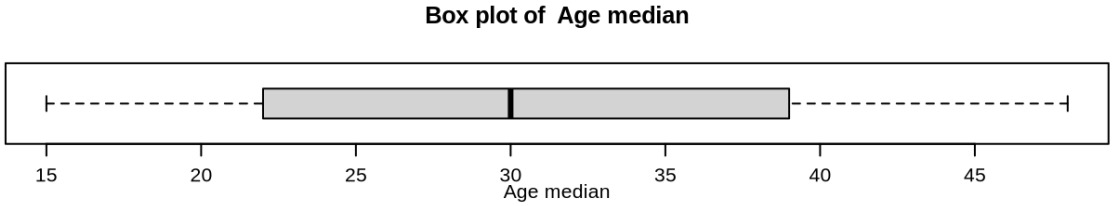




```
## [1] "Min of  Population (in millions) : 8.01e-05"
## [1] "Max of  Population (in millions) : 144.0297825"
## [1] "Median of  Population (in millions) : 0.5460109"
## [1] "Quantile of  Population (in millions) :"
##          0%          25%          50%          75%         100%
##   0.00008010   0.03994905   0.54601090   2.06716995 144.02978250
## [1] "Decile of  Population (in millions) :"
##        10%        20%        30%        40%        50%        60%        70%
## 0.00506432 0.01956692 0.07758982 0.27942520 0.54601090 0.96063792 1.67073338
##        80%        90%
## 3.01846600 5.97821300
## [1] "Expected value of  Population (in millions) : 3.32274442808511"
## [1] "Geometric mean of  Population (in millions) : 0.28132078145728"
## [1] "Harmonic mean of  Population (in millions) : 0.00510649911300835"
## [1] "Mode value of  Population (in millions) : 144.0297825"
## [1] "Median value of  Population (in millions) : 0.5460109"
## [1] "Variance value of  Population (in millions) : 183.070204691465"
## [1] "Root mean square of  Population (in millions) : 13.9043810749782"
## [1] "Coefficient of variation of  Population (in millions) : 407.203868445292"
## [1] "Range of  Population (in millions) : 144.0297825"
## [1] "Distribution concentration interval of  Population (in millions) : ( -37.2682871210538 ,  43.913775977224 )"
## [1] "Skewness of  Population (in millions) : 9.1098017555071 [ Positive skew ]"
## [1] "Kurtosis of  Population (in millions) : 89.0271720252694 [ More sharpness that normal distribution ]"
```

Результати попередньої обробки даних про середній вік:

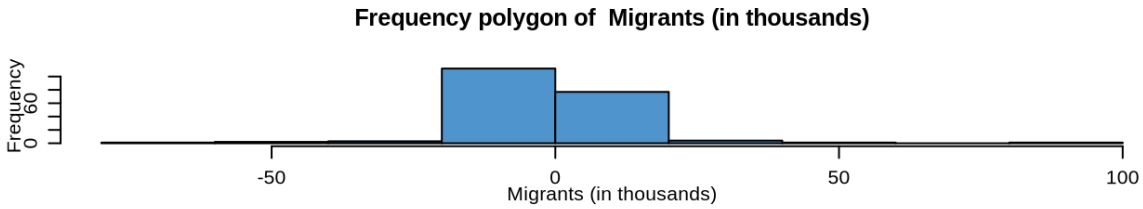


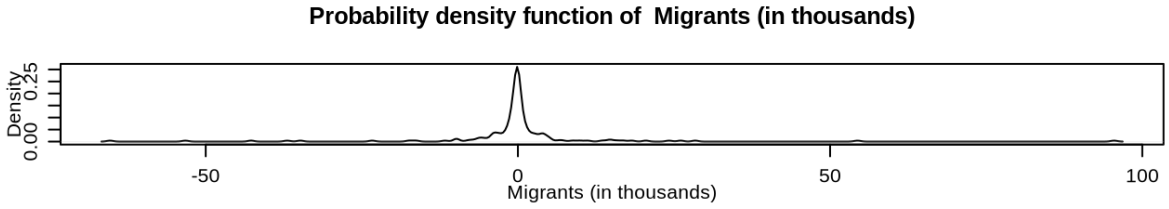



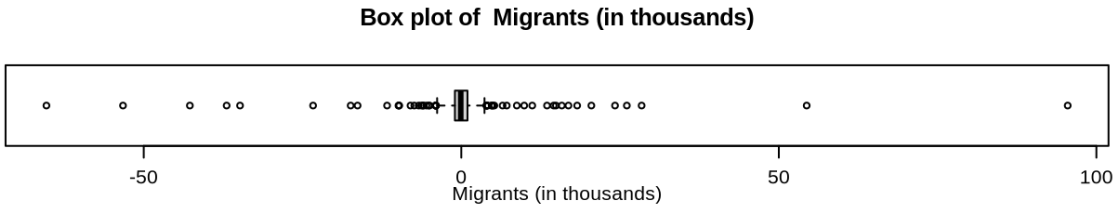




```
## [1] "Min of  Age median : 15"
## [1] "Max of  Age median : 48"
## [1] "Median of  Age median : 30"
## [1] "Quantile of  Age median :"
##   0%  25%  50%  75% 100%
##   15   22   30   39   48
## [1] "Decile of  Age median :"
## 10% 20% 30% 40% 50% 60% 70% 80% 90%
##  19  20  24  28  30  33  37  41  43
## [1] "Expected value of  Age median : 30.6069651741294"
## [1] "Geometric mean of  Age median : 29.2038966365248"
## [1] "Harmonic mean of  Age median : 27.8011410481483"
## [1] "Mode value of  Age median : 19"
## [1] "Median value of  Age median : 30"
## [1] "Variance value of  Age median : 83.3197512437811"
## [1] "Root mean square of  Age median : 31.9326093874046"
## [1] "Coefficient of variation of  Age median : 29.823163740099"
## [1] "Range of  Age median : 48"
## [1] "Distribution concentration interval of  Age median : ( 3.2230691548623 ,  57.9908611933964 )"
## [1] "Skewness of  Age median : 0.109844356267583 [ Positive skew ]"
## [1] "Kurtosis of  Age median : -1.26878207902588 [ Less sharpness that normal distribution ]"
```

Результати попередньої обробки даних про кількість мігрантів:

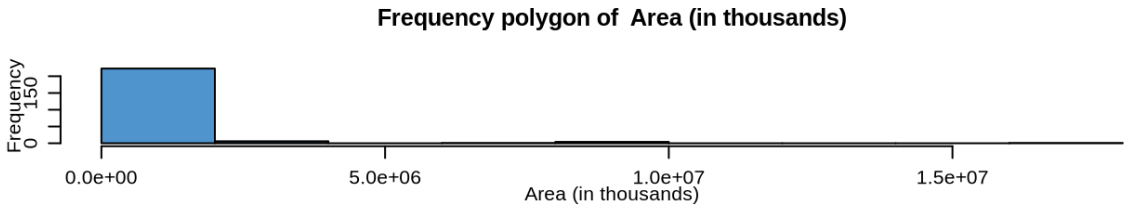


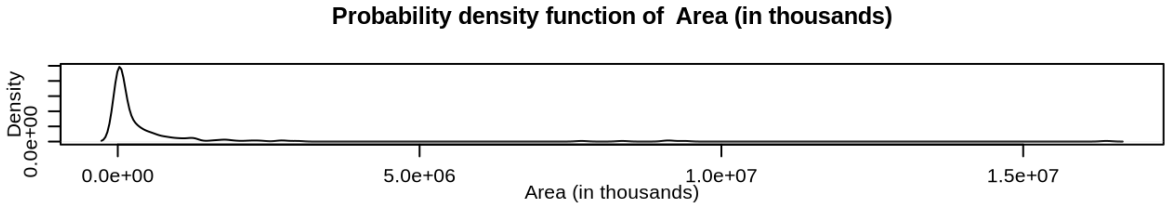



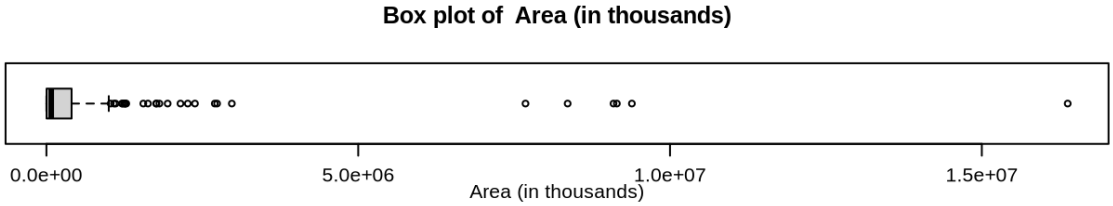




```
## [1] "Min of  Migrants (in thousands) : -65.3249"
## [1] "Max of  Migrants (in thousands) : 95.4806"
## [1] "Median of  Migrants (in thousands) : -0.0852"
## [1] "Quantile of  Migrants (in thousands) :"
##       0%      25%      50%      75%     100%
## -65.3249  -1.0047  -0.0852   0.9741  95.4806
## [1] "Decile of  Migrants (in thousands) :"
##     10%     20%     30%     40%     50%     60%     70%     80%     90%
## -5.0000 -1.8000 -0.8001 -0.4000 -0.0852  0.0320  0.4200  1.6000  4.7800
## [1] "Expected value of  Migrants (in thousands) : 0.000628358208955159"
## [1] "Geometric mean of  Migrants (in thousands) : NaN"
## [1] "Harmonic mean of  Migrants (in thousands) : -0.802308684905996"
## [1] "Mode value of  Migrants (in thousands) : 0"
## [1] "Median value of  Migrants (in thousands) : -0.0852"
## [1] "Variance value of  Migrants (in thousands) : 152.008895351842"
## [1] "Root mean square of  Migrants (in thousands) : 12.2984809056127"
## [1] "Coefficient of variation of  Migrants (in thousands) : 1962127.42653836"
## [1] "Range of  Migrants (in thousands) : 95.4806"
## [1] "Distribution concentration interval of  Migrants (in thousands) : ( -36.9869379062342 ,  36.9881946226521 )"
## [1] "Skewness of  Migrants (in thousands) : 1.30743669021505 [ Positive skew ]"
## [1] "Kurtosis of  Migrants (in thousands) : 24.4240323409915 [ More sharpness that normal distribution ]"
```

Результати попередньої обробки даних про середній вік життя:



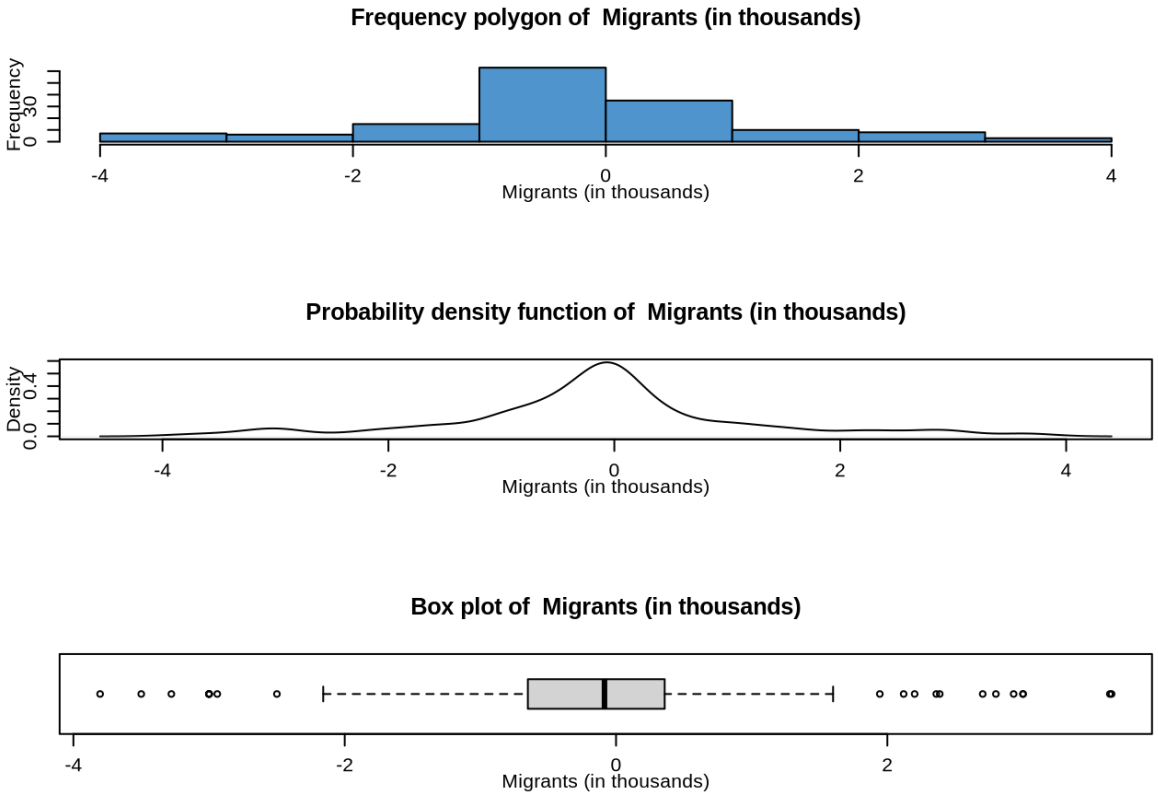




```
## [1] "Min of  Age median : 15"
## [1] "Max of  Age median : 48"
## [1] "Median of  Age median : 30"
## [1] "Quantile of  Age median :"
##   0%  25%  50%  75% 100%
##   15   22   30   39   48
## [1] "Decile of  Age median :"
## 10% 20% 30% 40% 50% 60% 70% 80% 90%
##  19  20  24  28  30  33  37  41  43
## [1] "Expected value of  Age median : 30.6069651741294"
## [1] "Geometric mean of  Age median : 29.2038966365248"
## [1] "Harmonic mean of  Age median : 27.8011410481483"
## [1] "Mode value of  Age median : 19"
## [1] "Median value of  Age median : 30"
## [1] "Variance value of  Age median : 83.3197512437811"
## [1] "Root mean square of  Age median : 31.9326093874046"
## [1] "Coefficient of variation of  Age median : 29.823163740099"
## [1] "Range of  Age median : 48"
## [1] "Distribution concentration interval of  Age median : ( 3.2230691548623 ,  57.9908611933964 )"
## [1] "Skewness of  Age median : 0.109844356267583 [ Positive skew ]"
## [1] "Kurtosis of  Age median : -1.26878207902588 [ Less sharpness that normal distribution ]"
```

