APPLIED MULTIVARIATE STATISTICS

Data Set 4: Car Data Evaluation

FAL001 // 17 June 2024

Supervisor: Martin Waltz

Professor: Prof. Dr. Ostap Okhrin

Contents

1. Study objective + data description
2. Correspondence Analysis
   a. Factor Map 1
   b. Factor Map 2
   c. Factor Map 3
   d. Factor Map 4
   e. Factor Map 5
   f. Factor Map 6
3. Original study: Criteria Tree
4. Multidimensional Scaling
   a. Configuration Plot
5. Discriminant Analysis
6. Conclusion

Study objective + data description

Objective: use dimensionality reduction techniques to identify relationships between categories of each descriptive variable with the categories of car's evaluation level, and visualize them for pattern recognition.

Database: contains examples with the structural information removed, i.e., directly relates car acceptability *(evaluation level)* to the six input attributes

<u>Variable Type:</u> Categorical   <u>Instances:</u> 1728   <u>Missing Values:</u> No

- Categories of car evaluation (target variable in the original study):

  class: *unacceptable, acceptable, good, very good*

- Categories of input attributes (feature variables in the original study):
  buying price: *vhigh, high, med, low*
  price of the maintenance: *vhigh, high, med, low*
  number of doors: *2, 3, 4, 5more*
  capacity in terms of persons to carry: *2, 4, more*
  the size of luggage boot: *small, med, big*
  estimated safety of the car: *low, med, high*

Table 1: Car evaluation and Buying price

Contingency Table:

	high	low	med	vhigh
acc	108	89	115	72
good	0	46	23	0
unacc	324	258	268	360
vgood	0	39	26	0

CA output Table

C. Eval	Inertia	Ctr 1	Ctr 2
acc	6.812	0.272	74.967
good	48.804	47.815	7.198
unacc	13.275	11.666	17.832
vgood	40.625	40.247	0.004

Price	Inertia	ctr1	ctr2
high	21.139	19.295	18.693
low	50.245	48.792	13.582
Med	8.948	6.161	30.155
vhigh	29.184	25.752	37.570

- **P-value for Pearson Chi-square test ~ 0**
- **We reject H0: dependency exists**
- **Variance explained by first 2 eigenvectors: 99.6%**
- **Price: Inertia is very high for “Low” (Σr = 109.516)**
- **Car Evaluation: Inertia is very high for “good” and “vgood”**
- **Conclusion: Negative Relationship**

