ISSN 2078-6115. Вісник Львівського університету. Серія економічна. 2021. Випуск 61. С. 137–147
Visnyk of the Lviv University. Series Economics. 2021. Issue 61. P. 137–147

DOI: http://dx.doi.org/10.30970/ves.2021.61.0.6111

УДК 311.2
JEL C8

ПРО ПІДГОТОВКУ ДАНИХ ДО АНАЛІЗУ У СОЦІАЛЬНО-ЕКОНОМІЧНИХ ДОСЛІДЖЕННЯХ

Оксана Марець, Ольга Сисан

Львівський національний університет імені Івана Франка,
79008, м. Львів, просп. Свободи, 18
e-mail: oksana.marets@lnu.edu.ua; ORCID: 0000-0002-4044-7443
e-mail: olha.sysan@lnu.edu.ua; ORCID: 0000-0002-8097-3149

Анотація. *Соціально-економічні дослідження охоплюють питання рівня та якості життя, економічної активності, доступності та якості соціальних послуг, освіти тощо. Крім населення, об'єктом соціально-економічних досліджень є діяльність підприємств, регіональний і муніципальний розвиток, суспільна діяльність державних та громадських організацій тощо.*

Для проведення будь-якого дослідження потрібні певні джерела інформації – дані. Від правильно сформованої бази даних залежить не тільки легкість подальшої роботи, а й зміст кінцевих результатів. Під час аналізу даних відбуваються такі процеси: отримання даних, опрацювання, аналіз та інтерпретація результатів. Саме на опрацювання вхідних даних аналітик витрачає найбільше часу, адже це один із найважливіших і найбільш трудомістких процесів.

Отже, було поставлене питання, чи впливає попереднє опрацювання даних на поліпшення змісту та надійності статистичних звітів на підставі цих даних. Таке дослідження провели, щоб визначити етапи перетворення мікроданих вибіркового обстеження домогосподарств та узагальнити статистичні методи підготовки «сирих» даних у технічно коректні дані, які придатні до аналізу. Для цього застосовано різні способи очищення та опрацювання даних, зокрема методи виявлення та усунення нетипових значень, коректного імпорту даних, сортування рядків, модифікації типів даних та їхніх якісних складових. У статті описано технічні та предметні аспекти очищення даних. Для безпосереднього опрацювання даних використано мову програмування R, що є одним з найкращих інструментів для статистичних обчислень, аналізу та зображення даних у графічному вигляді. Технічні аспекти охоплюють зчитування даних, перетворення типів даних, зіставлення рядків і різні маніпуляції. Аспекти, які пов'язані з предметом, охоплюють такі теми: перевірка даних, локалізація помилок та імпутація значень.

За результатами дослідження сформульовано висновки, що такі дії, як видалення непотрібних значень, їх фільтрування, групування чи заміна некоректних даних, суттєво впливають на результати статистичного аналізу, адже опрацювання даних безпосередньо спрямоване на поліпшення змісту та надійності статистичних звітів на підставі цих даних.

Ключові слова: *перетворення даних, очищення даних, статистичний аналіз, опрацювання даних, методи підготовки даних.*

Постановка проблеми. В Україні, як і в будь-якій іншій країні світу, оцінюють ефективність й результативність соціальної та економічної політики, програм соціально-економічного розвитку на національному, регіональному та місцевому рівнях. За допомогою аналізу даних фахівці оцінюють вплив реалізованих і запланованих заходів політики на якість життя населення.

Проте з кожним роком ми накопичуємо все більше і більше інформації, тому зростає потреба у вмінні її правильно використовувати й інтерпретувати. Йдеться не лише про коректний аналіз даних, а й про їх якісну попередню підготовку перед дослідженням. Це стало визначальним у формулюванні мети цього дослідження. Мета дослідження – визначити етапи перетворення мікроданих вибіркового обстеження домогосподарств, узагальнення статистичних методів підготовки «сирих» даних у технічно коректні дані, які придатні до аналізу, задля поліпшення змісту та надійності статистичних звітів на підставі цих даних.

