

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА
Факультет прикладної математики та інформатики
Кафедра програмування



Індивідуальне завдання №1
Вступ до R

Виконала:
студентка групи ПМОм-11
Кравець Ольга

Львів 2025

Хід роботи

Варіант - 3

Встановлюю значення змінної `variant`: для цього 0 (номер групи ПМОм-11)*25 + 3 (порядковий номер в списку групи) = 3.

```
> variant=3  
> variant  
[1] 3
```

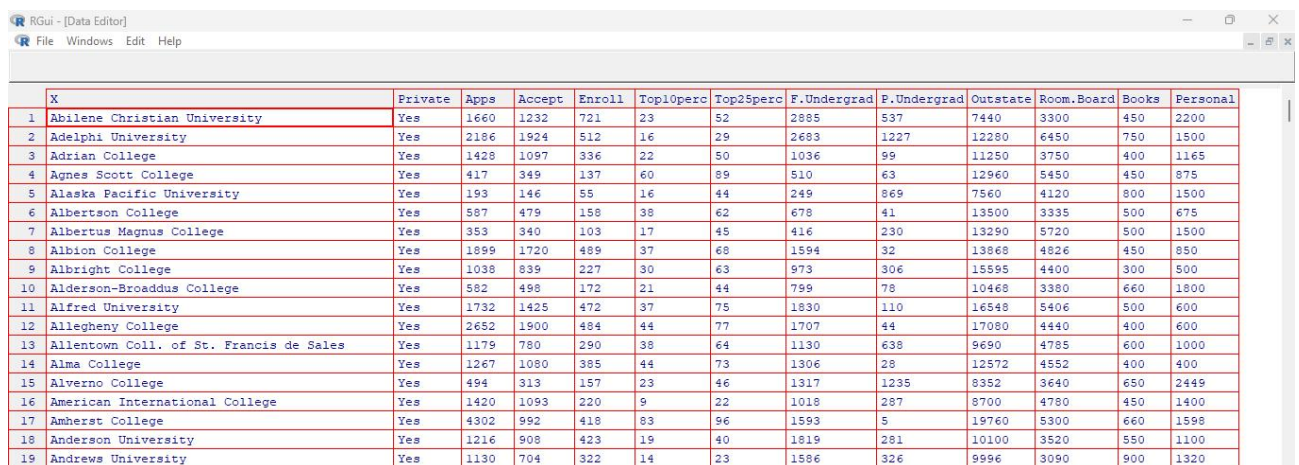
Далі встановлюю `set.seed(variant)` та генерую значення змінної `redundant` як заокруглене до цілого випадкове число з рівномірного на інтервалі (5, 25) розподілу. Для заокруглення використовую функцію `floor`, а для вибору випадкового числа з інтервалу функцію `runif`.

```
> set.seed(variant)  
> redundant=floor(runif(1,5,25))  
> redundant  
[1] 8
```

Завдання 1.

Для того, щоб завантажити дані з файлу `College.csv` спочатку використовую функцію `setwd()`, яка визначає шлях до необхідного файлу. Для зчитування даних використовую функцію `read.csv()`

```
> setwd('D:\\Навчання\\Магістратура\\Моделі статистичного навчання\\01')  
> College=read.csv('College.csv')  
> fix(College)
```



	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	F.Undergrad	Outstate	Room.Board	Books	Personal
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500
3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165
4	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875
5	Alaska Pacific University	Yes	193	146	85	16	44	249	869	7560	4120	800	1500
6	Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675
7	Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290	5720	500	1500
8	Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850
9	Albright College	Yes	1038	839	227	30	63	973	306	15595	4400	300	500
10	Alderson-Broadus College	Yes	582	498	172	21	44	799	78	10468	3380	660	1800
11	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600
12	Allegheny College	Yes	2652	1900	484	44	77	1707	44	17080	4440	400	600
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000
14	Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552	400	400
15	Alverno College	Yes	494	313	157	23	46	1317	1235	8352	3640	650	2449
16	American International College	Yes	1420	1093	220	9	22	1018	287	8700	4780	450	1400
17	Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300	660	1598
18	Anderson University	Yes	1216	908	423	19	40	1819	281	10100	3520	550	1100
19	Andrews University	Yes	1130	704	322	14	23	1586	326	9996	3090	900	1320

