

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

Перевірів:  
Проф. Заболоцький Т. М.

Львів – 2021

# Хід виконання

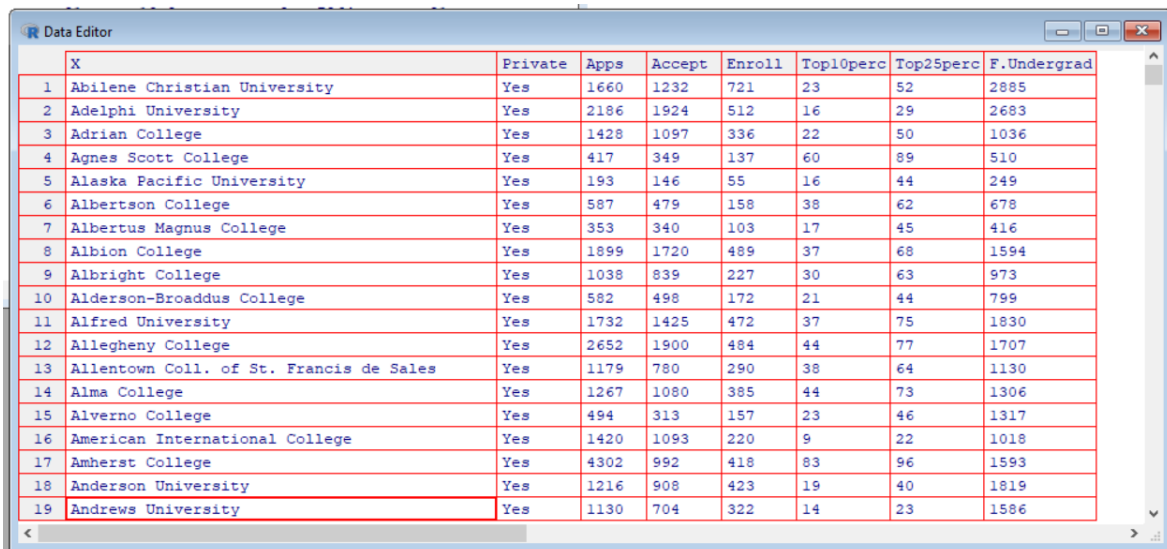
## Пункт 1 (College)

### 1.1 Завантажені та викликані дані College.

```
> setwd('D:\\LNU\\MSN\\MSNLabs\\Lab1\\task')
> college = read.csv('College.csv', header = T, na.string = '?')
> print(college)
```

		X	Private	Apps	Accept	Enroll
1	Abilene Christian University	Yes	1660	1232	721	
2	Adelphi University	Yes	2186	1924	512	
3	Adrian College	Yes	1428	1097	336	
4	Agnes Scott College	Yes	417	349	137	
5	Alaska Pacific University	Yes	193	146	55	
6	Albertson College	Yes	587	479	158	
7	Albertus Magnus College	Yes	353	340	103	
8	Albion College	Yes	1899	1720	489	
9	Albright College	Yes	1038	839	227	
10	Alderson-Broadus College	Yes	582	498	172	
11	Alfred University	Yes	1732	1425	472	
12	Allegheny College	Yes	2652	1900	484	
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	
14	Alma College	Yes	1267	1080	385	
15	Alverno College	Yes	494	313	157	
16	American International College	Yes	1420	1093	220	
17	Amherst College	Yes	4302	992	418	
18	Anderson University	Yes	1216	908	423	
19	Andrews University	Yes	1130	704	322	
20	Angelo State University	No	3540	2001	1016	
21	Antioch University	Yes	713	661	252	
22	Appalachian State University	No	7313	4664	1910	
23	Aquinas College	Yes	619	516	219	
24	Arizona State University Main campus	No	12809	10308	3761	
25	Arkansas College (Lyon College)	Yes	708	334	166	
26	Arkansas Tech University	No	1734	1729	951	
27	Assumption College	Yes	2135	1700	491	
28	Auburn University-Main Campus	No	7548	6791	3070	
29	Augsburg College	Yes	662	513	257	
30	Augustana College IL	Yes	1879	1658	497	

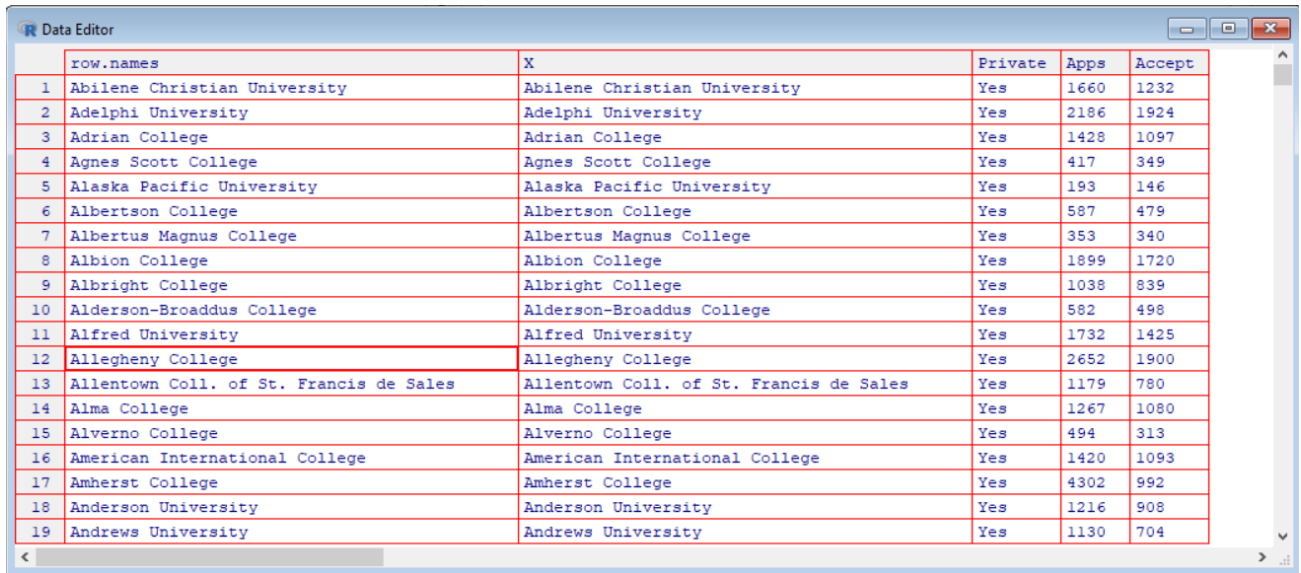
### 1.2 Дані переглянуті за допомогою функції fix.



