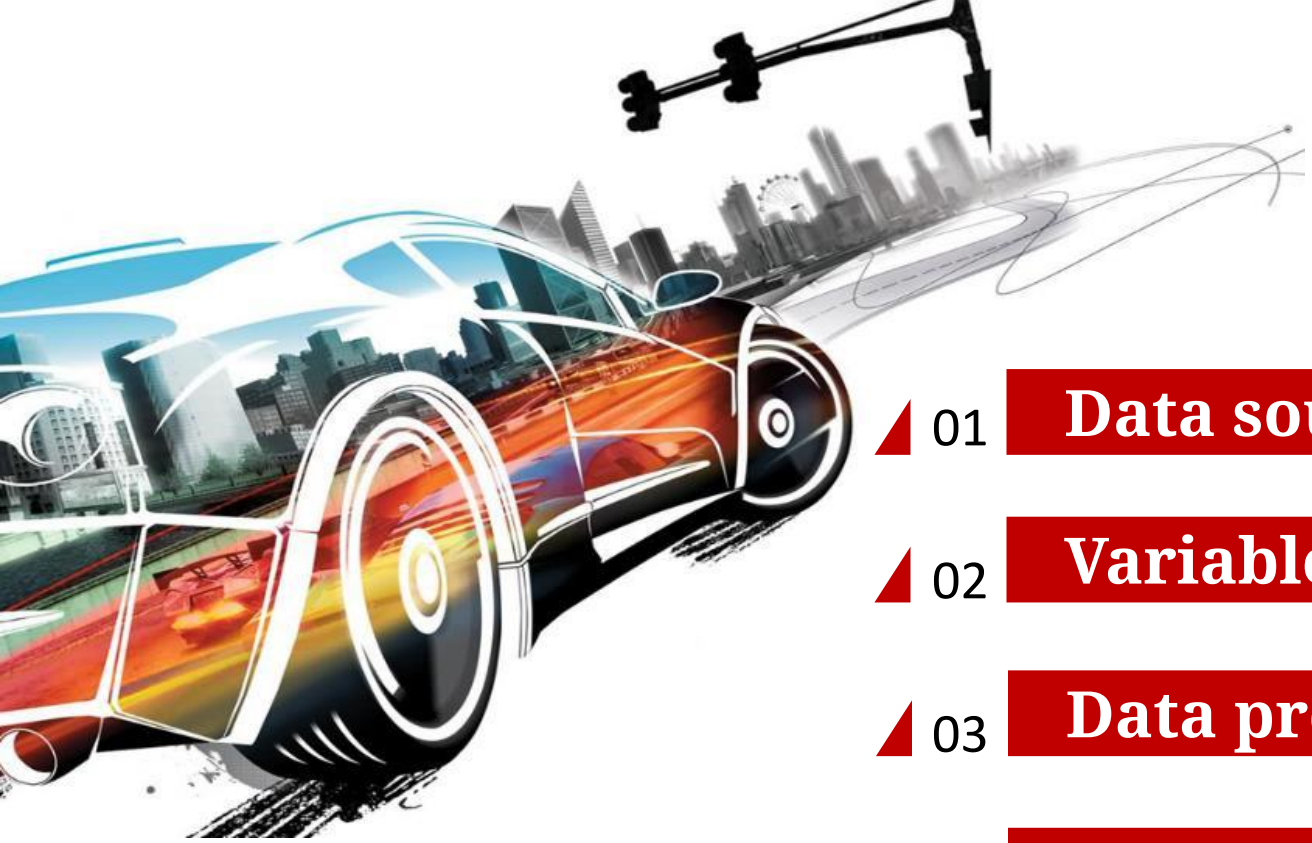
金金計較

組員：李奇學、蔡宗宏、曾士育、冷勇全、王愛茹



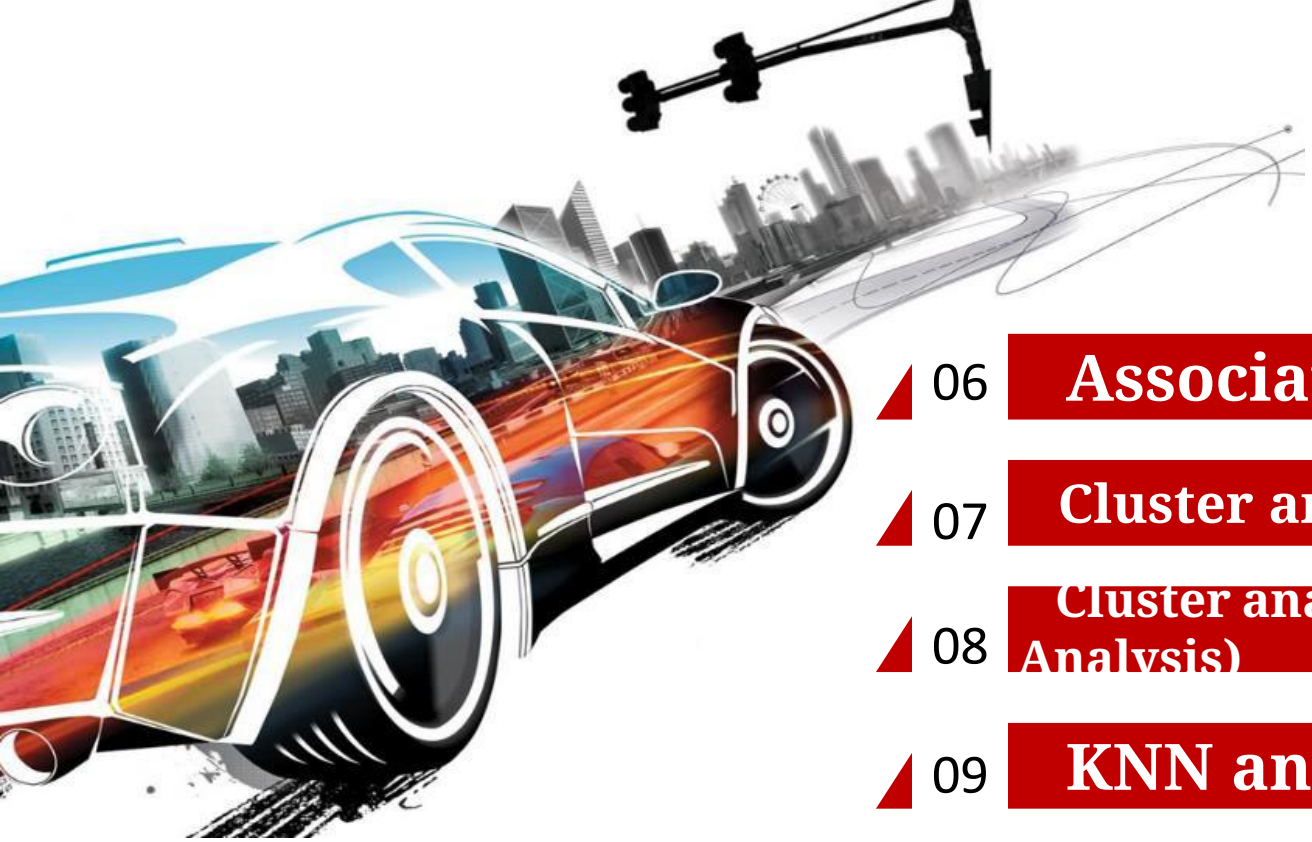
01 Data source

02 Variable description

03 Data preprocessing

04 Variable selection

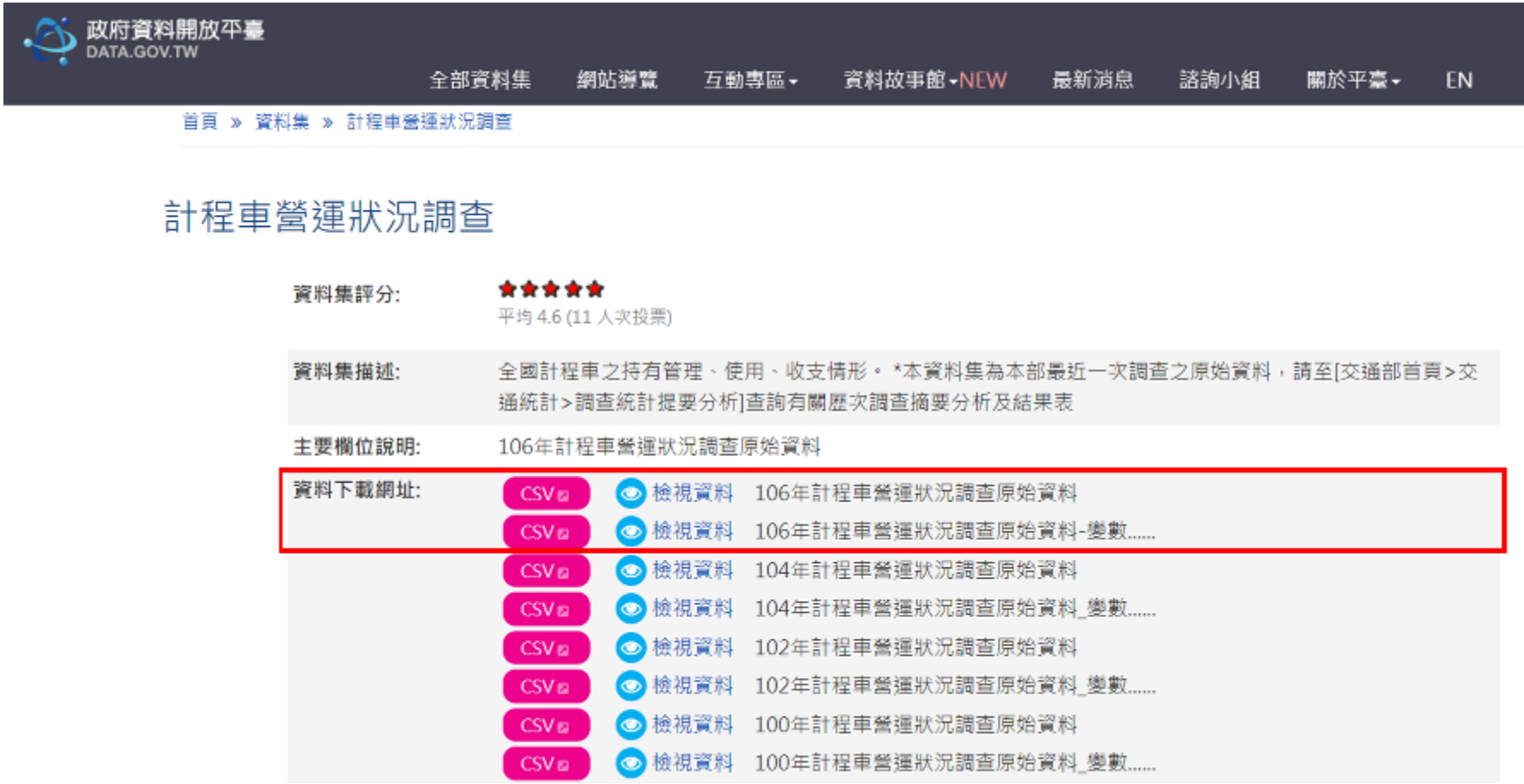
05 Decision tree analysis



01 Part One

Data source

01 Data source



02 Part Two