Далі використовую функцію `sample()` для модифікації завантажених даних `College` - видалення `redundant` (% спостережень).

```
> College_new=
+ College[-
+ sample(1:length(College[,1]),round((redundant/100)*length(College[,1]))),]
> fix(College_new)
```

	row.names	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
1	1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450
2	2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750
3	3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400
4	4	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450
5	5	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800
6	6	Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500
7	7	Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290	5720	500
8	8	Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826	450
9	9	Albright College	Yes	1038	839	227	30	63	973	306	15595	4400	300
10	10	Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10468	3380	660
11	11	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406	500
12	13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690	4785	600
13	14	Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552	400
14	16	American International College	Yes	1420	1093	220	9	22	1018	287	8700	4780	450
15	17	Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300	660
16	18	Anderson University	Yes	1216	908	423	19	40	1819	281	10100	3520	550
17	19	Andrews University	Yes	1130	704	322	14	23	1586	326	9996	3090	900
18	20	Angelo State University	No	3540	2001	1016	24	54	4190	1512	5130	3592	500
19	21	Antioch University	Yes	713	661	252	25	44	712	23	15476	3336	400

Створюю змінну rownames, щоб помістити в неї перший стовпець таблиці – назви університетів. Далі ініціалізую нову таблицю без вищезгаданого стовпця та переглядаю дані, використовуючи функцію fix().

```
> rownames=College_new[,1]
> College_new=College_new[,-1]
> fix(College_new)
```

	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni
1	1	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12
2	2	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16
3	3	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30
4	4	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37
5	5	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2
6	6	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11
7	7	Yes	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26
8	8	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37
9	9	Yes	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23
10	10	Yes	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15
11	11	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600	82	88	11.3	31
12	13	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000	60	84	13.3	21
13	14	Yes	1267	1080	385	44	73	1306	28	12572	4552	400	400	79	87	15.3	32
14	16	Yes	1420	1093	220	9	22	1018	287	8700	4780	450	1400	78	84	14.7	19
15	17	Yes	4302	992	418	83	96	1593	5	19760	5300	660	1598	93	98	8.4	63
16	18	Yes	1216	908	423	19	40	1819	281	10100	3520	550	1100	48	61	12.1	14
17	19	Yes	1130	704	322	14	23	1586	326	9996	3090	900	1320	62	66	11.5	18
18	20	No	3540	2001	1016	24	54	4190	1512	5130	3592	500	2000	60	62	23.1	5
19	21	Yes	713	661	252	25	44	712	23	15476	3336	400	1100	69	82	11.3	35

Що тепер сталося з першим стовпцем? Він зник, бо фактично, друга команда у попередньому блоці коду переписала існуючі дані такими, які не містять назв університетів. тобто не містять першого стовпця.

Для чого змінна rownames? На випадок, якщо нас зацікавлять певні показники і ми захочемо дізнатись назву університету - ми

зможемо звернутись до цієї змінної і витягнути цю назву за допомогою стовпця rownames. Тобто, для полегшення аналізу даних.

