ВАЕВ035 Икономическа (Бизнес) статистика

Задачи за упражнения

гл. ас. д-р Слав Емилов Ангелов, НБУ

Типове данни

Задача:

Каква е разликата между основните два типа данни – категорни и числови? Дайте примери за всеки един от тях.

Типове данни

Задача:

Каква е разликата между основните два типа данни – категорни и числови? Дайте примери за всеки един от тях.

Решение:

- Разстоянията между всяка една числова стойност могат да бъдат точно определни – техният размер и позиция са известни. Примери за числови данни са цената на газта, скорост на автомобил, тегло на човек ....
- Категовните данни нямат точно разтояние по между си. Примери: цвят на коса, цвят на очи, вид растение, име на фирма и т.н. Важно е да се отбележи,че те се делят на вда типа = ординални и номинални. При ординалните независимо, че нямат точни разстояния помежду си те могат да се нареждат – дни от седмицата, месеца, оценки и т.н..

Типове данни

Задача 2: Дайте предложение как да трансформираме X1 и X2 към съответно числово представяне Y1 и Y2 така, че да са използваеми за произволен софтуер и статистически метод.

X1	X2	Y1	Y2
Мъж	Април		
Жена	Май		
Мъж	Юни		
Жена	Март		
Мъж	Август		
Жена	Октомври		
Жена	Септември		

Типове данни

Задача 2: Дайте предложение как да трансформираме X1 и X2 към съответно числово представяне Y1 и Y2 така, че да са използваеми за произволен софтуер и статистически метод.

X1	X2	Y1	Y2
Мъж	Април	1	4
Жена	Май	0	5
Мъж	Юни	1	6
Жена	Март	0	3
Мъж	Август	1	8
Жена	Октомври	0	10
Жена	Септември	0	9

Отговор:
За X1: Мъж =1; Жена = 0. За X2 на всеки месец отговаря число от 1 до 12. Подобна трансформация е необходима, защото повечето статистически методи изискват числов вход.

Липсващи данни

Задача: Попълнете чрез **средна** липсващите стойности.

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182		
мъж	178	1900	90
жена		2100	120
жена	165	2000	

Липсващи данни

Задача: Попълнете чрез **средна** липсващите стойности.

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1992	107
мъж	178	1900	90
жена	174	2100	120
жена	165	2000	107

Решение:

- Среден ръст = (180+166+185+…+165)/12=174.25;
- Средна заплата = 1991.67 приблизително 1992;
- Средно IQ = 107,273 приблизително 107.

Липсващи данни

Задача: Попълнете чрез **локално средно** липсващите стойности.

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182		
мъж	178	1900	90
жена		2100	120
жена	165	2000	

Липсващи данни

Задача: Попълнете чрез **локално средно** липсващите стойности.

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1860	94
мъж	178	1900	90
жена	167	2100	120
жена	165	2000	118

Решение:

- Среден ръст (мъж) = (180+185+175+…+178)/6=182;
- Среден ръст (жена) = (166+170+160+…+165)/6=167;

- Средна заплата (мъж) = 1860;
- Средна заплата (жена) = 2086;

- Средно IQ (мъж) = 94;
- Средно IQ (жена) = 118.

Липсващи данни

Задача: Попълнете чрез **медиана** липсващите стойности.

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182		
мъж	178	1900	90
жена		2100	120
жена	165	2000	

Липсващи данни

Задача: Попълнете чрез **медиана** липсващите стойности.

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1800	90
мъж	178	1900	90
жена	166	2100	120
жена	165	2000	115

Решение:

- Медиана Ръст (мъже) = 175,178,**180,182**,185,190=(180+182)/2=181;
- Медиана Ръст (жени) = 160,165,**165,166**,170,175=165.5=166;

- Медиана Заплата (мъж) = 1400,1700,**1800**,1900,2500=1800;
- Медиана Заплата (жена) = 1600,1700,2000,2000,2100,2200,3000=2000;

- Медиана IQ (мъж)= 70,90,**90**,100,120=90;
- Медиана IQ (жена)= 100,110,**110,120**,130,140=115.

Липсващи данни

Задача: Кой е най-лошо представилият се метод от трите?