	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885
2	Adelphi University	Yes	2186	1924	512	16	29	2683
3	Adrian College	Yes	1428	1097	336	22	50	1036
4	Agnes Scott College	Yes	417	349	137	60	89	510
5	Alaska Pacific University	Yes	193	146	55	16	44	249
6	Albertson College	Yes	587	479	158	38	62	678
7	Albertus Magnus College	Yes	353	340	103	17	45	416
8	Albion College	Yes	1899	1720	489	37	68	1594
9	Albright College	Yes	1038	839	227	30	63	973
10	Alderson-Broadus College	Yes	582	498	172	21	44	799
11	Alfred University	Yes	1732	1425	472	37	75	1830
12	Allegheny College	Yes	2652	1900	484	44	77	1707
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130
14	Alma College	Yes	1267	1080	385	44	73	1306
15	Alverno College	Yes	494	313	157	23	46	1317
16	American International College	Yes	1420	1093	220	9	22	1018
17	Amherst College	Yes	4302	992	418	83	96	1593
18	Anderson University	Yes	1216	908	423	19	40	1819
19	Andrews University	Yes	1130	704	322	14	23	1586

Після виконання команд

```
> rownames (college )=college [,1]  
> fix (college )
```



	row.names	X	Private	Apps	Accept
1	Abilene Christian University	Abilene Christian University	Yes	1660	1232
2	Adelphi University	Adelphi University	Yes	2186	1924
3	Adrian College	Adrian College	Yes	1428	1097
4	Agnes Scott College	Agnes Scott College	Yes	417	349
5	Alaska Pacific University	Alaska Pacific University	Yes	193	146
6	Albertson College	Albertson College	Yes	587	479
7	Albertus Magnus College	Albertus Magnus College	Yes	353	340
8	Albion College	Albion College	Yes	1899	1720
9	Albright College	Albright College	Yes	1038	839
10	Alderson-Broaddus College	Alderson-Broaddus College	Yes	582	498
11	Alfred University	Alfred University	Yes	1732	1425
12	Allegheny College	Allegheny College	Yes	2652	1900
13	Allentown Coll. of St. Francis de Sales	Allentown Coll. of St. Francis de Sales	Yes	1179	780
14	Alma College	Alma College	Yes	1267	1080
15	Alverno College	Alverno College	Yes	494	313
16	American International College	American International College	Yes	1420	1093
17	Amherst College	Amherst College	Yes	4302	992
18	Anderson University	Anderson University	Yes	1216	908
19	Andrews University	Andrews University	Yes	1130	704

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).



	row.names	Private	Apps	Accept
1	Abilene Christian University	Yes	1660	1232
2	Adelphi University	Yes	2186	1924
3	Adrian College	Yes	1428	1097
4	Agnes Scott College	Yes	417	349
5	Alaska Pacific University	Yes	193	146
6	Albertson College	Yes	587	479
7	Albertus Magnus College	Yes	353	340
8	Albion College	Yes	1899	1720
9	Albright College	Yes	1038	839
10	Alderson-Broaddus College	Yes	582	498
11	Alfred University	Yes	1732	1425
12	Allegheny College	Yes	2652	1900
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780
14	Alma College	Yes	1267	1080
15	Alverno College	Yes	494	313
16	American International College	Yes	1420	1093

**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).

```
> summary(college)
```

Private	Apps	Accept	Enroll
Length:777	Min. : 81	Min. : 72	Min. : 35
Class :character	1st Qu.: 776	1st Qu.: 604	1st Qu.: 242
Mode :character	Median : 1558	Median : 1110	Median : 434
	Mean : 3002	Mean : 2019	Mean : 780
	3rd Qu.: 3624	3rd Qu.: 2424	3rd Qu.: 902
	Max. : 48094	Max. : 26330	Max. : 6392

Top10perc	Top25perc	F.Undergrad	P.Undergrad
Min. : 1.00	Min. : 9.0	Min. : 139	Min. : 1.0
1st Qu.:15.00	1st Qu.: 41.0	1st Qu.: 992	1st Qu.: 95.0
Median :23.00	Median : 54.0	Median : 1707	Median : 353.0
Mean :27.56	Mean : 55.8	Mean : 3700	Mean : 855.3
3rd Qu.:35.00	3rd Qu.: 69.0	3rd Qu.: 4005	3rd Qu.: 967.0
Max. :96.00	Max. :100.0	Max. :31643	Max. :21836.0

Outstate	Room.Board	Books	Personal
Min. : 2340	Min. :1780	Min. : 96.0	Min. : 250
1st Qu.: 7320	1st Qu.:3597	1st Qu.: 470.0	1st Qu.: 850
Median : 9990	Median :4200	Median : 500.0	Median :1200
Mean :10441	Mean :4358	Mean : 549.4	Mean :1341
3rd Qu.:12925	3rd Qu.:5050	3rd Qu.: 600.0	3rd Qu.:1700
Max. :21700	Max. :8124	Max. :2340.0	Max. :6800

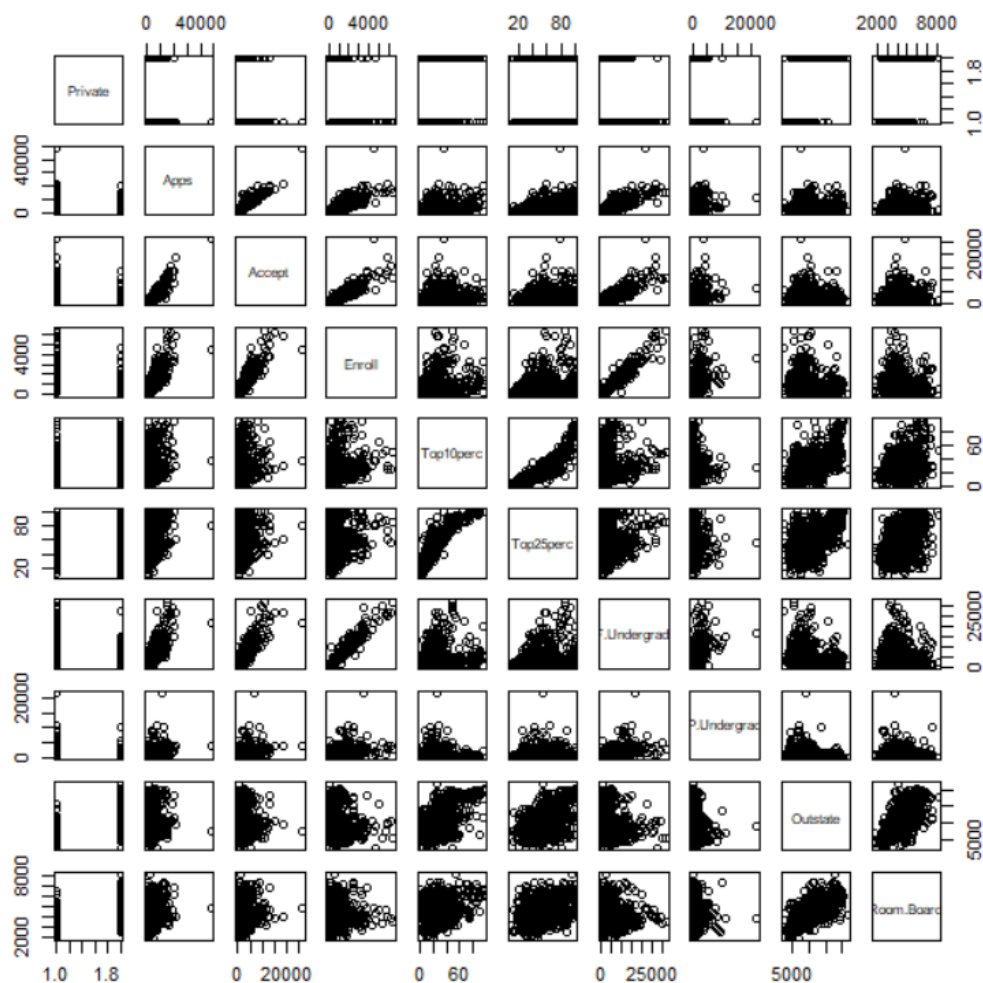
  

PhD	Terminal	S.F.Ratio	perc.alumni
Min. : 8.00	Min. : 24.0	Min. : 2.50	Min. : 0.00
1st Qu.: 62.00	1st Qu.: 71.0	1st Qu.:11.50	1st Qu.:13.00
Median : 75.00	Median : 82.0	Median :13.60	Median :21.00
Mean : 72.66	Mean : 79.7	Mean :14.09	Mean :22.74
3rd Qu.: 85.00	3rd Qu.: 92.0	3rd Qu.:16.50	3rd Qu.:31.00
Max. :103.00	Max. :100.0	Max. :39.80	Max. :64.00

