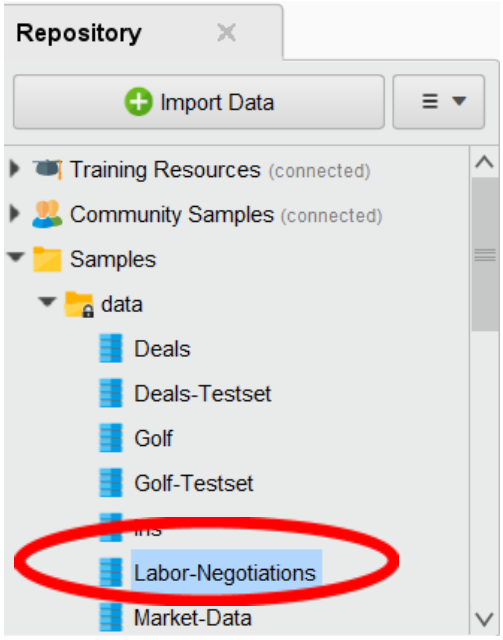
นาย DAYUTH THY 66122420126

ใบงานที่ 1
Data Cleaning -- missing value

ให้นักศึกษาตรวจสอบค่า missing จากชุดข้อมูล (Dataset) **Labor-Negotiations** ใน sample data ของ RapidMiner Repository และดำเนินการดังต่อไปนี้



1. จงตอบคำถามต่อไปนี้

a. จำนวนข้อมูลใน dataset Labor-Negotiations มีจำนวนเท่าไหร่

Answer : 40 Data

b. จำนวน Attribute ใน dataset Labor-Negotiations มีจำนวนเท่าไหร่ ประกอบไปด้วย Attribute อะไรบ้าง

Answer : 16 attributes :

- **contrib-to-health-plan**
- **bereavement-assistance**
- **contrib-to-dental-plan**

- **longterm-disability-assistance**
- **vacation**
- **statutory-holidays**
- **education-allowance**
- **shift-differential**
- **standby-pay**
- **pension**
- **working-hours**
- **col-adj**
- **wage-inc-3rd**
- **wage-inc-2nd**
- **wage-inc-1st**
- **duration**

c. ประเภทข้อมูลใน dataset Labor-Negotiations มีกี่ประเภท ประกอบไปด้วยประเภทอะไรบ้าง

Answer : 3 types :

- **Nominal**
- **integer**
- **Real**

d. Attribute ที่มีค่า missing มีกี่ Attribute ประกอบไปด้วย Attribute อะไรบ้าง และมีจำนวนเท่าไหร่

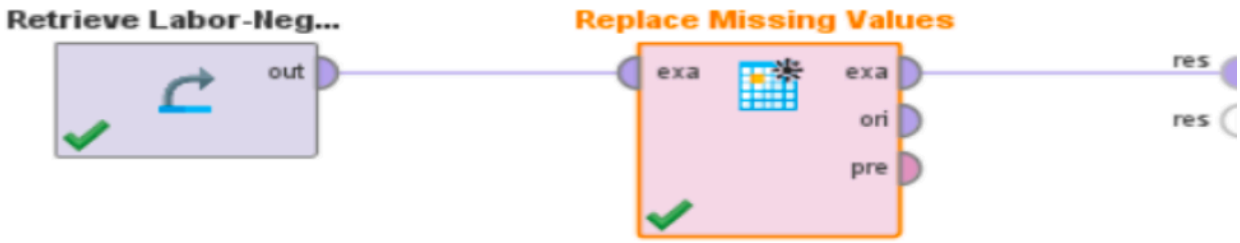
Answer :

