МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА

Факультет прикладної математики та інформатики

ЗВІТ

до індивідуального завдання №1

з дисципліни «Моделі статистичного навчання»

Виконали

студенти групи ПМіМ-12:

Бордун Михайло

Зелінський Олександр

Перевірив:

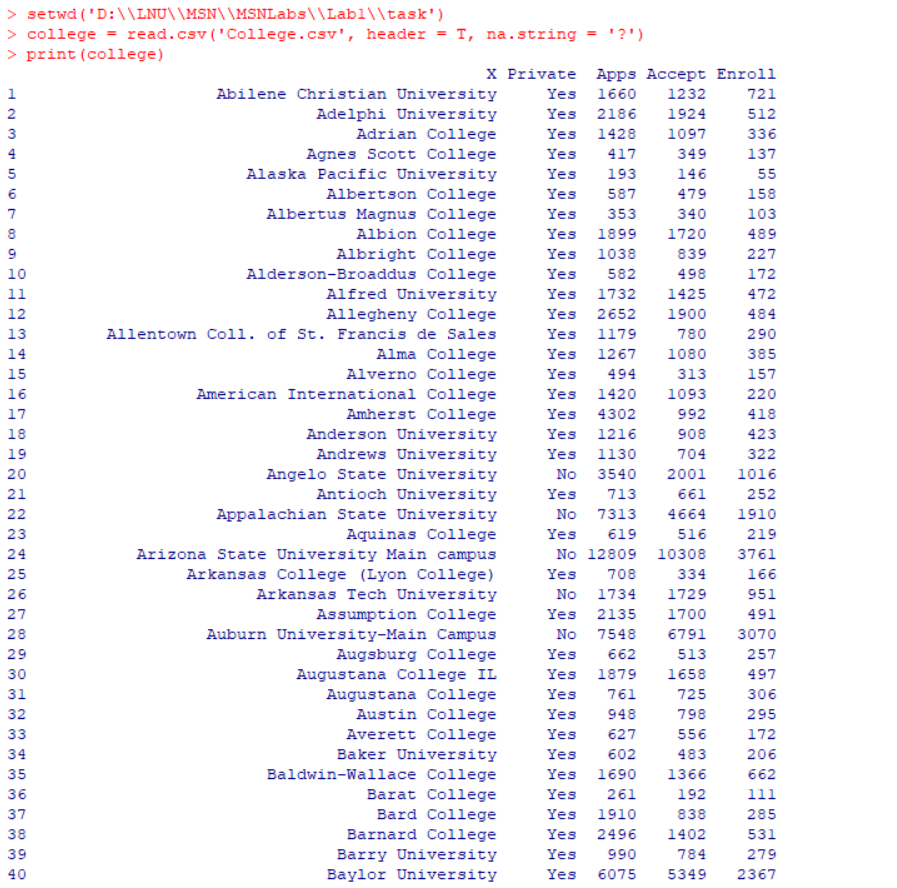
Проф. Заболоцький Т. М.

Львів – 2021

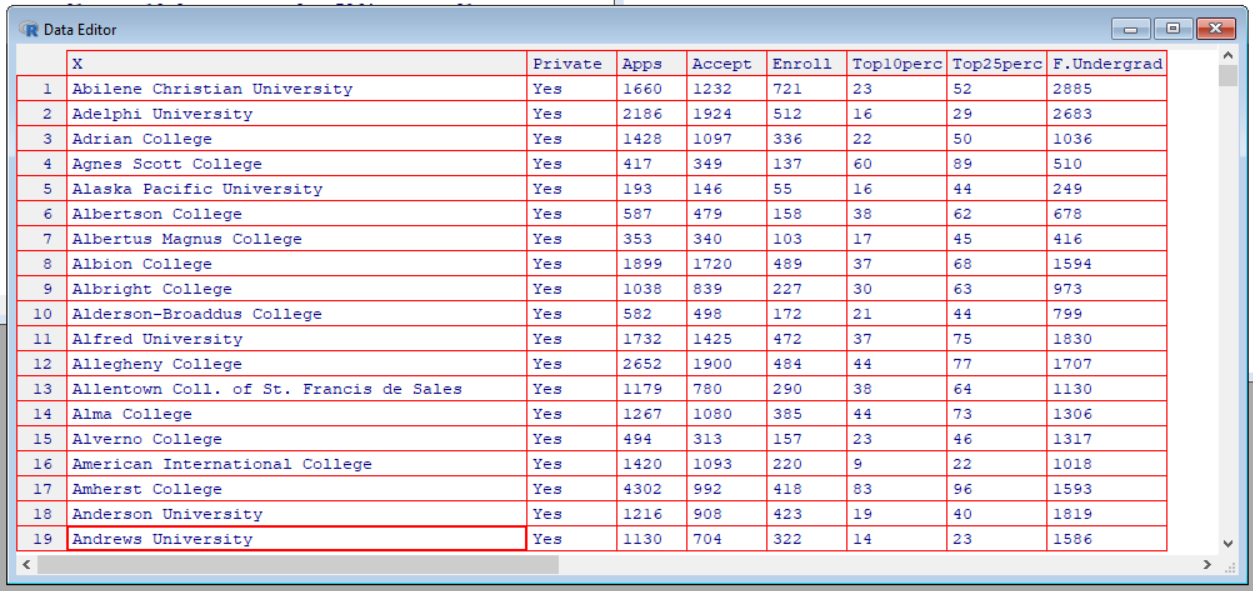
**Хід виконання**

**Пункт 1 (College)**

* 1. Завантажені та викликані дані College.