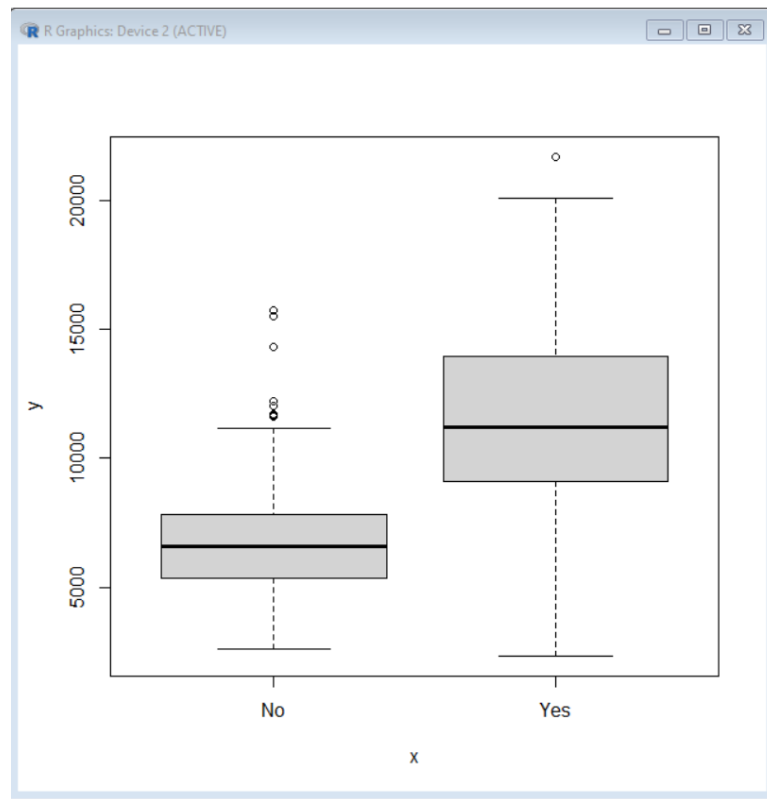
Expend	Grad.Rate
Min. : 3186	Min. : 10.00
1st Qu.: 6751	1st Qu.: 53.00
Median : 8377	Median : 65.00
Mean : 9660	Mean : 65.46
3rd Qu.:10830	3rd Qu.: 78.00
Max. :56233	Max. :118.00

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



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.

```

> summary(college)
Private      Apps      Accept      Enroll      Top10perc
No :212   Min.   :   81   Min.   :   72   Min.   :   35   Min.   : 1.00
Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.:  242   1st Qu.:15.00
        Median : 1558   Median : 1110   Median :  434   Median :23.00
        Mean   : 3002   Mean   : 2019   Mean   :  780   Mean   :27.56
        3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.:  902   3rd Qu.:35.00
        Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00

Top25perc   F.Undergrad   P.Undergrad   Outstate
Min.   : 9.0   Min.   : 139   Min.   : 1.0   Min.   : 2340
1st Qu.:41.0   1st Qu.: 992   1st Qu.: 95.0   1st Qu.: 7320
Median :54.0   Median :1707   Median : 353.0   Median : 9990
Mean   :55.8   Mean   :3700   Mean   : 855.3   Mean  :10441
3rd Qu.:69.0   3rd Qu.:4005   3rd Qu.: 967.0   3rd Qu.:12925
Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700

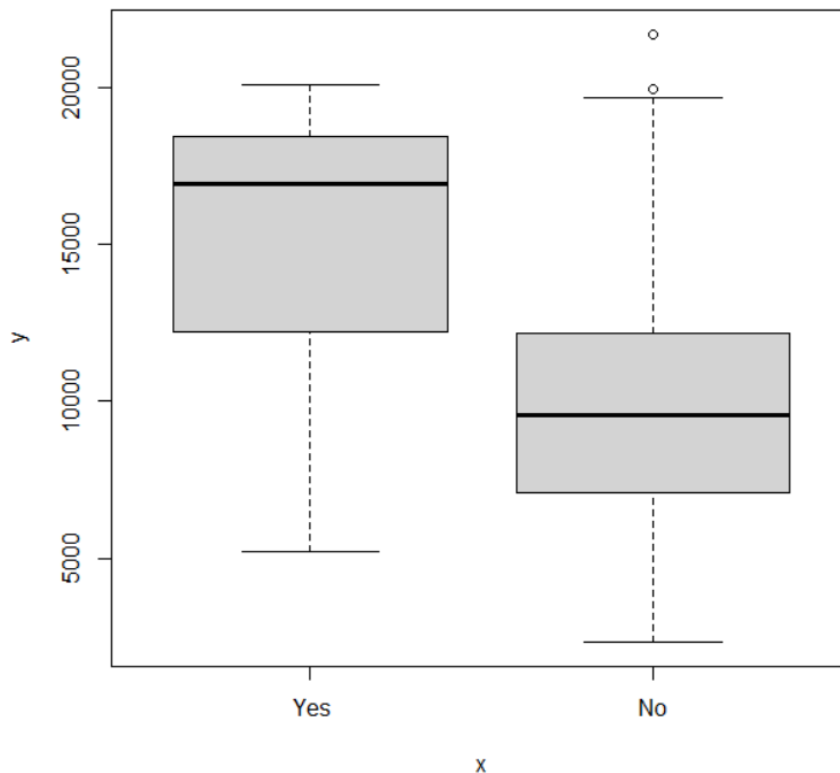
Room.Board   Books      Personal      PhD
Min.   :1780   Min.   : 96.0   Min.   : 250   Min.   : 8.00
1st Qu.:3597   1st Qu.:470.0   1st Qu.: 850   1st Qu.:62.00
Median :4200   Median :500.0   Median :1200   Median :75.00
Mean   :4358   Mean   :549.4   Mean   :1341   Mean  :72.66
3rd Qu.:5050   3rd Qu.:600.0   3rd Qu.:1700   3rd Qu.:85.00
Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00

Terminal     S.F.Ratio   perc.alumni   Expend
Min.   :24.0   Min.   : 2.50   Min.   : 0.00   Min.   :3186
1st Qu.:71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.:6751
Median :82.0   Median :13.60   Median :21.00   Median :8377
Mean   :79.7   Mean   :14.09   Mean   :22.74   Mean  :9660
3rd Qu.:92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233

Grad.Rate     Elite
Min.   :10.00   Yes: 78
1st Qu.:53.00   No :699
Median :65.00
Mean   :65.46
3rd Qu.:78.00
Max.   :118.00
> plot(college$Elit, college$Outstate)

