ROAD INVESTMENT ANALYSIS

Advanced Multivariate Statistics

David Fernandez

January 23, 2023

Università degli studi di Milano

Outline

Structure

1. Objectives
2. Data Description & Data Pre-processing
3. MDS, PCA & Clustering
4. Feature Selection & Modelling
5. Conclusions & Further Analysis

Objectives

- Identify similarities amongst countries
- Find key features that explain the behavior of the data
- Model road investment per capita

Data Description & Data Pre-processing

What kind of data are we talking about?

- OECD Indicators: Transport Infrastructure, Infrastructure Usage, Safety, Economic and Social and Environment topics.
- Data is from 2017.
- y = *road investment per capita*
- $X \in R^{30}$, $n = 43$ (countries). All features are numeric
- E.g.: *road density, share of urban roads in total road network, number of passanger cars, number of road fatalities, CO2 emissions from transport*, etc.

Data Description & Data Pre-processing

- 13.5% datapoints missing

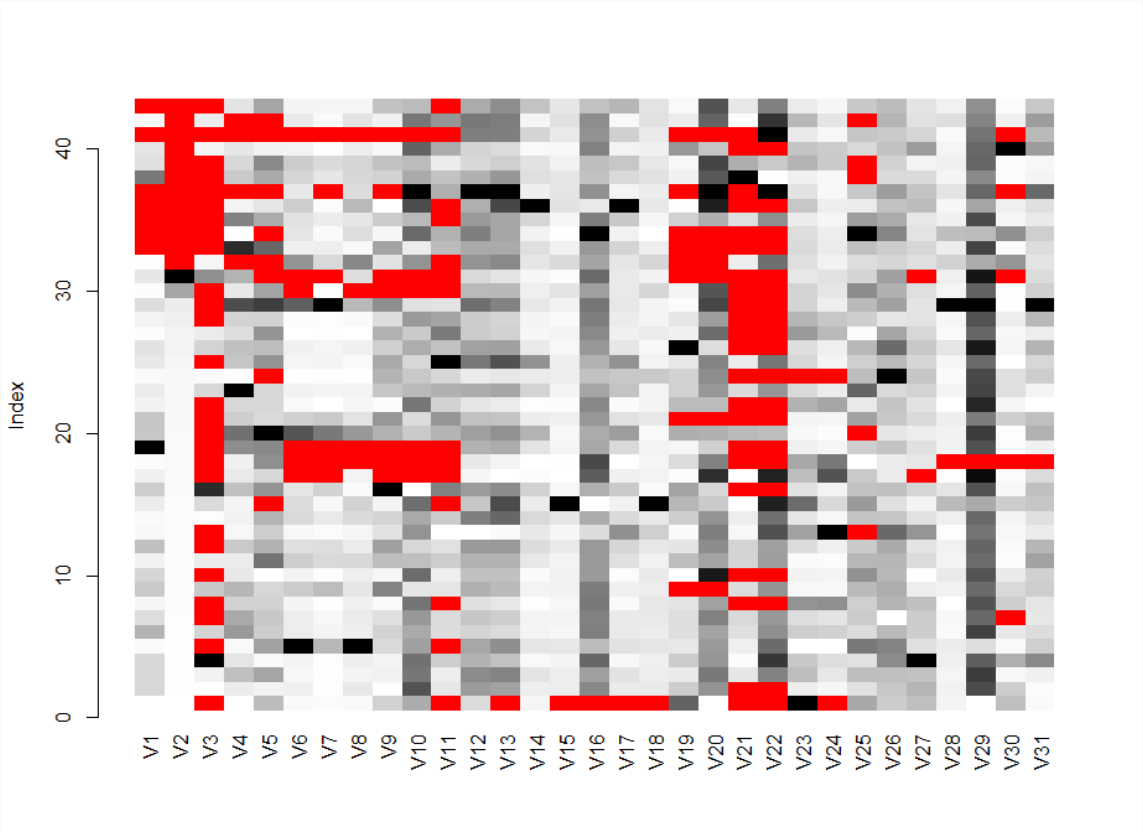

Figure: Data pattern, raw data

Data Description & Data Pre-processing

- Removed features with >30% missing values (4)
- Still 8.4% datapoints missing
- Mice imputation (pmm method)