Variable description

02 Variable description

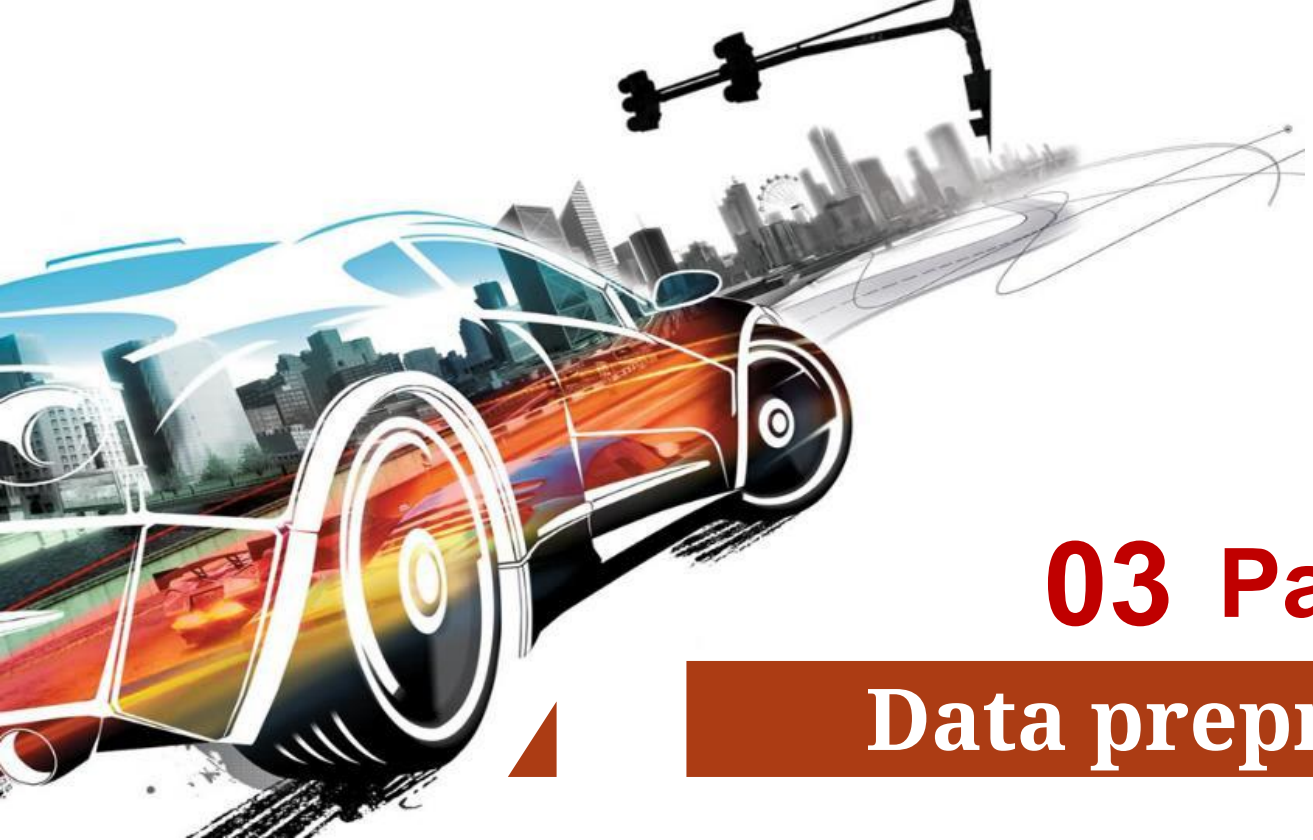
- **Questionnaire survey**
- **Operating type**

  → Way to find customer, Valuation method, Car's condition, etc.
- **Working habits**

  → Working time period, Number of day off, Reason for driving taxi, etc.
- **Expenditure**

