MONASH University

MONASH
BUSINESS
SCHOOL

Introduction and Motivation

High Dimensional Data Analysis

Anastasios Panagiotelis
Lecture 1

Housekeeping

Welcome to HDDA!

- Lecturer for the unit
  - Anastasios Panagiotelis
- Use the moodle forum to ask questions
  - Feel free to answer each others questions.
  - I will also provide answers.
- Email anastasios.panagiotelis@monash.edu only for issues that are personal.
- For details on consultation see Moodle

High-Dimensional Data?

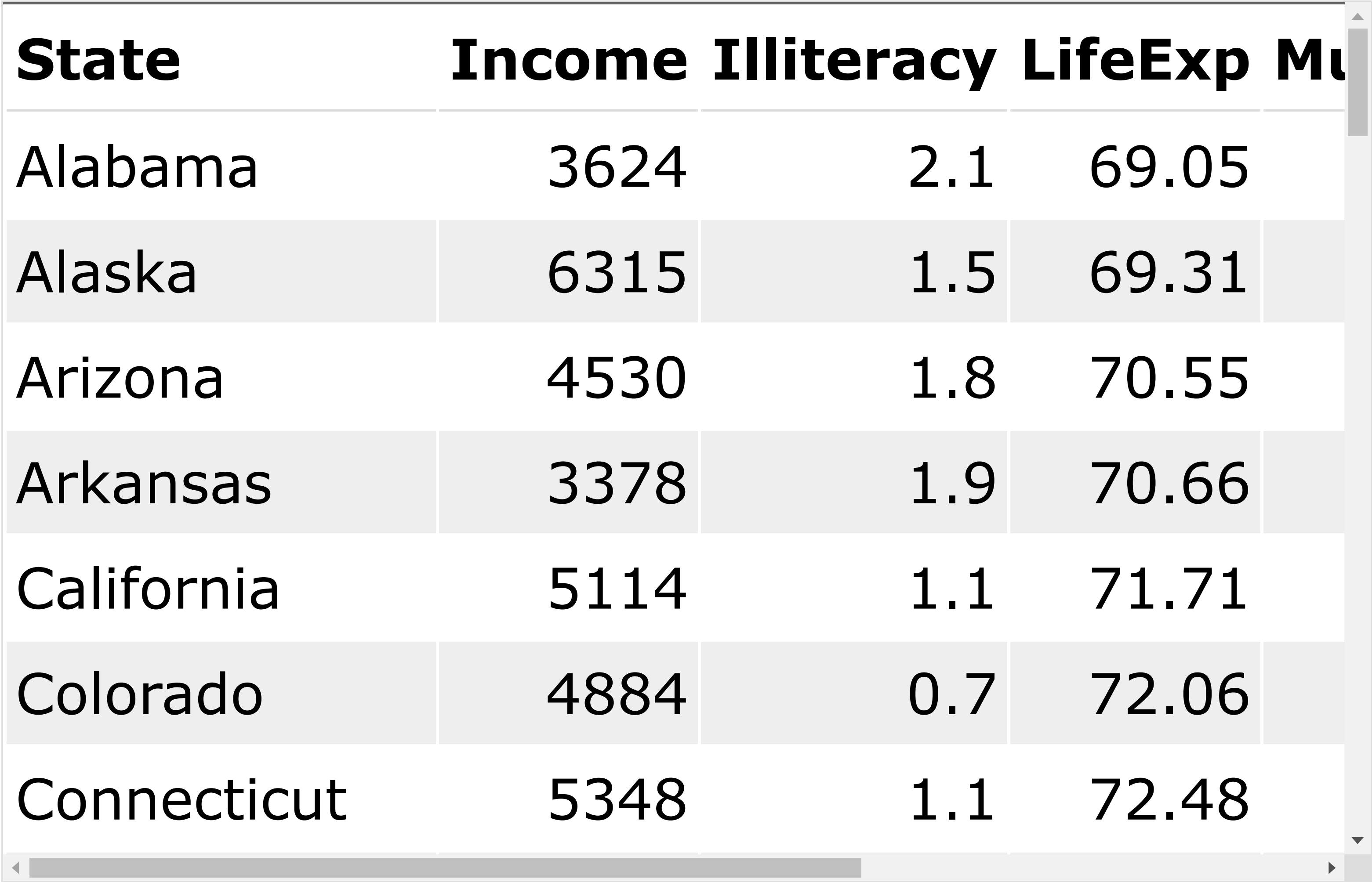
- First what do we mean by *High Dimensional*?
- The data we look at will have:
  - Observations
  - Variables
- Generally *High Dimensional* implies that the number of *variables* is large.
- The term, high-dimensional also relates thinking about and visualising data as points in space.

A Data Story

US States

- Five indicators of the quality of life in the 50 States of the USA in 1977.
  - Income,
  - Illiteracy rate,
  - High school graduation rate,
  - Life expectancy,
  - Murder rates.
- Let's explore!

