ПО 01. Аналіз даних
Лабораторна робота №1: Розвідковий аналіз даних

Данило Тавров

1 Завдання на роботу

1.1 Ідея роботи

У лабораторній роботі студентам пропонується здійснити **розвідковий аналіз даних** (exporatory data analysis, **EDA**) для раніше вибраного набору даних та попередньо сформульованих дослідницьких питань.

Виконуючи EDA, студенти мають **повну творчу свободу** і можуть використовувати **будь-які ідеї**, як висловлені в лекційних матеріалах та відповідній літературі, так і знайдені в інших джерелах інформації. **Головне**, щоб результати аналізу були дотичні до висунутих раніше дослідницьких питань, проливали на них світло, уточнювали наявні питання та створювали нові.

Студенти обов'язково повинні використати принаймні два види візуалізацій, окрім тих, що згадані в Лекції 2. Масу цікавих ідей можна знайти в таких джерелах:

- галерея графіків R;
- R Graphics Cookbook;
- офіційний сайт `ggplot2`.

Виконуючи EDA, студенти обов'язково повинні виконати такі основні кроки:

- перевірка даних на відсутність очевидних помилок, одруків, неправильного кодування тощо;
- перевірка на наявність пропущених даних (у тому числі специфічно закодованих) та їхній потенційний вплив на результати аналізу;
- побудова одновимірних графіків (гістограм, вусатих скриньок, стовпчикових діаграм тощо) для виявлення особливостей розподілів окремих змінних;
- побудова двовимірних графіків (діаграм розсіювання, в окремих випадках — лінійних графіків тощо) для виявлення залежностей між окремими змінними;
- обчислення та візуалізація кореляцій між змінними;
- візуалізація на одному графіку декількох змінних за допомогою кольорів, форм та розмірів елементів графіків;
- трансформація координатних осей (зокрема, логаритмування) для ліпшої візуалізації особливостей даних.

1.2 Звітність щодо результатів роботи

Як і в усіх наступних лабораторних роботах, студенти повинні надати викладачу:

- програмний код в R, за допомогою якого проводився аналіз. Код повинно бути написано так, **щоб він успішно виконався на комп'ютері викладача**. Окремо потрібно подати:

  - код, використаний для очищення даних та приведення їх у вигляд, із яким будуть виконуватися всі інші роботи;
  - код власне EDA;
- звіт із лабораторної роботи у **форматі PDF** у **довільній формі** (див. нижче вимоги);
- презентацію до звіту у **форматі PDF** (див. нижче вимоги).

Презентацію результатів студенти повинні здійснити на одному з лабораторних занять або в інший час за домовленістю з викладачем. Тривалість презентації повинна складати **орієнтовно 15 хвилин**. Презентацію **буде записано** та **поширено серед студентів потоку**. На основі презентацій студентів **буде сформульовано окремі питання на МКР**.

1.3 Загальні вимоги до наповнення звіту

Звіт із лабораторної роботи повинен містити такі обов'язкові елементи:

- назву та перелік учасників колективу виконавців;
- вступ із мотивацією проведення дослідження:
  - на які дослідницькі питання була спроба дати відповідь. Ці питання повинні мати реальний сенс і становити цікавість для читачів, на кшталт «Чи існує зв'язок між характеристиками діамантів та їхньою ціною?»
  - важливо вказати гіпотези, очікування, які має дослідник;
- опис даних та кроків із їх підготовки, зокрема:
  - походження даних, звідки їх було взято і за допомогою чого їх було зібрано;
  - що саме було виконано для очищення даних і в яких спосіб;
  - основні характеристики очищених даних (**щонайменше** розмір набору даних, кількість змінних, кількість пропущених даних, дескриптивні статистики, викиди тощо; словом, **усе, про що говорилося в Лекції 2** і більше, виходячи з особливостей набору даних);
- результати EDA у вигляді графіків, таблиць тощо;
- висновки та огляд можливих обмежень проведеного дослідження (вплив викидів, значного числа пропущених даних тощо). Під висновками мається на увазі не фраза «Було виконано роботу», а **конкретні** висновки щодо поставлених на початку питань і висунутих гіпотез, у тому числі ідеї для дальших досліджень, які було сформульовано в процесі EDA;
- посилання на використані джерела (ідеї щодо графіків, фрагментів коду тощо).