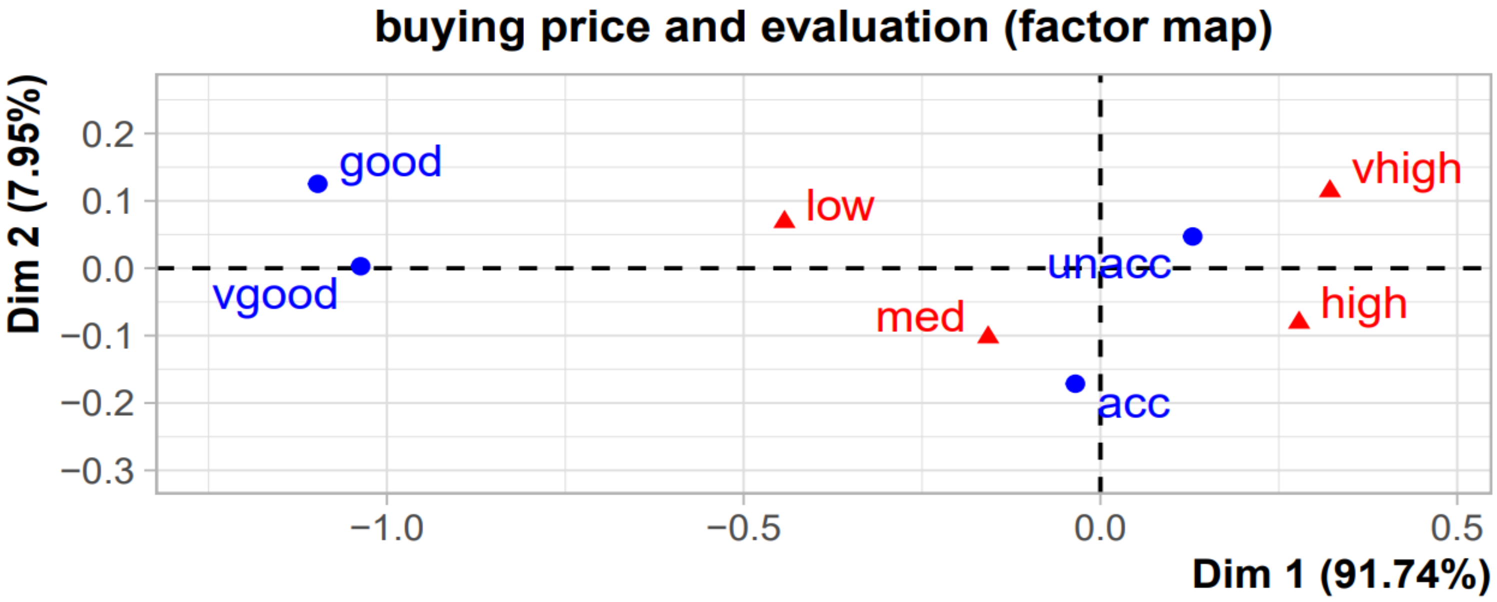


Table 2: Car evaluation and price of maintenance

Contingency Table:

	high	low	med	vhigh
acc	105	92	115	72
good	0	46	23	0
unacc	314	268	268	360
vgood	13	26	26	0

CA output Table

C. Eval	Inertia	Ctr 1	Ctr 2
acc	6.233	1.704	43.706
good	48.804	64.344	25.847
unacc	11.132	13.653	12.244
vgood	16.551	20.299	18.204

Maint	Inertia	ctr1	ctr2
high	11.1	11.85	23.15
low	33.488	43.553	21.447
Med	8.948	8.248	26.752
vhigh	29.184	36.349	28.651

- **P-value for Pearson Chi-square test ~ 0**
- **We reject H0: dependency exists**
- **Variance explained by first 2 eigenvectors: 100%**
- **Maint: Inertia is high for "Low and vhigh" (Σr = 82.72)**
- **Car Evaluation: Inertia is very high for "good"**
- **Conclusion: Negative Relationship**