A dataset



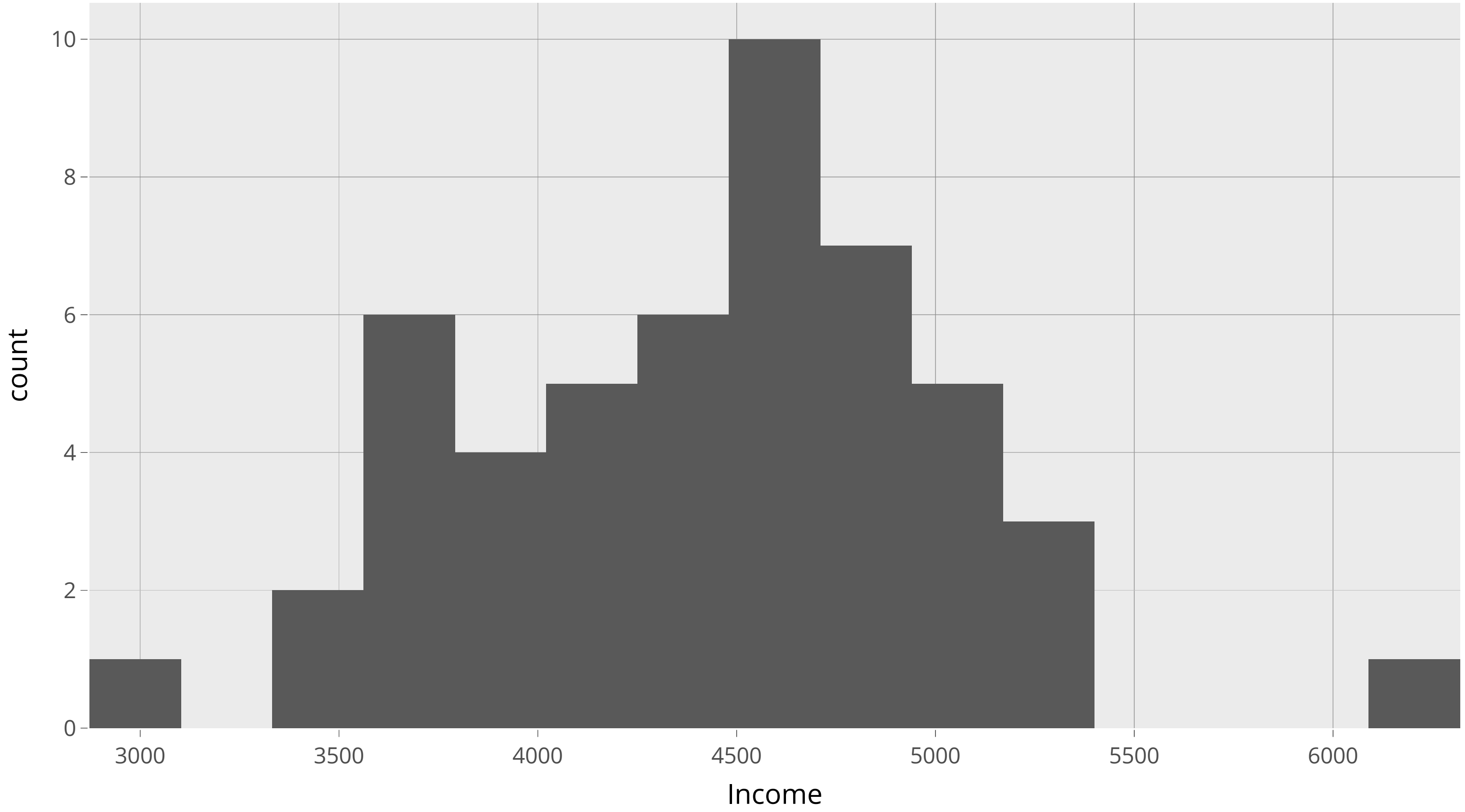
State	Income	Illiteracy	LifeExp	Mu
Alabama	3624	2.1	69.05	
Alaska	6315	1.5	69.31	
Arizona	4530	1.8	70.55	
Arkansas	3378	1.9	70.66	
California	5114	1.1	71.71	
Colorado	4884	0.7	72.06	
Connecticut	5348	1.1	72.48	