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1992	107
мъж	178	1900	90
жена	174	2100	120
жена	165	2000	107

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1860	94
мъж	178	1900	90
жена	167	2100	120
жена	165	2000	118

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1800	90
мъж	178	1900	90
жена	166	2100	120
жена	165	2000	115

Липсващи данни

Задача: Попълнете чрез **регресия** липсващите стойности, ако моделът е:

Пол	Ръст	IQ	Заплата
мъж	180	100	1800
жена	166	100	1600
мъж	185	120	2500
жена	170	110	2000
мъж	175	90	1400
жена	160	130	2200
жена	175	140	
мъж	190	70	1700
жена	165	110	1700
мъж	182	90	
мъж	178	90	1900
жена	166	120	2100
жена	165	115	2000

Regression Statistics	
Multiple R	0,920372
R Square	0,847084
Adjusted R Square	0,781548
Standard Error	143,2985
Observations	11

	Coefficients	Standard Error	t Stat	P-value
Intercept	-6239,4	1968,855	-3,16905	0,015729
Пол	-210,755	189,6257	-1,11143	0,303094
Ръст	34,77061	10,89939	3,190141	0,015274
IQ	21,23205	3,52491	6,023431	0,00053

Липсващи данни

Задача: Попълнете чрез **регресия** липсващите стойности, ако моделът е:

Пол	Ръст	IQ	Заплата
мъж	180	100	1800
жена	166	100	1600
мъж	185	120	2500
жена	170	110	2000
мъж	175	90	1400
жена	160	130	2200
жена	175	140	2817,94
мъж	190	70	1700
жена	165	110	1700
мъж	182	90	1788,98
мъж	178	90	1900
жена	166	120	2100
жена	165	115	2000

Regression Statistics	
Multiple R	0,920372
R Square	0,847084
Adjusted R Square	0,781548
Standard Error	143,2985
Observations	11

	Coefficients	Standard Error	t Stat	P-value
Intercept	-6239,4	1968,855	-3,16905	0,015729
Пол	-210,755	189,6257	-1,11143	0,303094
Ръст	34,77061	10,89939	3,190141	0,015274
IQ	21,23205	3,52491	6,023431	0,00053

Решение:

Заплата=-210.755*Пол+34,77061*Ръст+21,23205*IQ-6239.4

Нормализация

Задача: Нормализирайте данните с десетично скалиране, мин-макс и по стандартно отклонение?

Пол	Ръст	Заплата	IQ
мъж	180	1800	100
жена	166	1600	100
мъж	185	2500	120
жена	170	2000	110
мъж	175	1400	90
жена	160	2200	130
жена	175	3000	140
мъж	190	1700	70
жена	165	1700	110
мъж	182	1800	90
мъж	178	1900	90
жена	166	2100	120
жена	165	2000	115

Пол‘	Ръст‘	Заплата‘	IQ‘
1			
0			
1			
0			
1			
0			
0			
1			
0			
1			
1			
0			
0			

Описателни характеристики

Задача: Как да интерпретираме, че дадена променлива има очакване 1 и стандартно отклонение 10?

Описателни характеристики

Задача: Как да интерпретираме, че дадена променлива има очакване 1 и стандартно отклонение 10?

Отговор:

Очакване 1 показва каква е стойността, която ще получим средно след множество опити, например ако това е 1 % печалба от инвестиции ще означава, че след N опита, където N е достатъчно голямо число, нашата печалба ще е средно 1%. Стандартно отклонение 10 показва колко средно ще се отклони тази величина от очакването си, в случая с 10%, т.е. може да имаме печалби средно в интервала [-9%, 11%].

Описателни характеристики

Задача 2: Намерете медианата и средната стойност на [0,1,0,1,-2,3,-2,-1]?

Описателни характеристики

Задача 2: Намерете медианата и средната стойност на [0,1,0,1,-2,3,-2,-1]?

Отговор:

- Mean = (0+1+0+1-2+3-2-1)/8=0;
- Median (-2,-2,-1,0,0,1,1,3) = 0, медианата се пада в средата, в случая е между две еднакви стойности(иначе осредняваме).

Описателни характеристики

Задача 3:

Средната възраст в случая на Age=[10,15,12,45,64,11,47,38,39,87,21,459] е по-уместно с медиана или средно аритметично(очакването) да се намери?