Аналіз останніх досліджень і публікацій. Актуальність цієї теми важко переоцінити, бо дані – це так звана сировина для бізнесу та будь-якої сфери. Без аналізу даних компанії не знатимуть чи в тому напрямі вони рухаються, що потрібно поліпшити чи змінити, як впливають ті чи інші чинники на розвиток бізнесу тощо. Тому не дивно, що на інтернет ресурсі Scopus – база даних, що містить понад 50 млн реферативних записів – за ключовими словами «data manipulation» знайдено 4 147 статей, а за результатами пошуку по «data cleaning» – понад 3 852.

Дані, зібрані з різних ресурсів, дослідники називають «безладними» (messy) [7]. Їх використання за принципом «сміття на вході – сміття на виході» дає спотворені результати. Обсяг даних, які збирають організації, збільшується з кожним роком, тому треба постійно вдосконалювати методи опрацювання даних [6]. Науковці не тільки пояснюють важливість правильного опрацювання даних за допомогою R, а й наводять приклади реального застосування очищення та аналізу великих баз даних з метою генерування нової інформації, зокрема для різних сфер: медичної, аграрної, хімічної, економічної, соціальної тощо [5]. Також автори наголошують на важливості виправлення нетипових значень (так званих викидів), які відіграють вирішальну роль в аналізі та моделюванні часових рядів. Вони порівнюють ефективність різних методів виявлення викидів та створювали свої метрики їхньої оцінки [3]. Проте варто зазначити, що методи та етапи опрацювання даних можуть бути різними, оскільки вони залежать від досліджуваного об’єкта. Тому перед попереднім опрацюванням даних доцільно вивчати їхній зміст, значення, одиниці виміру, взаємозв’язки тощо [7].

Постановка завдання. Відповідно до досліджень попередників і власних спостережень можна дійти висновку, що процес перетворення вхідних (так званих сирих) даних у технічно коректні дані, які можна проаналізувати, є необхідним під час аналізу соціально-економічних даних. Часто початкові дані, з якими працює людина, мають неправильний формат, відсутні значення, потрібні для аналізу, неправильно написані заголовки або містять помилки. Тому аналітик даних витрачає більшу частину свого часу на підготовку даних перед виконанням будь-якої статистичної операції. Мета статті – описати вищезазначені етапи підготовки даних до аналізу, показати, як

ISSN 2078-6115. Вісник Львівського університету. Серія економічна. 2021. Випуск 61

їх виконати у середовищі R та застосувати на реальних даних вибіркового обстеження домогосподарств [1].

Опрацювання даних спрямоване на поліпшення змісту та надійності статистичних звітів на підставі цих даних. Такі дії, як видалення непотрібних значень, їх фільтрування, групування чи заміна некоректних даних, суттєво впливають на результати статистичного аналізу. Кожного року можливості R значно розширюються додатковими бібліотеками, тому таке середовище чудово підходить для різних маніпуляцій з базами даних. Наприклад, пакети dplyr [8] та tidyr [11] за авторством відомого науковця-програміста Хедлі Вікхема сьогодні є одними з найпопулярніших пакетів мови R. Обидва пакети призначені для оперування даними: фільтрація, вибірка, сортування, перебудова таблиць. Вони мають доволі простий і зрозумілий синтаксис та характеризуються значною швидкодією.

На рис. 1 показано огляд типових аналітичних даних. Кожен прямокутник подає дані в певному стані, а кожна стрілка представляє діяльність, яку треба виконати, щоб перейти до наступного етапу.

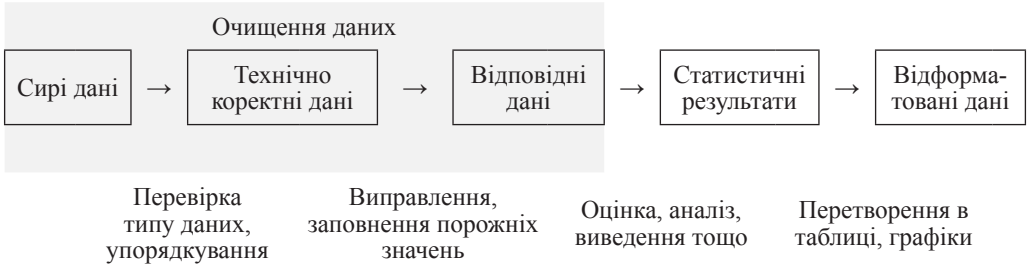


Рис. 1. Етапи роботи з даними для статистичного аналізу
Джерело: побудовано автором за [4].