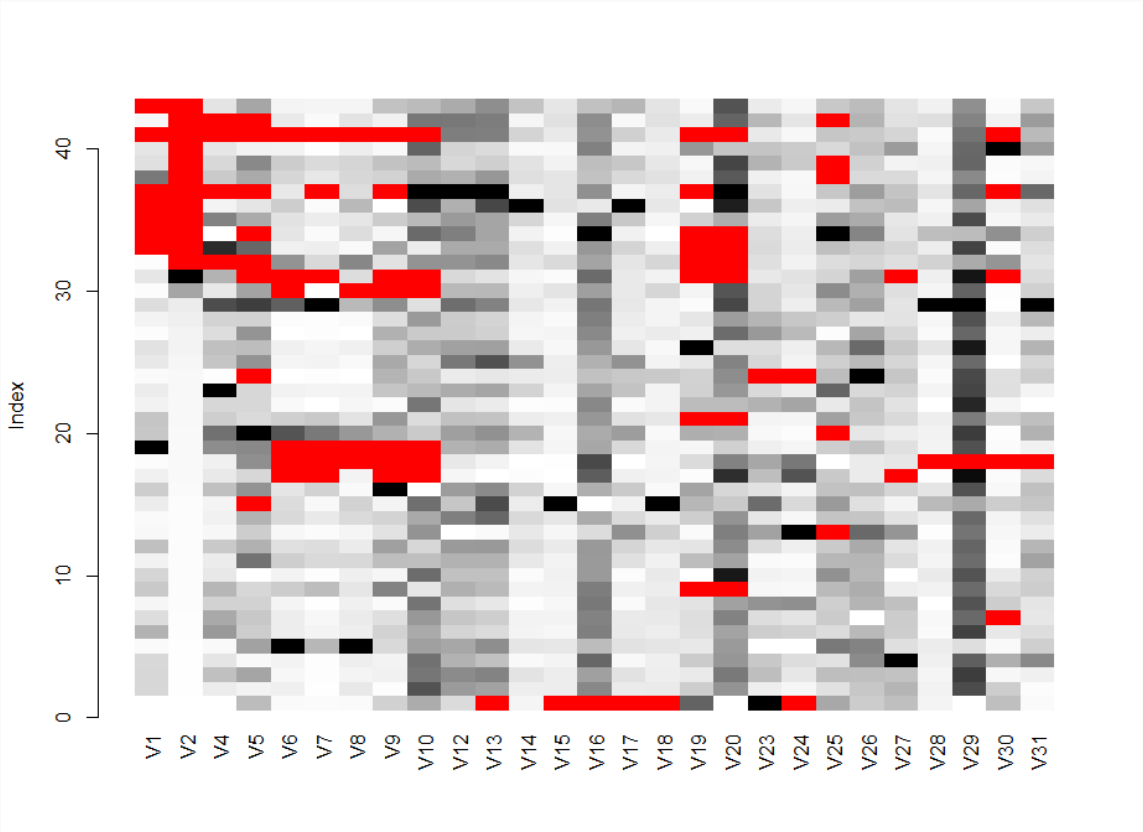

Figure: Data pattern, reduced raw data

Data Description & Data Pre-processing

- Imputed data

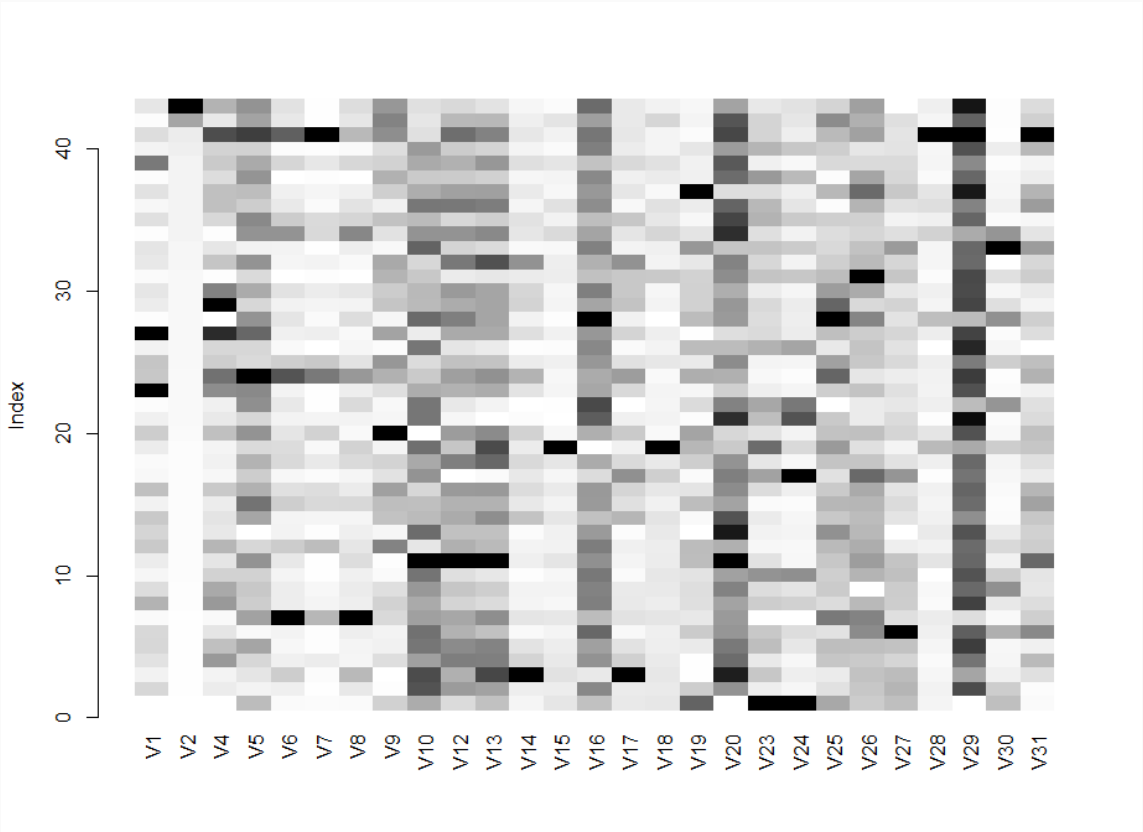

Figure: Data pattern, imputed data

Data Description & Data Pre-processing