Варто пам'ятати, що у звіті **не потрібно** зазначати **абсолютно всіх** видів аналізу, які було виконано. Потрібно фокусуватися тільки на тих речах, які були справді результативні та дають можливість дати відповідь на заявлені дослідницькі питання.

1.4 Загальні вимоги до презентацій

Захист кожної лабораторної роботи повинен супроводжуватися відповідною презентацією. Потрібно розуміти, що метою презентації є **не детальний опис** усіх кроків аналізу тощо, які можна прочитати у звіті чи навіть у коді, **а саме презентація** дослідницького **питання** та **основних результатів** аналізу. Приблизне наповнення презентації може бути таке:

- короткий опис дослідницького питання, у чому полягає його суть та цікавість;
- опис даних, які було використано для аналізу, та як їх було попередньо опрацьовано;

- опис результатів EDA та їхній зв'язок із початковим дослідницьким питанням;
- відповідні висновки (чи було знайдено відповіді на питання, які саме, які додаткові питання виникли тощо).

Під час підготування презентації корисно додержуватися таких порад [1]:

- варто використовувати крупні шрифти та контрастну кольорову гаму, щоб зміст слайду можна було легко прочитати;
- на слайдах повинно бути мінімум тексту (це не лекція), а сам текст повинно бути організовано у вигляді маркованих списків;
- слайди повинно бути пронумеровано, щоб глядачі могли робити осмислені нотатки і ставити питання до конкретних слайдів;
- усі графіки повинні бути з крупними читовними підписами осей;
- під час презентації кожний графік потрібно детально описати (у чому його ідея, на що звернути увагу, який висновок можна зробити), інакше немає сенсу його показувати;
- усі позичені картинки повинні мати посилання на свої джерела.

2 Деякі корисні поради

2.1 Загальні рекомендації щодо проведення повторюваних досліджень

Корисною практикою є зберігання даних, коду та різного роду документації в окремих каталогах, наприклад:

- каталог для даних може містити початкові дані та опрацьовані;
- каталог із рисунками може містити проміжні рисунки (яких в окремих випадках може бути декілька десятків, а то й сотень) та остаточні рисунки, готові до публікації;
- каталог із кодами може містити окремі каталоги для попереднього опрацювання даних, розвідкового аналізу, різних статистичних моделей та, звісно, остаточного коду з усіма коментарями;
- каталог зі звітами та презентаціями.

Для контролю версій доцільно користуватися спеціальними ресурсами на кшталт Github для написання коду (і звітів у форматі R Markdown), або ж Overleaf чи навіть Google Docs для спільного написання звітів і презентацій.

Для проведення статистичного аналізу даних майже напевно виникатиме потреба генерувати випадкові числа. Тому завжди **обов'язково** потрібно явно задавати початкове значення **генератора псевдовипадкових чисел**, щоб усі результати були повторювані.

Варто фіксувати версії використовуваних програмних засобів, щоб за потреби відтворити результати і мати впевненість, що всі коди буде виконано в ідентичний спосіб.

Перед здачею лабораторної роботи доцільно **прогнати** остаточний код **на незалежній машині**, щоб пересвідчитися, що на комп'ютері викладача не виникнуть помилки, які унеможливлять перевірку результатів.

2.2 Поради щодо організації попередньої підготовки даних

На етапі попередньої підготовки та очищення початкових даних (raw data) та приведення їх в охайний вигляд (tidy data) доцільно слідувати таким порадам [1].

У проєкті потрібно завжди зберігати початкові дані й у жодному разі не модифікувати їх. Опрацьовуючи такі дані, щоразу потрібно зберігати модифіковані версії в окремих файлах. Потрібно пам'ятати, що помилки може бути вчинено вже на етапі трансформації початкових даних, тому варто турбуватися про їх недоторканність.