Перший етап – сирі дані (raw data) – це дані у початковому вигляді, в якому надходять. Сирі дані можуть не мати заголовків, містити неправильні типи даних (наприклад, числа, які класифікуються як текст), невідомі чи незрозумілі кодування символів тощо. Тобто, читати такі файли у R безпосередньо або важко, або неможливо без попереднього опрацювання. Перетворення сирих даних у технічно коректні дані складається з [7]:

1) правильного завантаження («зчитування») текстових даних у R data.frame;
2) перетворення типів даних;
3) перетворення, пов'язані з якісними даними.

Після попереднього опрацювання дані можна вважати технічно правильними або коректними. Тобто, в такому стані дані можна без зайвих проблем прочитати в R-data.frame (таблицю з даними) з правильними назвами, форматом і заголовками. Однак це ще не означає, що дані вже можна аналізувати. Наприклад, значення вікової змінної може бути від'ємним або дані можуть бути просто відсутні. Такі невідповідності, очевидно,

залежать від предмета, до якого належать дані. Потрібно усунути ці недоліки для того, щоб майбутній статистичний висновок був коректним.

Відповідні дані – це етап, коли дані готові до аналізу і створення статистичних висновків. Саме їх використовують як вихідну точку в статистичних дослідженнях чи аналітичних роботах. Процес отримання відповідних даних передбачає такі три кроки.

1. Виявлення невідповідностей. Тут аналітик визначає, які обмеження порушені. Наприклад, географічні координати об'єктів у певному географічному регіоні помилково вказують на об'єкти з інших регіонів.
2. Вибір поля або полів, що спричиняють невідповідність. Цей процес може бути більш громіздким, особливо коли передбачається виявлення взаємозв'язків між змінними. Наприклад, сімейний стан дитини не може бути «одружений». У разі порушення не одразу зрозуміло, де саме помилка: у змінній вік, сімейний статус чи обидвох.
3. Виправлення полів, які визнані помилковими.

За допомогою спеціальних пакетів R можна легко і швидко виконати всі вищезгадані кроки. Статистичне середовище R містить хороший інструментарій для роботи з даними, бо всі дії з очищення можуть бути не тільки написані і збережені в коді, а й відтворені або редаговані з іншим масивом даних.

Після всіх маніпуляцій отримані дані можна зберігати для повторного використання і, нарешті, результати можна відформатувати (перетворені у таблиці, графіки) для статистичних звітів, публікацій, аналітичних досліджень тощо.

Виклад основного матеріалу дослідження. Робота з перетворенням даних розпочинається з їх завантаження та дослідження. Для того, щоб розпочати працювати у середовищі R, треба завантажити відповідні пакети. Працюватимемо з бібліотекою dplyr [8] та readr [9].

Потрібно дізнатись скільки рядків, стовпців містить наш набір даних, формат даних, а якщо дані кількісні (наприклад, річний заробіток домогосподарства) – глянути мінімальне, максимальне, NA значення, медіану, середнє значення, перший і третій квартилі.

Отож за допомогою функції read_delim завантажуємо дані [1] і присвоюємо їм назву:

```
microdata20<- read_delim(«Households_microdani_anonimni_2020.txt»)
```

Далі, щоб мати уявлення про вигляд наших даних, їх треба дослідити. Зробимо це за допомогою функцій summary() і str(). Результат функції summary() зображений на рис. 2. У консолі отримуємо таку інформацію:

```
Rows: 7849 Columns: 123
```

Це означає, що кожне з 7849 домогосподарств описали за 123 характеристиками (рік обстеження, тип населеного пункту, загальний дохід, сукупні витрати, наявність гарячого водопостачання тощо).