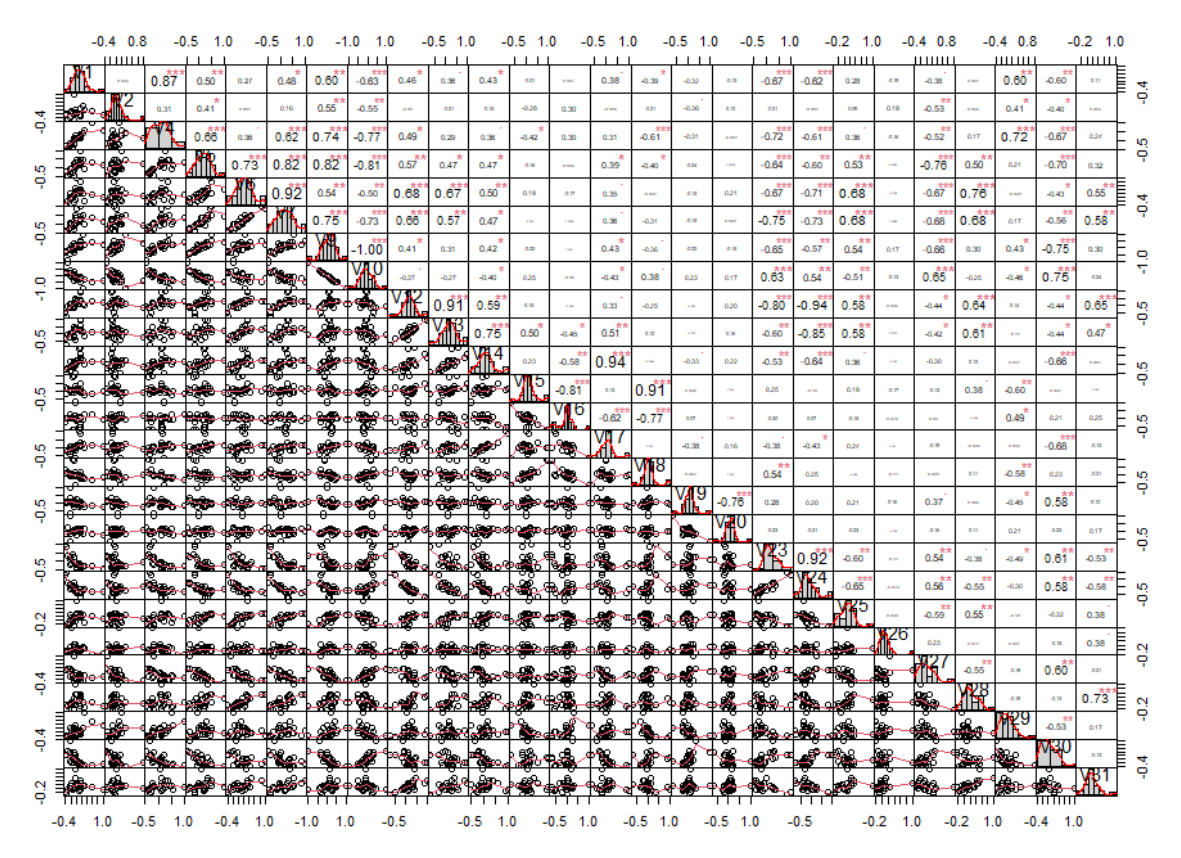

Figure: Correlation Matrix

MDS, PCA AND CLUSTERING

Multidimensional Scaling

Correlations

- D1: road fatalities per vehicle (-0.80), passenger cars per inhabitant (+0.68)