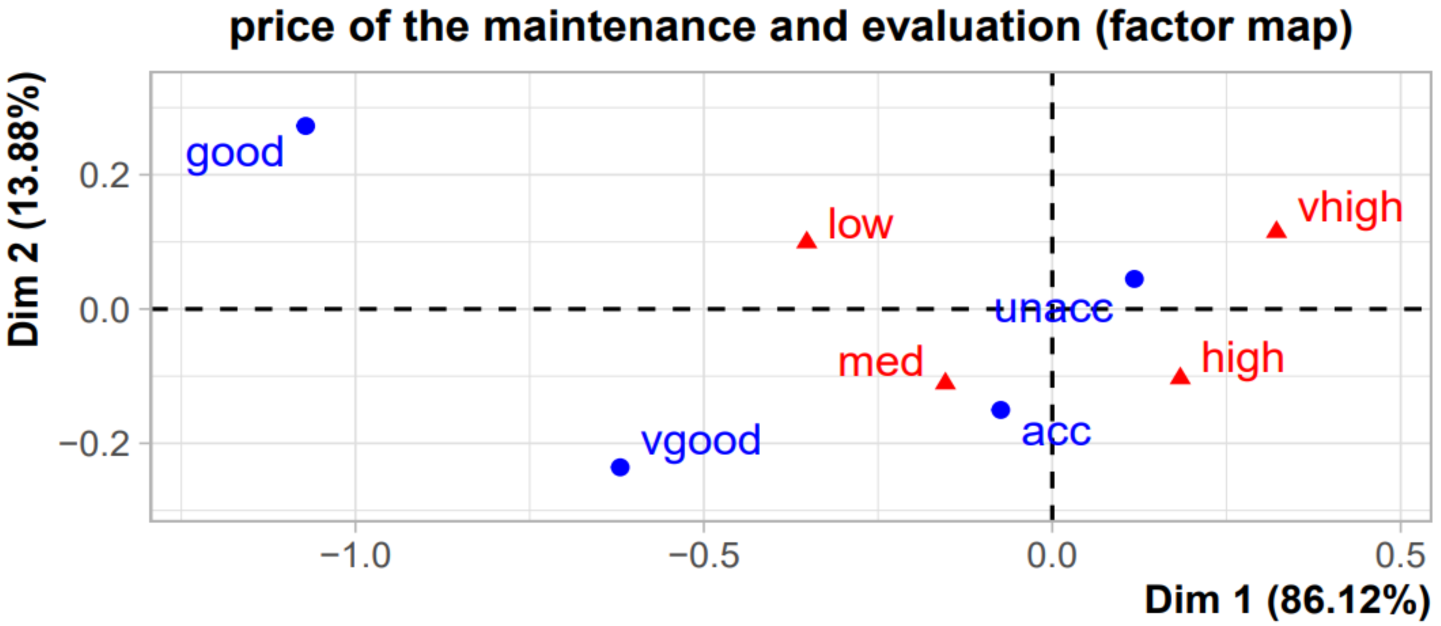


Table 3: Car evaluation and Doors

Contingency Table:

	2	3	4	5more
acc	81	99	102	102
good	15	18	18	18
unacc	326	300	292	292
vgood	10	15	20	20

CA output Table

C. Eval	Inertia	Ctr 1	Ctr 2
acc	1.845	30.746	29.023
good	0.226	3.522	11.59
unacc	1.49	25.434	4.543
vgood	2.448	40.297	54.844

doors	Inertia	ctr1	ctr2
2	3.974	67.978	7.022
3	0.141	0.062	74.938
4	0.948	15.98	9.02
5more	0.948	15.98	9.02

- **P-value for Pearson Chi-square test ~ 0.32**
- **We cannot reject H0: dependency may exist**
- **Variance explained by first 2 eigenvectors: 100%**
- **Doors: Inertia is high for “2 doors” (Σr = 6.0)**
- **Car Evaluation: Inertia is high for “vgood”**
- **Conclusion: Positive Relationship**