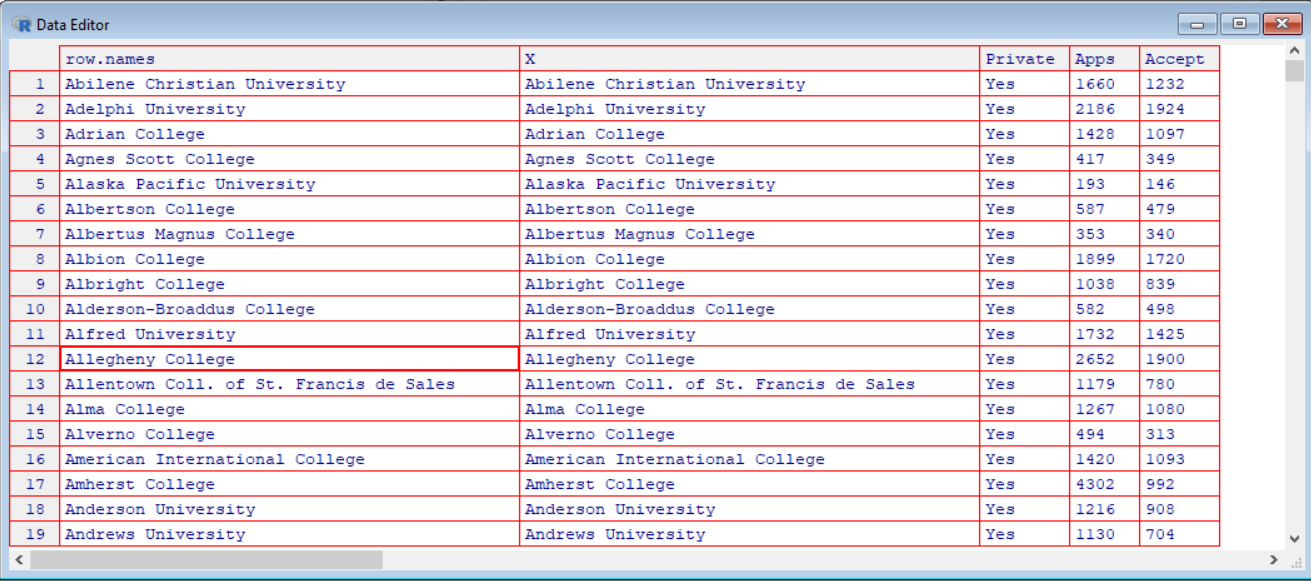
* 1. Дані переглянуті за допомогою функції fix.



Після виконання команд

> rownames (college )=college [,1]

> fix (college )

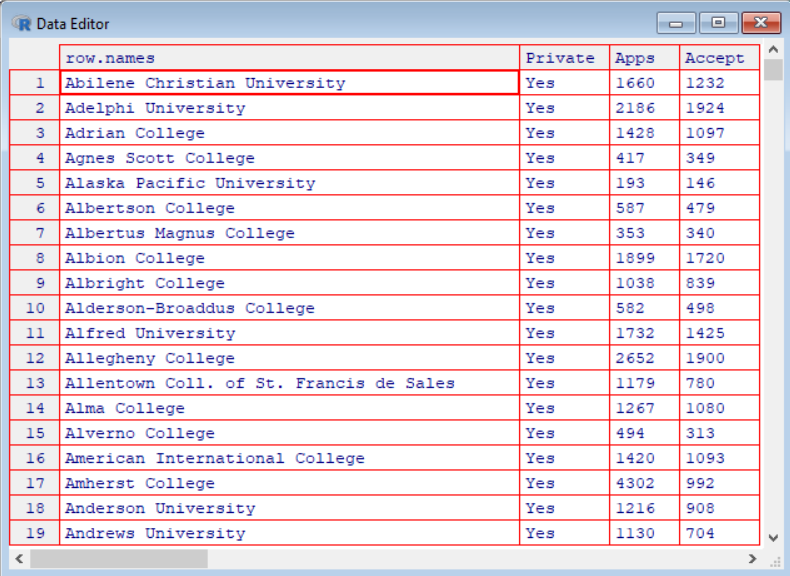


Можемо помітити, що назви факультетів здублювались в row.names, а після виконання команд

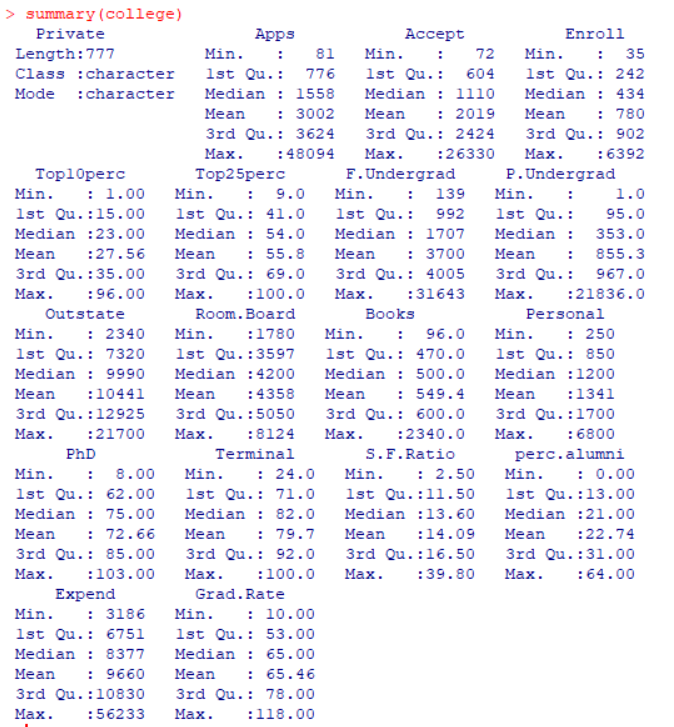
> rownames (college )=college [,1]

> fix (college )

рядок з іменами університетів зник з даних і залишився лише у вигляді назв рядків (row.names).



**1.3.1** За допомогою функції summary можемо побачити деякі статистичні величини по кожному з рядків. Також можемо побачити, що всі значення числові окрім Private (приватний університет чи ні).