  → Fuel, Maintenance, Insurance, Parking fee, etc.
- **Basic information**

  → Age, Education, Working area, Seniority, Gender



03 Part Three

Data preprocessing

03 Data preproce

Attributes

Too many missing value → Delete

Attributes

Chi-squared test of independence → Between tw … p-val … → signific …

Attributes

More than 8 interval → Compa … with a …

Attributes

Working area : 23 intervals → Simplify

Attributes

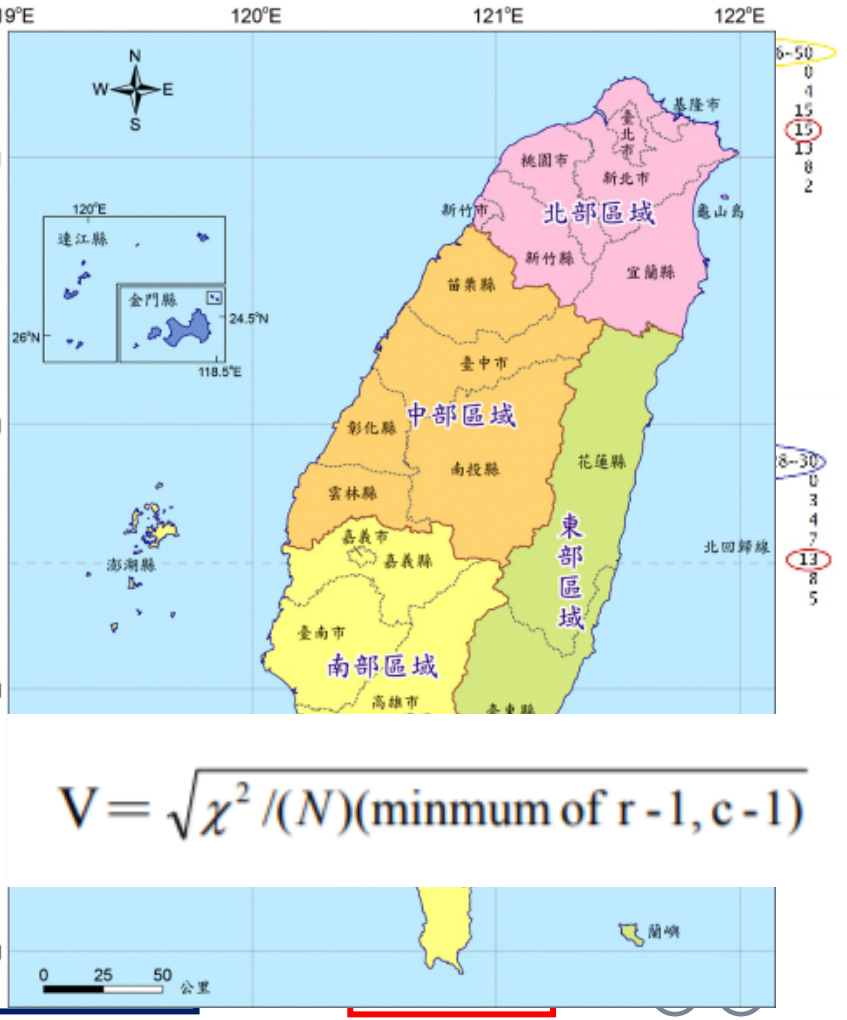
Cramer's V coefficient → **Between X and Y** Find the attribu … are more correla …

```
> t1 <- table(df$Y..Tota
> t1
                (01)1-5
(1)<600             262
(2)600-1200         209
(3)1200-1800         33
(4)1800-2400         12
(5)2400-3000          2
(6)>3000              2
Refuse               12

                (11)51-55
(1)<600
(2)600-1200
(3)1200-1800
(4)1800-2400
(5)2400-3000
(6)>3000
Refuse

> t1 <- table(df$Y..Tot
> t1
                (01)1-3
(1)<600             249
(2)600-1200         119
(3)1200-1800         21
(4)1800-2400          9
(5)2400-3000          3
(6)>3000              7
Refuse                8

                (11)>31
(1)<600               1
(2)600-1200           1
(3)1200-1800          3
(4)1800-2400          1
(5)2400-3000          0
(6)>3000              3
Refuse                1
```



$$V = \sqrt{\chi^2 / (N)(\text{minmum of } r-1, c-1)}$$

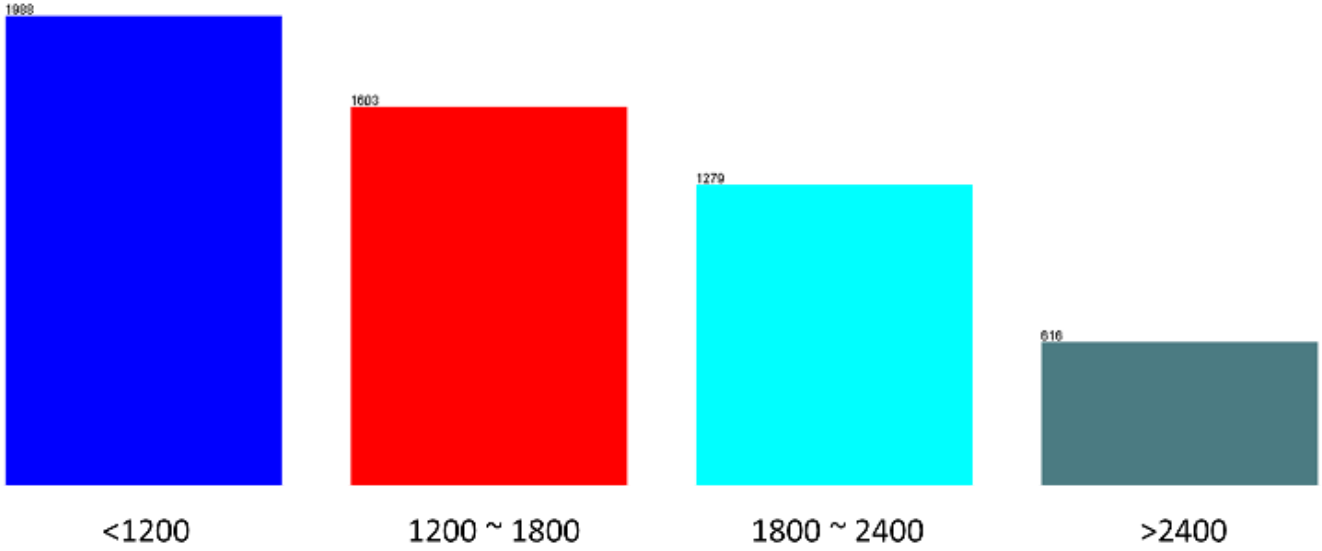
04 Part Four

Variable selection

Attribute selected	Intervals
Fulltime or Part time	Full time job , Part time job
Taxi operating type	Private car , Company's car , Cooperative's car
Mileage per day(km)	<100 , 100~130 , 130~190 , 190~250 , >250
Business Hours per Day	<4hr , 4~6hr , 6~8hr , 8~12hr , >12hr
Turns with passenger per day	1~3 , 4~9 , 10~12 , 13~24 , 25~30 , >31
Average of passengers per day	1~5 , 6~10 , 11~20 , 21~50 , 51~60 , 61~75 , >76
Expenditure of fuel per day(NT$)	<200 , 200~400 , 400~600 , 600~800 , >800
Maintenance fees per year($NT)	<10k , 10k~20k, 20k~30k , 30k~40k , 40k~50k , 50k~60k
Cooperative arrange service fee	Y , N
Radio arrange service fee	Y , N
Parking fee per month(NT$)	0 , <2000 , 2000~3000 , 3000~4000 , >4000
Age	20~25 , 25~40 , 40~50 , 50~60 , >60
Education	Elementary school , Junior high school , Senior high school , Junior collage , University
Reason for driving taxi	Free working time , Easy to enter , have no other option then looking for a job , Do as part time job as for extra income , other reason
Working area	Northern area , Central area , Southern area , Eastern area , Offshore island
Y->Total income per day	<1200 , 1200~2400 , 1800~2400 , >2400

04 Variable selection

Target attribute : Total income per day



05 Part Five

Decision tree analysis

05 Decision tree analysis

CFS

PERSENTAGE SPILT=80%

```
: 3,4,5,6,7,9,10,12 : 8
  Milage per day(km)
  Business Hours per Day
  Number of turns with passenger per day
  Average of passengers per day
  Expenditure of fuel per day($NT)
  Cooperative arrange service fee
  Radio arrange service fee
  Age
```