ISSN 2078-6115. Вісник Львівського університету. Серія економічна. 2021. Випуск 61

```
> summary(microdata20)
    rik_fa_1       kvart_kd    code_fam           w_q           tp_ns_p         cod_obl          hsize
 Min.   :2020   Min.   :4   Min.   :100001   Min.   :  100   Min.   :0.000   Min.   : 0.00   Min.   :0.000
 1st Qu.:2020   1st Qu.:4   1st Qu.:102241   1st Qu.:  497   1st Qu.:1.000   1st Qu.:23.00   1st Qu.:1.000
 Median :2020   Median :4   Median :104444   Median : 1109   Median :2.000   Median :48.00   Median :2.000
 Mean   :2020   Mean   :4   Mean   :104442   Mean   : 1884   Mean   :2.092   Mean   :44.57   Mean   :2.013
 3rd Qu.:2020   3rd Qu.:4   3rd Qu.:106656   3rd Qu.: 2251   3rd Qu.:3.000   3rd Qu.:63.00   3rd Qu.:3.000
 Max.   :2020   Max.   :4   Max.   :108976   Max.   :12030   Max.   :3.000   Max.   :80.00   Max.   :5.000
    type_dom          gnd          cashinc           totalinc          totalres          cashexp
 Min.   :1.000   Min.   :1.000   Length:7849        Length:7849        Length:7849        Length:7849
 1st Qu.:1.000   1st Qu.:2.000   Class :character   Class :character   Class :character   Class :character
 Median :2.000   Median :3.000   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :1.747   Mean   :3.611
 3rd Qu.:2.000   3rd Qu.:5.000
 Max.   :2.000   Max.   :6.000
   totalexp           d_tin            d_tex           h0111              h0112              h0113
 Length:7849        Min.   :0.0000   Min.   :0.0000   Length:7849        Length:7849        Length:7849
 Class :character   1st Qu.:0.0000   1st Qu.:0.0000   Class :character   Class :character   Class :character
 Mode  :character   Median :0.0000   Median :0.0000   Mode  :character   Mode  :character   Mode  :character
                    Mean   :0.2775   Mean   :0.2221
                    3rd Qu.:1.0000   3rd Qu.:0.0000
                    Max.   :1.0000   Max.   :1.0000
```

Рис. 2. Стовпці набору даних «Households_microdani_anonimni_2020» та їхні характеристики
Джерело: власні розрахунки.

Перегляд результатів дає змогу зробити такі висновки щодо даних.

Всі дані мають або якісний, або числовий тип. Хоча у нас є стовпці (наприклад, Q_mcwav – наявність у домогосподарстві мікрохвилівки), які містять тільки 1 або 0, тобто так або ні. Для подальшої зручності треба змінити тип таких стовпців на «фактор». Це означає, що ці дані будуть записані у таблиці з варіантами та їх частотою. Також стовпці, що містять дані про доходи, мають якісний тип даних. В подальшому ми не зможемо проаналізувати середній дохід домогосподарств, найбільший, найменший прибуток тощо. Тому тип цих змінних потрібно змінити на числовий.

Спробуємо знову імпортувати дані, але цього разу самі призначимо відповідні типи даних для стовпців.

Це можна зробити за допомогою коду:

```
microdata20<- read_delim(«Households_microdani_anonimni_2020.txt»,
col_types = cols(rik_fa_1 = col_character(), kvart_kd = col_factor(),
code_fam = col_character(), w_q = col_factor(), tp_ns_p = col_factor(),
cod_obl = col_character(), hsize = col_factor(), type_dom = col_factor(),
gnd = col_factor(), cashinc = col_number(), totalinc = col_number(),
totalres = col_number(), cashexp = col_number(), totalexp = col_number(),
d_tin = col_factor(), d_tex = col_factor(), h0111 = col_character(), h0112
= col_character(), h0113 = col_character(), h0114 = col_character()))
```