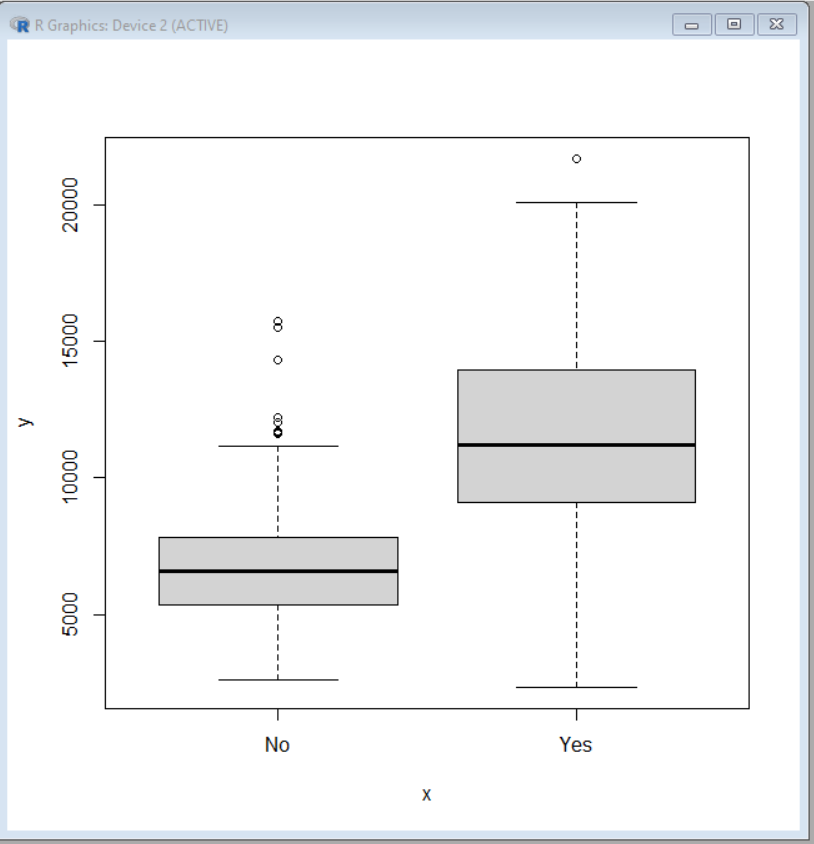
**1.3.2** Для побудови матриці графіків перших 10 стовпців використано функцію pairs(college[1:10]).

Зображення, що містить текст

Автоматично згенерований опис

Детальний опис цього графіка можна глянути в пункті 1.3.6.

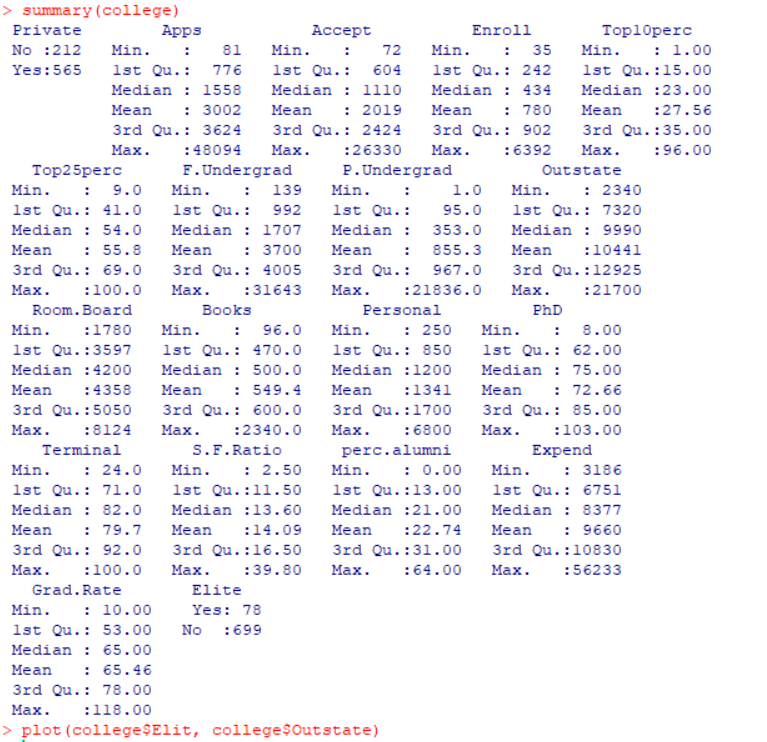
**1.3.3** Для побудови графіку залежності college$Private та college$Outstate використану функцію plot.