- D2: share of passenger cars in motor vehicles (-0.77), goods road motor vehicles per inhabitant (+0.85)

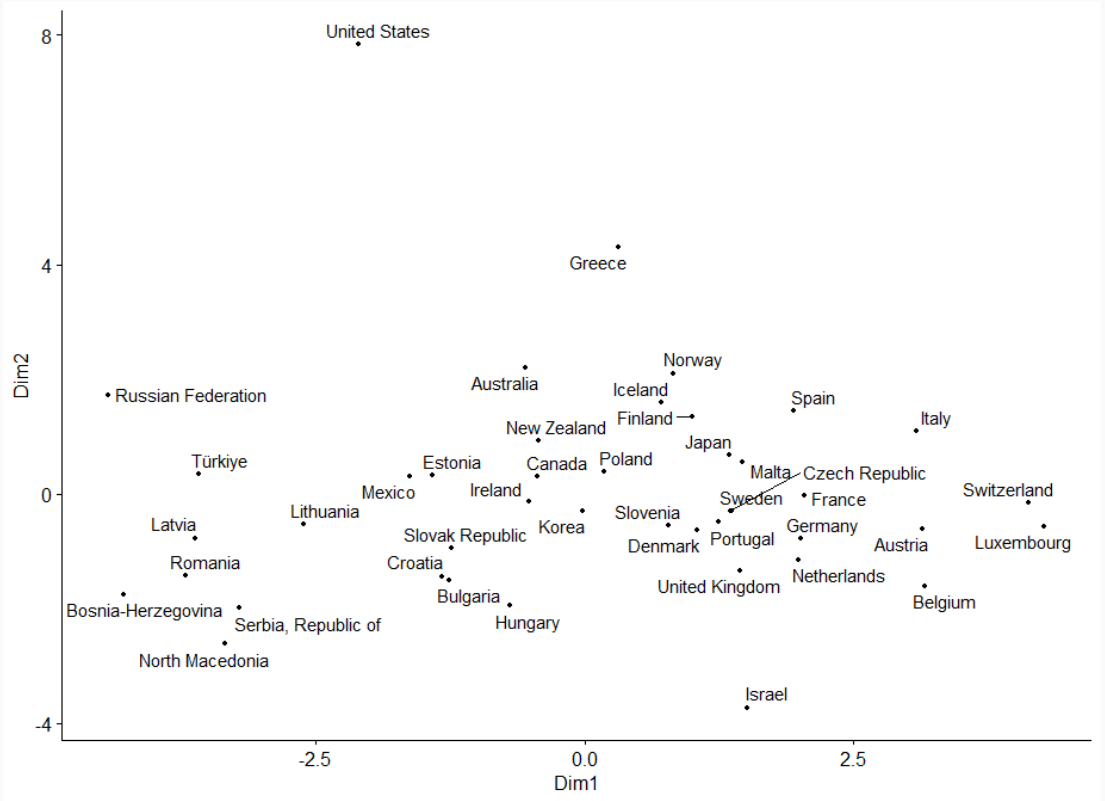

Figure: MDS dimensions - Euclidean distance

Model Based Clustering

Do these clusters explain our independent variable?