Дані для аналізу повинні бути приведені в охайний формат [2], тобто:

- одній змінній відповідає один стовпець;
- одному спостереженню відповідає один рядок;
- кожне значення повинно міститися в окремій комірці.

До того ж корисно, щоб охайний набір даних мав такі властивості:

- перший рядок займають назви змінних, які максимально зрозумілі досліднику;
- усі дані повинні зберігатися в максимально простому форматі на кшталт CSV (comma-separated values).

Критично важливо для кожного набору даних, із яким працюватиме аналітик, готувати кодову книжку (codebook), у якій потрібно в обов'язковому порядку зазначати:

- інформацію про кожну змінну, у т.ч. одиниці виміру;
- пояснення процесу очищення даних, особливо якщо це пов'язано з відбором даних чи їх перекодуванням;
- додаткову інформацію, пов'язану з особливостями збору відповідних даних (особливо якщо це дані, які дослідник збирав у рамках деякого дослідження особисто).

Код, за допомогою якого очищувалися дані, повинен бути загальнодоступним і повторюваним. Будь-який інший дослідник повинен мати можливість застосувати його до початкових даних і дістати точно таку саму очищену версію.

Типовими помилками в організації процесу є:

- об'єднання декількох змінних в один стовпець;
- об'єднання в одному файлі непов'язаних даних (напр., про фінансовий стан осіб і про їхні медичні показники);
- відсутність коду для очищення даних, або, у випадку наявности тільки опису загального алгоритму, відсутність деталей реалізації окремих кроків.

2.3 Поради щодо попередньої підготовки даних

Навіть якщо початкові дані мають більш-менш охайний вигляд, із ними все одно потрібно провести певну роботу [1].

Дуже часто в початкових даних містяться елементарні помилки: описки або помилково уведені дані. Знайти такі помилки можна, здійснивши швидку візуалізацію змінних або підрахувавши дескриптивні статистики. Наприклад, якщо в деякій комірці не там стоїть крапка, то число може бути на порядок більше від інших, і це буде сильно впадати у вічі.

Різні змінні в наборі даних можуть мати різну природу:

- неперервні (continuous) змінні, наприклад, вага людини, можуть набувати теоретично будь-яких дійсних значень на деякому проміжку;

- порядкові (ordinal) дані можуть бути як числовими, так і символьними, але вони позначають упорядковані категорії, наприклад, оцінки студентів. Такі дані не можна розглядати як числові, навіть якщо вони закодовані числами. На етапі опрацювання даних їх потрібно перетворити, напр., у факторні змінні;

- категорійні (categorical) дані схожі на порядкові, але між ними немає відношення порядку, напр., стать особи або пора року;

- пропущені (missing) дані повинно бути закодовано в однаковий спосіб (напр., через `NA`), і їх у жодному разі не можна залишати порожніми, щоб не плутати з коректними порожніми даними. Наприклад, в особи може бути відсутнє по батькові, але це не означає, що воно пропущене в тому розумінні, що воно є, але ми його не знаємо;

- цензуровані (censored) дані подібні до пропущених, але в цьому випадку ми розуміємо, чому вони відсутні. Типовим прикладом може бути вік пацієнта, якщо він перевищує деяке значення: напр., усі пацієнти, старші від 95 років, можуть мати вік `95`. У цьому випадку доречно додати окрему логічну змінну, яка б указувала, чи є значення в наборі даних справжнім, чи тільки цензурованим. Статистичні методи для роботи з пропущеними і цензурованими даними суттєво відрізняються.

У різних наборах даних пропущені дані може бути закодовано по-різному: `NA`, `.`, `999`, `-1` тощо. Важливо відразу зрозуміти, який підхід використовували автори даних і перекодувати відповідним чином. Особливо це стосується числових кодів на кшталт `-1`, які можна не помітити і через це дістати некоректні результати аналізу. Наприклад, якщо в змінній «Вік» присутнє значення `-1`, яке зустрічається достатньо часто, щоб не вважати, що це помилково уведене число 1, то це, радше за все, код пропущеного значення.

Часто початкові дані мають непослідовну політику щодо кодування своїх значень. В одному наборі даних для кодування чоловічої статі можуть зустрічатися і `male`, і `M`, і такі змінні потрібно перекодувати в послідовній формі.