Структура даних показана на рис. 3.

```
> summary(microdata20)
   rik_fa_1         kvart_kd   code_fam              w_q          tp_ns_p    cod_obl          hsize    type_dom gnd
 Length:7849        4:7849   Length:7849        100    : 296   1:2458   Length:7849        1:2861   2:5864   3:2522
 Class :character            Class :character   12000  :  71   2:2155   Class :character   2:2612   1:1985   6:1223
 Mode  :character            Mode  :character   488    :  52   3:3218   Mode  :character   4: 686             2:1932
                                                439    :  30   0:  18                      0: 268             1: 272
                                                3944   :  30                               3:1071             5:1701
                                                776    :  29                               5: 351             4: 199
                                                (Other):7341
    cashinc          totalinc        totalres         cashexp          totalexp      d_tin     d_tex       h0111
 Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  3.00   Min.   :  3.0   0:5671   0:6106   Length:7849
 1st Qu.: 56.0   1st Qu.: 61.0   1st Qu.: 61.0   1st Qu.: 48.00   1st Qu.: 54.0   1:2178   1:1743   Class :character
 Median : 91.0   Median : 99.0   Median :100.0   Median : 74.00   Median : 82.0                     Mode  :character
 Mean   :109.9   Mean   :117.4   Mean   :119.1   Mean   : 87.94   Mean   : 95.9
 3rd Qu.:144.0   3rd Qu.:153.0   3rd Qu.:154.0   3rd Qu.:112.00   3rd Qu.:122.0
 Max.   :845.0   Max.   :845.0   Max.   :809.0   Max.   :803.00   Max.   :885.0
```

Рис. 3. Стовпці набору даних «Households_microdani_anonimni_2020» та їхні характеристики після зміни типу даних
Джерело: власні розрахунки.

Отже, тепер тип даних коректний. Проте на цьому етапі ще потребують виправлення заголовки стовпців. На рис. 4 можна побачити, що не всі назви стовпців записані з малої літери. Тому для зручності варто перетворити їх у однаковий регістр. За допомогою коду ми змінюємо регістр «to lower» – тобто на нижній:

```
names(microdata20) <- tolower(names(microdata20))
```

gasbal	elektrpl	bath	hteleph	landplot	poultry	Q_refr	Q_froz	Q_wash	Q_vacum	Q_tv_cl	Q_compu	Q_mcwav
2	1	2	1	1	2	1	2	1	1	1	2	1

gasbal	elektrpl	bath	hteleph	landplot	poultry	q_refr	q_froz	q_wash	q_vacum	q_tv_cl	q_compu	q_mcwav
2	1	2	1	1	2	1	2	1	1	1	2	1

Рис. 4. Стовпці набору даних «Households_microdani_anonimni_2020» до та після зміни регістру заголовків
Джерело: власні розрахунки.

Отже, дані уже придатні до роботи, проте можуть містити нетипові значення, що призведуть до викривлення майбутнього результату. Тому переходимо до наступного етапу: виявлення і видалення нетипових значень. Найшвидший спосіб перевірити дані на наявність нетипових значень – побудувати графік. Завантажуємо ggplot2 [10] – пакет для створення графіків і візуалізації даних. Код матиме такий вигляд:

```
library(ggplot2)
ggplot(data, aes(tp_ns_p, cashexp_th)) +
  geom_point (alpha = 0.1) +
  geom_jitter(color = “gray”, size = 0.5) +
  geom_boxplot() +
  coord_flip() +
  xlab(«»)+
```

ISSN 2078-6115. Вісник Львівського університету. Серія економічна. 2021. Випуск 61

```
ylab("Грошові витрати, тис. грн") +
theme_light()
```

Діаграма зображена на рис. 5.