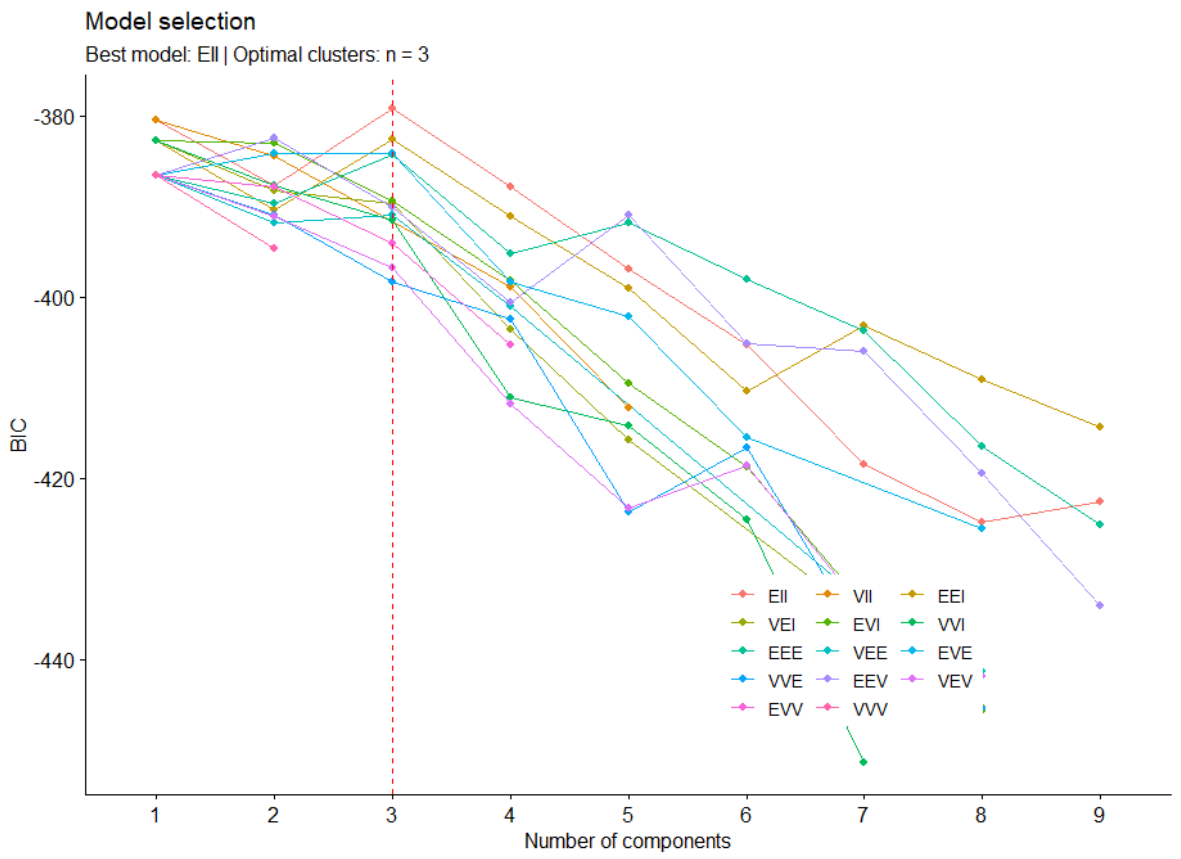

Figure: Model Selection: EEI (3)