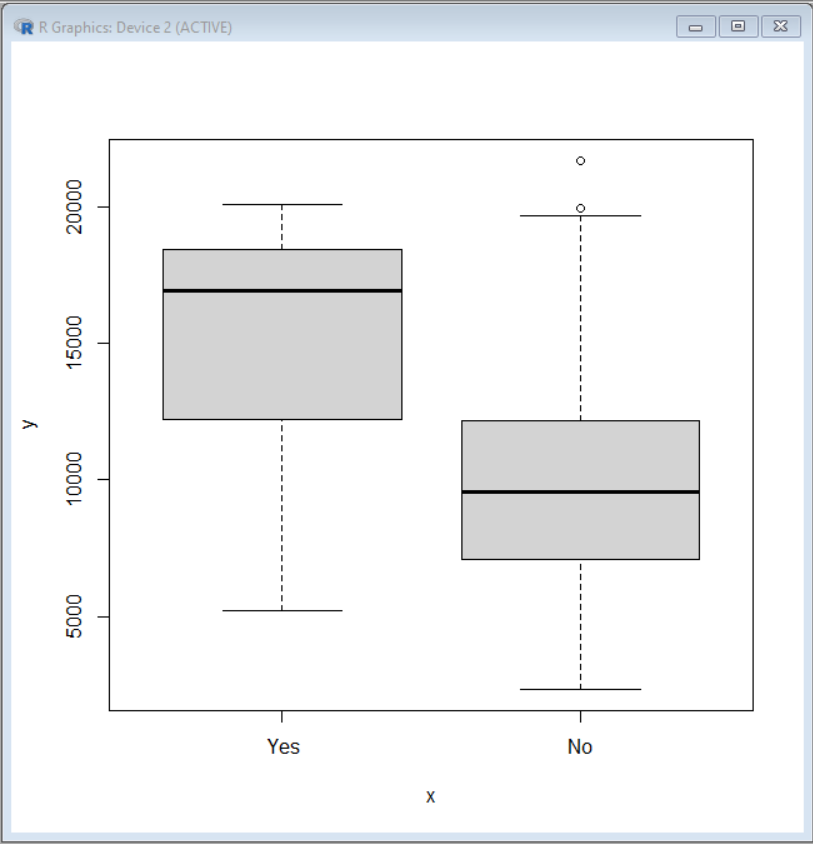
Зважаючи на наведений вище графік можна сказати, що вартість навчання для іноземних студентів в приватних коледжах вища ніж у державних, проте існують деякі державні коледжі в яких вартість навчання для іноземних студентів більша ніж середня вартість навчання в приватних, проте таких не багато.

**1.3.4** Створено новий показник Elit використовуючи Top10perc. Поділимо всі університети на дві групи в залежності чи перевищує відсоток студентів з топ 10% шкіл 50% чи ні.

За допомогою функції summary було визначено, що елітних шкіл 78, а не елітних 699.



Також було побудовано графік залежності college$Elit від college$Outstate.



Зважаючи на цей графік можна стверджувати, що навчання в елітних коледжах дорожче для іноземних студентів ніж в не елітних, про те є два винятки, дорогих не елітних коледжів.

**1.3.5** За допомогою функції hist було побудовано 4 гістограми: Витрати на навчання на одного студента, вартість навчання іноземних студентів, відсоток факультетів з PhD та Відсоток професорсько-викладацького складу.

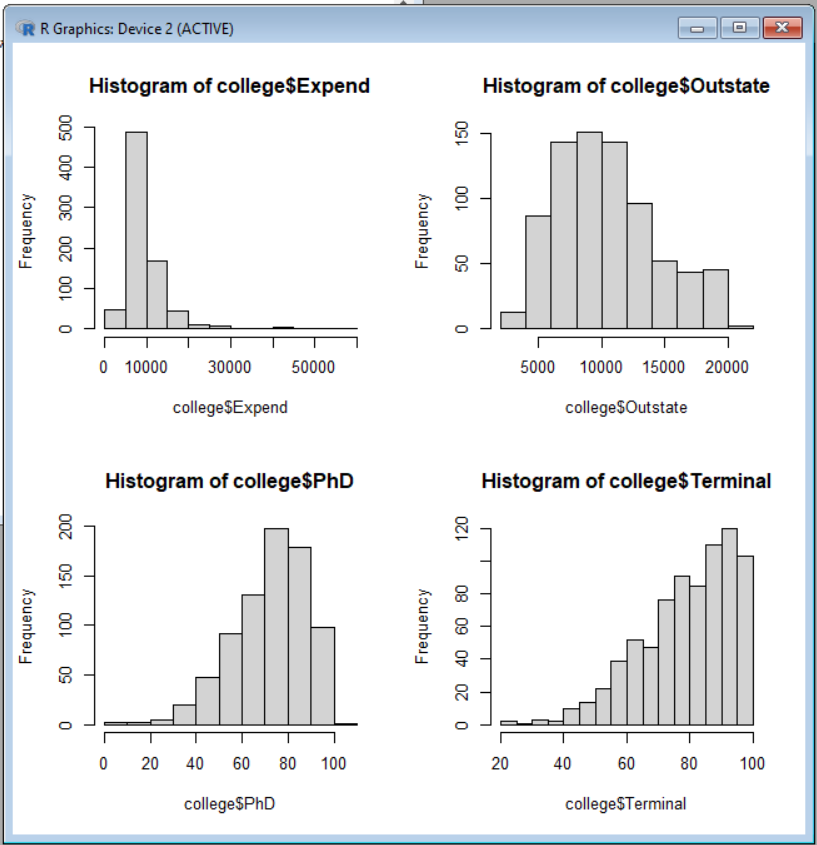
> par(mfrow=c(2,2))

> hist(college$Expend)

> hist(college$Outstate)

> hist(college$PhD)

> hist(college$Terminal)



Зважаючи на ці гістограми можна сказати, що в даних, щодо PhD є помилка, тому що % не може бути більшим за 100.

**1.3.6** Отже спираючись на дані можна зробити такі висновки:

Зважаючи на дані можна сказати що одна змінна Private – якісна змінна, а всі решта змінні – кількісні.

Також можна виділити чіткі попарні лінійні залежності між Apps, Accept, Enroll, F.Undergrad та між Top10perc та Top25perc. Ці залежності є досить логічними та інтуїтивно зрозумілими.

На додачу видно, що в даних є помилка, а саме в колонці PhD, бо 104% це нереальна цифра.