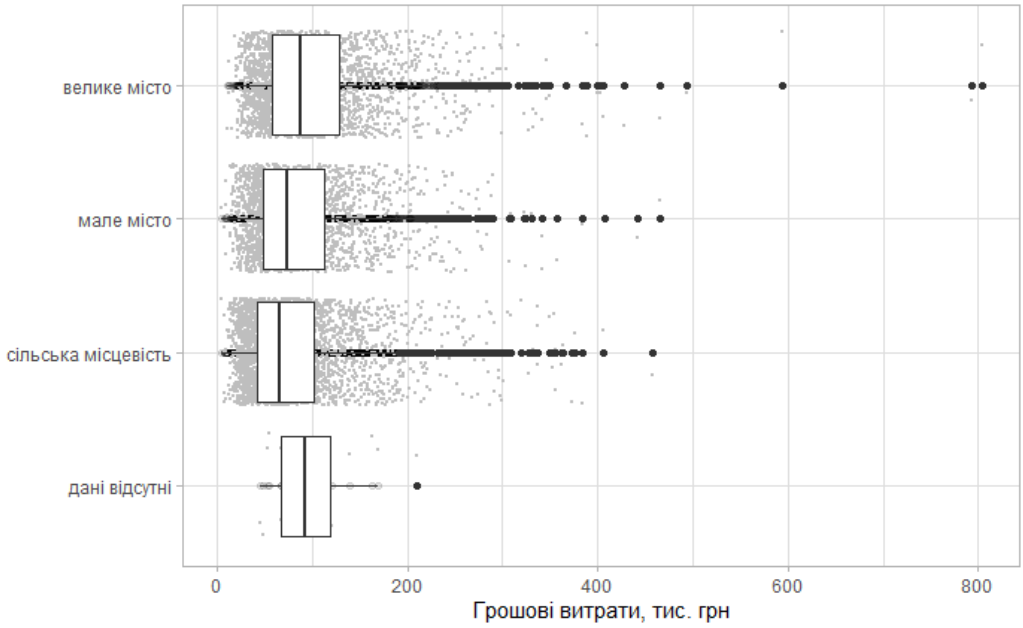


Рис. 5. Розподіл грошових витрат домогосподарств у 2000 році залежно від типу населеного пункту
Джерело: власні розрахунки.

Ми використали останні відредаговані дані і за допомогою geom_boxplot створили боксплот на підставі даних зі стовпців tp_ns_p та cashexp (рис. 5).

Також додали geom_jitter() – це зручний спосіб виділити кожну точку. Цей крок полегшує візуалізацію точок, які можуть накладатись одна на одну.

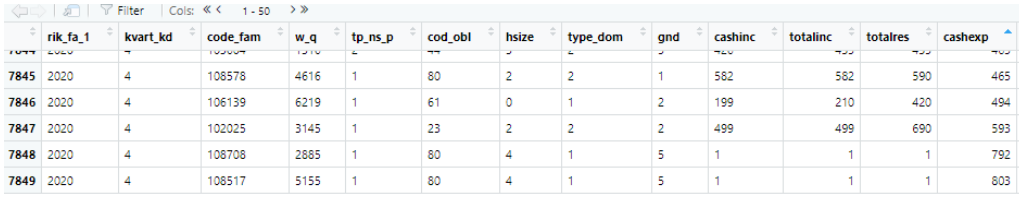
З боксплотів можна зробити висновок, що стовпець cashexp з даними грошових витрат домогосподарств містить доволі багато екстремальних значень, особливо за категорією 1 – домогосподарства у великих містах. Дослідимо найбільші значення та з'ясуємо, чи не є вони помилковими.

```
microdata20$cashexp <- sort(microdata20$cashexp, decreasing = T)
head(microdata20$cashexp, 4)
```

У консолі отримуємо таку інформацію:

```
[1] 803 792 593 494
```

Тобто, найбільші значення грошових витрат 4 домогосподарств за рік коливаються в межах 494-803 тис. грн.



	rik_fa_1	kvart_kd	code_fam	w_q	tp_ns_p	cod_obl	hsize	type_dom	gnd	cashinc	totalinc	totalres	cashexp
7845	2020	4	108578	4616	1	80	2	2	1	582	582	590	465
7846	2020	4	106139	6219	1	61	0	1	2	199	210	420	494
7847	2020	4	102025	3145	1	23	2	2	2	499	499	690	593
7848	2020	4	108708	2885	1	80	4	1	5	1	1	1	792
7849	2020	4	108517	5155	1	80	4	1	5	1	1	1	803

Рис. 6. Стовпці набору даних «Households_microdani_anonimni_2020» після сортування
Джерело: власні розрахунки.

З рис. 6 видно, що у домогосподарства №7849 грошові витрати становлять 803 тис. грн. на рік при доході у 1 тис. грн. Це наштовхує на сумніви щодо правильності даних. Проте варто спершу перевірити, чи R добре перетворив тип даних на числовий, оскільки часто пропуск або «.» розпізнаються програмою як «,», тобто як знак, що відділяє десяткові значення.

Завантажимо знову масив даних, але замість `cashinc = col_number()` зазначимо `cashinc = col_character()`.