Model Based Clustering

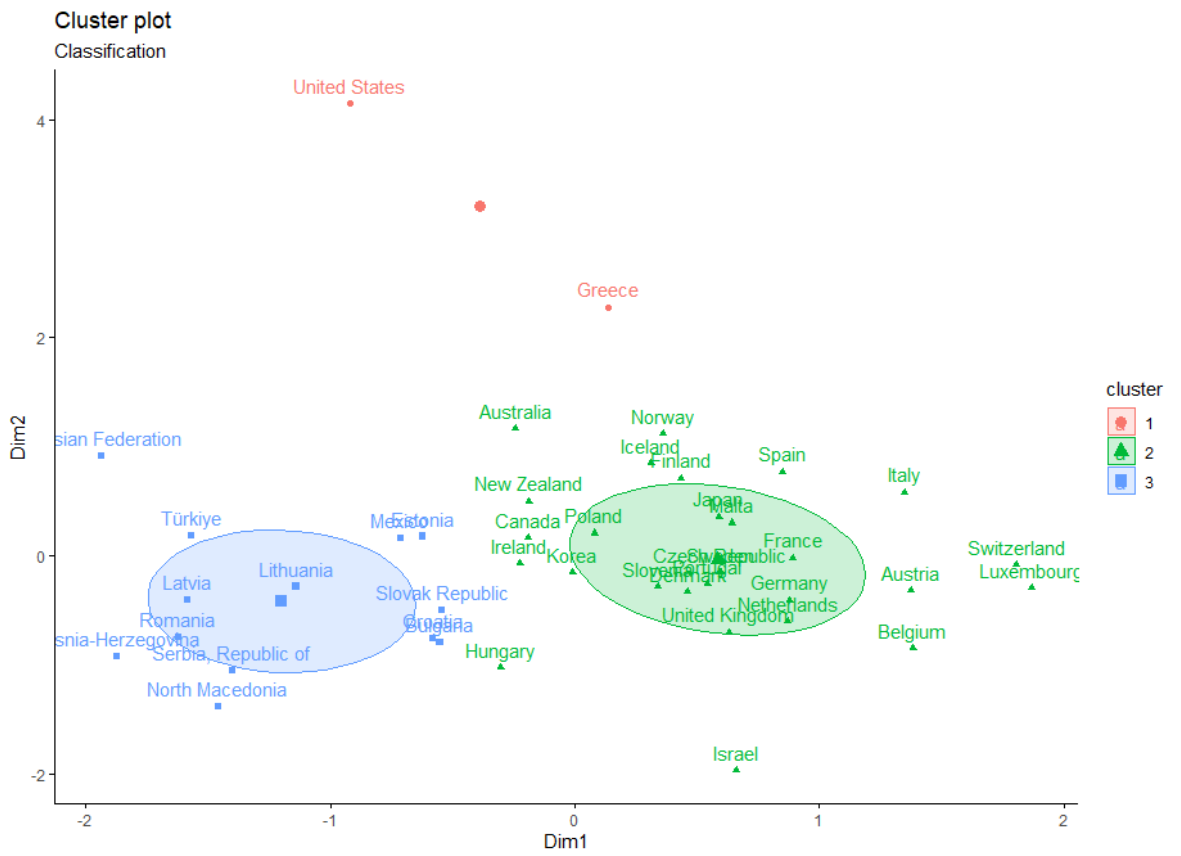

Figure: Cluster Plot

Model Based Clustering

Anova test p-value: 0.00575**

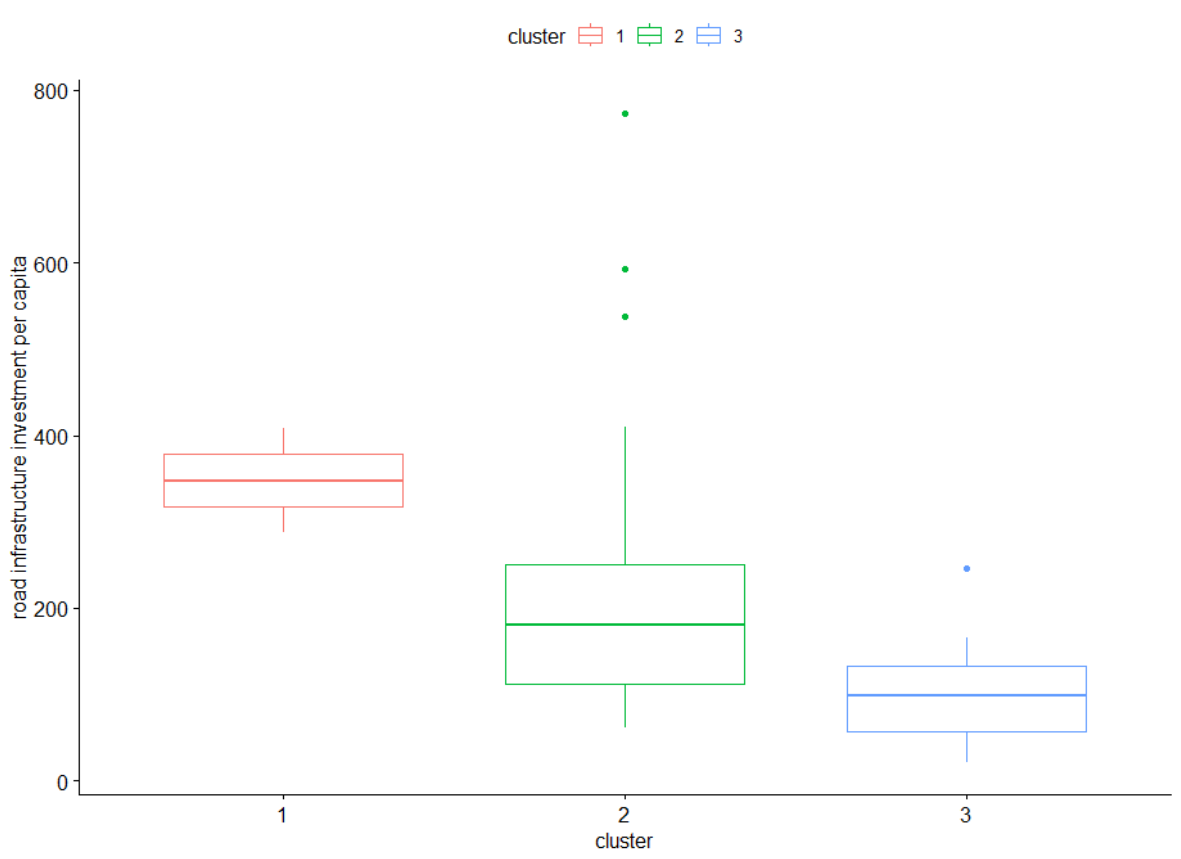

Figure: Box Plot: y against model based clusters

PCA and Clustering

But, what about other clustering methods? and what about PCA?

Dimensionality reduction	Clustering	AOV p-value
MDS (2)	3-Means	0.34
	EEI (3)	0.006**
	3-Medoids	0.54
PCA (9)	4-Means	0.55
	VEI (2)	0.11
	DBSCAN (e = 0.7)	0.37

Note: KMO test score for PCA was 0.37

Table: Clustering and AOV

FEATURE SELECTION & MODELLING