```

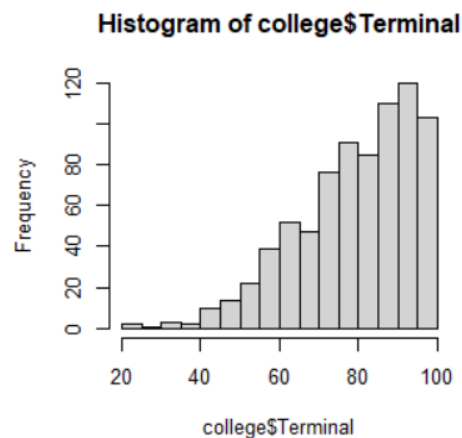
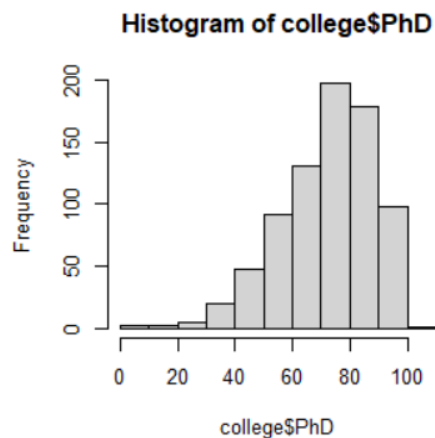
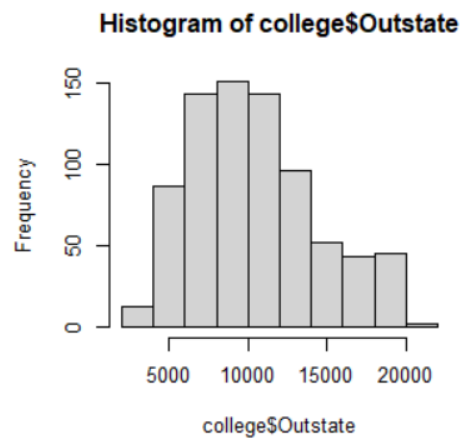
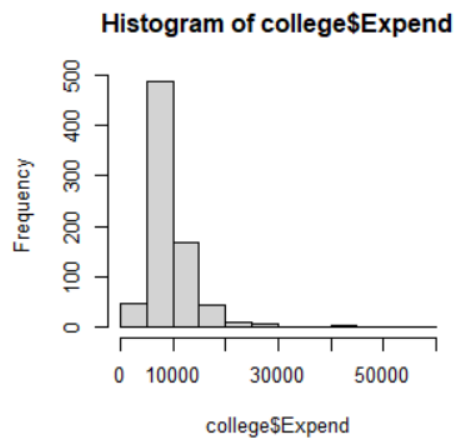
Також було побудовано графік залежності college\$Elit від college\$Outstate.



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.

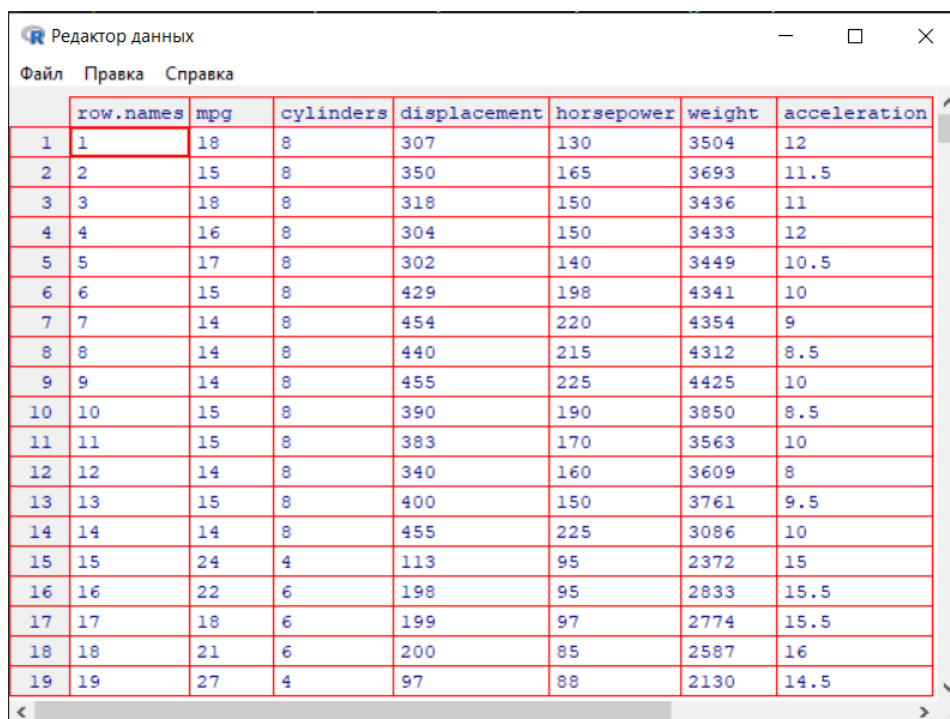
## Пункт 2 (Auto)

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)
```



	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration
1	1	18	8	307	130	3504	12
2	2	15	8	350	165	3693	11.5
3	3	18	8	318	150	3436	11
4	4	16	8	304	150	3433	12
5	5	17	8	302	140	3449	10.5
6	6	15	8	429	198	4341	10
7	7	14	8	454	220	4354	9
8	8	14	8	440	215	4312	8.5
9	9	14	8	455	225	4425	10
10	10	15	8	390	190	3850	8.5
11	11	15	8	383	170	3563	10
12	12	14	8	340	160	3609	8
13	13	15	8	400	150	3761	9.5
14	14	14	8	455	225	3086	10
15	15	24	4	113	95	2372	15
16	16	22	6	198	95	2833	15.5
17	17	18	6	199	97	2774	15.5
18	18	21	6	200	85	2587	16
19	19	27	4	97	88	2130	14.5

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[[ ]]))
```

mpg	cylinders	displacement	horsepower	weight
Min. : 9.00	3: 4	Min. : 68.0	Min. : 46.0	Min. :1613
1st Qu.:17.00	4:199	1st Qu.:105.0	1st Qu.: 75.0	1st Qu.:2225
Median :22.75	5: 3	Median :151.0	Median : 93.5	Median :2804
Mean :23.45	6: 83	Mean :194.4	Mean :104.5	Mean :2978
3rd Qu.:29.00	8:103	3rd Qu.:275.8	3rd Qu.:126.0	3rd Qu.:3615
Max. :46.60		Max. :455.0	Max. :230.0	Max. :5140

acceleration	year	origin	name
Min. : 8.00	73 : 40	1:245	amc matador : 5
1st Qu.:13.78	78 : 36	2: 68	ford pinto : 5
Median :15.50	76 : 34	3: 79	toyota corolla : 5
Mean :15.54	75 : 30		amc gremlin : 4
3rd Qu.:17.02	82 : 30		amc hornet : 4
Max. :24.80	70 : 29		chevrolet chevette: 4
	(Other):193		(Other) :365

**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])) }
```

```
9.0 46.6
68 455
46 230
1613 5140
8.0 24.8
```

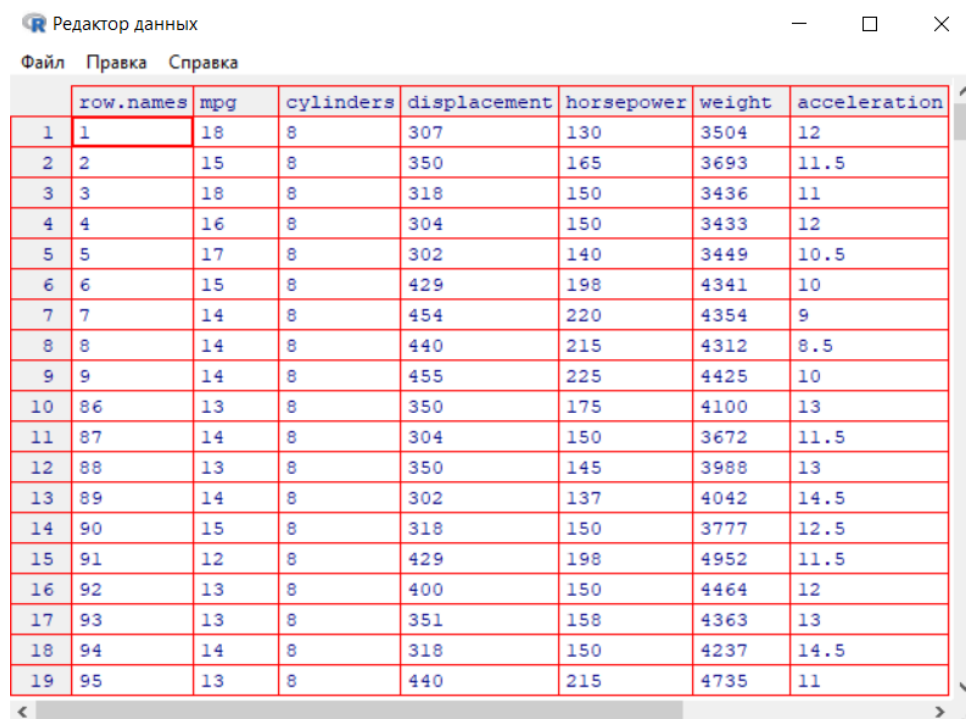
**2.3.** Обчислив середнє та стандартне відхилення для всіх кількісних показників за допомогою функцій `mean()` та `sd()` відповідно.