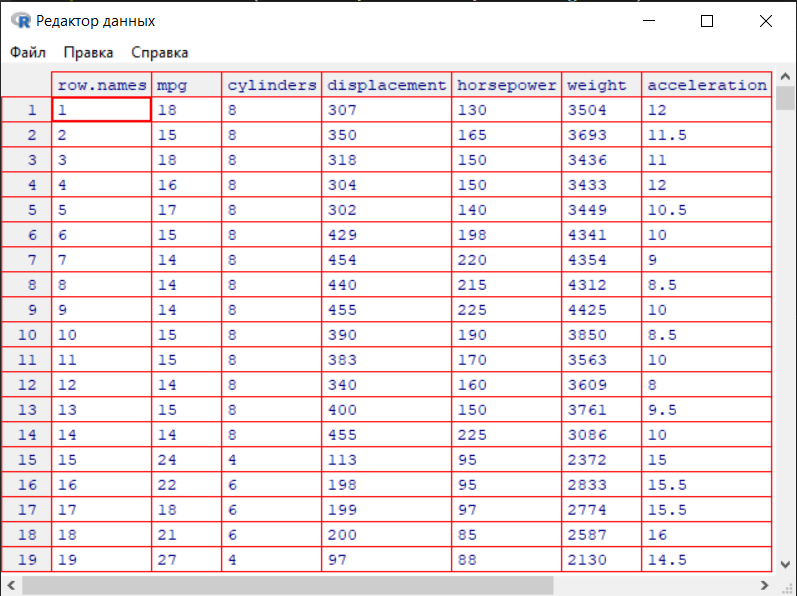
**Пункт 2 (Auto)**

**2**. Переконався, що в даних Auto видалені пропущені значення за допомогою функції na.omit() та переглянув дані за допомогою функції fix().

>autos = read.csv('Auto.csv', header = T, na.string = '?')

>autos = na.omit(autos)

>fix(autos)



**2.1.** Знизу наведено розподіл показників за критерієм кількісні/якісні

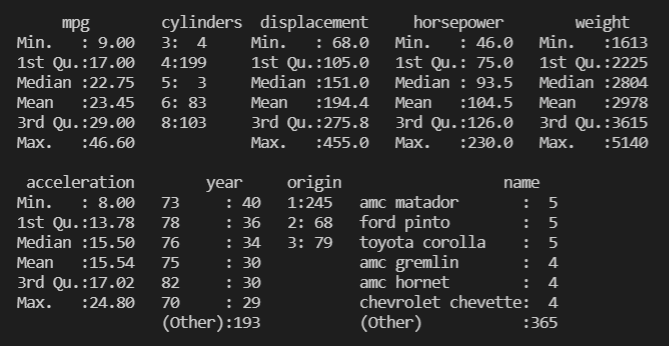
* Якісні: origin, year, cylinders, name
* Кількісні: mpg, displacement, horsepower, weight, acceleration

Якісні показники я визначив в середовищі програмування за допомогою функції as.factor() та вивів підсумок для кожної змінної з таблиці за допомогою функції summary().

>qualitative = c(2, 7, 8, 9)

>for (val in qualitative) { autos[, val] = as.factor(autos[, val]) }

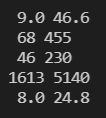
>print(summary(autos[]))



**2.2.** Використовуючи функцію range() я визначив межі для кожного кількісного показника. Вивід поданий в такому порядку показників: mpg, displacement, horsepower, weight, acceleration.

>quantitative = c(1, 3, 4, 5, 6)

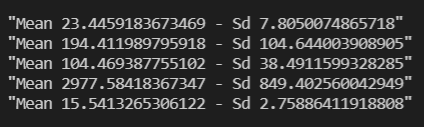
>for (val in quantitative) { print(range(autos[, val])) }



**2.3.** Обчислив середнє та стандартне відхилення для всіх кількісних показників за допомогою функцій mean() та sd() відповідно.

> for (val in quantitative) { print(paste("Mean", mean(autos[, val]),

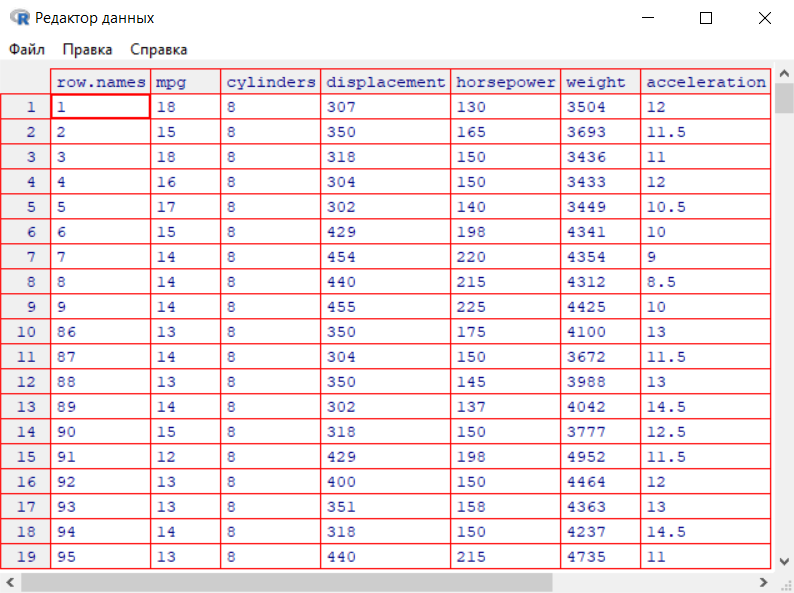
"- Sd", sd(autos[, val]))) }



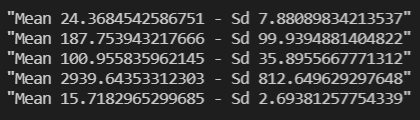
**2.4.** Видалив спостереження з 10-го по 85-те з допомогою вилучення масиву заданого діапазону з нашої вибірки. Також переглянув дані за допомогою функції fix().

>autos\_clipped = autos[-c(10:84),]

>fix(autos\_clipped)

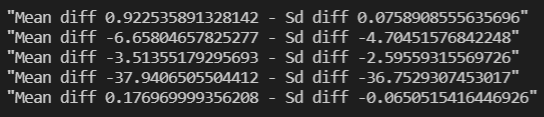


Обчислив середнє та стандартне відхилення для всіх кількісних показників за допомогою функцій mean() та sd() відповідно.