Справді, під значенням «1» R мав на увазі цілий мільйон (рис. 7). Це трапилось через пропуски в числах: програма зберегла значення як числові, натомість пропуск розпізнала як десятковий знак.

	rik_fa_1	kvart_kd	code_fam	w_q	tp_ns_p	cod_obl	hsize	type_dom	gnd	cashinc	totalinc	totalres	cashexp
7845	2020	4	108578	4616	1	80	2	2	1	582 591,10	582	590	465
7846	2020	4	106139	6219	1	61	0	1	2	199 157,10	210	420	494
7847	2020	4	102025	3145	1	23	2	2	2	499 838,21	499	690	593
7848	2020	4	108708	2885	1	80	4	1	5	1 046 982,10	1	1	792
7849	2020	4	108517	5155	1	80	4	1	5	1 028 527,10	1	1	803

Рис. 7. Стовпці набору даних «Households_microdani_anonimni_2020»
Джерело: власні розрахунки.

Отже, тепер цілком реально уявити ситуацію, що при грошовому доході у розмірі понад 1 млн грн. домогосподарство витрачає 803 тис. грн. на рік. Тому можна зробити висновки, що дані не є помилковими.

Висновки та перспективи подальших досліджень. Отже, перетворення даних – це необхідний процес перед їх безпосереднім аналізом. Завдяки попередньому опрацюванню аналітик може не тільки виявити певні невідповідності, а й усунути їх. Це значно полегшує майбутню роботу, поліпшує зміст і надійність статистичних звітів на підставі цих даних. Якщо нехтувати такими діями, як перетворення типів даних, усунення невідповідностей, аналіз нетипових значень тощо, то можна отримати неправильні або викривлені висновки щодо того чи іншого соціально-економічного явища. Доцільно також зазначити, що середовище R задовольняє всі вимоги користувача-

статистика: зрозумілий інтерфейс, легкість роботи та широкий спектр можливостей для опрацювання даних завдяки різноманітності пакетів і бібліотек. Оскільки методи та етапи опрацювання даних залежать від досліджуваного об'єкта, то вони можуть бути різними. Тому перед попереднім опрацюванням даних доцільно вивчати їхній зміст, значення, одиниці виміру, взаємозв'язки тощо.

Перспективи подальших досліджень передбачають детальніше вивчення нетипових значень, а саме аномалій даних, їхнього впливу на статистичні висновки, методи виявлення, інтерпретації та усунення.

Список використаних джерел

1. Анонімні мікродані за основними показниками щодо доходів, витрат та умов життя домогосподарств за 2020 рік [Anonymous microdata on key indicators of household income, expenditure and living conditions for 2020]. URL: http://www.ukrstat.gov.ua/
2. Гринькевич О., Островерх П., Гринькевич В. Джерела даних в аналітиці бізнес-середовища та їх трансформація у цифровій економіці. *Вісник Львівського університету. Серія економічна*. 2021. Випуск 60. С. 77–85.
3. Jain N., Suman S., Prusty R., Performance Comparison of Two Statistical Parametric Methods for Outlier Detection and Correction. *IFAC-PapersOnLine*. Volume 54. Issue 16. 2021. Pages 168–174. URL: https://doi.org/10.1016/j.ifacol.2021.10.089
4. Jonge E. de, Loo M. van der An introduction to data cleaning with R. Statistics Netherlands. 2013. 53 P. URL: https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf (дата звернення 10.04.2021).
5. Lee J.-S., Jun S.-P., Privacy-preserving data mining for open government data from heterogeneous sources. *Government Information Quarterly*. Volume 38, Issue 1, January 2021. URL: https://doi.org/10.1016/j.giq.2020.101544.
6. Ridzuan F., Zainon W., A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science*. Volume 161, 2019, Pages 731–738. URL: https://doi.org/10.1016/j.procs.2019.11.177.
7. Wickham H, Grolemund G. R for Data Science. URL: https://r4ds.had.co.nz/
8. Wickham H., Francois R, Henry L., Muller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.0. URL: https://CRAN.R-project.org/package=dplyr.
9. Wickham H., Hester J., Francois R. readr: Read Rectangular Text Data. R package version 1.3.1. URL: https://CRAN.R-project.org/package=readr.
10. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
11. Hadley Wickham, Maximilian Girlich (2022). tidyr: Tidy Messy Data. R package version 1.2.0. https://cran.r-project.org/web/packages/tidyr/index.html