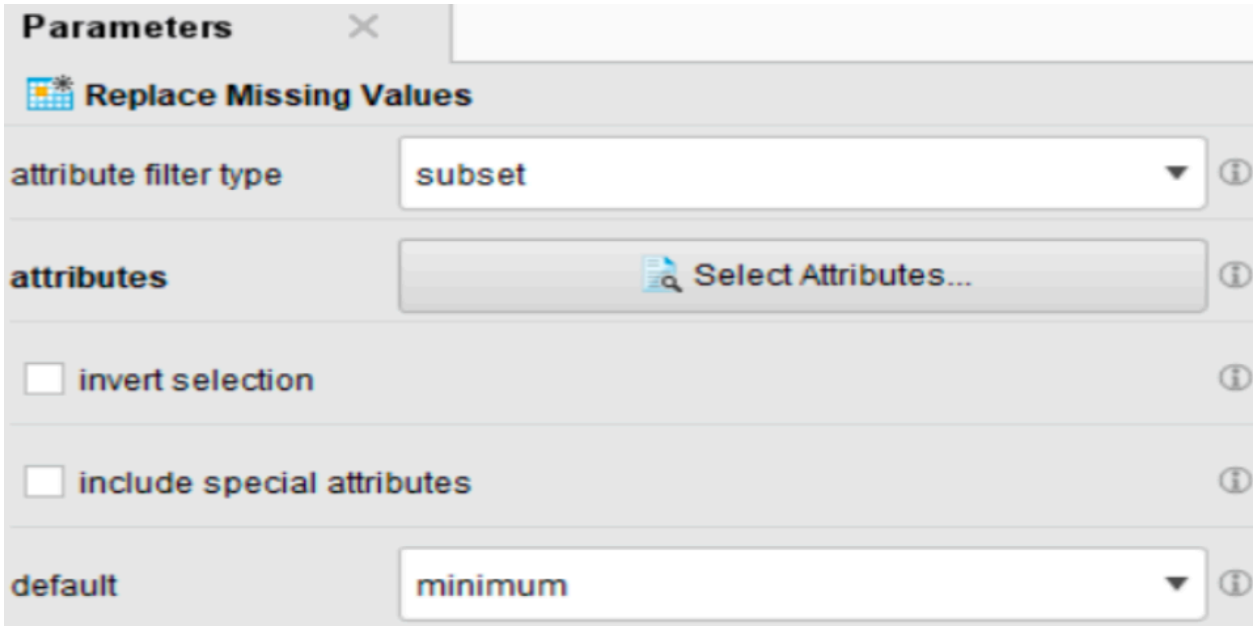
- **contrib-to-health-plan (16)**
- **bereavement-assistance (20)**
- **contrib-to-dental-plan (15)**
- **longterm-disability-assistance (24)**
- **vacation (3)**
- **statutory-holidays (2)**
- **education-allowance (22)**

- **shift-differential (16)**
- **standby-pay (33)**
- **pension (22)**
- **working-hours (3)**
- **col-adj (16)**
- **wage-inc-3rd (28)**
- **wage-inc-2nd (10)**
- **wage-inc-1st (1)**
- **duration (1)**

---

2. **จงจัดการกับค่า missing ที่ปรากฏ ตามข้อกำหนดต่อไปนี้ และ capture หน้า design และ result เมื่อดำเนินการตามข้อกำหนดเรียบร้อยแล้ว**
   a. Replace ค่าของ Attribute ที่มีประเภทเป็น Real ที่ปรากฏค่า missing ให้เปลี่ยนโดยใช้ค่าน้อยที่สุด (minimum)
   Answer :