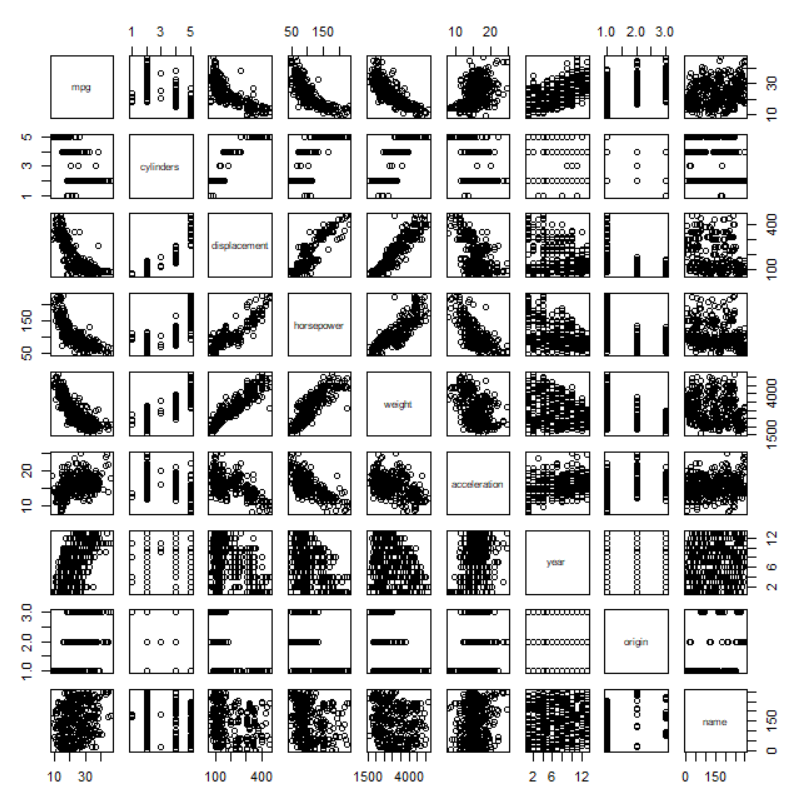


Результати свідчать про те, що для наступних показників відбулося

* зростання середнього 🡪 mpg, acceleration
* спадання середнього 🡪 displacement, horsepower, weight
* зростання стандартного відхилення 🡪 mpg
* спадання стандартного відхилення 🡪 displacement, horsepower, weight, acceleration

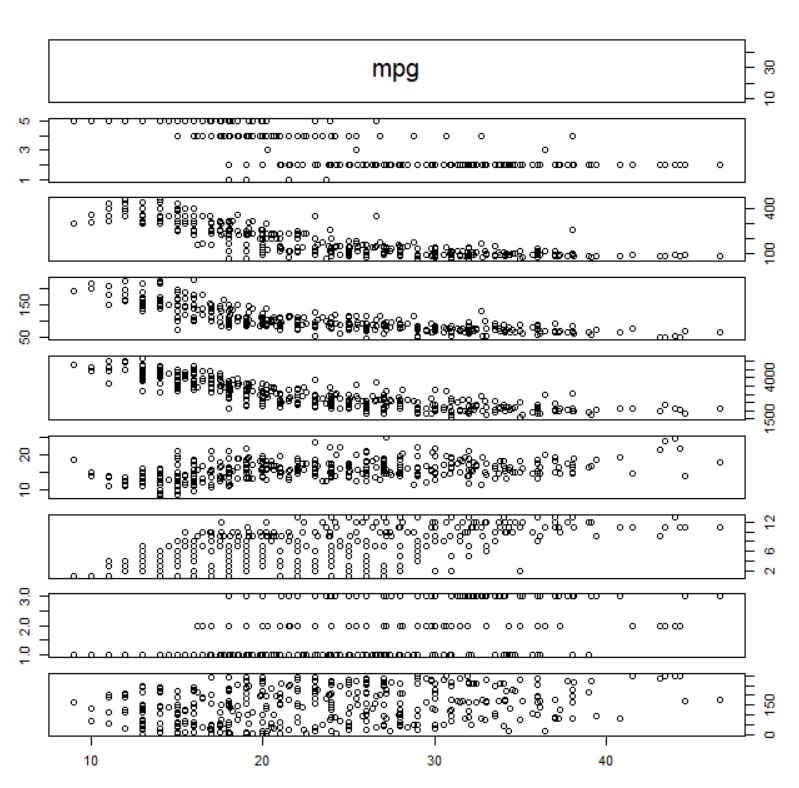


**2.5.** Використавши функцію pairs() я вивів матрицю графіків набору даних.



Підсумовуючи наведені вище графіки, можна сказати, що серед кількісних показників є досить чітко виражена лінійна залежність між собою.

**2.6.** Для аналізу розходу пального (mpg) від інших наявних показників використано функцію pairs() з аргументом verInd=1 для виведення графічної залежності тільки для першої змінної.



Загалом з наведених вище графіків можна сказати, що рівень розходу пального обернено пропорційно залежить від показників displacement, hoursepower та weight. Також можна дійти до однозначного висновку щодо розхід пального має тенденцію до зростання відповідно до зростання показника year.

**Пункт 3 (Boston)**

**3.1** Завантажено дані Boston з бібліотеки MASS.

Зображення, що містить стіл

Автоматично згенерований опис

За допомогою ?Boston можна переглянути інформацію про дані. Видно що дані містять 506 рядків та 14 колонок з їхнім описом.