Feature Selection & Modelling

- How to select features? How to model data?
- Weak linear correlation of y and X; strong presence of outliers
- Strategy: visual inspection; simple and robust models (IRLS)

Feature Selection & Modelling

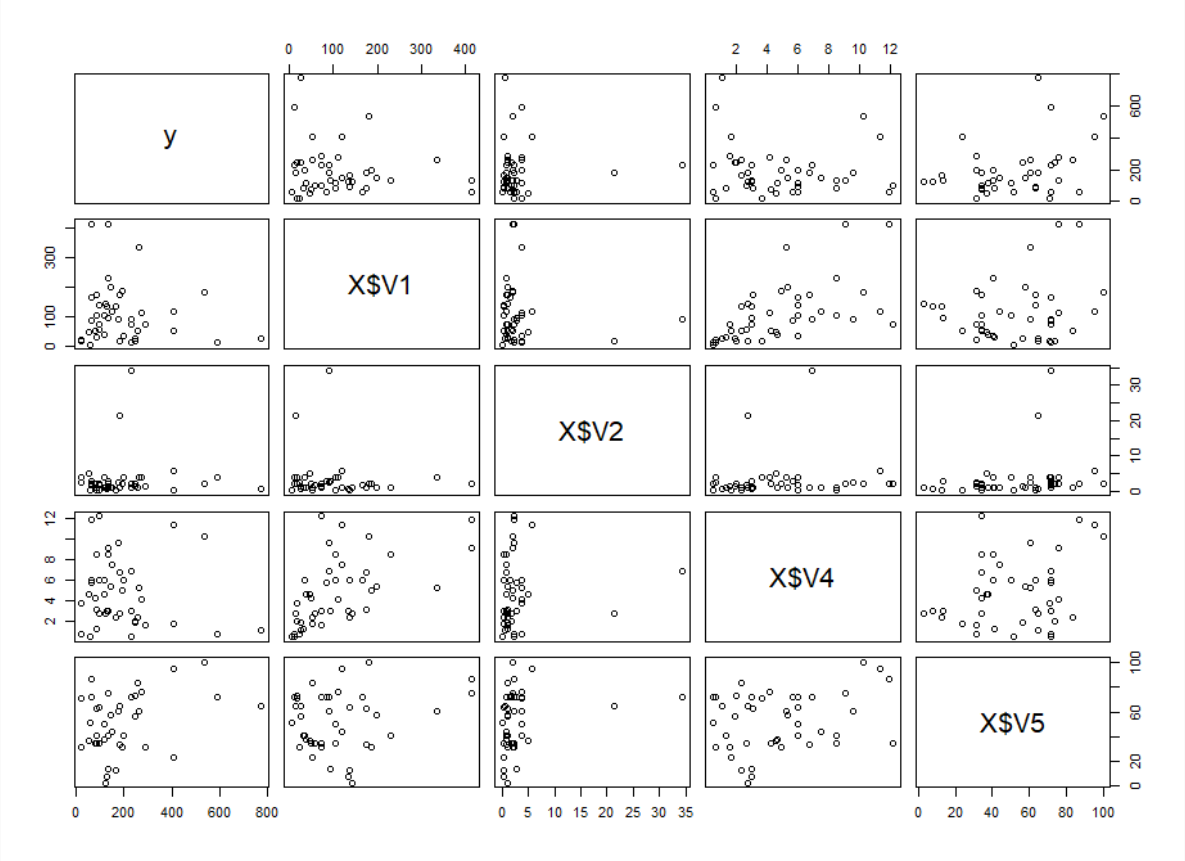

Figure: Visual analysis: y vs first five features of X

Feature Selection & Modelling

Models tested

y = road investment per capita

$$y_i = \beta 1 \cdot (\text{Dim1 MDS}) + \epsilon_i \tag{1}$$

$$y_i = \alpha + \beta 1 \cdot (\text{road motor vehicles}) + \epsilon_i \tag{2}$$