Observations and Variables

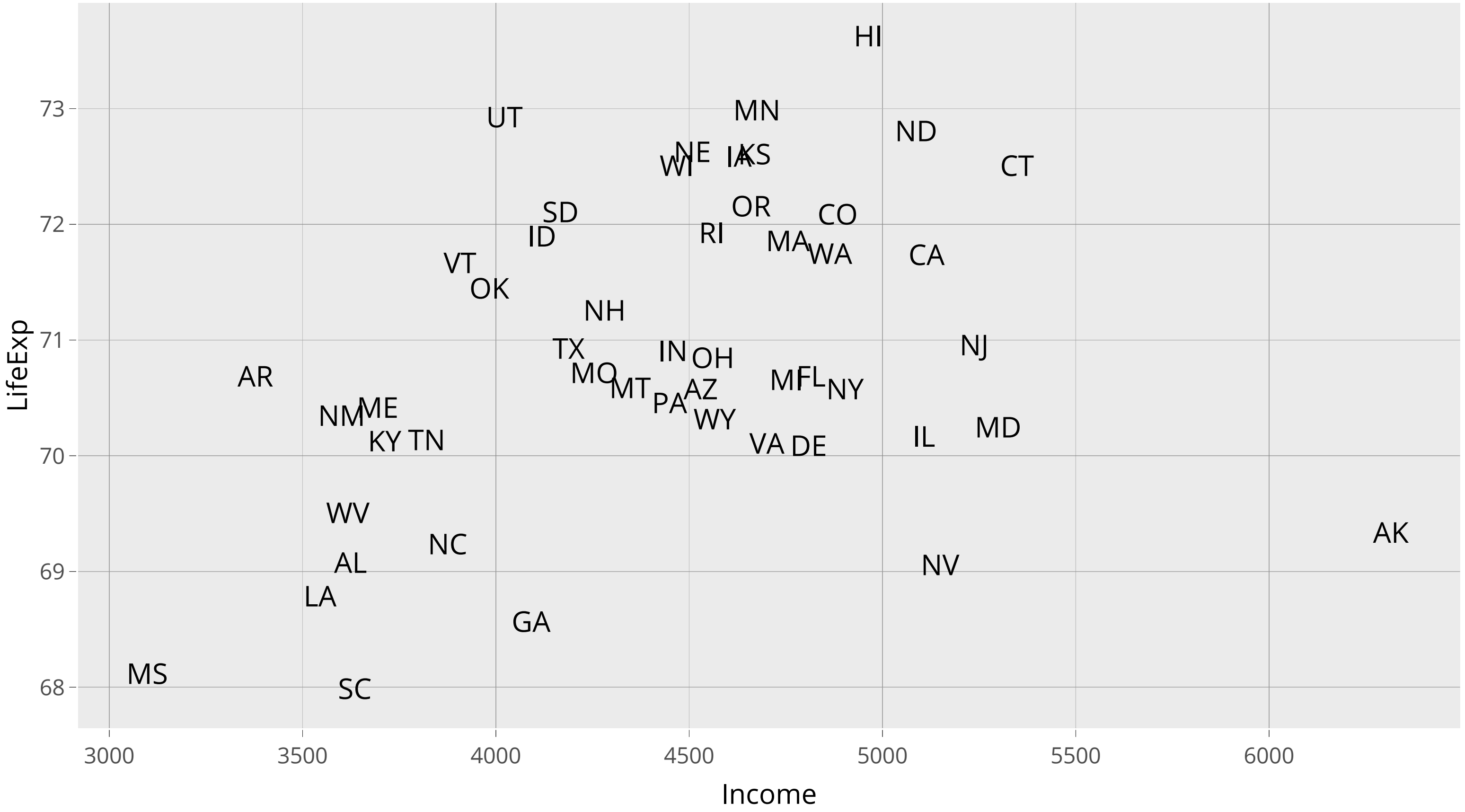
On the previous slide and in general:

- Each row corresponds to an *observation*
  - In this example that is a State.
- Each column corresponds to a *variable*
  - In this example that is an attribute of each State.