```
> for (val in quantitative) { print(paste("Mean", mean(autos[, val]),
"- Sd", sd(autos[, val]))) }
```

```
"Mean 23.4459183673469 - Sd 7.8050074865718"  
"Mean 194.411989795918 - Sd 104.644003908905"  
"Mean 104.469387755102 - Sd 38.4911599328285"  
"Mean 2977.58418367347 - Sd 849.402560042949"  
"Mean 15.5413265306122 - Sd 2.75886411918808"
```

**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)
```



Редактор данных

Файл Правка Справка

	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration
1	1	18	8	307	130	3504	12
2	2	15	8	350	165	3693	11.5
3	3	18	8	318	150	3436	11
4	4	16	8	304	150	3433	12
5	5	17	8	302	140	3449	10.5
6	6	15	8	429	198	4341	10
7	7	14	8	454	220	4354	9
8	8	14	8	440	215	4312	8.5
9	9	14	8	455	225	4425	10
10	86	13	8	350	175	4100	13
11	87	14	8	304	150	3672	11.5
12	88	13	8	350	145	3988	13
13	89	14	8	302	137	4042	14.5
14	90	15	8	318	150	3777	12.5
15	91	12	8	429	198	4952	11.5
16	92	13	8	400	150	4464	12
17	93	13	8	351	158	4363	13
18	94	14	8	318	150	4237	14.5
19	95	13	8	440	215	4735	11

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

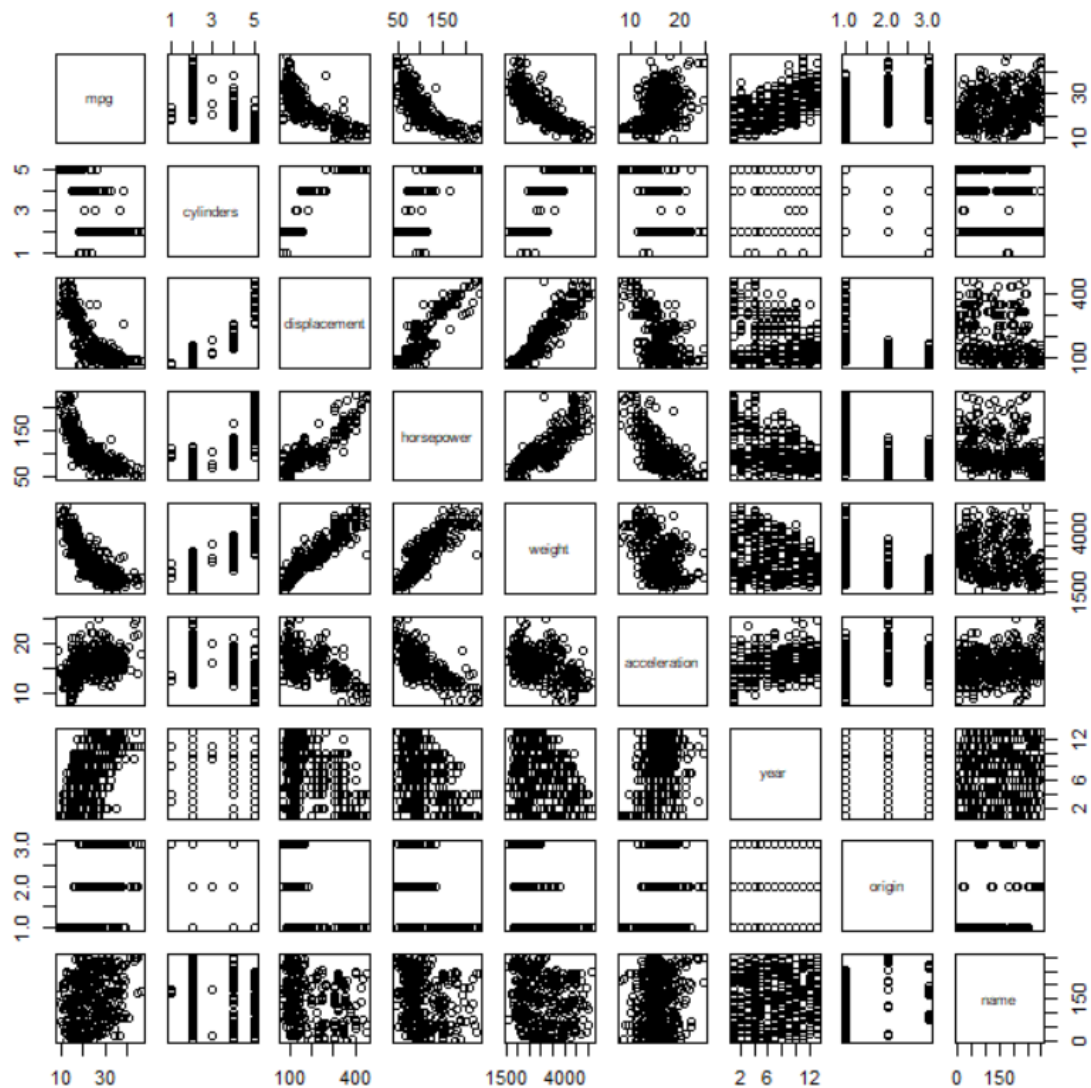
```
"Mean 24.3684542586751 - Sd 7.88089834213537"  
"Mean 187.753943217666 - Sd 99.9394881404822"  
"Mean 100.955835962145 - Sd 35.8955667771312"  
"Mean 2939.64353312303 - Sd 812.649629297648"  
"Mean 15.7182965299685 - Sd 2.69381257754339"
```