Використовую функцію summary() і отримала різні статистичні дані, такі як мінімум, максимум, перший квартиль та третій квартиль, медіану та середнє статистичне для усіх числових значень. Оскільки Private має не факторизовані значення - для цієї групи даних отримала тільки кількість елементів та клас "character":

```
> summary(College)
```

X	Private	Apps	Accept	Enroll
Length:777	Length:777	Min. : 81	Min. : 72	Min. : 35
Class :character	Class :character	1st Qu.: 776	1st Qu.: 604	1st Qu.: 242
Mode :character	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

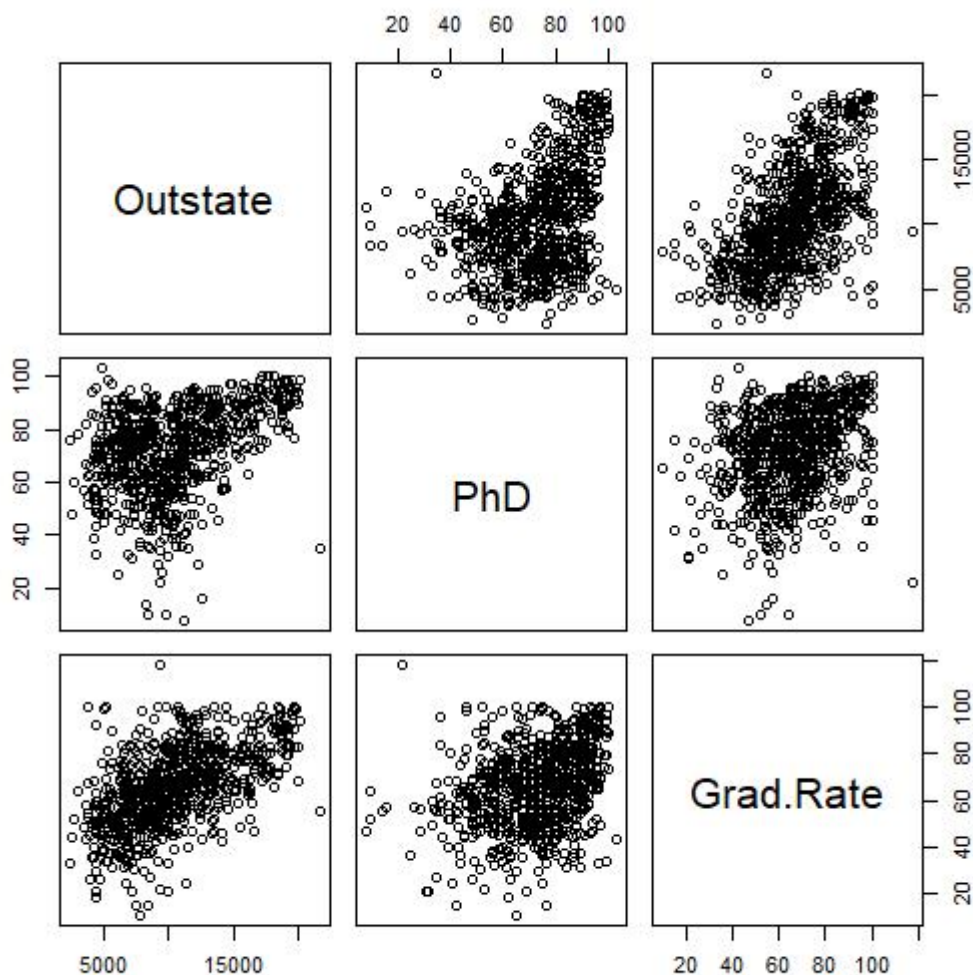
Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Min. : 1.00	Min. : 9.0	Min. : 139	Min. : 1.0	Min. : 2340
1st Qu.:15.00	1st Qu.: 41.0	1st Qu.: 992	1st Qu.: 95.0	1st Qu.: 7320
Median :23.00	Median : 54.0	Median : 1707	Median : 353.0	Median : 9990
Mean :27.56	Mean : 55.8	Mean : 3700	Mean : 855.3	Mean :10441
3rd Qu.:35.00	3rd Qu.: 69.0	3rd Qu.: 4005	3rd Qu.: 967.0	3rd Qu.:12925
Max. :96.00	Max. :100.0	Max. :31643	Max. :21836.0	Max. :21700