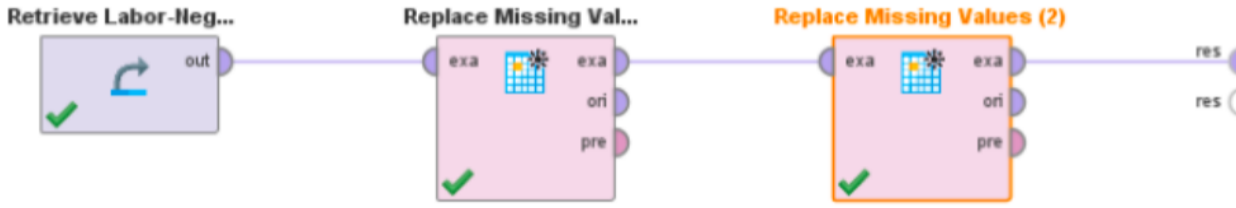


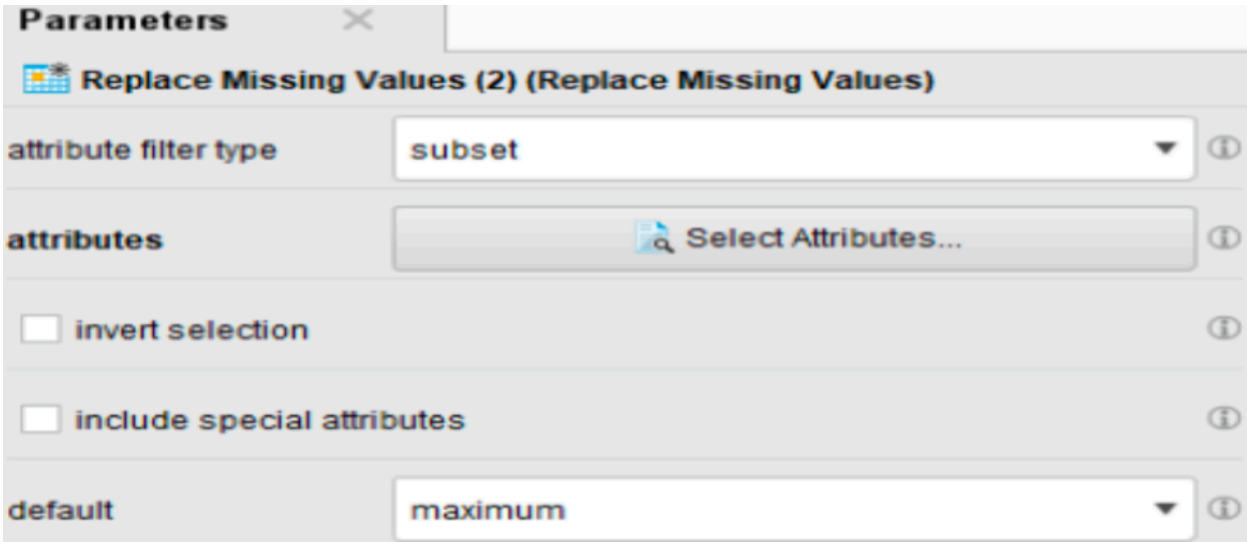
Row No.	class	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	duration	col-adj	working-hou...	pension	standby-pay	shift-differe...	education-al...	stat
1	good	5	2	2	1	?	40	?	?	2	?	11
2	good	4.500	5.800	2	2	?	35	ret_allw	?	?	yes	11
3	good	2	2	2	?	?	38	empl_contr	?	5	?	11
4	good	3.700	4	5	3	tc	?	?	?	?	yes	?
5	good	4.500	4.500	5	3	?	40	?	?	?	?	12
6	good	2	2.500	2	2	?	35	?	?	6	yes	12
7	good	4	5	5	3	tc	?	empl_contr	?	?	?	12
8	good	6.900	4.800	2.300	3	?	40	?	?	3	?	12
9	good	3	7	2	2	?	38	?	12	25	yes	11
10	good	5.700	2	2	1	none	40	empl_contr	?	4	?	11
11	good	3.500	4	4.600	3	none	36	?	?	3	?	13
12	good	6.400	6.400	2	2	?	38	?	?	4	?	15
13	bad	3.500	4	2	2	none	40	?	?	2	no	10
14	good	3.500	4	5.100	3	tcf	37	?	?	4	?	13
15	good	3	2	2	1	none	36	?	?	10	no	11
16	good	4.500	4	2	2	none	37	empl_contr	?	?	?	11
17	good	2.800	2	2	1	?	35	?	?	2	?	12
18	bad	2.100	2	2	1	tc	40	ret_allw	2	3	no	9
19	bad	2	2	2	1	none	38	none	?	?	yes	11
20	good	4	5	2	2	tcf	35	?	13	5	?	15
21	good	4.300	4.400	2	2	?	38	?	?	4	?	12
22	bad	2.500	3	2	2	?	40	none	?	?	?	11
23	good	3.500	4	4.600	3	tcf	27	?	?	?	?	?
24	good	4.500	4	2	2	?	40	?	?	4	?	10

25	good	6	2	2	1	?	38	?	8	3	?	9
26	bad	2	2	2	3	none	40	none	?	?	?	10
27	good	4.500	4.500	2	2	tcf	?	?	?	?	yes	10
28	good	3	3	2	2	none	33	?	?	?	yes	12
29	good	5	4	2	2	none	37	?	?	5	no	11
30	bad	2	2.500	2	3	?	35	none	?	?	?	10
31	good	4.500	4.500	5	3	none	40	?	?	?	no	11
32	bad	3	2	2.500	3	tc	40	none	?	5	no	10
33	bad	2.500	2.500	2	2	?	38	empl_contr	?	?	?	10
34	bad	4	5	2	2	none	40	none	?	3	no	10
35	bad	2	2.500	2.100	3	tc	40	none	2	1	no	10
36	bad	2	2	2	2	none	40	none	?	?	no	11
37	bad	2	2	2	1	tc	40	ret_allw	4	0	no	11
38	bad	2.800	2	2	1	none	38	empl_contr	2	3	no	9
39	bad	2	2.500	2	3	?	37	empl_contr	?	?	?	10
40	good	4.500	4	2	2	none	40	?	?	4	?	12

b. Replace ค่าของ Attribute ที่มีประเภทเป็น Integer ที่ปรากฏค่า missing ให้เปลี่ยนโดยใช้ค่าสูงสุด (maximum)