M	PRUNED	UNPRUNED
2	56.4266 %	55.6062 %
3	56.5178 %	55.3327 %
4	56.8824 %	55.6062 %
5	57.247 %	55.8797 %
6	57.3382 %	56.4266 %
7	57.3382 %	56.3355 %
8	57.5205 %	56.9736 %
9	57.5205 %	57.9763 %
10	57.4294 %	57.9763 %
11	57.4294 %	57.7028 %
12	57.6117 %	57.794 %
13	57.794 %	57.794 %
14	57.794 %	57.794 %
15	57.8851 %	58.0675 %
16	57.8851 %	58.3409 %
17	57.5205 %	58.0675 %
18	57.5205 %	58.2498 %
19	57.5205 %	58.2498 %
20	57.1559 %	57.794 %
21	57.1559 %	57.4294 %
22	57.1559 %	57.5205 %
23	57.247 %	57.247 %
24	57.3382 %	57.7028 %
25	57.3382 %	57.7028 %
26	57.4294 %	57.6117 %
27	57.247 %	57.1559 %
28	56.7001 %	56.8824 %

05 Decision tree analysis

INFOGAIN

PERSENTAGE SPILT=80%

```
Ranked attributes:
 0.29434    5 Number of turns with passenger per day
 0.28109    3 Milage per day(km)
 0.23018    6 Average of passengers per day
 0.20694    7 Expenditure of fuel per day($NT)
 0.18835    4 Business Hours per Day
 0.08136   10 Radio arrange service fee
 0.05566    1 Fulltime or Partime
 0.055     15 Working area
 0.04217   12 Age
 0.0289    11 Parking fee per month($NT)
 0.02882   14 Reason for driving taxi
 0.02878    8 Maintenance fees per year($NT)
 0.02466   13 Education
 0.02257    2 Taxi operating type
 0.00514    9 Cooperative arrange service fee

Selected attributes: 5,3,6,7,4,10,1,15,12,11,14,8,13,2,9 : 15
```

M	PRUNED	UNPRUNED
2	57.247 %	55.4239 %
3	57.247 %	55.3327 %
4	57.247 %	55.3327 %
5	57.4294 %	55.6974 %
6	57.6117 %	55.6062 %
7	57.6117 %	55.8797 %
8	57.7028 %	55.8797 %
9	57.794 %	55.8797 %
10	57.794 %	55.8797 %
11	57.794 %	55.6974 %
12	57.794 %	55.7885 %
13	57.794 %	55.7885 %
14	57.794 %	55.7885 %
15	57.794 %	56.2443 %
16	57.794 %	56.1531 %
17	58.1586 %	56.6089 %
18	57.9763 %	56.6089 %
19	57.9763 %	56.5178 %
20	57.9763 %	56.6089 %
21	57.9763 %	56.6089 %
22	57.9763 %	56.6089 %

05 Decision tree analysis

FULL

PERSENTAGE SPILT=80%

M	PRUNED	UNPRUNED
2	57.1559%	52.3245%
3	57.247 %	53.0538%
4	57.611%	54.1477%
5	57.7028%	55.4239%
6	57.9763%	57.3382%
7	57.8851%	57.247 %
8	58.2498%	57.3382%
9	57.8851%	58.1586%
10	57.1559%	57.247 %
11	57.247 %	56.8824%
12	57.247 %	57.064%
13	57.0647%	56.3355%
14	57.1559%	56.7001%
15	57.064%	57.247 %
16	57.0647%	58.2498%
17	57.0647%	58.1586%
18	56.9736%	58.0675%
19	57.0647%	58.7056%
20	56.7912%	58.0675%
21	56.4266%	57.1559%
22	56.3355%	57.247 %
23	56.6089%	57.611%
24	56.7001%	57.7028%
25	57.0647%	58.3409%

05 Decision tree analysis

Compare decision methods

	CFS	INFOGAIN	FULL
M	16	17	19
ACCURACY	58.3409 %	58.1586 %	58.7056 %

Approach	CFS M=16	INFOGAIN M=17	FULL M=19
OPTIMISTIC	41.6591%	41.8414%	41.2944%
PESSIMISTIC	56.8368%	46.0346%	55.7429%

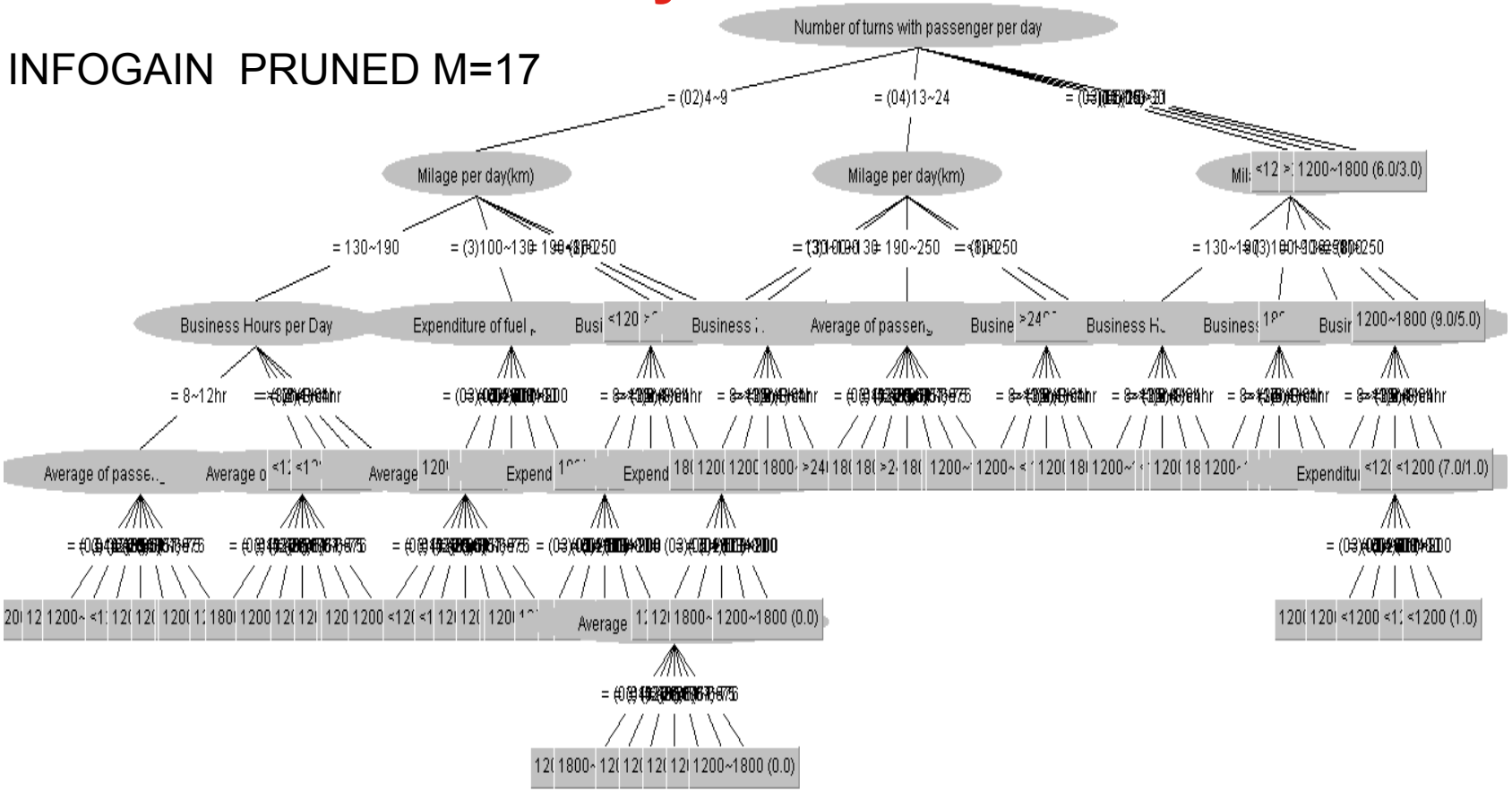
Although we have 58.7% of accuracy in full model