References

1. Anonimni mikrodani za osnovnymy pokaznykamy shchodo dokhodiv, vytrat ta umov zhyttia domohospodarstv za 2020 rik, URL: http://www.ukrstat.gov.ua/
2. Hrynkevych O., Ostroverkh, P., Hrynkevych V. (2021) Dzherela danykh v analityci biznes-seredovyshha ta jikh transformaciya u cyfrovij ekonomici [Data sources in business environment analytics and their transformation in the digital economy]. Visnyk Lvivskogo

universytetu. Seriya ekonomichna [Visnyk of the Lviv university. Series economics], Vypusk 60, 77-85. [in Ukrainian].
3. Jain, N., Suman, S., & Prusty, B. R. (2021). Performance Comparison of Two Statistical Parametric Methods for Outlier Detection and Correction. IFAC-PapersOnLine, 54(16), 168–174. https://doi.org/10.1016/j.ifacol.2021.10.089
4. Jonge E., Loo M. (2013) An introduction to data cleaning with R. Retrieved from https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
5. Lee, J.-S., & Jun, S.-P. (2021). Privacy-preserving data mining for open government data from heterogeneous sources. Government Information Quarterly, 38(1), 101544. https://doi.org/10.1016/j.giq.2020.101544
6. Ridzuan, F., & Wan Zainon, W. M. N. (2019). A Review on Data Cleansing Methods for Big Data. Procedia Computer Science, 161, 731–738. https://doi.org/10.1016/j.procs.2019.11.177
7. Wickham H, Grolemund G. (2017) R for Data Science. Retrieved from https://r4ds.had.co.nz/
8. Wickham H., Francois R, Henry L., Muller K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.0. https://CRAN.R-project.org/package=dplyr
9. Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
10. Wickham H. (2016). gplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
11. Hadley Wickham, Maximilian Girlich (2022). tidyr: Tidy Messy Data. R package version 1.2.0. https://cran.r-project.org/web/packages/tidyr/index.html

ON DATA PREPARATION FOR ANALYSIS IN SOCIO-ECONOMIC RESEARCH

Oksana Marets, Olha Sysan

Ivan Franko National University of Lviv,
18 Svobody Ave., Lviv, 79008
e-mail: oksana.marets@lnu.edu.ua; ORCID: 0000-0002-4044-7443
e-mail: olha.sysan@lnu.edu.ua; ORCID: 0000-0002-8097-3149

Abstract. Socio-economic research covers the issues of the level and quality of life, economic activity, accessibility and quality of social services, education, etc. In addition to population, activities of enterprises, regional and municipal development, social activities of state and public organizations, etc. constitute the object of socio-economic research.

In order to conduct any research scientists need certain sources of information – data. Therefore, both the ease of work and the final results depend on the database adequacy. The following processes take place during data analysis: data acquisition, processing, analysis and interpretation of the results. But it is processing of the input data that the analyst spends most of the time on, since it is one of the most important processes.

Thus, the aim of the article is to determine the stages of transforming microdata of household surveys and generalization of statistical methods of "raw" data transformation into technically correct data suitable for analysis for the sake of improving the content and reliability of statistical reports based on this data. To do this, various methods of data cleansing and processing, elimination of atypical values, filling in of

ISSN 2078-6115. Вісник Львівського університету. Серія економічна. 2021. Випуск 61

missing values are used. The article describes both technical and subject matter aspects of data cleansing. Technical aspects include data reading, data type conversion, string matching, and various manipulations. Subject-related aspects include data validation, error localization, and value imputation.

The research concludes that actions such as removing unnecessary values, their filtering, grouping or incorrect data replacing significantly affect the results of statistical analysis, as data processing is aimed at improving the content and reliability of statistical reports based on this data.

Keywords: data manipulation, data cleansing, statistical analysis, data processing, data preparation methods.

Стаття надійшла до редакції 30.11.2021
Прийнята до друку 29.12.2021