Описателни характеристики

Задача 3:

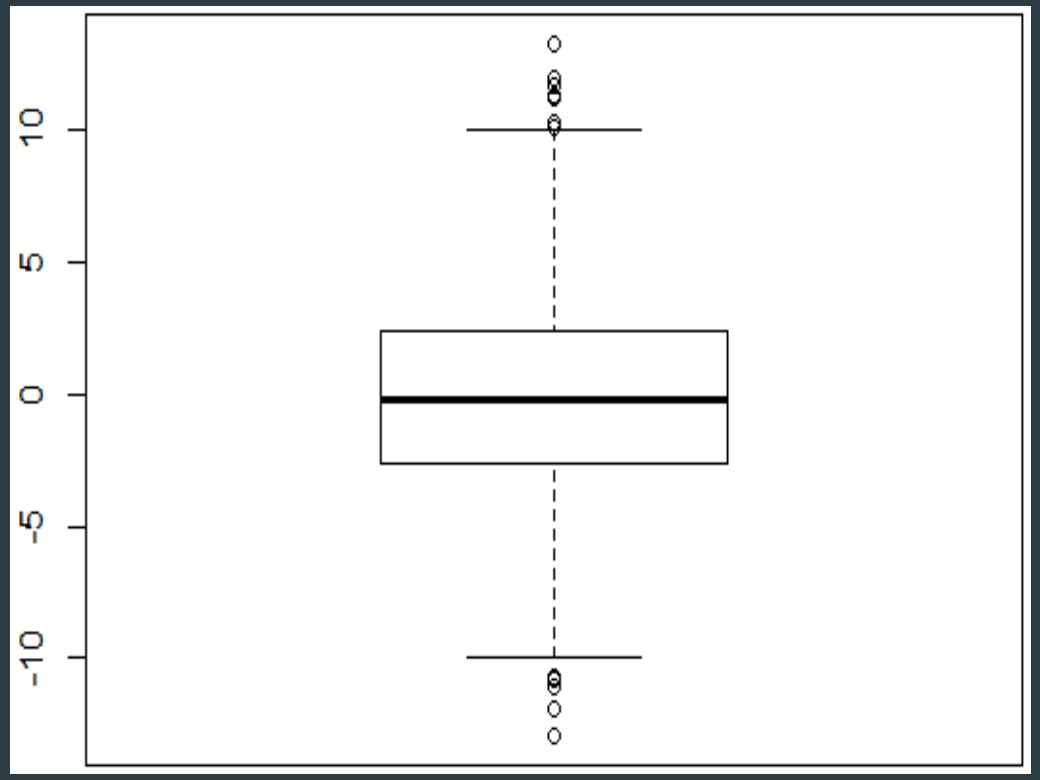
Средната възраст в случая на Age=[10,15,12,45,64,11,47,38,39,87,21,459] е по-уместно с медиана или средно аритметично(очакването) да се намери?

Отговор:

Медианата, защото имаме изключително наблюдение, което ще измести средната стойност.