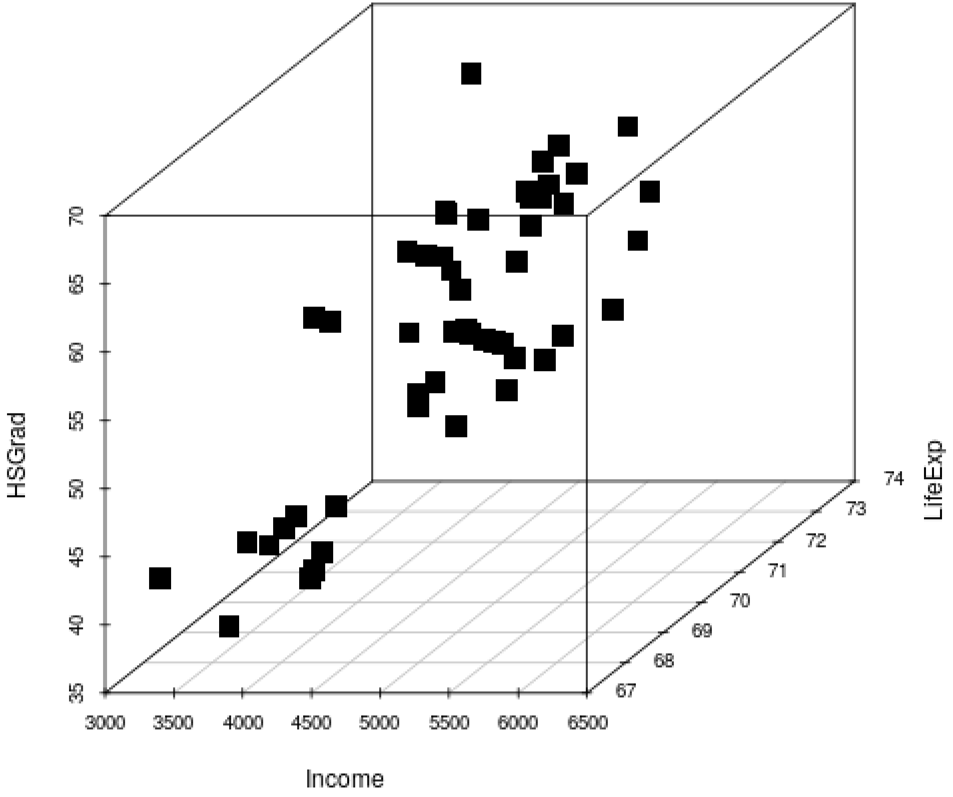
Histogram: Income



Scatter-plot: Income v Mortality



3D Scatter-plot



3D Scatter-plot

Click and drag to rotate

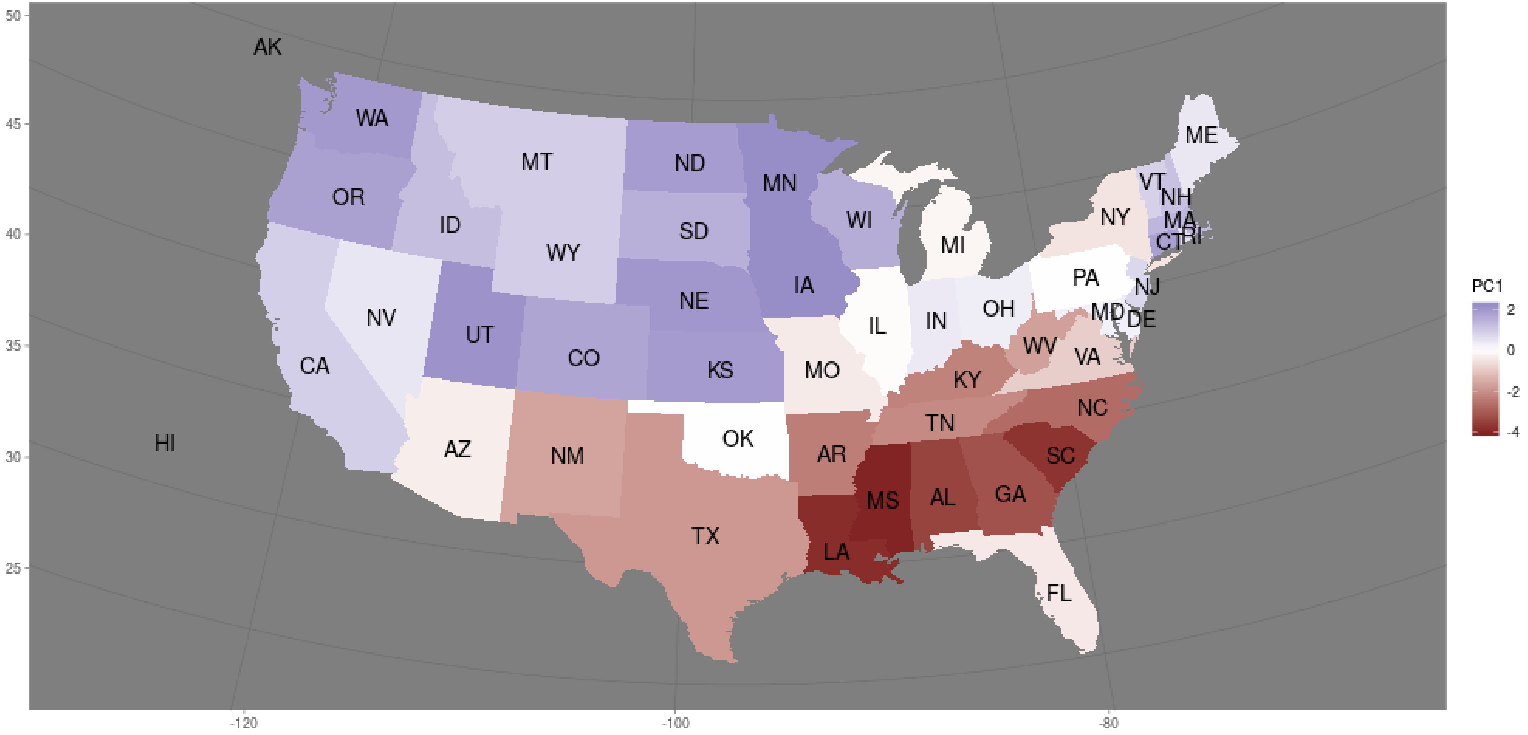
Lessons learnt

- With 2 variables we can do a 2-dimensional (2D) scatter plot.
  - This can be interpreted very easily
- With 3 variables we can do a 3D scatter plot
  - This doesn't look great on a flat screen
  - We get more insight by *rotating* the plot
- What about 5 variables? What about 100 variables?

Principal components

- Later on we will cover the method of *principal components*.
- This can be used to combine the variables into a single index.
- This single index explains most of the variation in the data.
- On the next slide we plot the first principal component on a map of the USA.