Room.Board	Books	Personal	PhD	Terminal
Min. :1780	Min. : 96.0	Min. : 250	Min. : 8.00	Min. : 24.0
1st Qu.:3597	1st Qu.: 470.0	1st Qu.: 850	1st Qu.: 62.00	1st Qu.: 71.0
Median :4200	Median : 500.0	Median :1200	Median : 75.00	Median : 82.0
Mean :4358	Mean : 549.4	Mean :1341	Mean : 72.66	Mean : 79.7
3rd Qu.:5050	3rd Qu.: 600.0	3rd Qu.:1700	3rd Qu.: 85.00	3rd Qu.: 92.0
Max. :8124	Max. :2340.0	Max. :6800	Max. :103.00	Max. :100.0

S.F.Ratio	perc.alumni	Expend	Grad.Rate
Min. : 2.50	Min. : 0.00	Min. : 3186	Min. : 10.00
1st Qu.:11.50	1st Qu.:13.00	1st Qu.: 6751	1st Qu.: 53.00
Median :13.60	Median :21.00	Median : 8377	Median : 65.00
Mean :14.09	Mean :22.74	Mean : 9660	Mean : 65.46
3rd Qu.:16.50	3rd Qu.:31.00	3rd Qu.:10830	3rd Qu.: 78.00
Max. :39.80	Max. :64.00	Max. :56233	Max. :118.00

Використовую функцію `pairs()` для перегляду графіків відношень стовбців `Outstate`, `PhD` та `Grad.Rate`.

```
> pairs(~ Outstate + PhD + Grad.Rate, data=College)
```

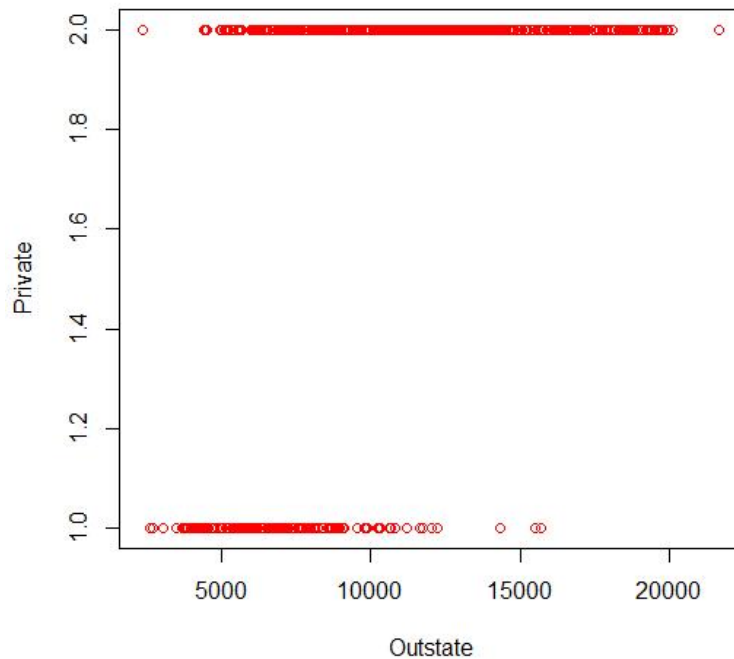


Із отриманих даних можемо сказати, що всі стовбці між собою певною мірою корелюють. Серед них найбільш помітна кореляція між `Outstate` та `Grad.Rate`.

Будую діаграму `Outstate` vs `Private`, використовуючи `plot()`.