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

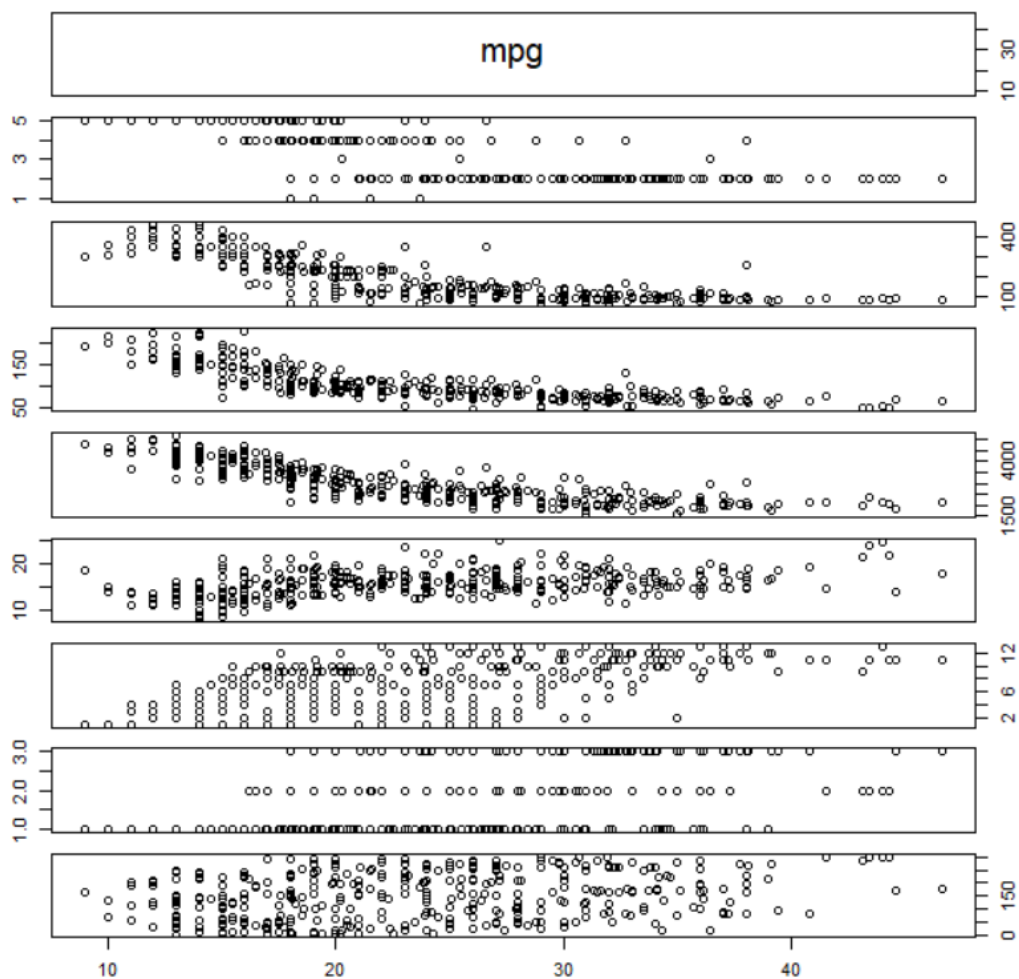
```
"Mean diff 0.922535891328142 - Sd diff 0.0758908555635696"  
"Mean diff -6.65804657825277 - Sd diff -4.70451576842248"  
"Mean diff -3.51355179295693 - Sd diff -2.59559315569726"  
"Mean diff -37.9406505504412 - Sd diff -36.7529307453017"  
"Mean diff 0.176969999356208 - Sd diff -0.0650515416446926"
```

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



Підсумовуючи наведені вище графіки, можна сказати, що серед кількісних показників є досить часто виражена лінійна залежність між собою. Також певні нелінійні зв'язки є між показником `mpg` та показниками `displacement`, `horsepower` та `weight`.

**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, horsepower 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.

```
> Boston
      crim      zn  indus  chas      nox      rm      age      dis rad tax ptratio  black
1  0.00632  18.0   2.31    0  0.5380  6.575  65.2  4.0900  1 296    15.3 396.90
2  0.02731   0.0   7.07    0  0.4690  6.421  78.9  4.9671  2 242    17.8 396.90
3  0.02729   0.0   7.07    0  0.4690  7.185  61.1  4.9671  2 242    17.8 392.83
4  0.03237   0.0   2.18    0  0.4580  6.998  45.8  6.0622  3 222    18.7 394.63
5  0.06905   0.0   2.18    0  0.4580  7.147  54.2  6.0622  3 222    18.7 396.90
6  0.02985   0.0   2.18    0  0.4580  6.430  58.7  6.0622  3 222    18.7 394.12
7  0.08829  12.5   7.87    0  0.5240  6.012  66.6  5.5605  5 311    15.2 395.60
8  0.14455  12.5   7.87    0  0.5240  6.172  96.1  5.9505  5 311    15.2 396.90
9  0.21124  12.5   7.87    0  0.5240  5.631 100.0  6.0821  5 311    15.2 386.63
10 0.17004  12.5   7.87    0  0.5240  6.004  85.9  6.5921  5 311    15.2 386.71
11 0.22489  12.5   7.87    0  0.5240  6.377  94.3  6.3467  5 311    15.2 392.52
12 0.11747  12.5   7.87    0  0.5240  6.009  82.9  6.2267  5 311    15.2 396.90
13 0.09378  12.5   7.87    0  0.5240  5.889  39.0  5.4509  5 311    15.2 390.50
14 0.62976   0.0   8.14    0  0.5380  5.949  61.8  4.7075  4 307    21.0 396.90
15 0.63796   0.0   8.14    0  0.5380  6.096  84.5  4.4619  4 307    21.0 380.02
16 0.62739   0.0   8.14    0  0.5380  5.834  56.5  4.4986  4 307    21.0 395.62
17 1.05393   0.0   8.14    0  0.5380  5.935  29.3  4.4986  4 307    21.0 386.85
18 0.78420   0.0   8.14    0  0.5380  5.990  81.7  4.2579  4 307    21.0 386.75
19 0.80271   0.0   8.14    0  0.5380  5.456  36.6  3.7965  4 307    21.0 288.99
20 0.72580   0.0   8.14    0  0.5380  5.727  69.5  3.7965  4 307    21.0 390.95
21 1.25179   0.0   8.14    0  0.5380  5.570  98.1  3.7979  4 307    21.0 376.57
22 0.85204   0.0   8.14    0  0.5380  5.965  89.2  4.0123  4 307    21.0 392.53
23 1.23247   0.0   8.14    0  0.5380  6.142  91.7  3.9769  4 307    21.0 396.90
24 0.98843   0.0   8.14    0  0.5380  5.813 100.0  4.0952  4 307    21.0 394.54
25 0.75026   0.0   8.14    0  0.5380  5.924  94.1  4.3996  4 307    21.0 394.33
26 0.84054   0.0   8.14    0  0.5380  5.599  85.7  4.4546  4 307    21.0 303.42
27 0.67191   0.0   8.14    0  0.5380  5.813  90.3  4.6820  4 307    21.0 376.88
28 0.95577   0.0   8.14    0  0.5380  6.047  88.8  4.4534  4 307    21.0 306.38
29 0.77299   0.0   8.14    0  0.5380  6.495  94.4  4.4547  4 307    21.0 387.94
30 1.00245   0.0   8.14    0  0.5380  6.674  87.3  4.2390  4 307    21.0 380.23
```

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

← → ↺

127.0.0.1:25613/library/MASS/html/Boston.html

☆

Facebook

Навчання

IT

Для Душі

Кіношки

Монети

Погода

Unity

DreamTeam

Vector

Бухгалтерія НР у ТТ - ...

// TODO: - Google До...

Volia Speed Test

Boston {MASS}

Housing Values in Suburbs of Boston

Description

The Boston data frame has 506 rows and 14 columns.

Usage

Boston



## Format

This data frame contains the following columns:

`crim`

per capita crime rate by town.

`zn`

proportion of residential land zoned for lots over 25,000 sq.ft.

`indus`

proportion of non-retail business acres per town.

`chas`

Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

`nox`

nitrogen oxides concentration (parts per 10 million).

`rm`

average number of rooms per dwelling.

`age`

proportion of owner-occupied units built prior to 1940.

`dis`

weighted mean of distances to five Boston employment centres.

`rad`

index of accessibility to radial highways.

`tax`

full-value property-tax rate per \$10,000.

`ptratio`

pupil-teacher ratio by town.

`black`

$1000(Bk - 0.63)^2$  where  $Bk$  is the proportion of blacks by town.

`lstat`

lower status of the population (percent).