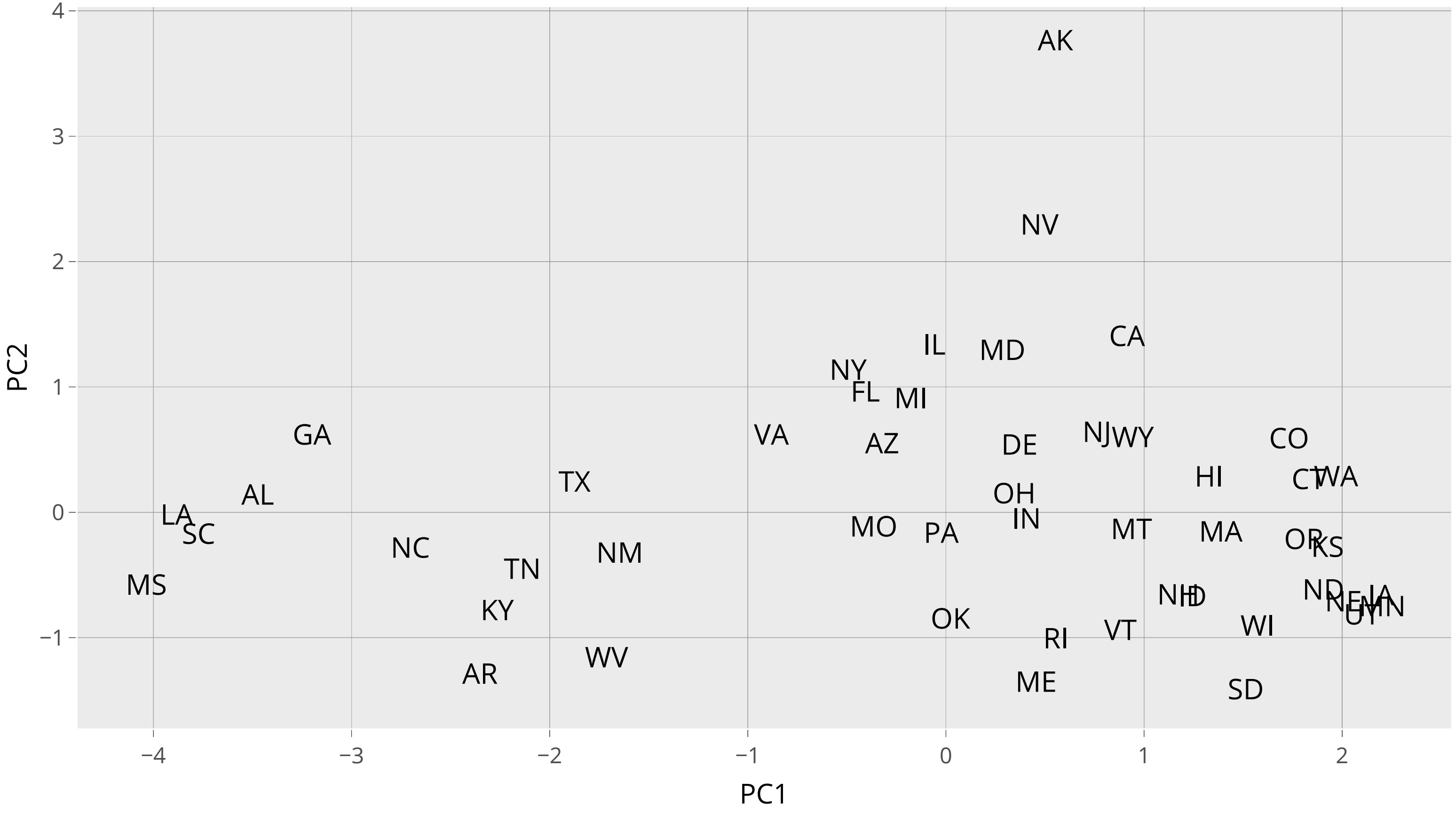
One PC on a map



Multidimensional Scaling

- Two states close to one another on the scatterplot had similar levels of income, and life expectancy.
- Can we do something similar but for all five variables.
- The method of *multidimensional scaling* finds two coordinates so that states close to one another on the scatterplot are close to one another across all five characteristics.