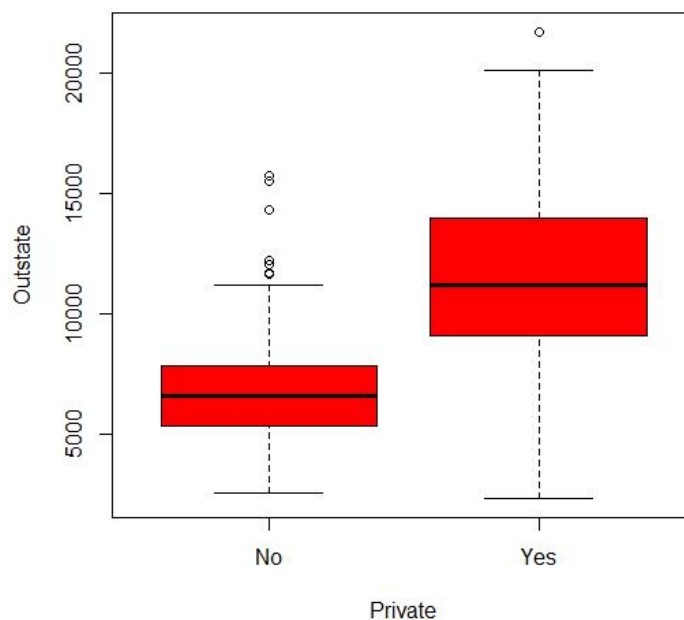
```
> x=College[,10]
> y=College[,2]
> y=as.factor(y)
> plot(x,y,,col="red",xlab='Outstate',ylab='Private')
```

У результаті вийшов графік.



Але такий графік мені не підходить, тому міняю значення x та y місцями і отримую блочний графік залежності.

```
> x=College[,2]
> y=College[,10]
> x=as.factor(x)
> plot(x,y,,col="red",xlab='Private',ylab='Outstate')
```



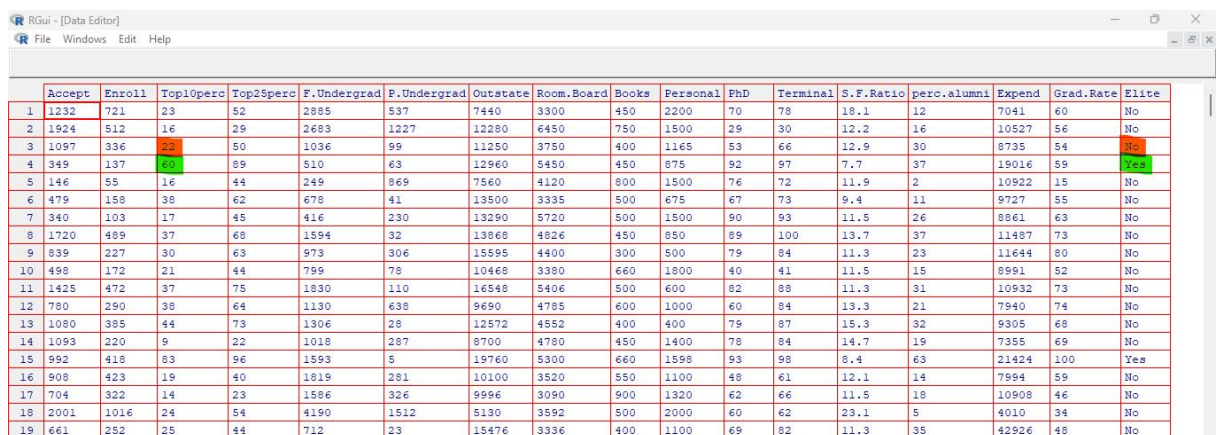
Діаграма чітко демонструє, що в приватних університетах вартість навчання значно вища, ніж у державних. У державних: ~5000-7500, у приватних: ~8000-14000.

Створюю новий якісний показник Elite, використовуючи Top10perc, тобто поділивши всі університети на дві групи в залежності чи перевищує відсоток студентів з топ 10% шкіл 50% чи ні.

```
> Elite=rep("No",nrow(College_new))
> Elite[College_new$Top10perc>50]="Yes"
> Elite=as.factor(Elite)
> College_new=data.frame(College_new,Elite)

> table(Elite)
Elite
No Yes
643 72
```

Якщо кількість студентів зі шкіл, які входять в Top10perc менше 50, то у створеній колонці відображається “No”, якщо навпаки - “Yes”.