we use **Infogain model** to explain

because of the **pessimistic approach**

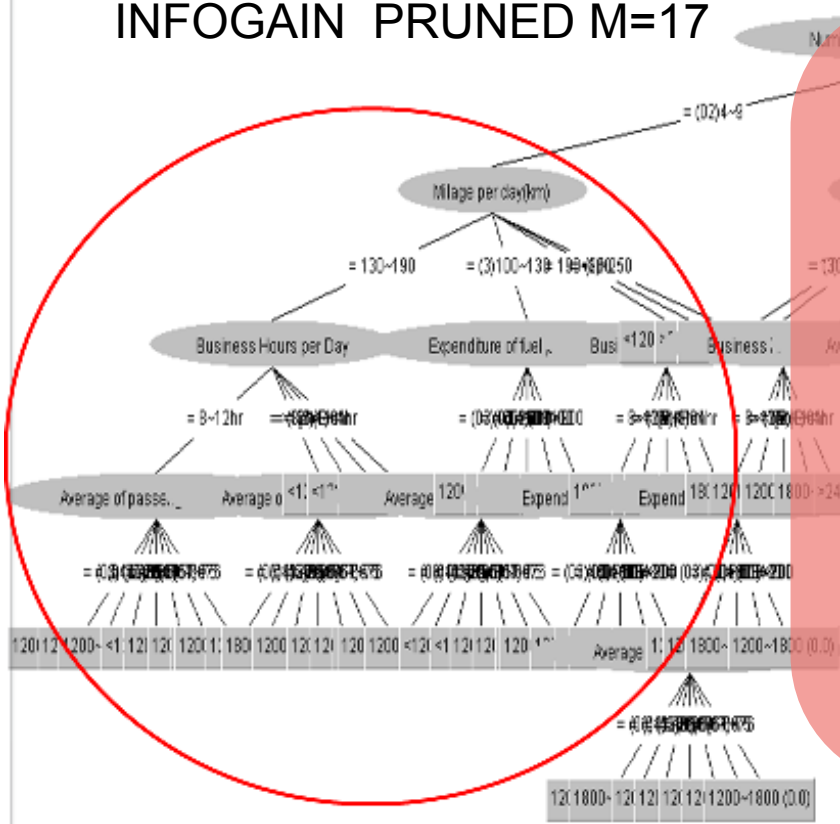
05 Decision tree analysis

INFOGAIN PRUNED M=17



05 Decision tree analysis

INFOGAIN PRUNED M=17



Number of turns with passenger per day 4~9

Mileage per day(km) 130~190

Average of passengers per day

6 out of 7 branches,

the average income per day is 1,200 to 1,800.

Business Hours per Day

When Business Hours per Day is lower than 8hrs, income will be lower than 1,200.