Інколи змінні мають не ту одиницю виміру, що заявлено в описі набору даних. Класичним прикладом є втрата космічного апарата Mars Climate Orbiter. Гістограми та інші графіки дадуть можливість швидко зорієнтуватися, які одиниці виміру насправді має відповідна змінна.

Типовими помилками на етапі попередньої підготовки даних є:

- невиконання цього кроку в принципі. Без детального аналізу даних на предмет наявности очевидних помилок жодний дальший аналіз не матиме сенсу;

- кодування категорійних значень числами замість факторів та їх використання в статистичних моделях у такому форматі є абсолютно помилковим;

- недостатнє використання засобів візуалізації для попереднього аналізу даних створює ризик залишити непоміченими деякі проблеми, напр., викиди.

2.4 Поради щодо оформлення графіків

Під час формування ілюстративного матеріалу для звітів і презентацій корисно додержуватися таких порад [1]:

- потрібно використовувати типи графіків, найбільш адекватні зображуваним змінним. Кругових діаграм потрібно **уникати** і надавати перевагу стовпчиковим, оскільки порівняння висот стовпчиків значно легше зробити, ніж величин кутів. Не варто нехтувати діаграмами розсіювання: вони показують кожне спостереження окремо і можуть бути інформативніші від вусатих скриньок тощо;

- якщо розподіл даних сильно скошений, то обов'язково потрібно логаритмувати шкалу вимірювання, інакше більшість даних буде зосереджено в одній ділянці графіка;

- підписи координатних осей повинні бути достатньо великі й читабельні (доброю рекомендацією є використання такого ж розміру шрифту, як і в тексті звіту), назви повинні бути зрозумілі людині, а не автозгенеровані програмою;

- якщо на графіку використовують різні кольори чи стилі ліній та точок для подання різної інформації, обов'язково потрібно подавати легенду або інші способи пояснення;
- заголовок графіка не повинен дублювати і так очевидну інформацію. Якщо в ньому є потреба, його варто присвятити короткій суті того, що зображено на графіку. Напр., «Куріння пов'язано з більшою ймовірністю раку», а не «Графік залежности ймовірности раку від куріння»;
- потрібно уникати графіків із низьким інформаційним навантаженням: в окремих випадках простіше подати числа в табличному форматі. Якщо в цьому є потреба, в окремих випадках текстову інформацію можна додавати в сам графік;
- якщо це не очевидно з контексту, у підписах осей та в легенду корисно додавати одиниці виміру, особливо якщо на одному графіку використано дві осі одночасно;
- підписи під графіками повинні бути самодостатні. Читач не повинен шукати по тексту звіту, що саме ілюструє цей графік;
- якщо це сприяє ліпшому розумінню, у графіки можна додавати різні кольори чи варіювати форми точок на графіку (напр., різними кольорами відмічати різні змінні на діаграмі розсіювання). Проте варто уникати надмірних дизайнерських рішень, які псують загальне враження від дослідження;
- інколи корисно будувати фацетовані графіки, щоб не перевантажувати основний графік розмаїттям кольорів чи форм точок. У цьому випадку окремі панелі повинні мати однакові шкали координатних осей, інакше дуже важко помітити різниці, якщо такі є. Панелі потрібно окремо пронумерувати або позначити окремими літерами. Помилкою є розміщення панелей графіка в рядок, коли порівняння потрібно робити в горизонтальному напрямку (напр., наскільки велика дисперсія розподілу), та у стовпець, коли порівняння потрібно робити у вертикальному напрямку (напр., порівнювати висоту стовпчиків);
- потрібно використовувати кольори, які суттєво різняться (особливо якщо презентацію потрібно показувати за допомогою проєктора на не зовсім білій поверхні).

Інші додаткові поради можна подивитися в цій презентації.

Література

[1] Leek J. The Elements of Data Analytic Style: A guide for people who want to analyze data. — Leanpub, 2015. — 98 p.

[2] Wickham H. Tidy Data / Wickham H. // Journal of Statistical Software 59(10). — 2014.