



Increasing utility of anonymised data

Improved banding techniques

Jasper Enderman

jasper.enderman@astrazeneca.com



The opinions in this presentation are my own. They do not necessarily reflect the views and policies of AstraZeneca.

Data in this presentation is based on randomly generated data and should be interpreted as such.



Today's agenda

- 1 What is anonymisation
- 2 How do we use banding?
- 3 What can you do to improve banding?
- 4 Considerations when implementing



Anonymised data

- Why anonymise?
- Anonymised data is transformed.
- Two main transformation techniques:
 - Randomisation
 - Generalisation
- Transformation lowers the amount of information available.
- Utility of data



Commonly used banding

Original

Name	Age
Reed	46
Iris	47
Camelia	48
Sorrel	50
Rose	50
Daisy	50
Lily	50

Pseudonymised

Subject	Age
54	46
83	47
57	48
64	50
22	50
2	50
97	50

Anonymised

Subject	Age (k=2)
54	46-48
83	46-48
57	46-48
64	50-50
22	50-50
2	50-50
97	50-50



Overlapping bands

- Retain utility with small amounts of outliers
- Benefits depend on data and selected k .
 - Smaller k values, or less uniform data will find most benefit from overlapping banding

Subject	Age ($k=3$)
1	40
2	50
3	50
4	50
5	50
6	50
7	50



Subject	Age ($k=3$)
1	40-50
2	40-50
3	40-50
4	50-50
5	50-50
6	50-50
7	50-50



Multiple identifying variables

Subject	Sex	Age
1	M	46
2	F	47
3	F	48
4	M	50
5	F	50
6	F	50
7	F	50



Subject	Sex	Age (k=2)
1	M	46-48
2	F	46-48
3	F	46-48
4	M	50-50
5	F	50-50
6	F	50-50
7	F	50-50

Subject	Sex	Age (k=2)
1	M	46-50
2	F	46-50
3	F	46-50
4	M	46-50
5	F	46-50
6	F	46-50
7	F	46-50



Multiple identifying variables

Subject	Sex	Age
1	M	46
2	F	47
3	F	48
4	M	50
5	F	50
6	F	50
7	F	50



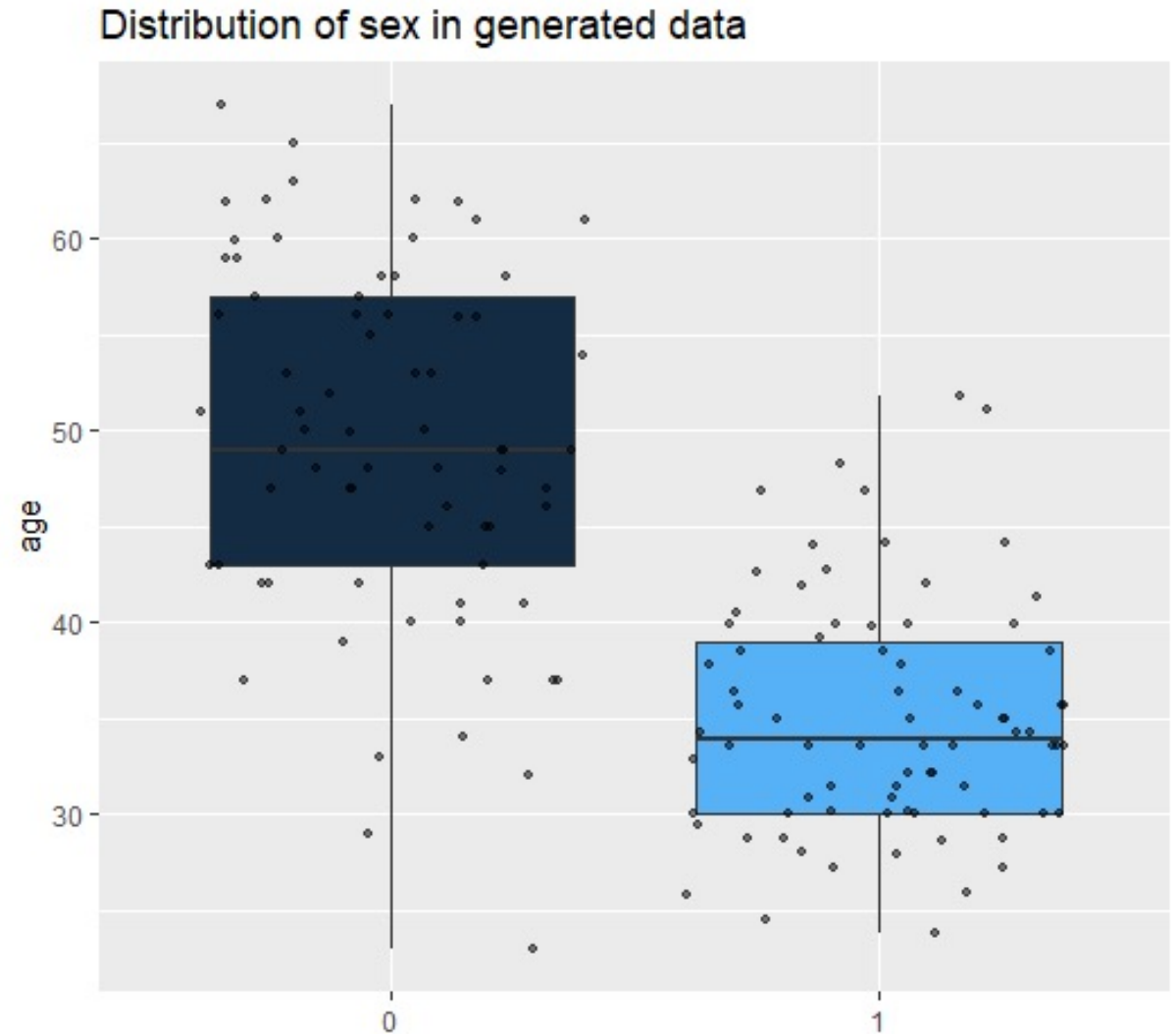
Subject	Sex	Age (k=2)
1	M	46-50
2	F	46-50
3	F	46-50
4	M	46-50
5	F	46-50
6	F	46-50
7	F	46-50