В результаті візуального огляду отриманих результатів було вирішено зробити видалення аномальних спостережень. Оскільки аномальність даних очевидна, як критерій аномальності спостереження було використано найлегший підхід, а саме вважати наступне: все, що не потрапляє в діапазон $\left[\left(x_{25} - 1.5 \cdot (x_{75} - x_{25})\right), \left(x_{75} + 1.5 \cdot (x_{75} - x_{25})\right)\right]$ є аномальним спостереженням, котре потрібно видалити з набору даних (а точніше змінити на NA – ключове слово, що вказує на відсутність даних). Для прикладу,

розглянемо графічне представлення кількості мігранітв, після видалення аномальних спостережень.




Резюмуючи попередню обробку даних, можна сказаати наступне: усі обрані змінні мають нормальний розподіл, деякі зі змінних місять аномальні спостереження, котрі ми успішно вилучили. Аналіз скошеності показав, що усі змінні є скошенні ліворуч, а аналіз гостроверхності показав, що дані про населення, площу та кількість мігрантів є більш гостроверхними в порівнянні з нормальним розподілом, в той час, як дані про середній вік показали меншу гостроверхність відносно нормального розподілу.

## Кореляціний аналіз

Постановка задачі: провести кореляційний аналіз обраних для обробки скалярних змінних, яких потрібно взяти не менше трьох. Для цього:

- на основі результатів попередньої обробки обраного набору даних визначитися, які характеристики статистичного зв'язку потрібно використати при подальшому їх кореляційному аналізі,

- провести аналіз істотності парних статистичних зв'язків для усіх пар скалярних змінних, навівши для них:
    - вибіркове значення відповідної парної характеристики статистичного зв'язку,
    - максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим,
    - впорядковану послідовність усіх пар скалярних змінних у порядку спадання істотності статистичного зв'язку між ними,
    - сформулювати висновки по кореляційному аналізу парних статистичних зв'язків для обраного набору скалярних змінних,
- провести аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною та множиною усіх інших скалярних змінних (які виступають у ролі незалежних змінних), навівши для них:
    - вибіркове значення відповідної множинної характеристики статистичного зв'язку,
    - максимальний рівень значущості при якому відповідний множинний статистичний зв'язок не є значимим,
    - впорядковану послідовність усіх скалярних змінних у порядку спадання істотності множинного статистичного зв'язку їх з множиною усіх інших скалярних змінних,
    - сформулювати висновки по кореляційному аналізу множинних статистичних зв'язків для обраного набору скалярних змінних.

Виконання:

Для проведення аналізу істотності парних статистичних зв’язків було обрано наступні пари: населення - площа, середній вік життя – кількість мігрантів, населення – середній вік життя, площа – середній вік життя. Для проведення аналізу істотності множинних статистичних зв’язків було обрано наступні пари: населення – {площа, середній вік життя, кількість мігрантів}, площа – {площа населення середній вік життя, кількість мігрантів}, середній вік життя – {площа, населення, кількість мігрантів}, кількість мігрантів– {площа, середній вік життя, населення}. Отримали насупні результати.