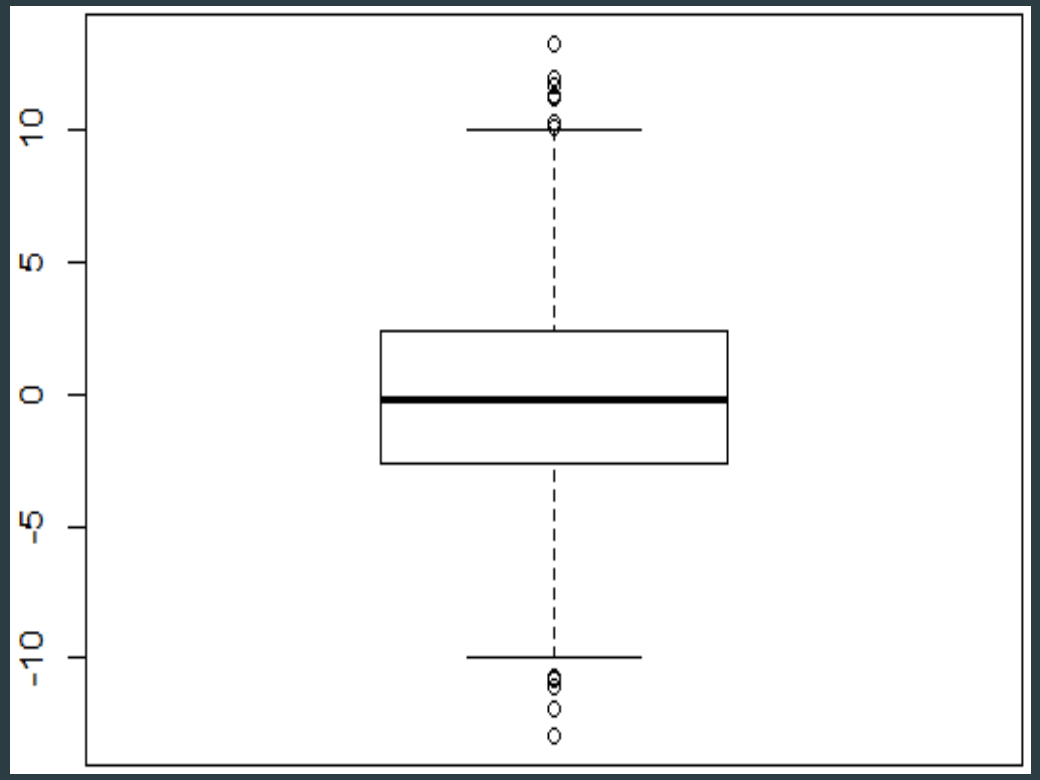
Описателни характеристики

Задача 4: Има ли изключителни точки според боксплота? Колко приблизително са те?



Описателни характеристики

Задача 4: Има ли изключителни точки според боксплота? Колко са те?



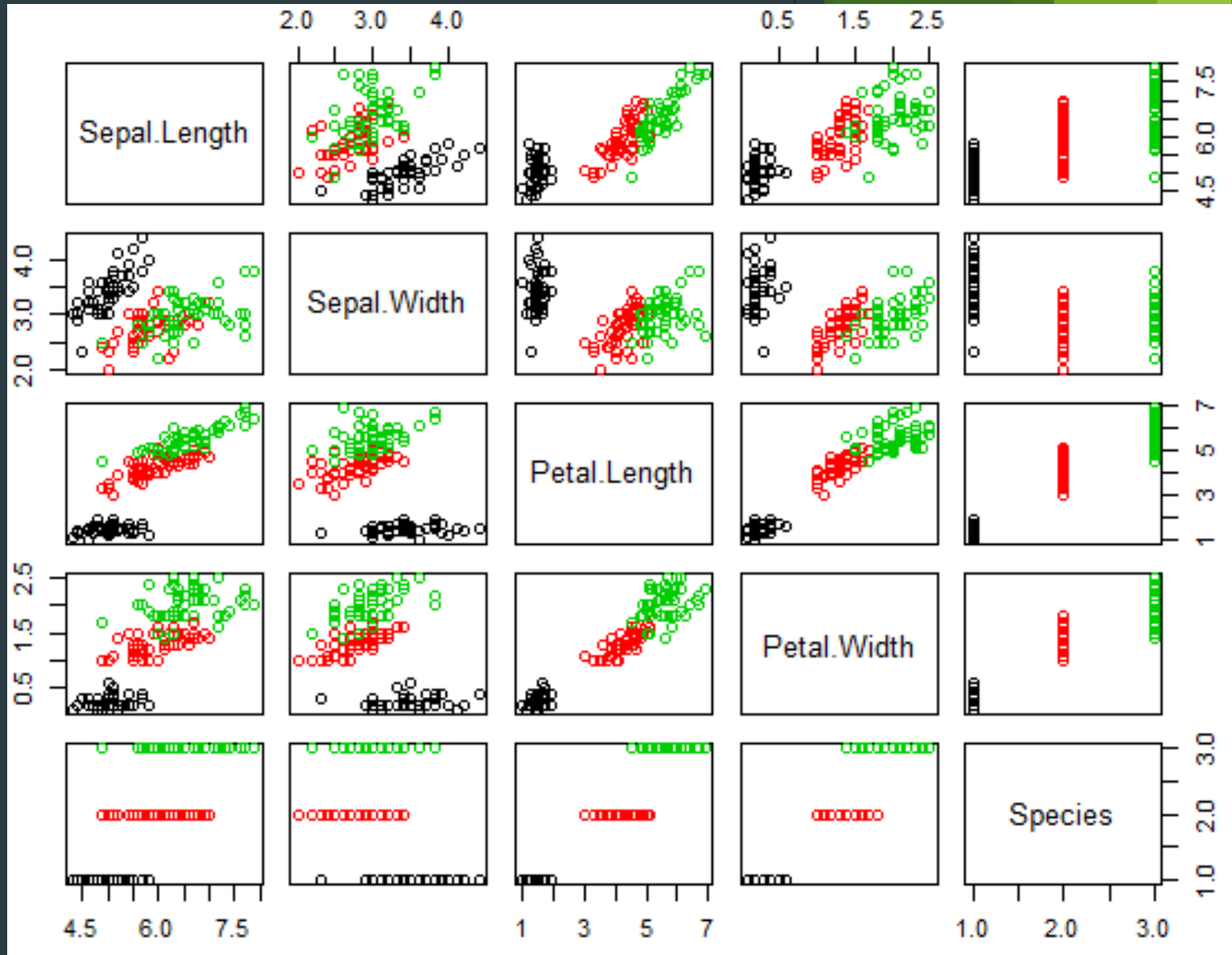
Отговор:

Да, защото има точки извън оградите. Различимите точки са 10.

Зависимости между променливите

Задача:

Вижте множетсвото от онагледени по двойки променливи. Има ли двойки променливи, между които ясно да се вижда линейна зависимост, кои са те?



Зависимости между променливите

Задача:

Има ли двойки променливи, между които ясно да се вижда линейна зависимост, кои са те?

Отговор:

Petal.Length и Petal.Width;

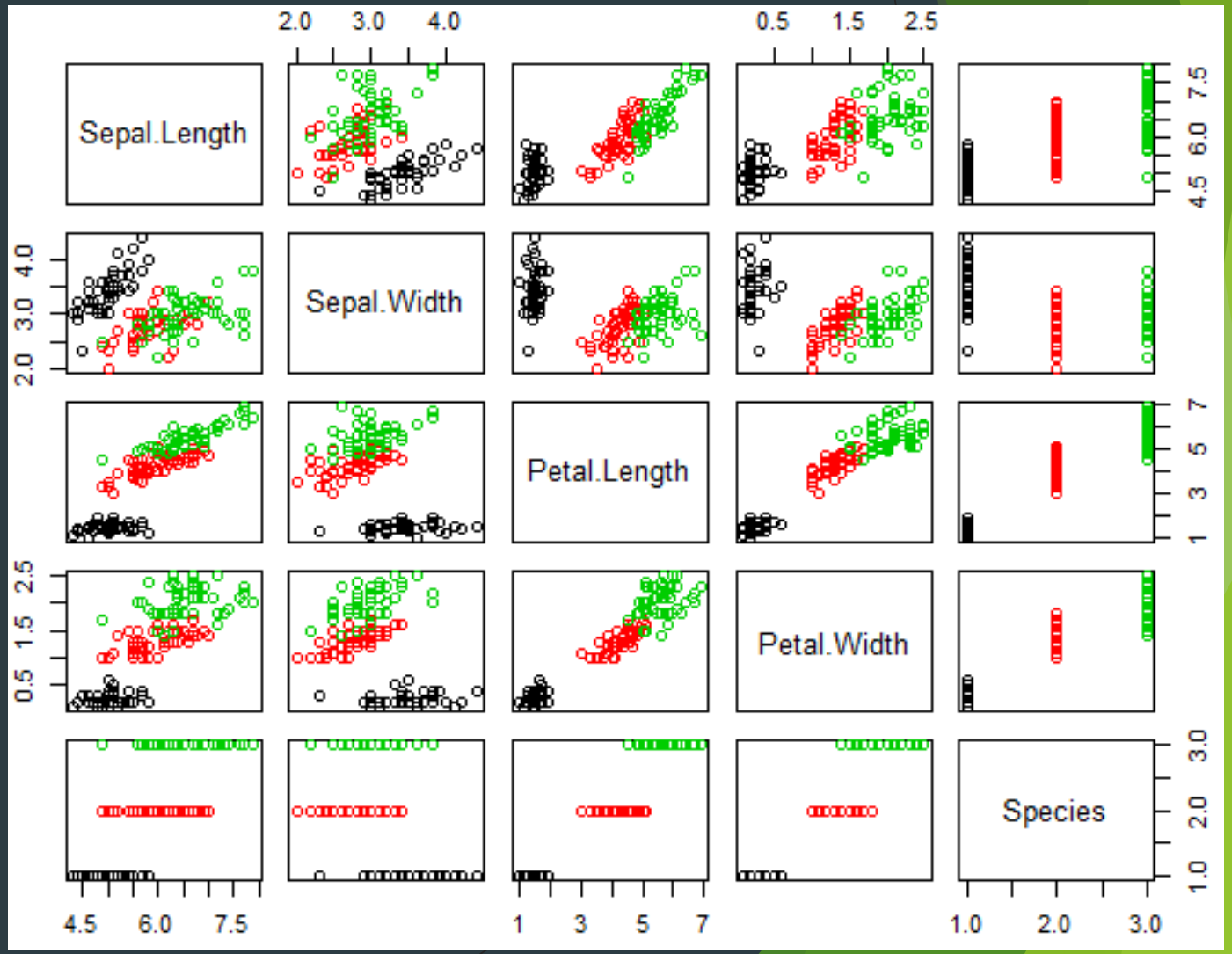
Petal. Length и Sepal.Length;