Multidimensional Scaling



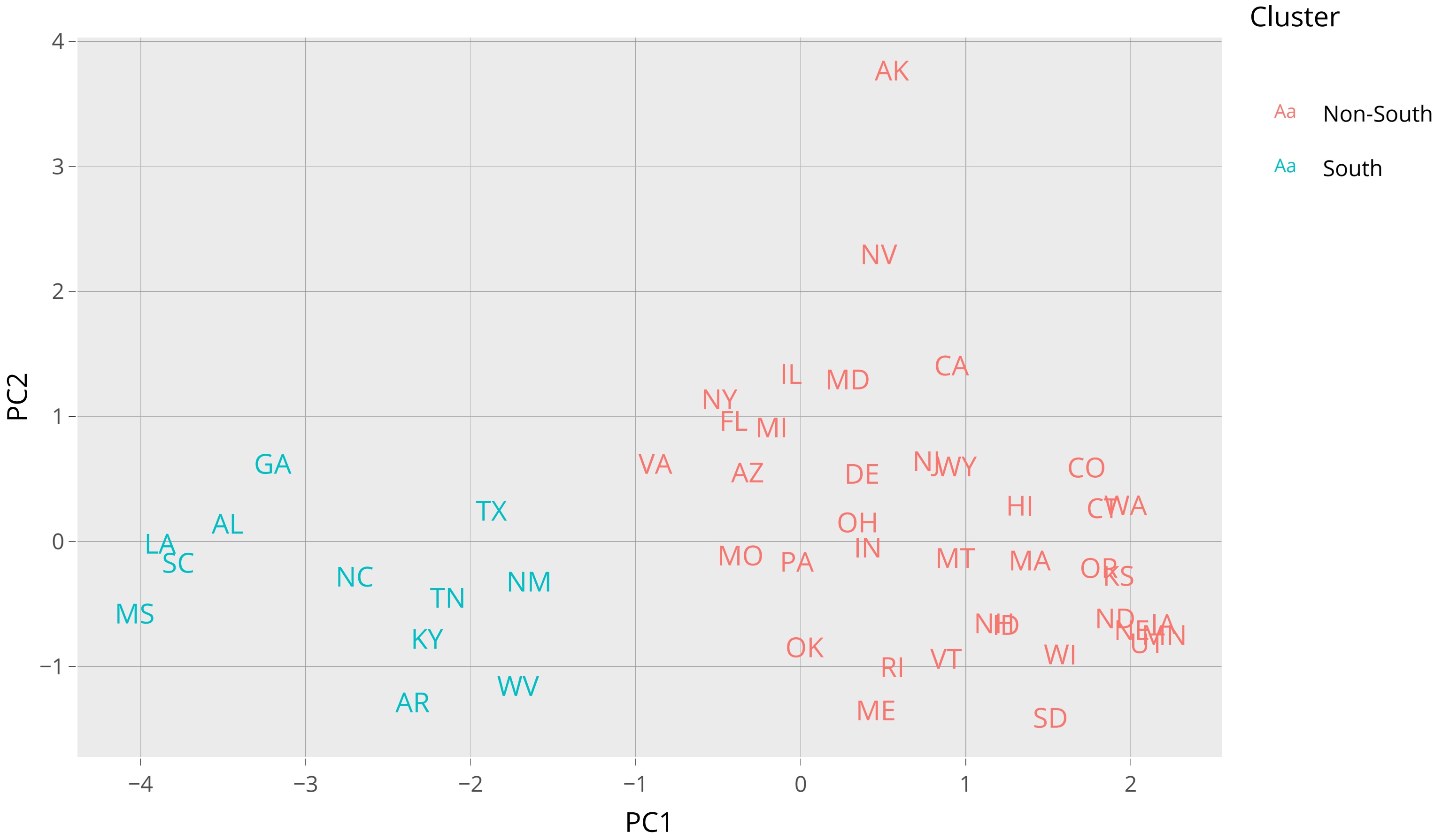
Factor Analysis

- Later on we will attempt to attach possible interpretations to these constructed variables.
- This is the objective of factor modelling.
- In this context *factor* refers to a latent construct that cannot be directly observed but can be measured via its correlation with observable data.

Cluster Analysis

- Even from the simple analysis so far, it appears that similar states can placed into a small number of groups.
- The use of algorithms that achieve this task is known as *cluster analysis*.
- It is extremely useful across a number of business disciplines.
- On the following slide we group the states into two clusters and present them in different colors.

Cluster Analysis: Example



A broad understanding of data.

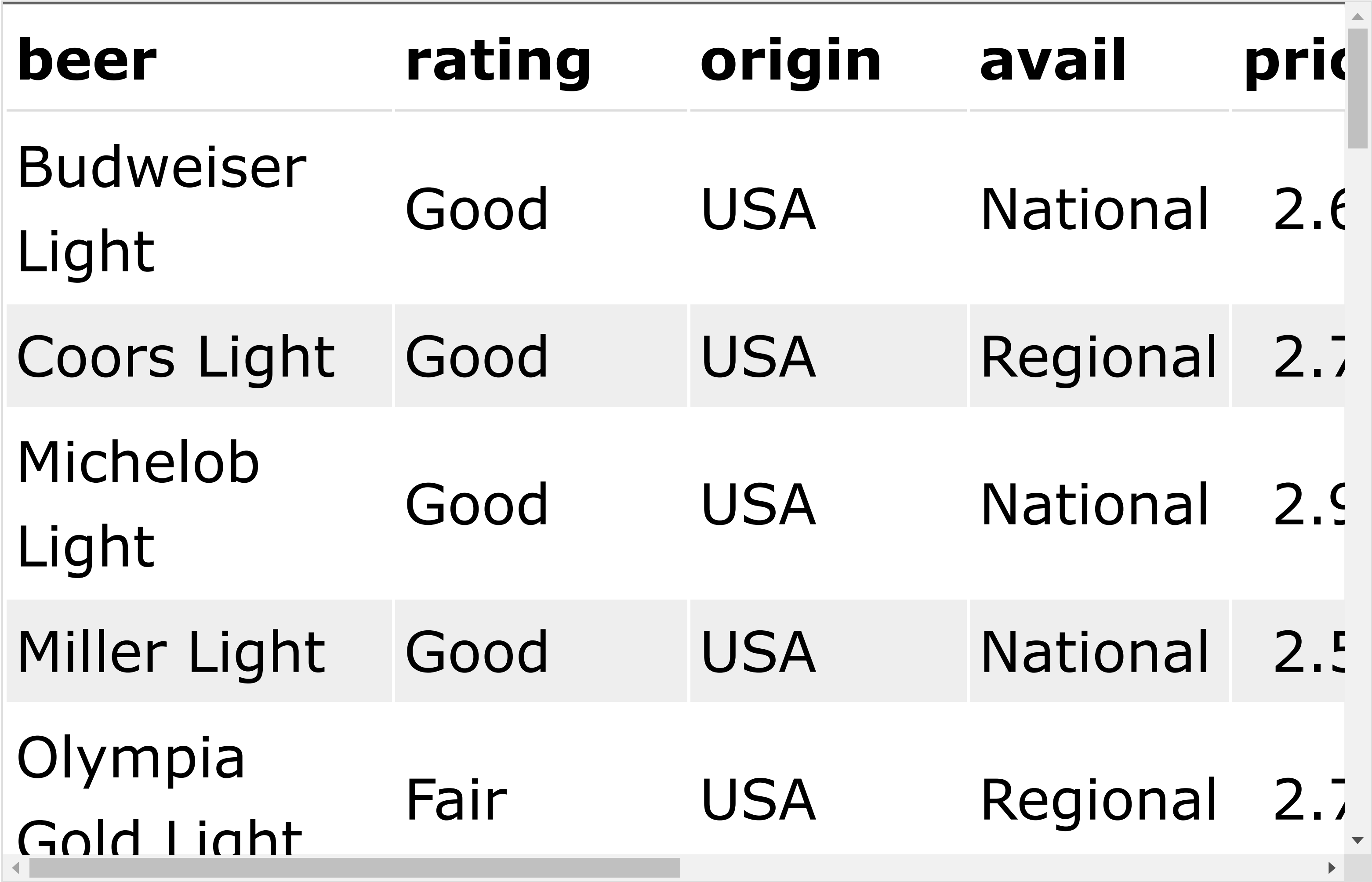
Numerical Data

- So far we looked at *numerical* data
  - This is also called *metric* data or *ratio* data
- The differences and ratios between values of the variable have some meaningful interpretation.
- A state with a mean income of $5000 has twice as much income as a state with a mean income of $2500.