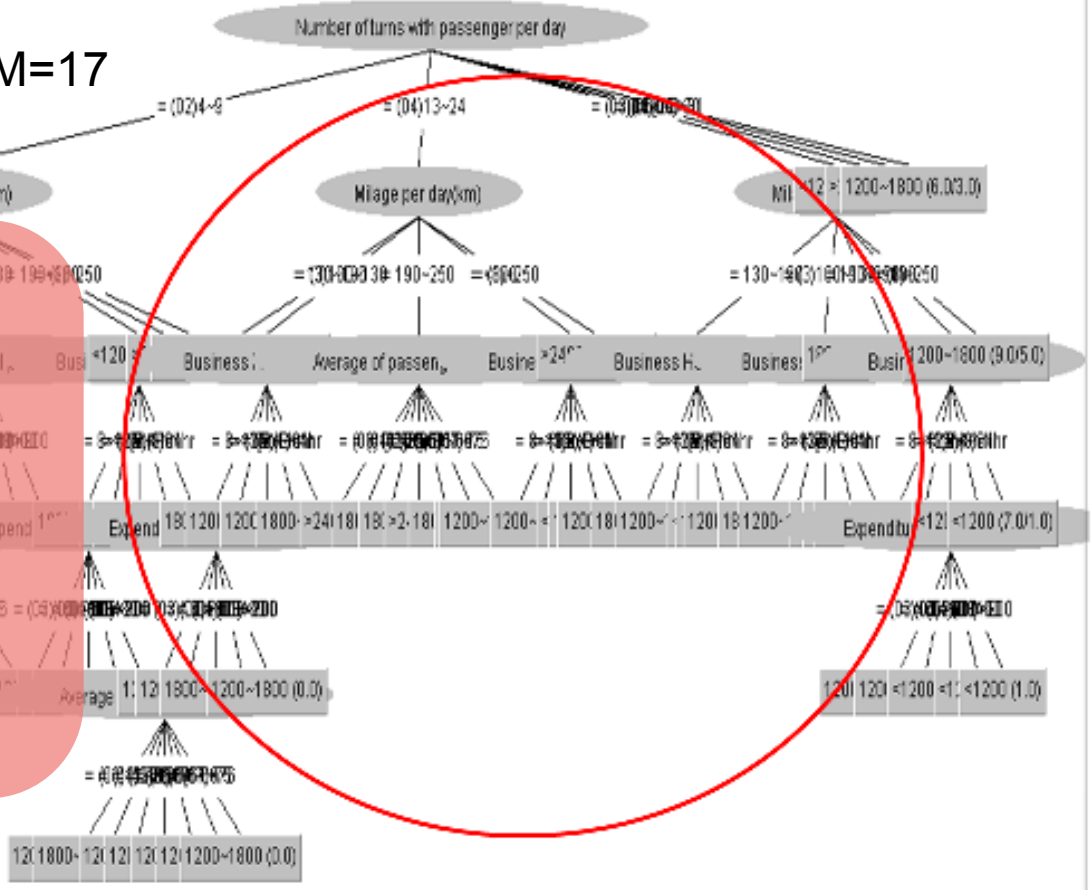
Mileage per day(km) 100~130

Expenditure of fuel per day($NT) <600 INCOME<1,2

Expenditure of fuel per day($NT) >600 INCOME 1,2~1,8

05 Decision tree analysis

INFOGAIN PRUNED M=17



Number of turns with passenger per day 13~24

Mileage per day(km) 190~250

Average of passengers per day 21~60

INCOME>2, 4<21 OR >60
INCOME 1, 8~2, 4

05 Decision tree analysis

Conclusions

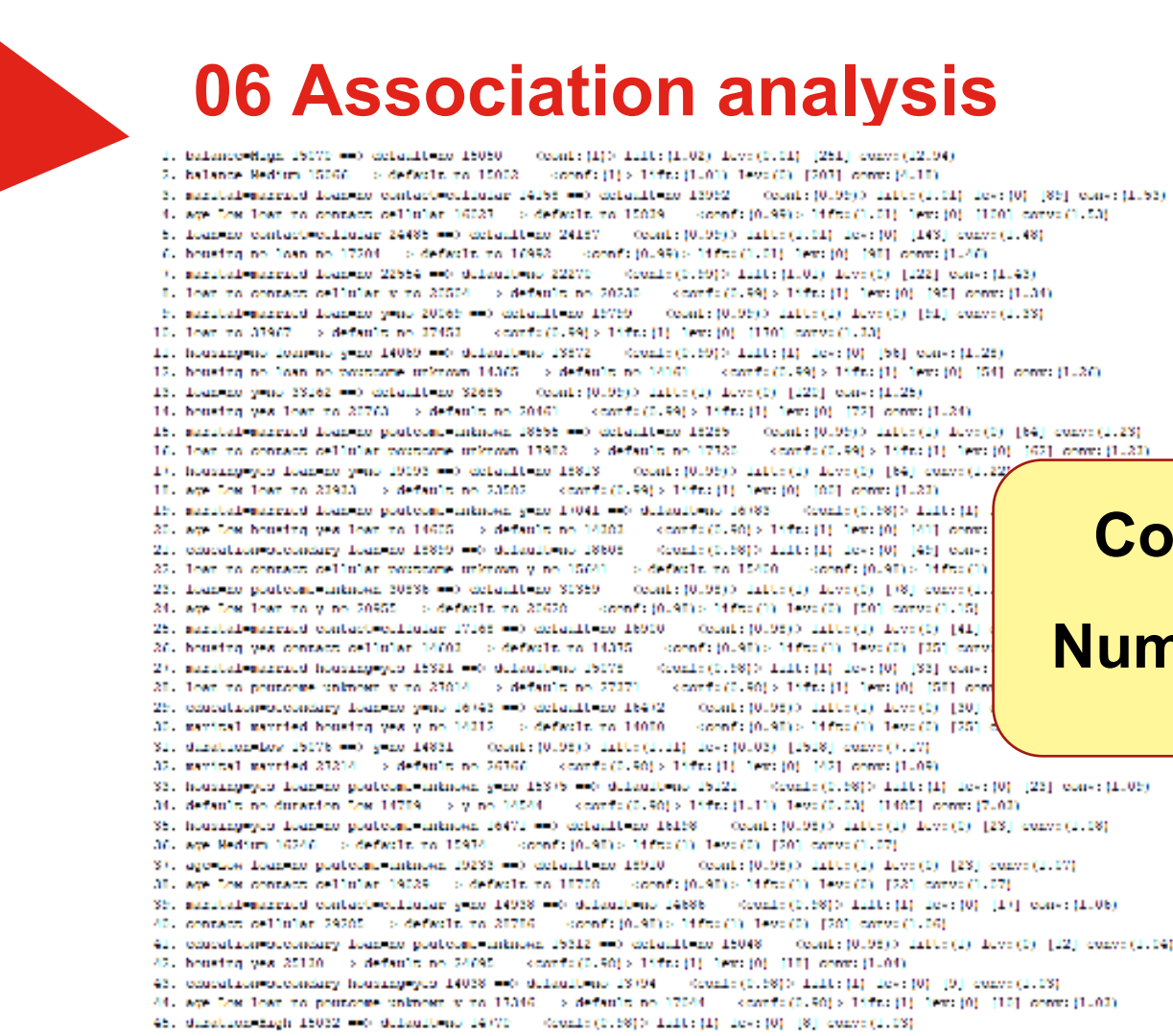
- **If the drivers have lower Number of turns,**

  their Average of passengers per day doesn't affect income.
- And then, they must work over time to get middle income.
- More fuel expenditure will make driver high income.
- **If the drivers have higher Number of turns**, their Mileage per day will be higher, surely.
- Especially, most drivers who has 21 to 60 capacities have high income.
- When Average of passengers is lower than 21 or higher than 60, the drivers will only get middle income.

06 Part Six

Association analysis

06 Association analysis



Confidence 0.9 -> 0.5

Numrules 100 -> 100000

06 Association analysis

Original data

- Mileage per day(km)=<100 ==>
  Y-> Total income per day=<1200 conf:(0.7)

- Average of passengers per day=6~10 ==>
  Y->Total income per day=<1200 conf:(0.63)

- Expenditure of fuel per day($NT)=200~4
  Y->Total income per day=<1200 conf:(0.

Resample

06 Association analysis

Resample data

- Mileage per day(km)=<100
  ==> Total income per day=<1200 conf:(0.62)

- Expenditure of fuel per day($NT)=200~400
  ==> Total income per day=<1200 conf:(0.52)

- Fulltime or Partime=FullTimeJob Radio arrange service fee=N Working area=Northern area ==> Total income per day=>2400 conf:(0.5)

07 Part Seven

Cluster analysis (Hierarchical)

07 Cluster analysis (Hierarchical)

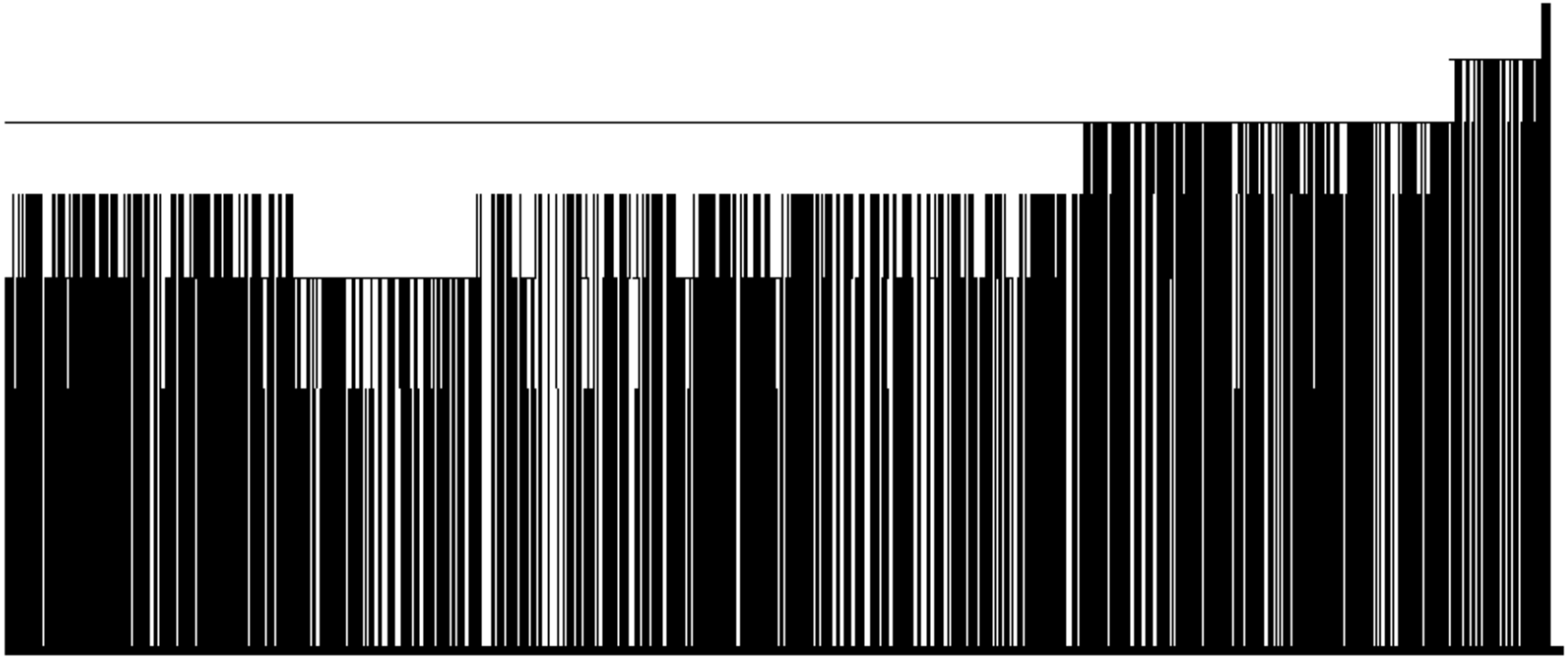
- Full sample: 5486
- 80% resampling → Weka can't run
- 70% resampling → Weka can't run
- 50% resampling → can't visualize the tree

- 30% resampling → can visualize , sample size = 1644

07 Cluster analysis

Weka Cluster Tree Visualizer: 14:11:00 - HierarchicalClusterer (NORDTUSCD-data processed-weka.filters.supervised.instance.Resample-B1.0-S1-Z30.0)