Малко по-слабо изразени:

Sepal.Length и Petal.Width;

Petal.Width и Species;

Petal.Length и Species.



Зависимости между променливите

Задача 2:

Ако ще оценявате линеен модел, коя корелационна мярка ще използвате?

Зависимости между променливите

Задача 2:

Ако ще оценявате линеен модел, коя корелационна мярка ще използвате?

Отговор:

Коефициентът на Пирсън, защото е специализиран в намиране на линейни връзки.

Зависимости между променливите

Задача 3:

Кой се оценява по-бързо Kendall или Spearman?

Отговор:

Зависимости между променливите

Задача 3:

Кой се оценява по-бързо Kendall или Spearman?

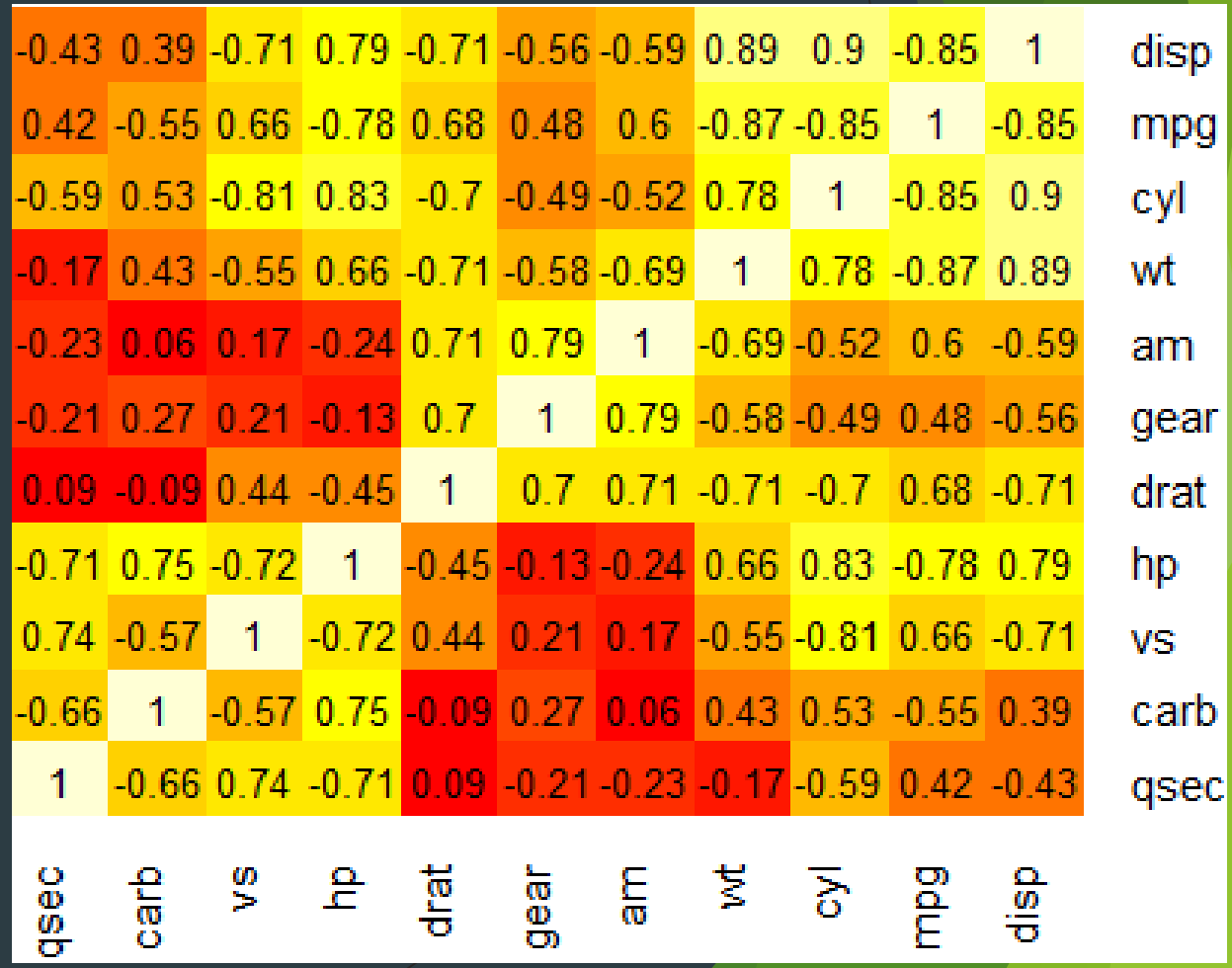
Отговор:

Spearman е със сложност O(n*log n), а Kendall със O(n*n), т.е. Spearman се оценява по-бързо.

Зависимости между променливите

Задача 4: Вижте heatmap изображението:

- Коя двойка променливи е най-силно корелирана?
- А коя е с най-ниска корелация?
- Ако искаме да моделираме, посредством обикновена линейна регресия, wt, коя променлива (предиктор) ще изберем за моделна?



Решение

- cor(disp,cyl) = 0.9;
- cor(hp,gear) = 0.06;
- Ако искаме да моделираме wt, ще изберем променливата, която е най-силно корелирана с wt. Това е disp.

Регресионен анализ

Задача: Имаме следния резултат от регресия:

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

A. Каква е моделната точност?
B. Колко променливи имаме?
C. Има ли незначими коефициенти?
D. Каква е максималната грешка допусната от модела?
E. Каква е очакваната моделна грешка?

Решение

A. 65%
B. 1 (със свободния коефициент 2)
C. Не
D. 43.201
E. 15.38

Регресионен анализ

Задача: Използваме модела от предходната задача. Ако знаем, че speed = 40 изчислете стойността, която модела предвижда за тази входна стойност.

Регресионен анализ

Задача: Използваме модела от предходната задача. Ако знаем, че speed = 40 изчислете стойността, която модела предвижда за тази входна стойност.

Решение:

-17.5791+3.9324*40 = 139.72

Регресионен анализ (за Контролно 2)

Задача: Имаме следните резултати от даден модел:

Наблюдавани стойности	**9**	**6**	**5**	**10**	**4**	**8**	**3**
Предвидени стойности	8	7	5	7	2	4	6

Намерете колко е средната процентна грешка (MAPE) на модела.

Решение

Грешката се намира като от наблюдаваната стойност се извади предвидената от модела. Процентната грешка се намира като съотнесем грешката към наблюдаваната стойност.

Грешка	1	-1	0	3	2	4	-3
% грешка	1/9	-1/6	0	3/10	2/4	4/8	-3/3

Формулата за MAPE може да се види от предходната лекция.

$$MAPE = \frac{11.11\% + 16.16\% + 30\% + 50\% + 50\% + 100\%}{7} = \mathbf{36.82}\ \%$$