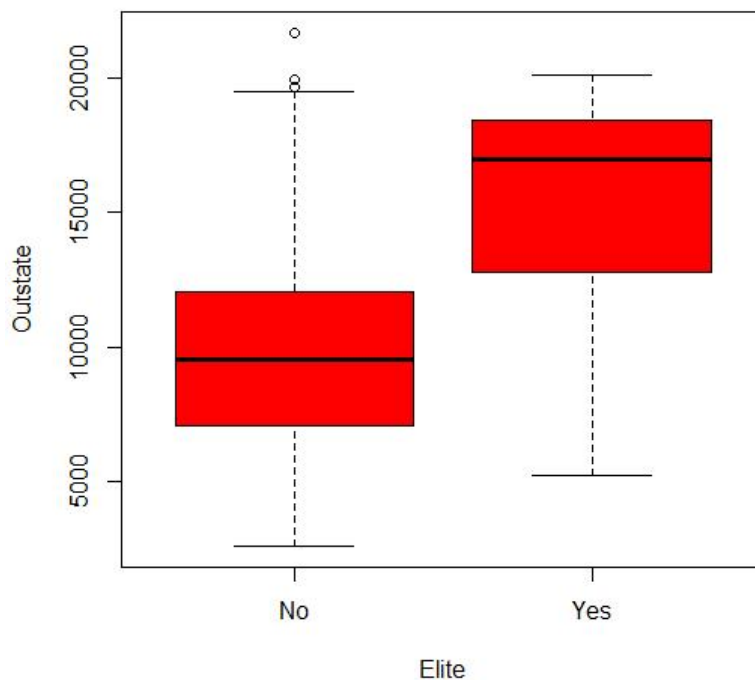
	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	Elite
1	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60	No
2	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56	No
3	1097	336	25	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54	No
4	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59	Yes
5	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15	No
6	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55	No
7	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26	8861	63	No
8	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37	11487	73	No
9	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23	11644	80	No
10	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15	8991	52	No
11	1425	472	37	75	1830	110	16548	5406	500	600	82	88	11.3	31	10932	73	No
12	780	290	38	64	1130	638	9690	4785	600	1000	60	84	13.3	21	7940	74	No
13	1080	385	44	73	1306	28	12572	4552	400	400	79	87	15.3	32	9305	68	No
14	1093	220	9	22	1018	287	8700	4780	450	1400	78	84	14.7	19	7355	69	No
15	992	418	83	96	1593	5	19760	5300	660	1598	93	98	8.4	63	21424	100	Yes
16	908	423	19	40	1819	281	10100	3520	550	1100	48	61	12.1	14	7994	59	No
17	704	322	14	23	1586	326	9996	3090	900	1320	62	66	11.5	18	10908	46	No
18	2001	1016	24	54	4190	1512	5130	3592	500	2000	60	62	23.1	5	4010	34	No
19	661	252	25	44	712	23	15476	3336	400	1100	69	82	11.3	35	42926	48	No

Для чого дві останні команди? Перша з двох останніх команд факторизує новоутворений показник Elite. Друга додає новий показник утворюючи новий датафрейм із існуючого College_new та щойно утвореного Elite.

Чи багато таких університетів? Ні, ~11%.

Будую діаграму Outstate vs Elite.

```
> x=College_new$Elite
> y=College_new$Outstate
> plot(x,y,,col="red",xlab='Elite',ylab='Outstate')
```



Розглядаю цей графік і можу сказати, що у не елітних університетах вартість навчання є нижчою ніж у елітних.

Завдання 2.

Завантажую бібліотеку MASS та переглядаю дані про об'єкт Boston.

```
> library(MASS)
> ?Boston
```

Boston {MASS}

R Documentation

Housing Values in Suburbs of Boston

Description

The Boston data frame has 506 rows and 14 columns.

Usage

Boston

Скільки рядків і стовпців міститься в множині? У цій множині є 506 рядків та 14 стовпців.