Зображення, що містить текст

Автоматично згенерований опис

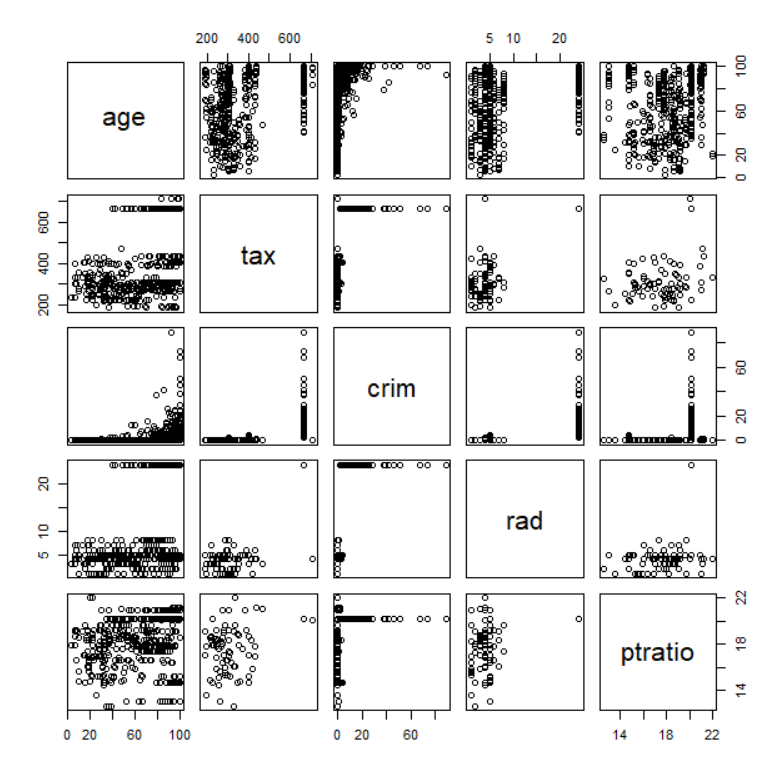
Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

**3.2** За допомогою функції pairs було побудовано попарні графіки для деяких величин (вік, повноцінна ставка податку на майно за \ 10000 доларів, рівень злочинності на душу населення, індекс доступності до радіальних магістралей, співвідношення вчитель-учень).



**3.3** Для того щоб перевірити чи пов’язаний якийсь показник із рівнем злочинності на душу населення було побудовано попарні графіки його з кожним іншим параметром.

Зображення, що містить стіл

Автоматично згенерований опис

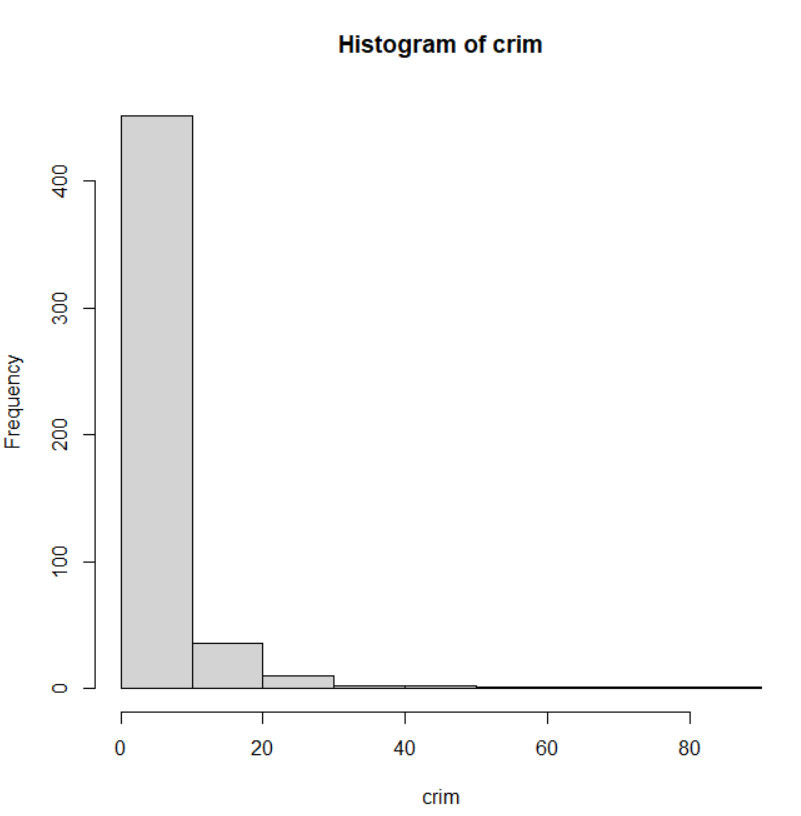
Загалом явних лінійних зв’язків не видно.

Можна сказати що рівень злочинності дещо зростає зі збільшенням відсотку бідного населення (lstat), а також дещо спадає зі збільшенням відстані до 5 центрів зайнятості (dis) і зростанням медіанної вартості нерухомості (medv). Також з графіка crim/chas видно що всі райони з особливо високим рівнем злочинності не межують з річкою Charles.

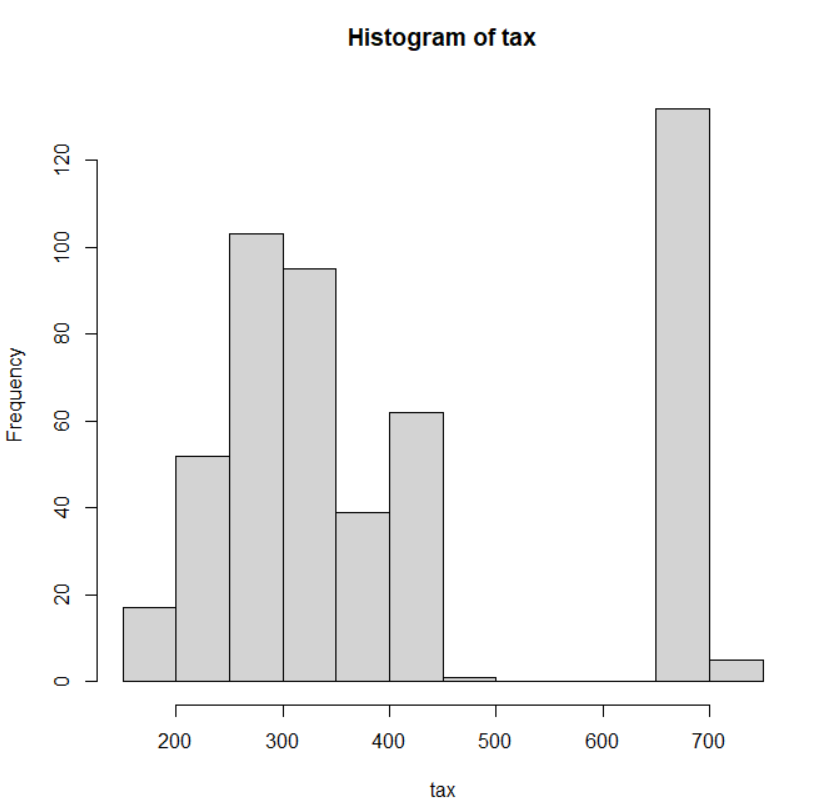
**3.4** За допомогою функції summary можемо побачити деякі статистичні величини по кожному з рядків.