Answer :



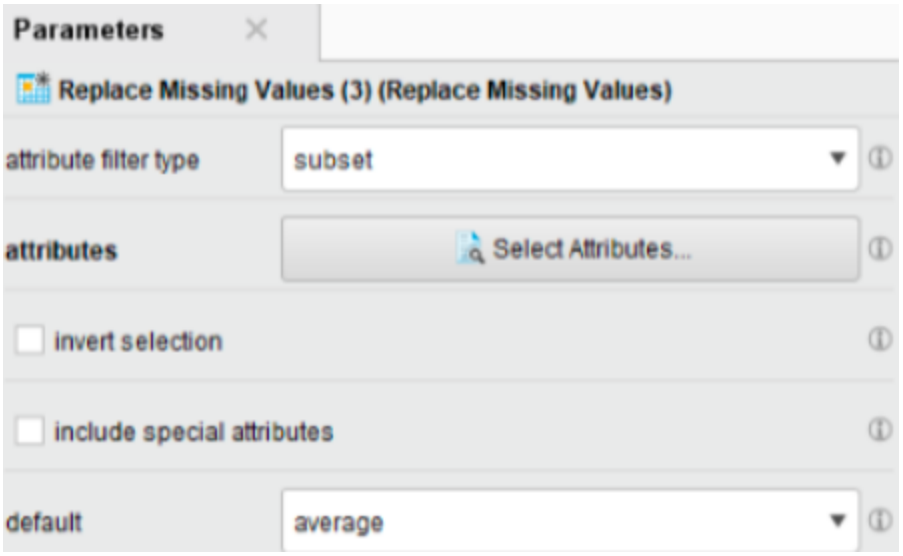


Row No.	class	duration	working-hou...	standby-pay	shift-differe...	statutory-ho...	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	col-adj	pension	ed
1	good	1	40	13	2	11	5	2	2	?	?	?
2	good	2	35	13	25	11	4.500	5.800	2	?	ret_allw	ye
3	good	3	38	13	5	11	2	2	2	?	empl_contr	?
4	good	3	40	13	25	15	3.700	4	5	tc	?	ye
5	good	3	40	13	25	12	4.500	4.500	5	?	?	?
6	good	2	35	13	6	12	2	2.500	2	?	?	ye
7	good	3	40	13	25	12	4	5	5	tc	empl_contr	?
8	good	3	40	13	3	12	6.900	4.800	2.300	?	?	?
9	good	2	38	12	25	11	3	7	2	?	?	ye
10	good	1	40	13	4	11	5.700	2	2	none	empl_contr	?
11	good	3	36	13	3	13	3.500	4	4.600	none	?	?
12	good	2	38	13	4	15	6.400	6.400	2	?	?	?
13	bad	2	40	13	2	10	3.500	4	2	none	?	no
14	good	3	37	13	4	13	3.500	4	5.100	tcf	?	?
15	good	1	36	13	10	11	3	2	2	none	?	no
16	good	2	37	13	25	11	4.500	4	2	none	empl_contr	?
17	good	1	35	13	2	12	2.800	2	2	?	?	?
18	bad	1	40	2	3	9	2.100	2	2	tc	ret_allw	no
19	bad	1	38	13	25	11	2	2	2	none	none	ye
20	good	2	35	13	5	15	4	5	2	tcf	?	?
21	good	2	38	13	4	12	4.300	4.400	2	?	?	?
22	bad	2	40	13	25	11	2.500	3	2	?	none	?
23	good	3	27	13	25	15	3.500	4	4.600	tcf	?	?
24	good	2	40	13	4	10	4.500	4	2	?	?	?
25	good	1	38	8	3	9	6	2	2	?	?	?
26	bad	3	40	13	25	10	2	2	2	none	none	?
27	good	2	40	13	25	10	4.500	4.500	2	tcf	?	ye
28	good	2	33	13	25	12	3	3	2	none	?	ye
29	good	2	37	13	5	11	5	4	2	none	?	no
30	bad	3	35	13	25	10	2	2.500	2	?	none	?
31	good	3	40	13	25	11	4.500	4.500	5	none	?	no
32	bad	3	40	13	5	10	3	2	2.500	tc	none	no
33	bad	2	38	13	25	10	2.500	2.500	2	?	empl_contr	?
34	bad	2	40	13	3	10	4	5	2	none	none	no
35	bad	3	40	2	1	10	2	2.500	2.100	tc	none	no
36	bad	2	40	13	25	11	2	2	2	none	none	no
37	bad	1	40	4	0	11	2	2	2	tc	ret_allw	no
38	bad	1	38	2	3	9	2.800	2	2	none	empl_contr	no
39	bad	3	37	13	25	10	2	2.500	2	?	empl_contr	?
40	good	2	40	13	4	12	4.500	4	2	none	?	?