Використовую функцію `sample()` для модифікації завантажених даних Boston - видалення redundant (% спостережень).

```
> Boston_new=Boston[-sample(1:length(Boston[,1]),round((redundant/100)*length(Boston[,1]))),]
> fix(Boston_new)
```


RGui - [Data Editor]															
File Windows Edit Help															
	row.names	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
2	2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
5	6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
6	7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
7	8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
8	9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
9	11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
10	12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
11	13	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
12	14	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
13	15	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
14	17	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
15	18	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
16	20	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2
17	21	1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6
18	22	0.85204	0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21	392.53	13.83	19.6
19	23	1.23247	0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21	396.9	18.72	15.2

Скільки кварталів в даній множині межують з річкою Charles?

У документації Boston є наступна інформація.

`chas`

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

Отже, потрібно знайти ті дані, в яких `chas = 1`.

```
> chas_bounds=which(Boston_new$chas==1)
> length(chas_bounds)
[1] 34
```

Відповідь: таких кварталів 34.

Обчисліть медіану для відношення учні-вчителі для міста загалом?

У документації Boston є наступна інформація.

`ptratio`

pupil-teacher ratio by town.

Отже, стовпець `ptratio` відповідає відношенню учні-вчителі.

```
> median(Boston_new$ptratio)
[1] 19.1
```

Відповідь: медіана 19.1.

Які квартали міста мають найменше та найбільше відношення учні-вчителі?

```
> ptratio_data=Boston_new$ptratio
> which(ptratio_data==min(ptratio_data))
[1] 176 177 178
> which(ptratio_data==max(ptratio_data))
[1] 321 322
```

Кwartали 176, 177 та 178 мають найменше відношення учні-вчителі, а 321 та 322 - найбільше.

В яких кварталах в середньому є більше 7 кімнат в помешканні? Більше 8?

У документації Boston є наступна інформація.

```
rm

average number of rooms per dwelling.
```

Отже, треба шукати по стовпцю rm. Виконую наступні команди.

```
> rm_data=Boston_new$rm
> which(rm_data>7)
[1] 3 4 35 49 57 78 79 86 87 142 143 144 147 161 163 166 169 172 175 176 177 178 180
[24] 182 183 184 203 204 205 206 207 210 211 212 216 230 233 234 235 237 238 239 240 241 242 247
[47] 250 254 256 257 258 265 273 275 277 309 331 337 342 417 444
> which(rm_data>8)
[1] 86 144 184 203 204 205 211 212 230 234 238 241 331
```

Більше 4, але менше 7?

```
> which(rm_data>4 & rm_data<7)
[1] 1 2 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
[23] 25 26 27 28 29 30 31 32 33 34 36 37 38 39 40 41 42 43 44 45 46 47
[45] 48 50 51 52 53 54 55 56 58 59 60 61 62 63 64 65 66 67 68 69 70 71
[67] 72 73 74 75 76 77 80 81 82 83 84 85 88 89 90 91 92 93 94 95 96 97
[89] 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
[111] 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141
[133] 145 146 148 149 150 151 152 153 154 155 156 157 158 159 160 162 164 165 167 168 170 171
[155] 173 174 179 181 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202
[177] 208 209 213 214 215 217 218 219 220 221 222 223 224 225 226 227 228 229 231 232 236 243
[199] 244 245 246 248 249 251 252 253 255 259 260 261 262 263 264 266 267 268 269 270 271 272
[221] 274 276 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297
[243] 298 299 300 301 302 303 304 305 306 307 308 310 311 312 313 314 315 316 317 318 319 320
[265] 321 322 323 324 325 326 327 328 329 330 333 335 336 338 339 340 341 343 344 345 346 347
[287] 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369
[309] 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391
[331] 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413
[353] 414 415 416 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436
[375] 437 438 439 440 441 442 443 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459
[397] 460 461 462 463 464 465 466
```