(Linktype = Single

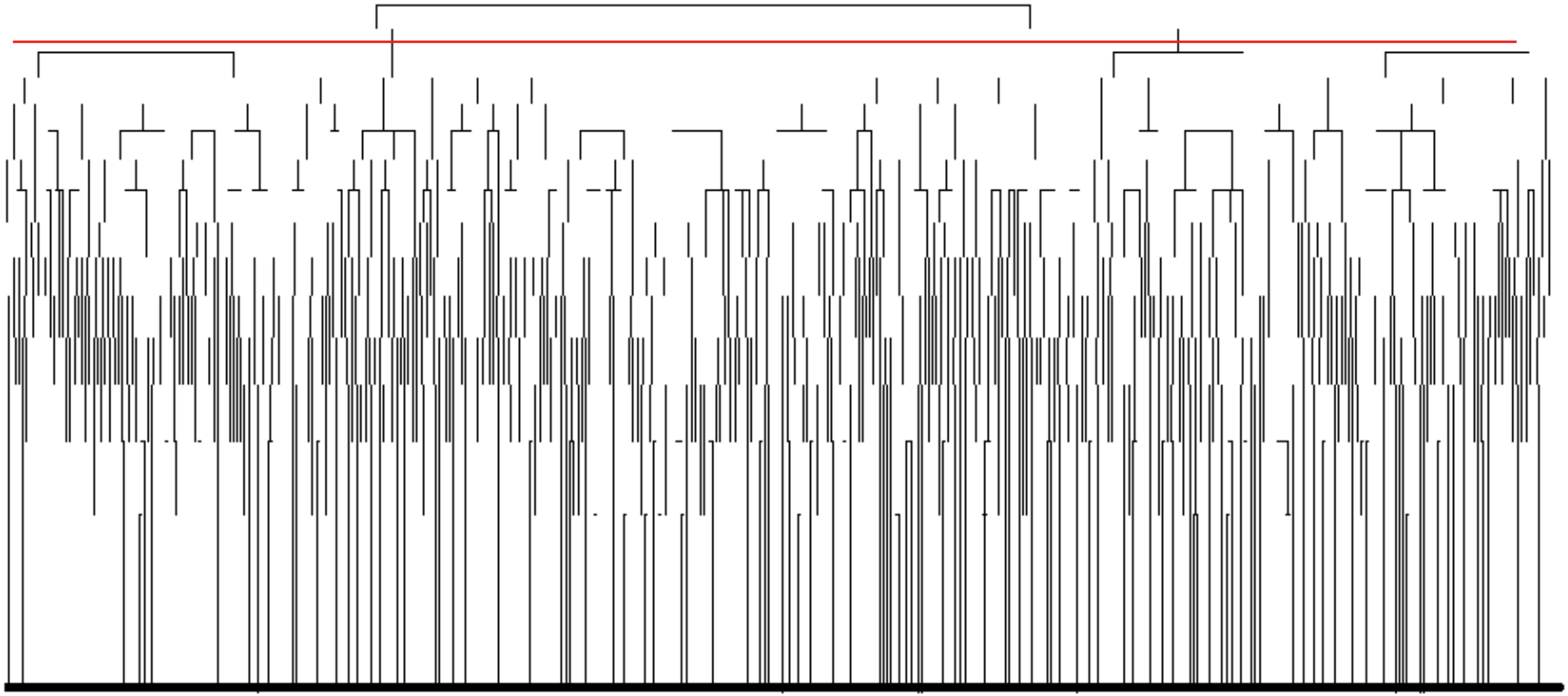


07 Cluster analysis

Weka Cluster Tree Visualizer: 14:44:05 - HierarchicalClusterer (NOBFEUSED) data processed weka.filters.supervised.instance.Resample -B1.0 -S1 -Z50.0)

Linktype = Complete

K=7



08 Part Eight

Cluster analysis(k-means, Cross Analysis)

08 Cluster analysis (K-means)

- **Seed = 7 , has the lowest SSE**

Seed	Cluster=7 , Sum of Squared Error
1	31723
2	31090
3	31802
4	31849
5	31655
6	32138
7	30972
8	31681
9	32363
10	31622

Attribute	Full Data(5486)	Cluster 0(1433)	Cluster 1(910)	Cluster 2(595)
Fulltime or Part time	Full Time Job	Full Time Job	Full Time Job	Full Time Job
Taxi operating type	Private Car	Private Car	Private Car	Private Car
Mileage per day(km)	<100	<100	130~190	<100
Business Hours per Day	8~12hr	6~8hr	8~12hr	8~12hr
Turns with passenger per day	4~9	4~9	10~12	4~9
Average of passengers per day	11~20	11~20	21~50	11~20
Expenditure of fuel per day(NT$)	400~600	200~400	400~600	200~400
Maintenance fees per year($NT)	10K~20K	20k~30k	50k~60k	10K~20K
Cooperative arrange service fee	Y	Y	Y	Y
Radio arrange service fee	Y	Y	Y	Y
Parking fee per month(NT$)	<2000	<2000	<2000	0
Age	50~60	>60	>60	50~60
Education	Senior high school	Senior high school	Junior high school	Junior high school
Reason for driving taxi	Free working time	Free working time	Free working time	Free working time
Working area	Northern area	Northern area	Northern area	Northern area
Y->Total income per day	<1200	<1200	1200~1800	<1200

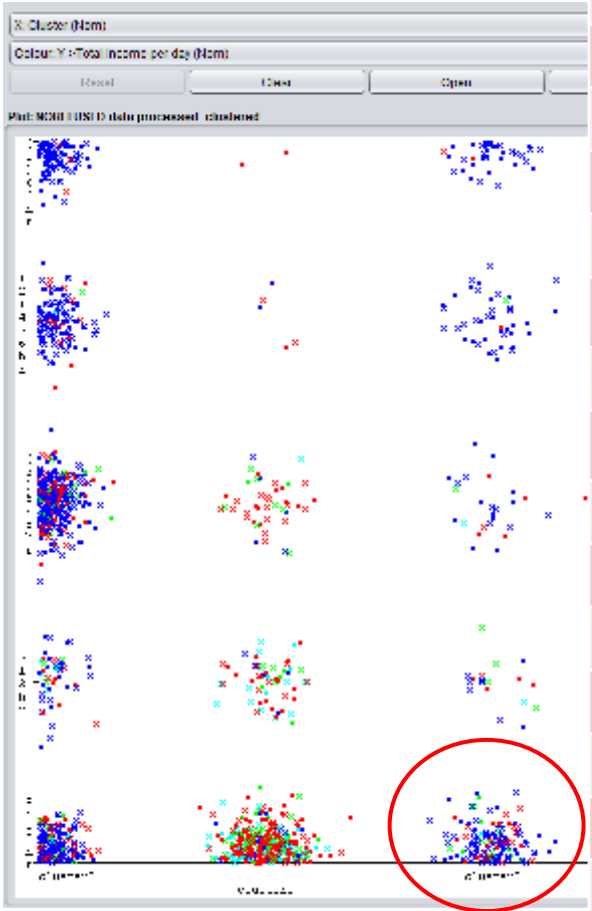
Attribute	Cluster 3(939)	Cluster 4(588)	Cluster 5(580)	Cluster 6(441)
Fulltime or Part time	Full Time Job	Full Time Job	Full Time Job	Full Time Job
Taxi operating type	Company's Car	Company's Car	Private Car	Company's Car
Mileage per day(km)	130~190	130~190	130~190	100~130
Business Hours per Day	8~12hr	8~12hr	8~12hr	8~12hr
Turns with passenger per day	13~24	10~12	13~24	4~9
Average of passengers per day	21~50	11~20	11~20	6~10
Expenditure of fuel per day(NT$)	400~600	400~600	400~600	200~400
Maintenance fees per year($NT)	10K~20K	40k~50k	10K~20K	10K~20K
Cooperative arrange service fee	Y	Y	Y	Y
Radio arrange service fee	Y	N	Y	Y
Parking fee per month(NT$)	<2000	<2000	<2000	<2000
Age	40~50	50~60	50~60	50~60
Education	Senior high school	Senior high school	Senior high school	Senior high school
Reason for driving taxi	Free working time	Free working time	Free working time	Free working time
Working area	Northern area	Northern area	Northern area	Northern area
Y->Total income per day	1800~2400	1800~2400	1200~1800	<1200