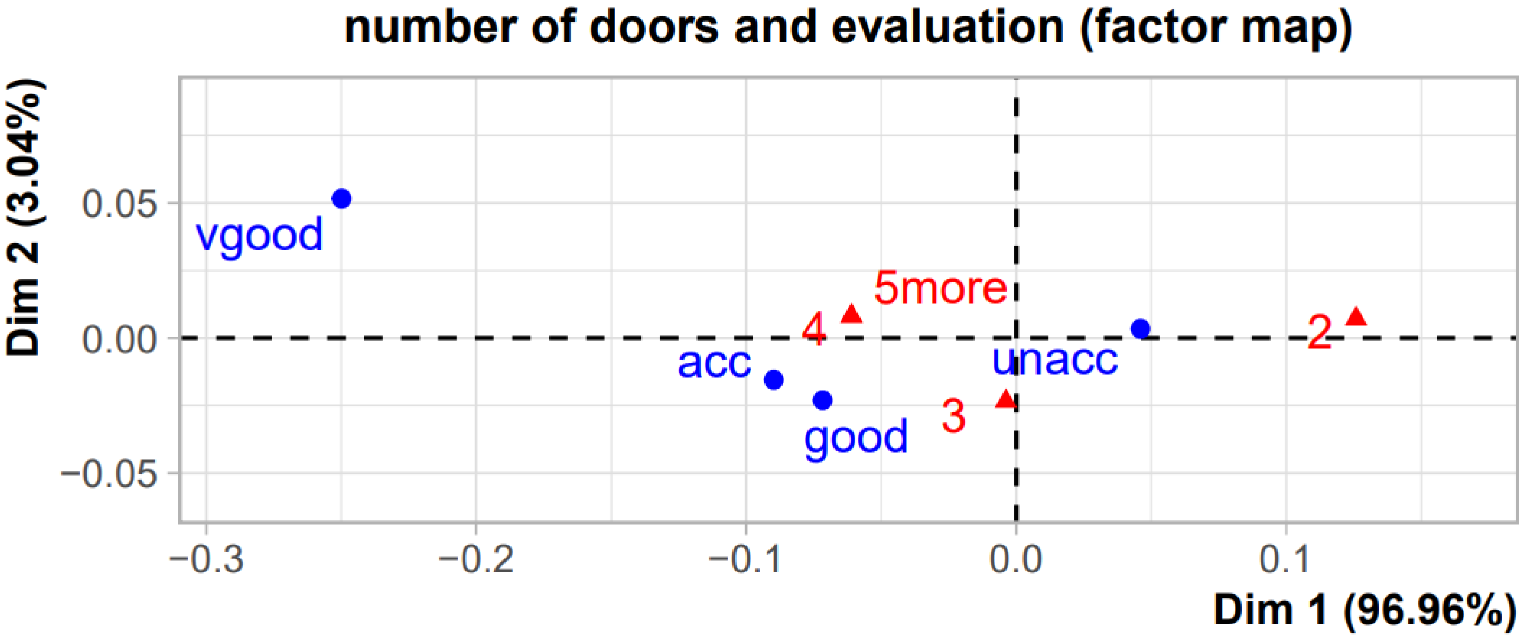

number of doors and evaluation (factor map)

Table 4: Car evaluation and Person capacity

Contingency Table:

	2	4	More
acc	0	198	186
good	0	36	33
unacc	576	312	322
vgood	0	30	35

CA output Table

C. Eval	Inertia	Ctr 1	Ctr 2
acc	111.437	51.981	7.789
good	20.079	9.354	5.759
unacc	64.237	29.977	0
vgood	19.142	8.689	86.452

capac.	Inertia	ctr1	ctr2
2	142.7	66.592	0.075
4	40.229	18.638	48.029
More	31.965	14.77	51.896

- **P-value for Pearson Chi-square test ~ 0**
- **We reject H0: dependency exists**
- **Variance explained by first 2 eigenvectors: 100%**
- **P.capacity: Inertia is very high for “2 person capacity” (Σr = 214.894)**
- **Car Evaluation: Inertia is very high for “acc”**
- **Conclusion: Positive Relationship**