Label			Least	Most	Values
class	Nominal	0	bad (14)	good (26)	good (26), bad (14)
duration	Integer	0	Min 1	Max 3	Average 2.125
working-hours	Integer	0	Min 27	Max 40	Average 37.975
standby-pay	Integer	0	Min 2	Max 13	Average 11.800
shift-differential	Integer	0	Min 0	Max 25	Average 12.750
statutory-holidays	Integer	0	Min 9	Max 15	Average 11.300
wage-inc-1st	Real	0	Min 2	Max 6.900	Average 3.580
wage-inc-2nd	Real	0	Min 2	Max 7	Average 3.435
wage-inc-3rd	Real	0	Min 2	Max 5.100	Average 2.530
col-adj	Nominal	16	Least tcf (4)	Most none (14)	Values none (14), tc (6), ...[1 more]
pension	Nominal	22	Least ret_allw (3)	Most none (8)	Values none (8), empl_contr (7), ...[1 more]
education-allowance	Nominal	22	Least yes (7)	Most no (11)	Values no (11), yes (7)

c. Replace ค่าของ Attribute ที่มีประเภทเป็น Nominal ที่ปรากฏค่า missing ให้เปลี่ยนโดยใช้ค่าเฉลี่ย (average)

Answer :



Row No.	class	col-adj	pension	education-al...	vacation	longterm-di...	contrib-to-d...	bereavemen...	contrib-to-h...	duration	working-hou...	sta
1	good	none	none	no	average	yes	half	yes	full	1	40	13
2	good	none	ret_allw	yes	below-average	yes	full	yes	full	2	35	13
3	good	none	empl_contr	no	generous	yes	half	yes	half	3	38	13
4	good	tc	none	yes	below-average	yes	half	yes	full	3	40	13
5	good	none	none	no	average	yes	half	yes	half	3	40	13
6	good	none	none	yes	average	yes	half	yes	full	2	35	13
7	good	tc	empl_contr	no	generous	yes	none	yes	half	3	40	13
8	good	none	none	no	below-average	yes	half	yes	full	3	40	13
9	good	none	none	yes	below-average	yes	half	yes	full	2	38	12
10	good	none	empl_contr	no	generous	yes	full	yes	full	1	40	13
11	good	none	none	no	generous	yes	half	yes	full	3	36	13
12	good	none	none	no	below-average	yes	full	yes	full	2	38	13
13	bad	none	none	no	below-average	no	half	yes	half	2	40	13
14	good	tcf	none	no	generous	yes	full	yes	full	3	37	13
15	good	none	none	no	generous	yes	half	yes	full	1	36	13
16	good	none	empl_contr	no	average	yes	full	yes	full	2	37	13
17	good	none	none	no	below-average	yes	half	yes	full	1	35	13
18	bad	tc	ret_allw	no	below-average	yes	half	yes	none	1	40	2
19	bad	none	none	yes	average	no	none	no	none	1	38	13
20	good	tcf	none	no	generous	yes	half	yes	full	2	35	13
21	good	none	none	no	generous	yes	full	yes	full	2	38	13
22	bad	none	none	no	below-average	yes	half	yes	full	2	40	13
23	good	tcf	none	no	below-average	yes	half	yes	full	3	27	13
24	good	none	none	no	generous	yes	half	yes	full	2	40	13