Non-metric data

- *Categorical* (or *nominal*) Data
  - The value of the variable does not measure the *size* of some characteristic.
- *Ordinal* data
  - Different values of the variable measure *more* or *less* of a characteristic but not *how much* more or *how much* less.

Beer Data



beer	rating	origin	avail	pri
Budweiser Light	Good	USA	National	2.6
Coors Light	Good	USA	Regional	2.7
Michelob Light	Good	USA	National	2.9
Miller Light	Good	USA	National	2.5
Olympia Gold Light	Fair	USA	Regional	2.7

Questions for you

- How many variables in the Beer dataset?
- Which are metric?
- Which are nominal?
- Which are ordinal?

Discussion

- Price is an example of a numerical variable.
- Country of Origin is an example of a nominal variable:
  - You can not have more or less *France-ness* or *Mexico-ness*
- Rating is an example of an ordinal variable:
  - A very *good beer* is better than a *good* beer but we do not know how much better.

Cross tab

- A useful tool for exploring non-metric variables is the cross tab.
- Cross tabs that are small can be very useful in providing some indication of the relationships between categorical variables.
- Since most Beers in our dataset are from the US, the following cross tab only looks at US beers against beers from all other countries combined.

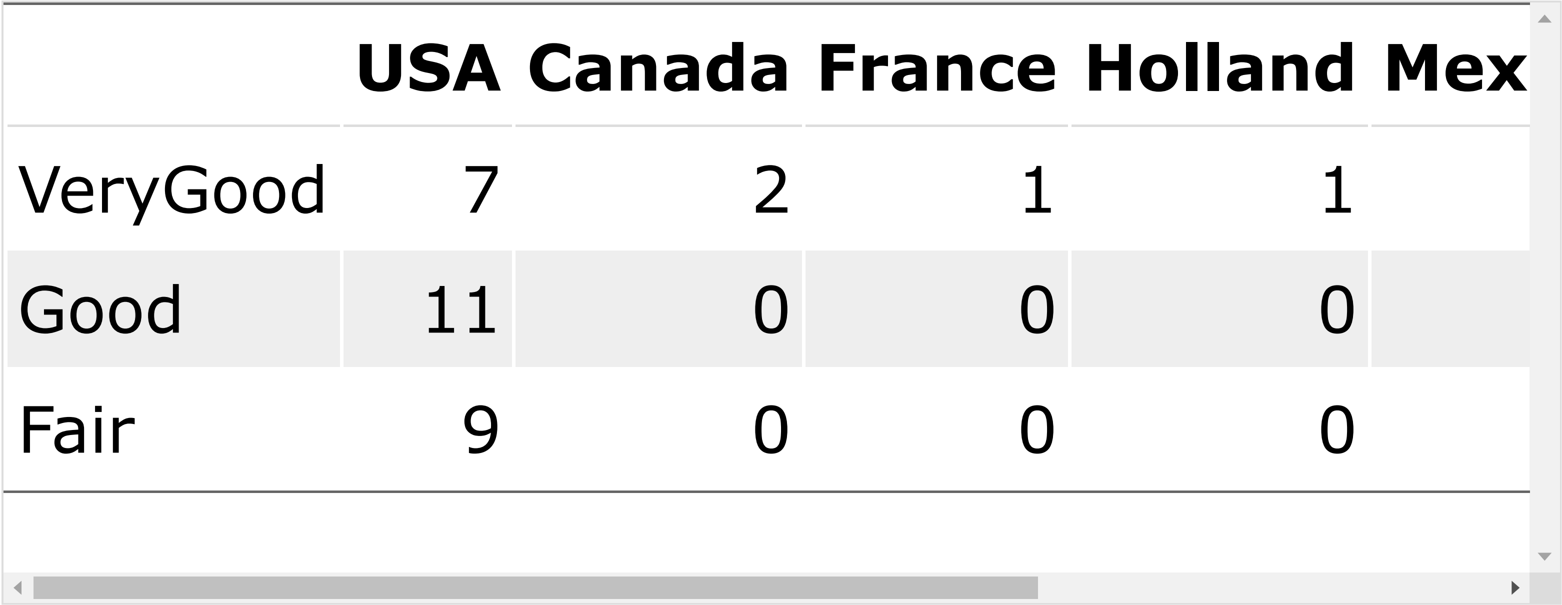
Cross Tab: Rating v Origin

International v US

	Int.	**US**
VeryGood	4	7
Good	3	11
Fair	1	9

Is there a relationship between origin and rating?