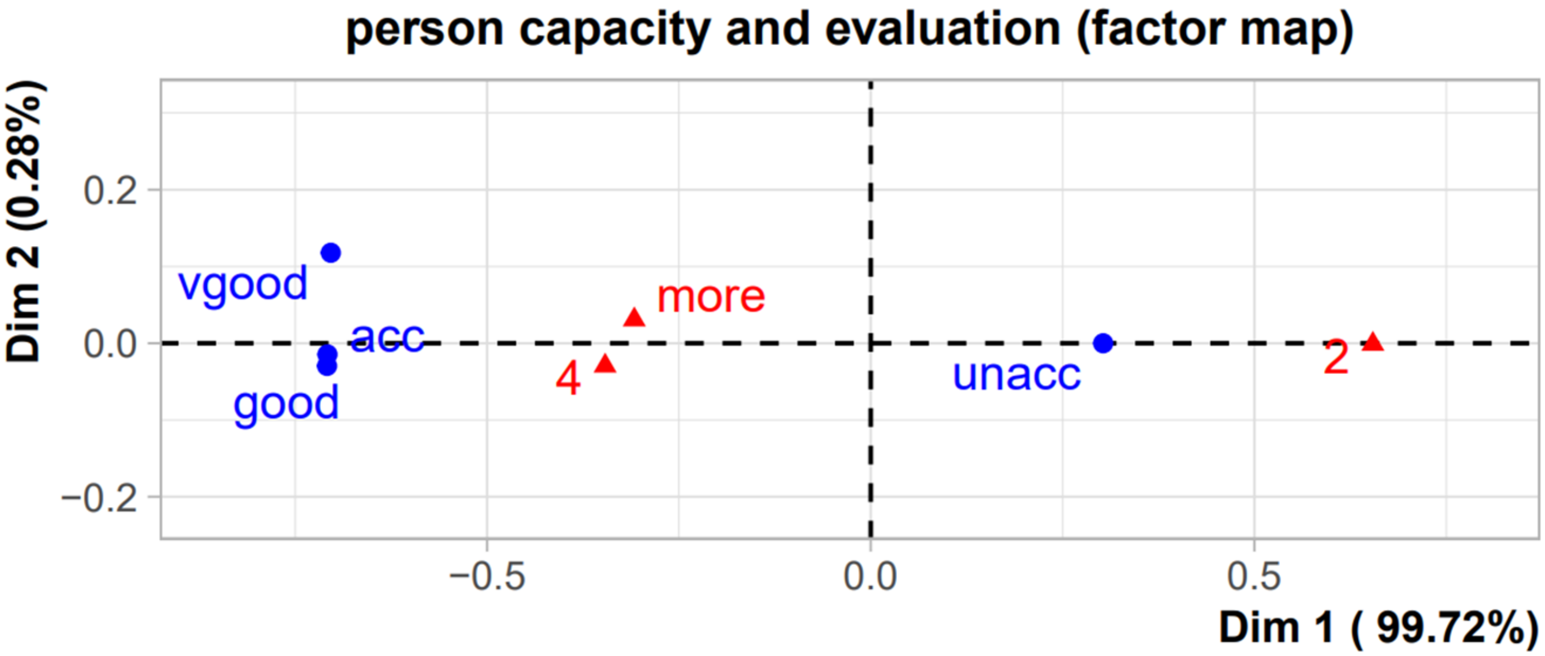


Table 5: Car evaluation and boot size

Contingency Table:

	big	med	small
acc	144	135	105
good	24	24	21
unacc	368	392	450
vgood	40	25	0

CA output Table

C. Eval	Inertia	Ctr 1	Ctr 2
acc	3.771	12.093	47.676
good	0.151	0.436	14.677
unacc	5.1	16.553	13.424
vgood	21.813	70.918	24.223

boot s.	Inertia	ctr1	ctr2
big	11.951	38.801	27.866
med	0.728	2.123	64.544
small	18.156	59.076	7.59

- **P-value for Pearson Chi-square test ~ 0**
- **We reject H0: dependency exists**
- **Variance explained by first 2 eigenvectors: 100%**
- **bootsize: Inertia is high for “small” boot size (Σr = 30.835)**
- **Car Evaluation: Inertia is high for “vgood”**
- **Conclusion: Positive Relationship**