`medv`

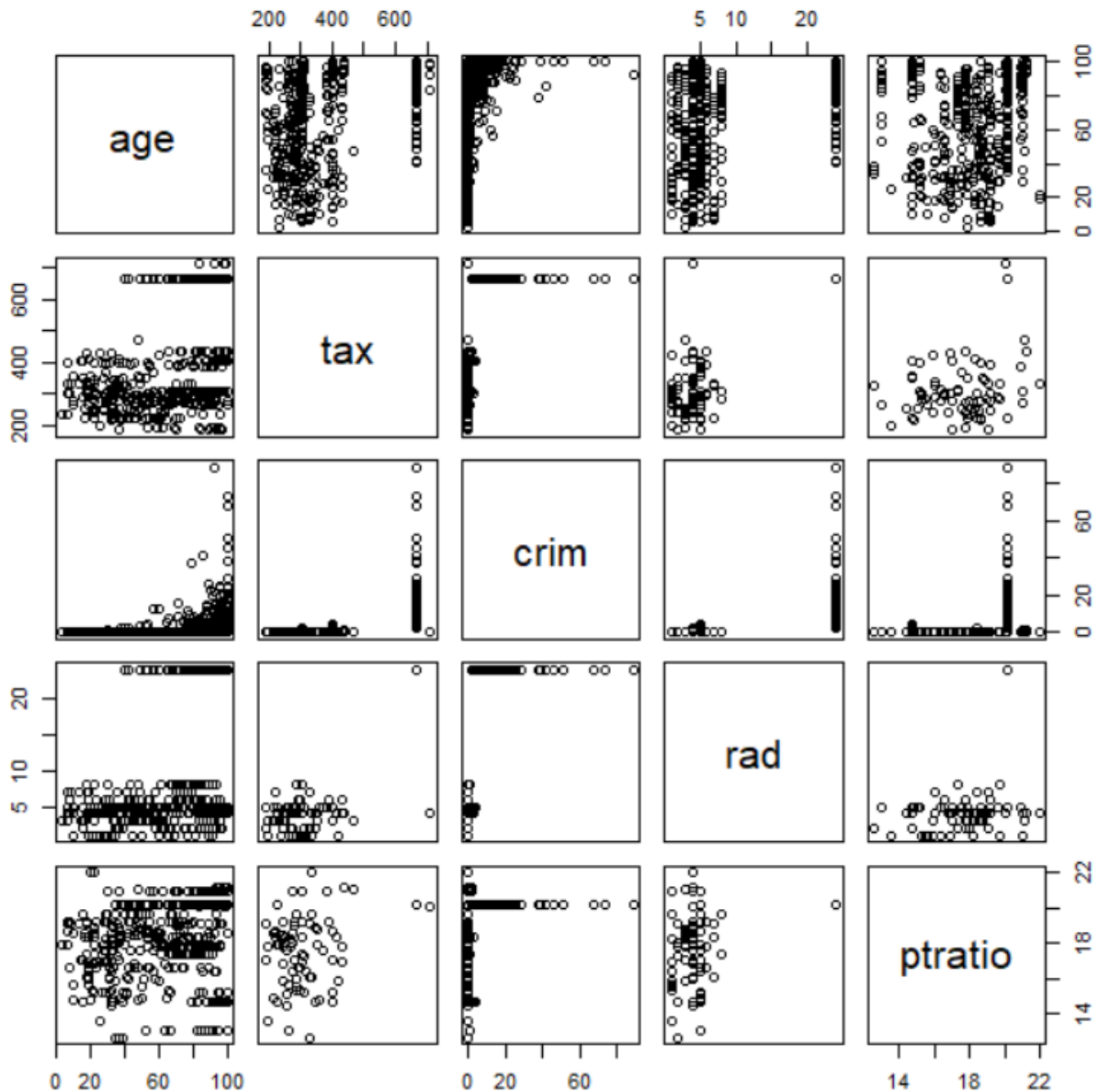
median value of owner-occupied homes in \$1000s.

## Source

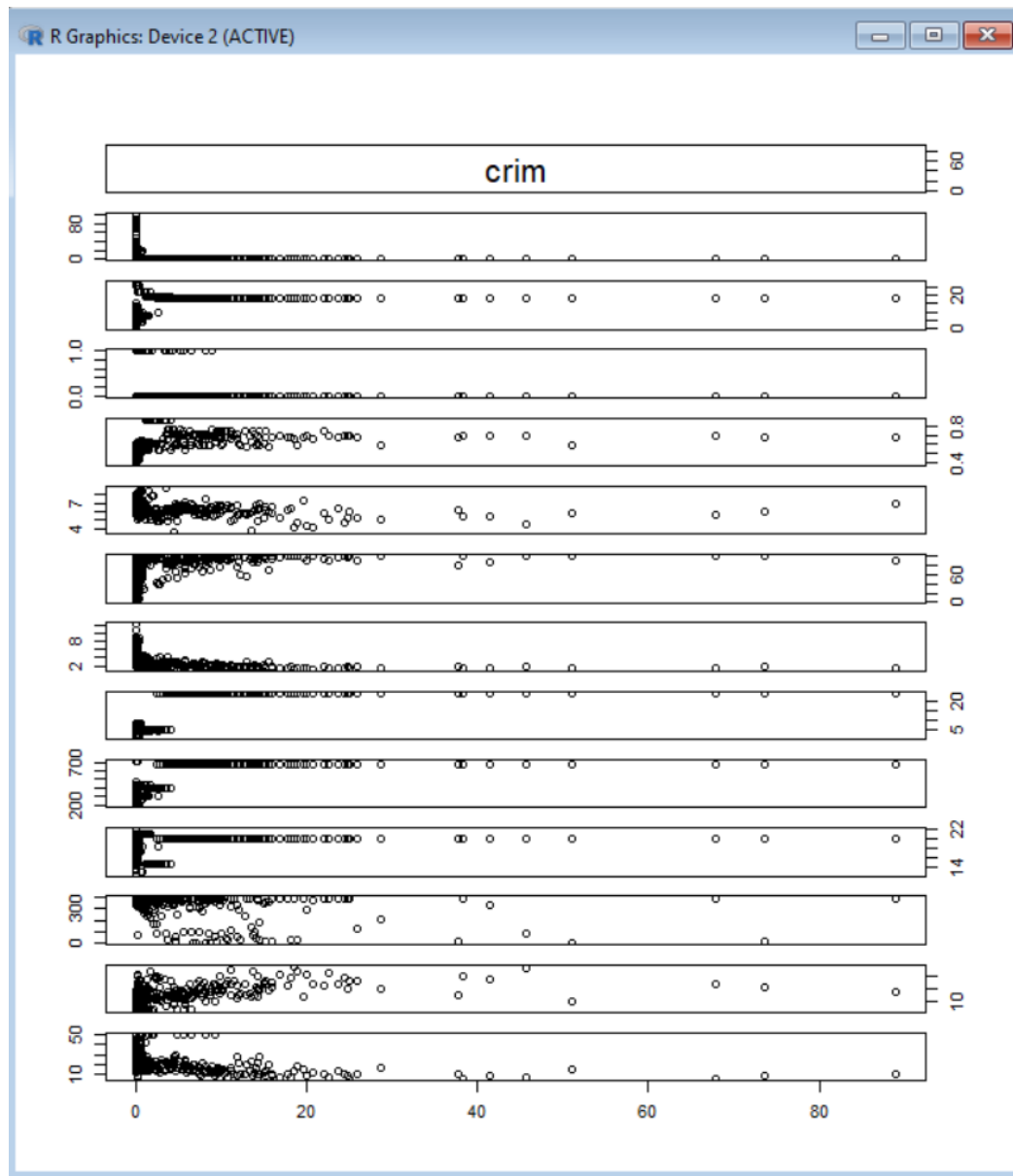
Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics and Management* **5**, 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