$$y_i = \alpha + \beta 1 \cdot (\text{share of road freight transpor}) + \epsilon_i \tag{3}$$

$$y_i = \frac{1}{\beta 1 \cdot (\text{road fatalities})} + \epsilon_i \tag{4}$$

$$y_i = \alpha + \beta 1 \cdot (\text{CO2 emissions}) + \epsilon_i \tag{5}$$

Feature Selection & Modelling

Weighted sum of squared residuals (WSSR)

$$WSSR_j = \sum_{n=0}^{n} w_i \cdot (y - x_i)^2 \tag{6}$$

R^2

$$R^2 = 1 - \frac{WSSR}{TSS} \tag{7}$$

Table: **Model evaluation**

Model	R^2
m1	0,535
m2	0,547
m3	0,564
m4	0,589
m5	0,713

Feature Selection & Modelling

$$y_i = 76,7 + 37,9 \cdot (\text{CO2 emissions}) + \epsilon_i \tag{8}$$

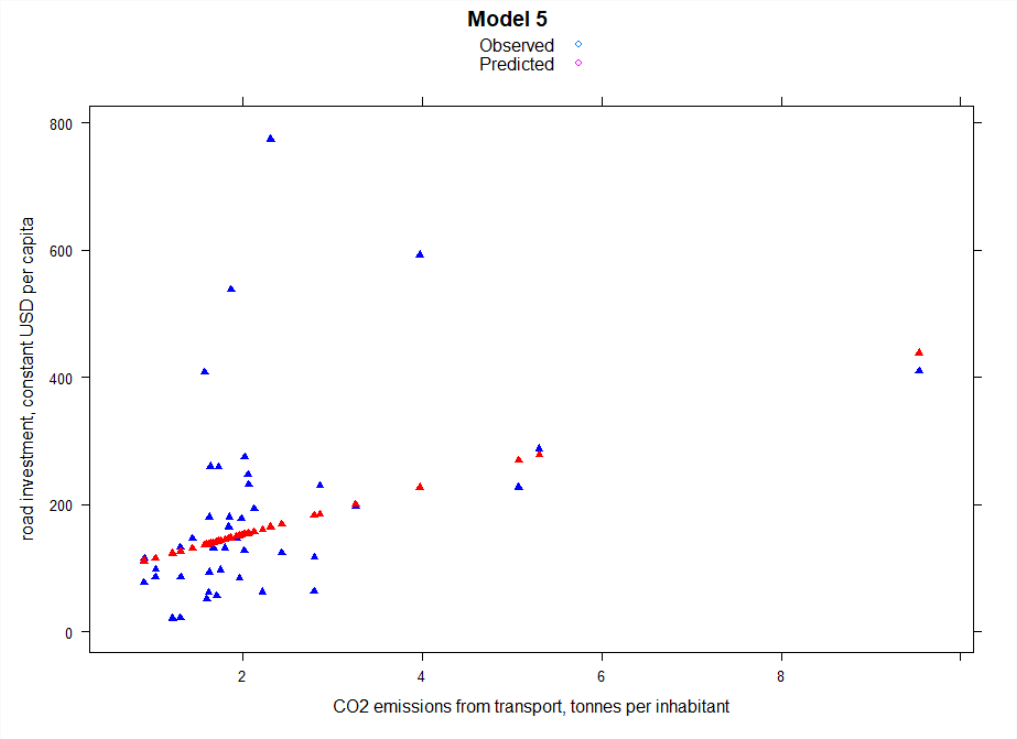

Figure: Model 5, Observed vs Fitted

CONCLUSIONS & FURTHER ANALYSIS

Conclusions & Further Analysis

- Measurements of safety and number of cars allow for clustering countries
- Although linear correlation among features is weak, CO2 emissions resulted being the best linear predictor by using IRLS
- Explore ccorrelation vs causation analysis
- To improve analysis: increase number of datapoints, review data imputation, evaluate log transformations to the data

Thank you