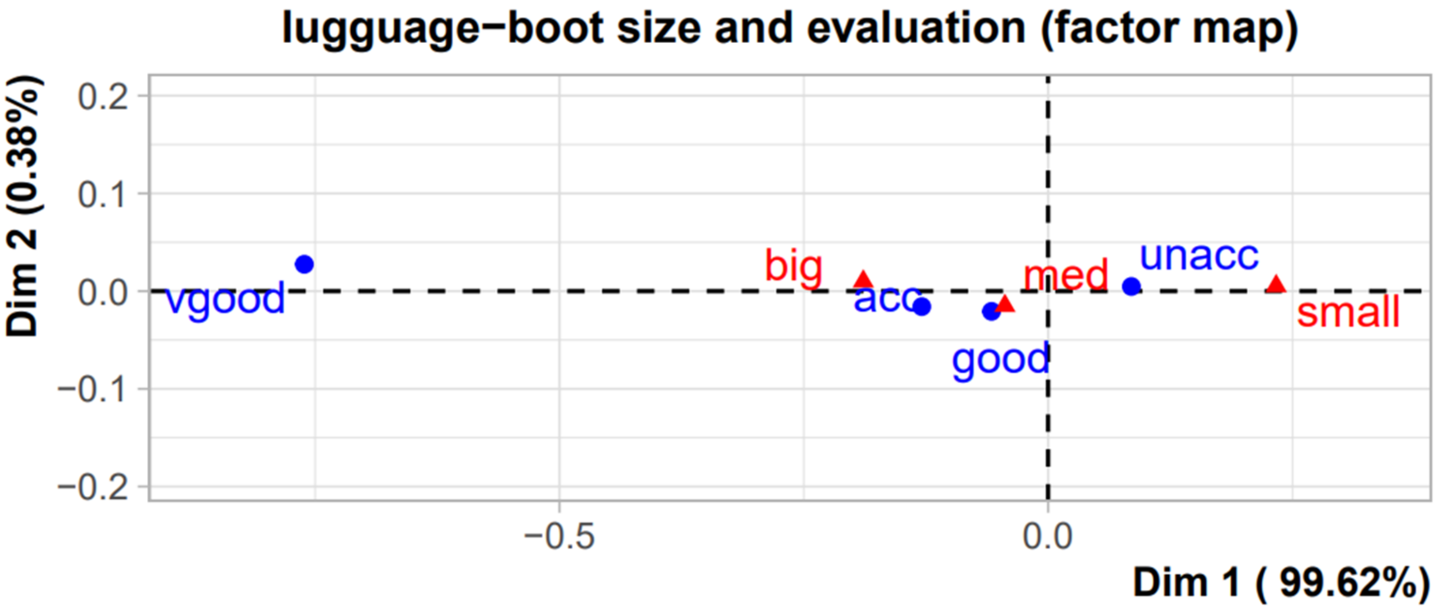


Table 6: Car evaluation and safety level

Contingency Table:

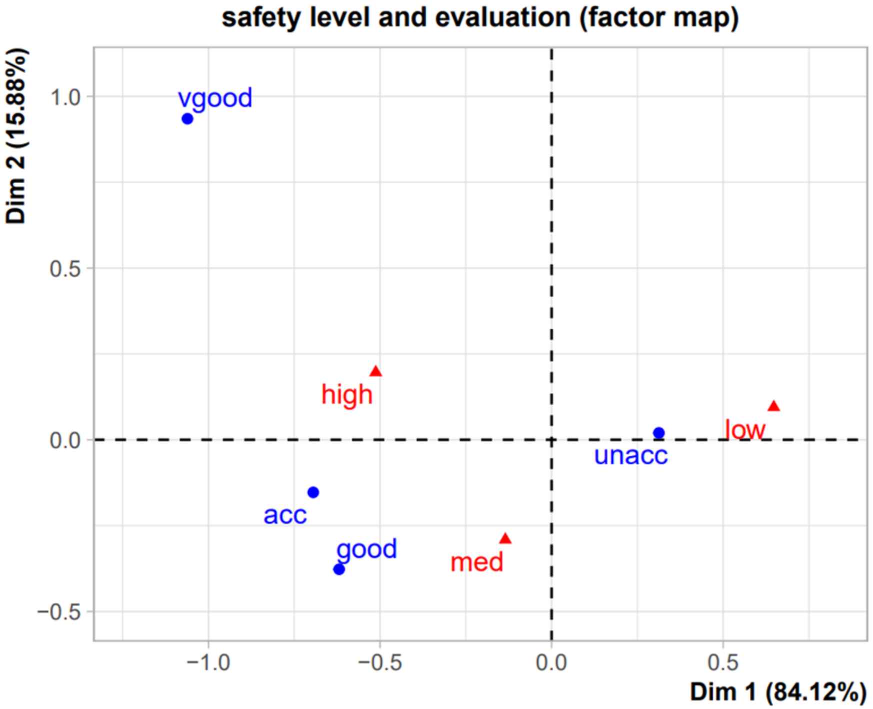
	high	low	med
acc	204	0	180
good	30	0	39
unacc	277	576	357
vgood	65	0	0

CA output Table

C. Eval	Inertia	Ctr 1	Ctr 2
acc	112.413	45.943	11.834
good	20.984	6.56	12.89
unacc	68.757	29.348	0.629
vgood	75.231	18.149	74.647