**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.



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.

```

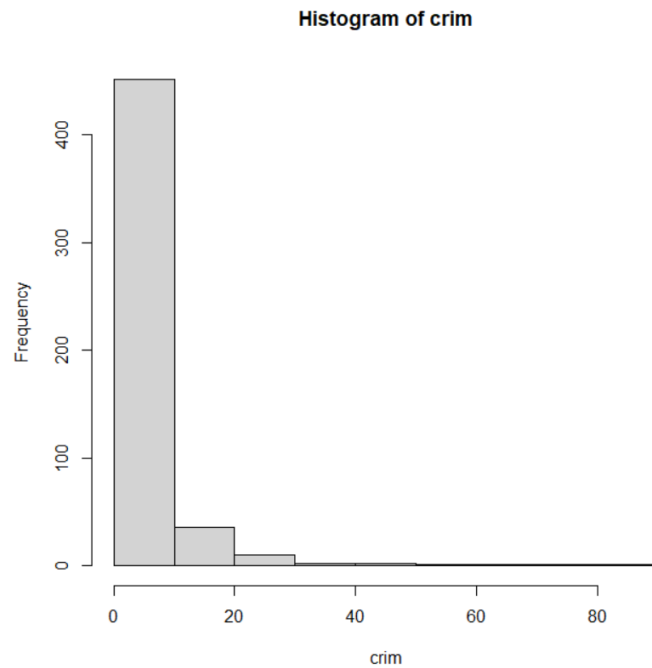
      crim          zn          indus      chas          nox
Min.   : 0.00632   Min.   : 0.00   Min.   : 0.46   0:471   Min.   :0.3850
1st Qu.: 0.08205   1st Qu.: 0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490
Median : 0.25651   Median : 0.00   Median : 9.69           Median :0.5380
Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547
3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240
Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710

      rm          age          dis          rad
Min.   :3.561     Min.   : 2.90   Min.   : 1.130   Min.   : 1.000
1st Qu.:5.886     1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000
Median :6.208     Median : 77.50   Median : 3.207   Median : 5.000
Mean   :6.285     Mean   : 68.57   Mean   : 3.795   Mean   : 9.549
3rd Qu.:6.623     3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000
Max.   :8.780     Max.   :100.00   Max.   :12.127   Max.   :24.000

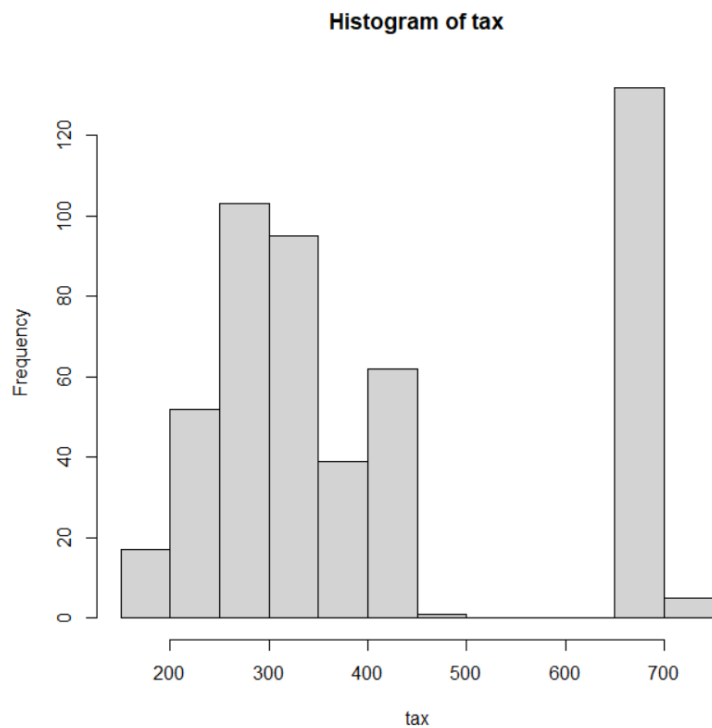
      tax          ptratio          black          lstat
Min.   :187.0     Min.   :12.60   Min.   : 0.32   Min.   : 1.73
1st Qu.:279.0     1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95
Median :330.0     Median :19.05   Median :391.44   Median :11.36
Mean   :408.2     Mean   :18.46   Mean   :356.67   Mean   :12.65
3rd Qu.:666.0     3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95
Max.   :711.0     Max.   :22.00   Max.   :396.90   Max.   :37.97

      medv
Min.   : 5.00
1st Qu.:17.02
Median :21.20
Mean   :22.53
3rd Qu.:25.00
Max.   :50.00

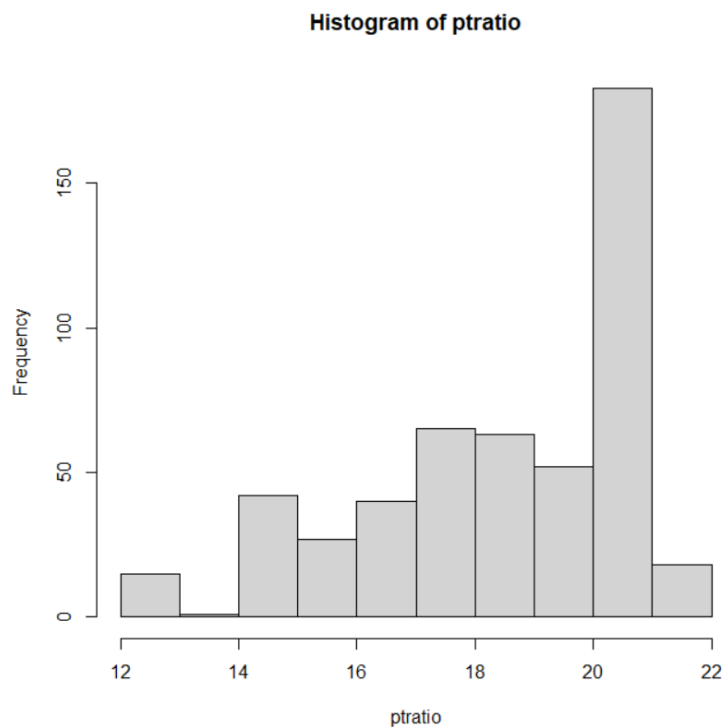
```



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 в значенні 21.

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

```
chas  
0: 471  
1: 35
```

Або запиту нижче

```
> print(paste("Bounds river:", sum(Boston$chas == 1)))  
[1] "Bounds river: 35"
```

З цього випливає що таких записів є 35.

### 3.6 Медіану визначено за допомогою функції median.

```
> print(paste("Median P/t ratio :", median(Boston$ptratio)))  
[1] "Median P/t ratio : 19.05"
```

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.

```
> min_medv = Boston[Boston$medv == min(Boston$medv), ]  
> print(min_medv)
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.90	30.59
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24	666	20.2	384.97	22.98

	medv
399	5
406	5

Тобто, було знайдено два квартали з найнижчою медіаною кількості зайнятих 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.

```
> print(min_medv[,-4] - apply(Boston[,-4], 2, mean))
```

	crim	zn	indus	nox	rm	age	dis
399	34.73828	-11.1367787	11.81537	-3.102043	-402.78415	-256.67403	-21.043206
406	56.55716	-0.5546951	-50.47490	-8.856407	-12.77253	87.34694	-2.188124

	rad	tax	ptratio	black	lstat	medv
399	12.63636	665.4453	-48.37490	387.35059	12.13447	-7.653063
406	12.86322	659.7154	16.40496	-23.26715	-333.69403	-17.532806

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.