Using all countries



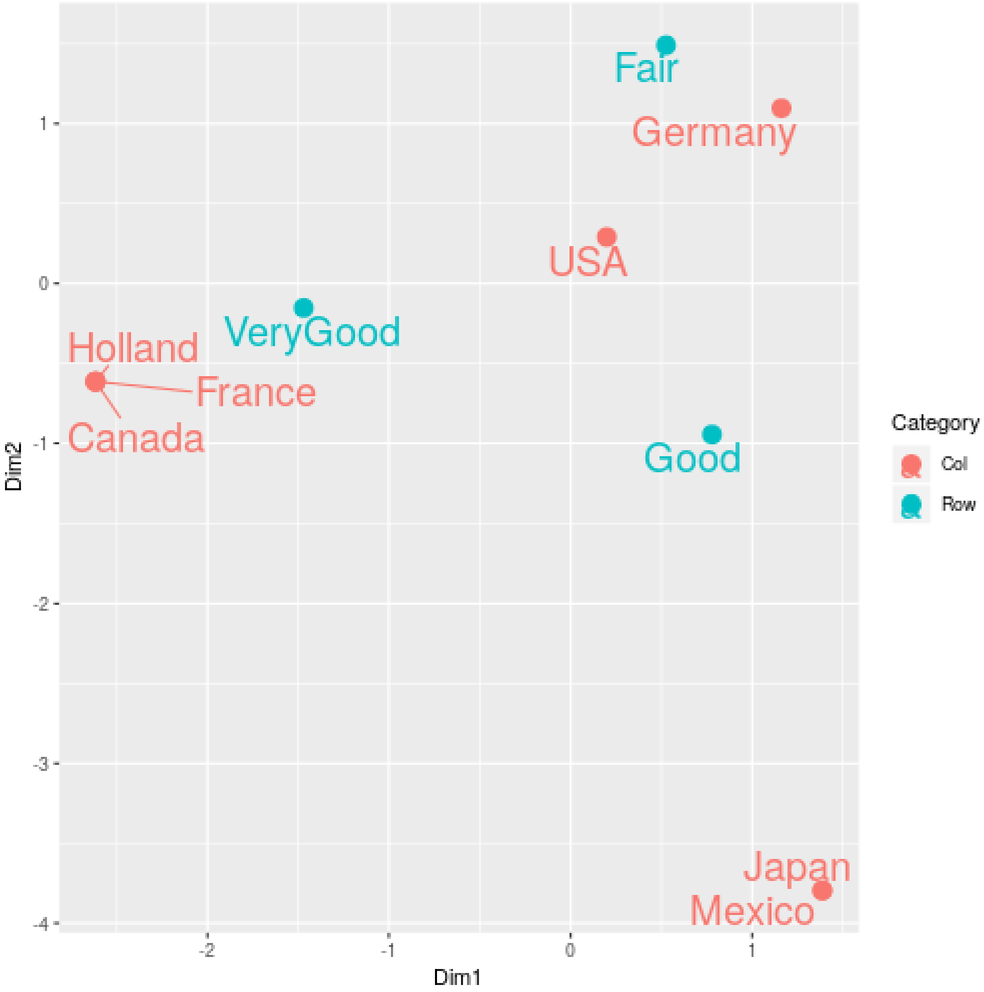
	USA	Canada	France	Holland	Mex
VeryGood	7	2	1	1	
Good	11	0	0	0	
Fair	9	0	0	0	

Is it as easy to find a relationship now?

Correspondence Analysis

- Large cross tabulations can be summarised and visualised with a technique known as *Correspondence Analysis*.
- This technique is mostly used to visualise the relationship between two variables.
- The problem is considered *high-dimensional* since the number of categories rather than the number of variables is large.
- On the next slide is the output from correspondence analysis

Correspondence Analysis



Other data

- Data comes in even more unusual forms.
  - The list of your favourite musicians on Spotify
  - The words used in online reviews of hotels
  - A ranking of pairs of products from most similar to most dissimilar
- All of these types of data can be analysed using methods covered in the unit.

What the unit involves

- The focus is on learning the methods and applying them to real business problems.
- To truly understand the methods requires us to learn about two things that students tend to be afraid of:
  - Programming
  - Maths
- This unit has plenty of both!

Why programming is easy

- Programming in this unit uses the *R language*
- You will be given lots of support in both lectures and tutorials
- Scripts will be provided on Moodle
- The best way to learn R is to:
  - Practice
  - Make mistakes
  - Use the Internet!
- In previous years students have found learning R extremely useful.

Why math is easy

- People who hate math tend to hate memorising formulas
- So do I…
- The emphasis in this unit is
  - Good intuition
  - Recognising patterns
  - Using your imagination.
- Most of the math we cover has a *geometric* interpretation which is great for visual learners.

How the unit is delivered

- There is no textbook. The lecture notes and tutorials make up the content of the unit.
- Lectures are recorded.
- However, you will be asked to work in the middle of lectures while I walk around answering questions.
- Install R on laptops and use in class.
- Forget to bring a laptop? Share with a friend.
- Tutorials will be a mix of reinforcing concepts from the lecture and using R.

How to succeed in this unit

- If your only motivation is to pass the unit
  - Follow lectures and attend tutorials.
  - Submit assignments on time.
  - Use common sense.
- If your motivation is to get a high distinction
  - Do all the above.
  - Attempt the practice questions between lectures and tutorials.
  - Try to teach yourselves how to do things in R. Treat it like a puzzle.
  - Understand matrix algebra.