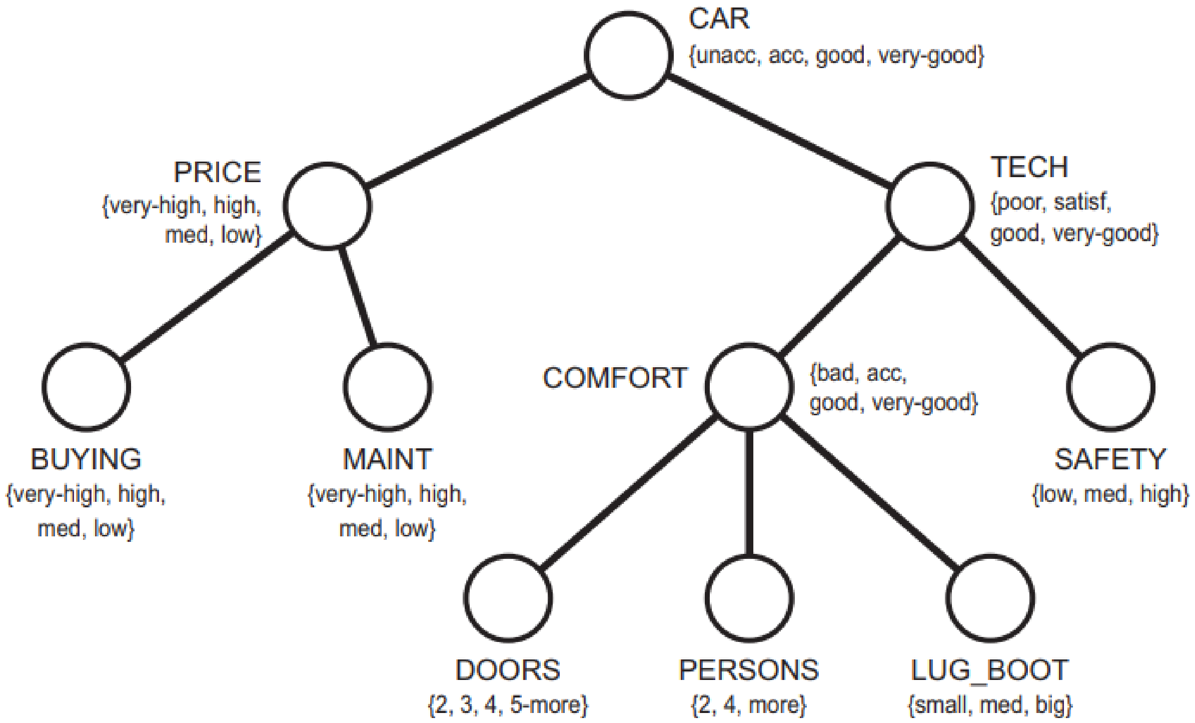
safety	Inertia	ctr1	ctr2
high	100.401	37.528	29.139
low	142.7	59.874	6.792
med	34.285	2.598	64.069

- **P-value for Pearson Chi-square test ~ 0**
- **We reject H0: dependency exists**
- **Variance explained by first 2 eigenvectors: 100%**
- **safety: Inertia is very high for “low” safety (Σr = 277.386)**
- **Car Evaluation: Inertia is very high for “acc”**
- **Conclusion: Positive Relationship**



Original Study

Criteria tree for the car selection problem



CAR
{unacc, acc, good, very-good}
PRICE
{very-high, high, med, low}
TECH
{poor, satisf, good, very-good}
BUYING
{very-high, high, med, low}
MAINT
{very-high, high, med, low}
COMFORT
{bad, acc, good, very-good}
SAFETY
{low, med, high}
DOORS
{2, 3, 4, 5-more}
PERSONS
{2, 4, more}
LUG_BOOT
{small, med, big}

Figure 2: KNOWLEDGE ACQUISITION AND EXPLANATION FOR MULTI-ATTRIBUTE DECISION MAKING
M. Bohanec, V. Rajkovič