25	good	none	none	no	generous	yes	half	yes	full	1	38	8
26	bad	none	none	no	below-average	yes	half	yes	full	3	40	13
27	good	tcf	none	yes	below-average	yes	none	yes	half	2	40	13
28	good	none	none	yes	generous	yes	half	yes	full	2	33	13
29	good	none	none	no	below-average	yes	full	yes	full	2	37	13
30	bad	none	none	no	average	yes	half	yes	full	3	35	13
31	good	none	none	no	average	yes	half	yes	full	3	40	13
32	bad	tc	none	no	below-average	yes	half	yes	full	3	40	13
33	bad	none	empl_contr	no	average	yes	half	yes	full	2	38	13
34	bad	none	none	no	below-average	no	none	yes	none	2	40	13
35	bad	tc	none	no	below-average	no	half	yes	full	3	40	2
36	bad	none	none	no	average	yes	none	yes	full	2	40	13
37	bad	tc	ret_allw	no	generous	no	none	no	none	1	40	4
38	bad	none	empl_contr	no	below-average	yes	half	yes	none	1	38	2
39	bad	none	empl_contr	no	average	yes	half	yes	none	3	37	13
40	good	none	none	no	average	yes	full	yes	half	2	40	13

Label			Least	Most	Values
class	Nominal	0	bad (14)	good (26)	good (26), bad (14)
col-adj	Nominal	0	tcf (4)	none (30)	none (30), tc (6), ...[1 more]
pension	Nominal	0	ret_allw (3)	none (30)	none (30), empl_contr (7), ...[1 more]
education-allowance	Nominal	0	yes (7)	no (33)	no (33), yes (7)
vacation	Nominal	0	average (11)	below-average (17)	below-average (17), generous (12), ...[1 more]
longterm-disability-assistance	Nominal	0	no (5)	yes (35)	yes (35), no (5)
contrib-to-dental-plan	Nominal	0	none (6)	half (26)	half (26), full (8), ...[1 more]
bereavement-assistance	Nominal	0	no (2)	yes (38)	yes (38), no (2)
contrib-to-health-plan	Nominal	0	none (6)	full (28)	full (28), half (6), ...[1 more]
duration	Integer	0	Min: 1	Max: 3	Average: 2.125
working-hours	Integer	0	Min: 27	Max: 40	Average: 37.975
standby-pay	Integer	0	Min: 2	Max: 13	Average: 11.800

d. จัดเก็บชุดข้อมูลที่ดำเนินการจัดการกับค่า missing ไว้ใน Local Repository ใน folder data โดยตั้งชื่อว่า Labor-Negotiations-cleanupmissing

Answer :

1	good	none	none	no	average	yes	half	yes	full	1	40	13
2	good	none	ret_allw	yes	below-average	yes	full	yes	full	2	35	13
3	good	none	empl_contr	no	generous	yes	half	yes	half	3	38	13
4	good	tc	none	yes	below-average	yes	half	yes	full	3	40	13
5	good	none	none	no	average	yes	half	yes	half	3	40	13
6	good	none	none	yes	average	yes	half	yes	full	2	35	13
7	good	tc	empl_contr	no	generous	yes	none	yes	half	3	40	13
8	good	none	none	no	below-average	yes	half	yes	full	3	40	13
9	good	none	none	yes	below-average	yes	half	yes	full	2	38	12
10	good	none	empl_contr	no	generous	yes	full	yes	full	1	40	13
11	good	none	none	no	generous	yes	half	yes	full	3	36	13
12	good	none	none	no	below-average	yes	full	yes	full	2	38	13
13	bad	none	none	no	below-average	no	half	yes	half	2	40	13
14	good	tcf	none	no	generous	yes	full	yes	full	3	37	13
15	good	none	none	no	generous	yes	half	yes	full	1	36	13
16	good	none	empl_contr	no	average	yes	full	yes	full	2	37	13
17	good	none	none	no	below-average	yes	half	yes	full	1	35	13
18	bad	tc	ret_allw	no	below-average	yes	half	yes	none	1	40	2
19	bad	none	none	yes	average	no	none	no	none	1	38	13
20	good	tcf	none	no	generous	yes	half	yes	full	2	35	13
21	good	none	none	no	generous	yes	full	yes	full	2	38	13
22	bad	none	none	no	below-average	yes	half	yes	full	2	40	13
23	good	tcf	none	no	below-average	yes	half	yes	full	3	27	13
24	good	none	none	no	generous	yes	half	yes	full	2	40	13