Аналіз істотності парних статистичних зв'язків:

```
# Correlation between Population and Area
cor.test(population, area, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  population and area
## t = 10.907, df = 188, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5268710 0.7026205
## sample estimates:
##       cor
## 0.6225331
```

```
# Correlation between Age media and Migrants
cor.test(age_med, migrants_net, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  age_med and migrants_net
## t = 2.764, df = 145, p-value = 0.006452
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06414851 0.37213360
## sample estimates:
##       cor
## 0.2237184
```

```
# Correlation between Population and Age median
cor.test(population, age_med, method = "pearson")
```

```
# Correlation between Population and Age median
cor.test(population, age_med, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  population and age_med
## t = -2.2067, df = 171, p-value = 0.02867
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.30795162 -0.01763426
## sample estimates:
##        cor
## -0.1663967
```

```
# Correlation between Area and Age median
cor.test(area, age_med, method = "pearson")
```

```
# Correlation between Area and Age median
cor.test(area, age_med, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  area and age_med
## t = -3.9793, df = 172, p-value = 0.0001017
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4209361 -0.1479704
## sample estimates:
##        cor
## -0.2903486
```

Аналіз істотності множинних статистичних зв'язків:

```
# Correlation between Population and {Area, Age median, Migrants}
xPopulation <- lm(population~area+age_med+migrants_net)
cor.test(xPopulation$model$population, xPopulation$fitted.values)
```

```
##
##  Pearson's product-moment correlation
##
## data:  xPopulation$model$population and xPopulation$fitted.values
## t = 8.6148, df = 127, p-value = 2.341e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4854194 0.7060509
## sample estimates:
##       cor
## 0.6073158
```

```
# Correlation between Area and {Population, Age median, Migrants}
xArea <- lm(area~population+age_med+migrants_net)
cor.test(xArea$model$area, xArea$fitted.values)
```

```
##
##  Pearson's product-moment correlation
##
## data:  xArea$model$area and xArea$fitted.values
## t = 9.1639, df = 127, p-value = 1.114e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5141142 0.7247226
## sample estimates:
##       cor
## 0.6309022
```

```
# Correlation between Age median and {Area, Population, Migrants}
xAge <- lm(age_med~population+area+migrants_net)
cor.test(xAge$model$age_med, xAge$fitted.values)
```

```
##
##  Pearson's product-moment correlation
##
## data:  xAge$model$age_med and xAge$fitted.values
## t = 4.6755, df = 127, p-value = 7.38e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2252796 0.5215205
## sample estimates:
##       cor
## 0.3832115
```

```
# Correlation between Migrants and {Area, Age median, Population}
xMigrants <- lm(migrants_net~population+area+age_med)
cor.test(xMigrants$model$migrants_net, xMigrants$fitted.values)
```

```
##
##  Pearson's product-moment correlation
##
## data:  xMigrants$model$migrants_net and xMigrants$fitted.values
## t = 3.3656, df = 127, p-value = 0.001011
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1192047 0.4373822
## sample estimates:
##       cor
## 0.2861627
```

Резюмуючи результати кореляційного аналізу, можна сказати наступне: існує сильна залежність між населенням та площею, помірна залежність між середнім віком життя та кількістю мігрантів, слабка обернена залежність між населенням та середнім віком життя, помірна обернена залежність між площею та середнім віком життя. Щодо істотності множинних статистичних зв’язків, то населення та площа у якоті незалежної змінної, показали сильну залежність, в свою чергу середній вік та кількість мігранітв показалу помірну залежність у якості незалежних змінних. Для того, щоб нульова гіпотеза щодо відсутності статичтичного зв’язку відкидалася, потрібно, щоб p-value було більшим за заданий наперед рівень значущості, отже в нашому випадку отримане p-value може вважатися максимальними рівнем значущості, при якому приймається гіпотеза про віднусність статистичного зв’язку. Щодо впорядкованості пар у порядку спадання істотності статистичного зв'язку між ними, отримали наступне:

1. населення – площа
2. площа – середній вік життя
3. середній вік життя – кількість мігрантів
4. населення – середні вік життя

та

1. населення – {площа, середній вік життя, кількість мігрантів}
2. площа – {площа населення середній вік життя, кількість мігрантів}
3. середній вік життя – {площа, населення, кількість мігрантів}
4. кількість мігрантів– {площа, середній вік життя, населення}

# Висновок

В результаті виконання лабораторної роботи було здійснено попередню обробку та кореляційний аналіз даних про населення країн світу. Отримані результаті не суперечать здоровому глузду та виглядають цілком коректно. Для прграмної реалізації було освоєно та використано мову програмування R та середовище розробки RStudio

# Список джерел


- https://www.wikipedia.org
- https://rdocumentation.org
- https://cran.r-project.org
- http://www.r-tutor.com