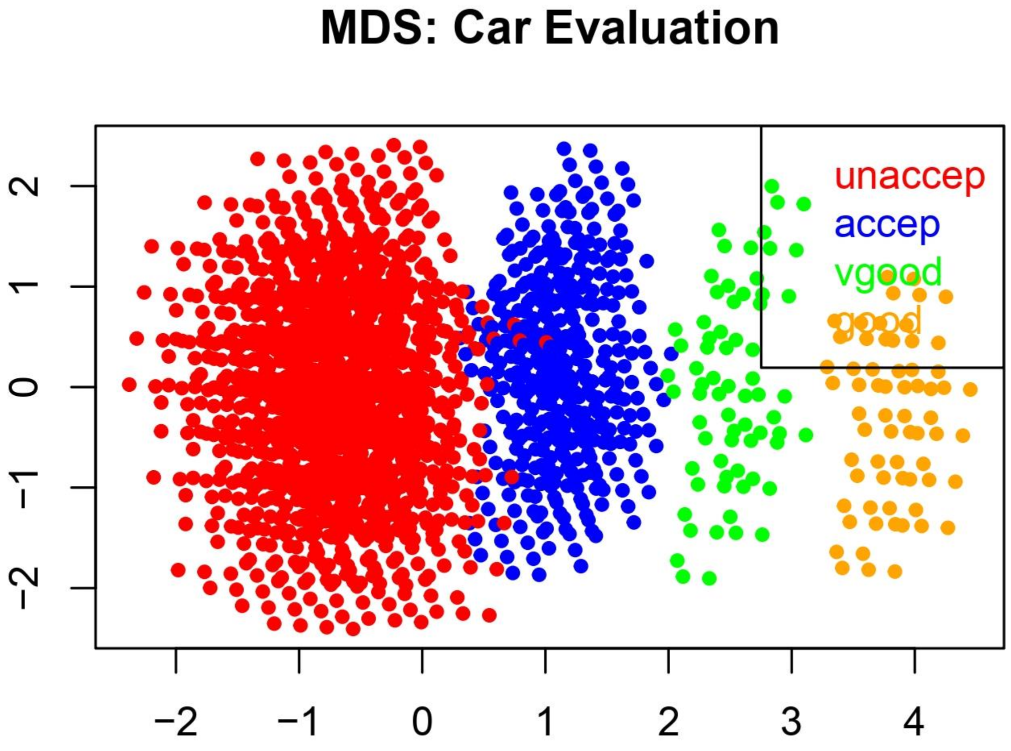
Multidimensional Scaling

Method:

- Original data is represented in 7 dimensions
- Used ordinal encoding to preserve order
- Scaled the data
- Calculated Euclidean distance for each instance
- Higher distance ~ more *dissimilarity*

Observations:

- Dimensionality reduction to 2 dimensions
- Instances of a particular *class* are closer



Configuration plot (a)

Discriminant Analysis

Assumptions:

- Each category in *class* ~ equal prior probability
- Each class ~ N_6
- $J = 4$
- Covariance matrix is equal for all classes
- Linear decision boundary (~linear classifier)

Observations:

- “Unacc” is the most wrongly classified category
- Misclassification rate between actual class and the predicted class is 0.232

Comparing true and predicted class in a table

	unacc_{pr}	acc_{pr}	good_{pr}	vgood_{pr}
unacc	933	182	81	14
acc	45	273	34	32
good	0	0	64	5
vgood	0	2	6	57

Conclusion

- Dependencies are also found in the original study (Except doors in our study)
- Using correspondence analysis after developing different contingency tables we have illustrated key relationships between a car's acceptability with six other car attributes
- Multidimensionality Scaling helps visualize the data in reduced dimensions while simultaneously preserving the *dissimilarity* based on **car acceptability**
- PCA: The first 4 PCs explain only 67% of total variation in our data
- FA:
  - ❖ First three factors explain only 47% of total variance in our data
  - ❖ The chi-square test shows that the first three factors are not sufficient
- DA: it works well to classify our class variable (response variable) with six predictive attributes