Зображення, що містить текст

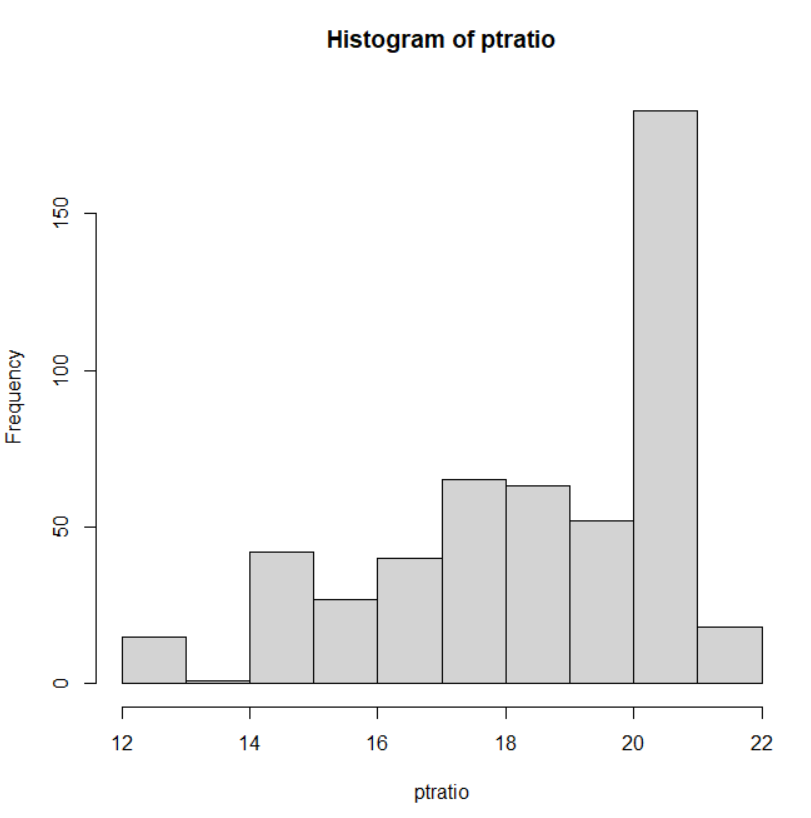
Автоматично згенерований опис



З огляду на гістограму для показника рівня злочинності можна сказати, що вона сильно нерівномірна і показує, що загалом рівень злочинності є низький, проте в окремих декількох кварталах цей рівень сильно зростає.



Щодо податкових ставок, то їхній діапазон становить 187-711$ на 10 000$. Аналізуючи гістограму податкових ставок можна дійти до висновку щодо певного розподілу кварталів на 2 класи, а саме з ставкою дешевшою за 500$ та дорочою за 650$.



Співвідношення учні-вчителі є в межах від 12.6 до 22 з модою зі значенням 21.

**3.5** Визначити скільки кварталів межують з річкою можна визначити за допомогою функції summary.

Зображення, що містить текст

Автоматично згенерований опис

Або запиту нижче

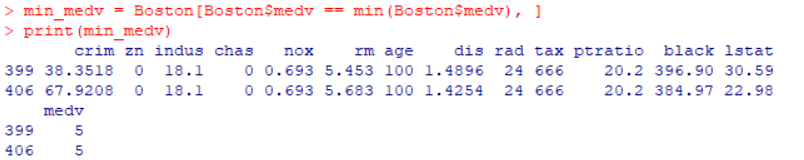


З цього випливає що таких записів є 35.

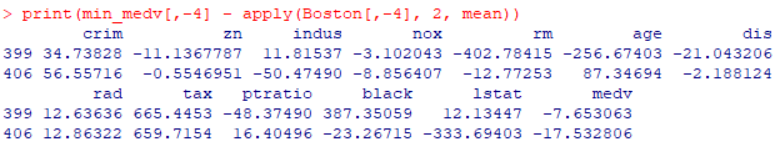
**3.6** Медіану визначено за допомогою функції median.



**3.7** Для знаходження кварталів міста з найнижчою медіаною кількості зайнятих помешкань та інших показників цих кварталів використано пошук за індексом з використанням функції min() серед набору даних Boston.



Тобто, було знайдено два квартали з найнижчою медіаною кількості занятих помешкань, medv=5. Для наведених кварталів якісна змінна chas=0, що свідчить про відсутність межування з річкою Charles. Аналіз співвідношення решти показників зі значеннями показників інших кварталів здійснено з використанням різниці з середнім значенням по всьому місту.



Результати показують, що значення показників

* crim, indus, nox, age, rad, tex, ptratio, black, lstat 🡪 більші ніж середні значення по всьому місту.
* zn, rm, dis, medv 🡪 менші ніж середні значення по всьому місту.