08 Cluster analysis (K-means)

Cluster name

- Cluster0 : The old man who should be retired
- Cluster1 : Same as cluster0 but more working hours
- Cluster2 : Leisure old uncle
- Cluster3 : Professional driver
- Cluster4 : Long-term driver
- Cluster5 : Exhausting uncle
- Cluster6 : Poor business uncle

08 Cluster an



Attribute	Cluster 2(595)
Fulltime or Part time	Full Time Job
Taxi operating type	Private Car
Mileage per day(km)	<100
Business Hours per Day	8~12hr
Turns with passenger per day	4~9
Average of passengers per day	11~20
Expenditure of fuel per day(NT$)	200~400
Maintenance fees per year($NT)	10K~20K
Cooperative arrange service fee	Y
Radio arrange service fee	Y
Parking fee per month(NT$)	0
Age	50~60
Education	Junior high school
Reason for driving taxi	Free working time
Working area	Northern area
Y->Total income per day	<1200

Leisure old uncle

09 Part Nine

KNN analysis (best K and weight)

09 KNN analysis (best K and weight) PERSENTAGE SPILT=80%

K	No Weight	1/d	1-d	Average	K	No Weight	1/d	1-d	Average
1	52.60%	52.60%	52.60%	52.60%	16	56.88%	57.61%	57.43%	57.31%
2	55.15%	54.60%	54.60%	54.79%	17	56.88%	57.61%	57.43%	57.31%
3	56.52%	56.15%	56.24%	56.31%	18	57.06%	57.61%	57.52%	57.40%
4	57.70%	57.16%	57.34%	57.40%	19	57.06%	57.61%	57.52%	57.40%
5	57.16%	56.79%	57.06%	57.00%	20	57.25%	57.70%	57.61%	57.52%
6	57.52%	57.43%	57.34%	57.43%	21	57.34%	57.70%	57.70%	57.58%
7	57.25%	57.43%	57.16%	57.28%	22	57.34%	57.79%	57.79%	57.64%
8	57.16%	57.43%	57.16%	57.25%	23	57.43%	58.07%	57.98%	57.82%
9	57.16%	57.61%	57.25%	57.34%	24	57.34%	57.98%	57.89%	57.73%
10	57.25%	57.98%	57.70%	57.64%	25	57.52%	58.16%	58.07%	57.92%
11	57.25%	57.89%	57.61%	57.58%	26	57.79%	58.43%	58.34%	58.19%
12	57.34%	58.07%	57.79%	57.73%	27	57.89%	58.43%	58.43%	58.25%
13	57.43%	58.16%	57.89%	57.82%	28	58.07%	58.52%	58.52%	58.37%
14	57.34%	58.07%	57.79%	57.73%	29	58.43%	58.80%	58.80%	58.68%
15	56.79%	57.70%	57.43%	57.31%	30	58.43%	58.80%	58.80%	58.68%
					Average	57.14%	57.53%	57.43%	

10 Part Ten

Compare the two supervised method

10 Compare the two supervised method

	Decision tree	KNN
Test method	80% percentage-split	80% percentage-split
Select	All attributes	All attributes
Accuracy	58.7056%	58.8000%
Scale	Can deal with continuous variable and categorical variable at the same time	Variable should be normalize to prevent distance measures from being dominated by one variable

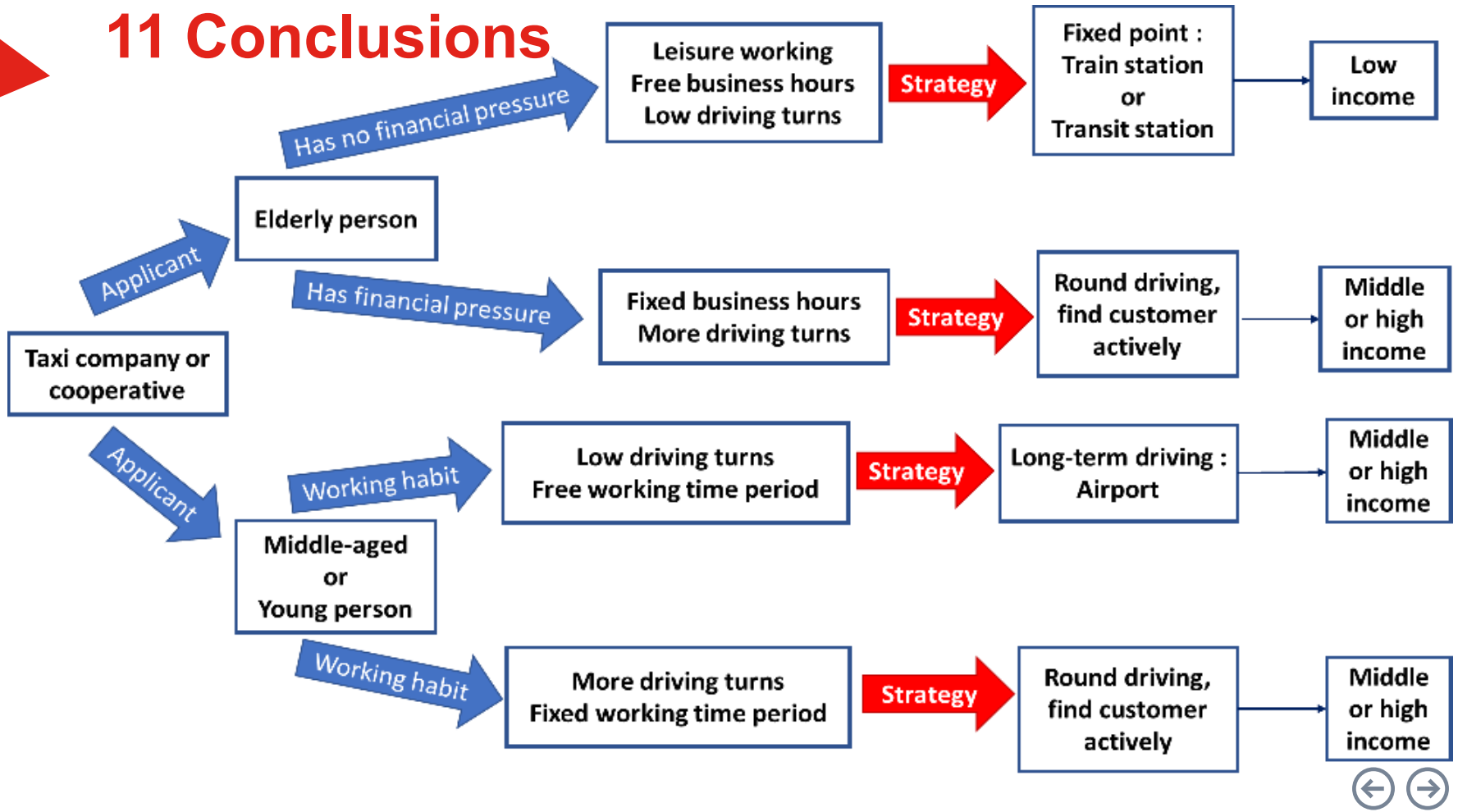
Target variable has no obvious pattern tend to use decision tree

Instances have highly similarity tend to use KNN

11 Part Eleven

Conclusions

11 Conclusions



Thanks

Thanks for your patience & listening