Subject	Sex	Age (k=2)
1	M	46-50
2	F	47-48
3	F	47-48
4	M	46-50
5	F	50-50
6	F	50-50
7	F	50-50

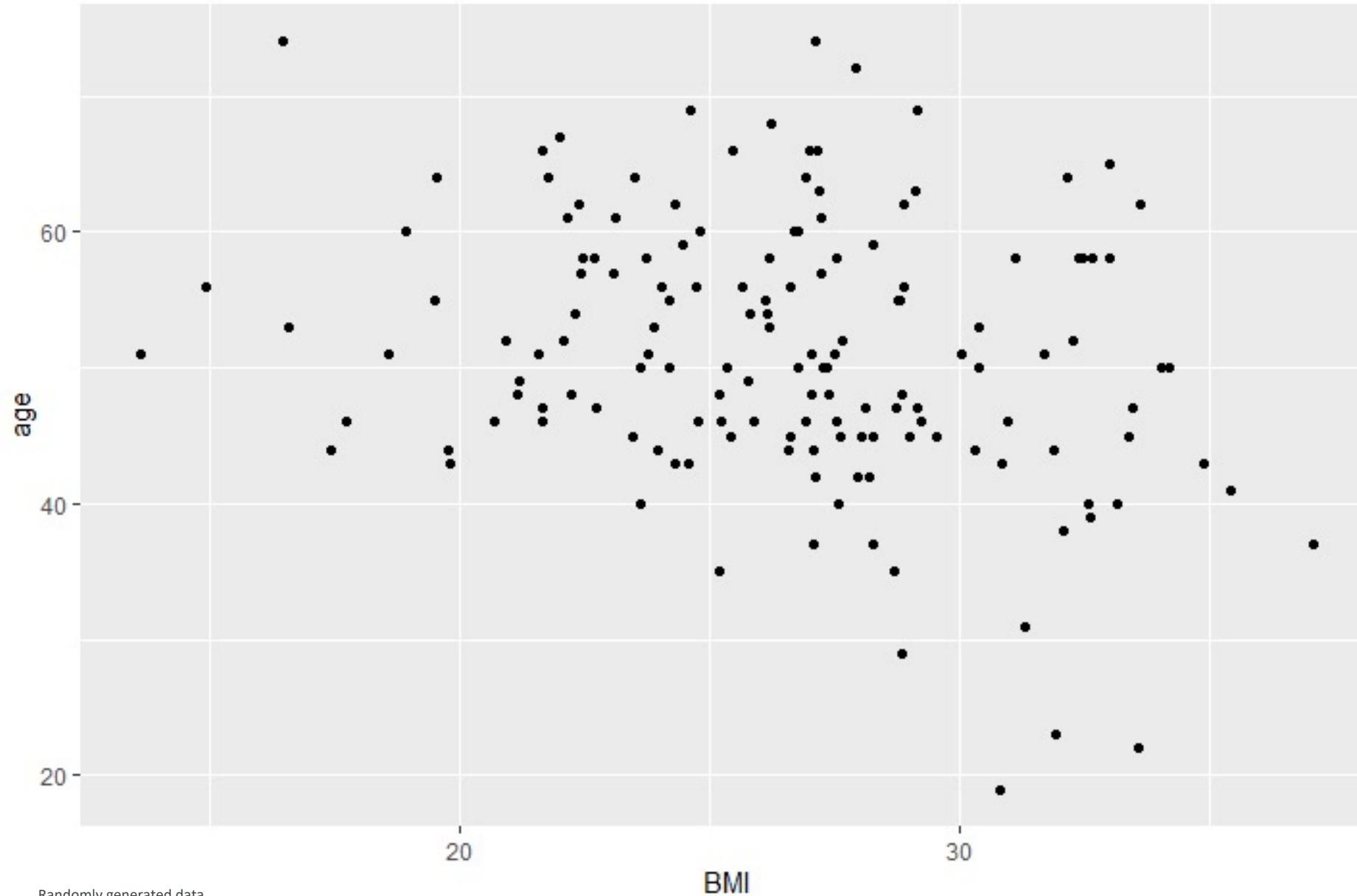


Multiple identifying variables

- Interactions between variables can cause non-homogeneity within the data or subpopulations
- Stratification before anonymization can help in these cases to increase data utility
- This is especially useful when stratifying by a variable such as sex

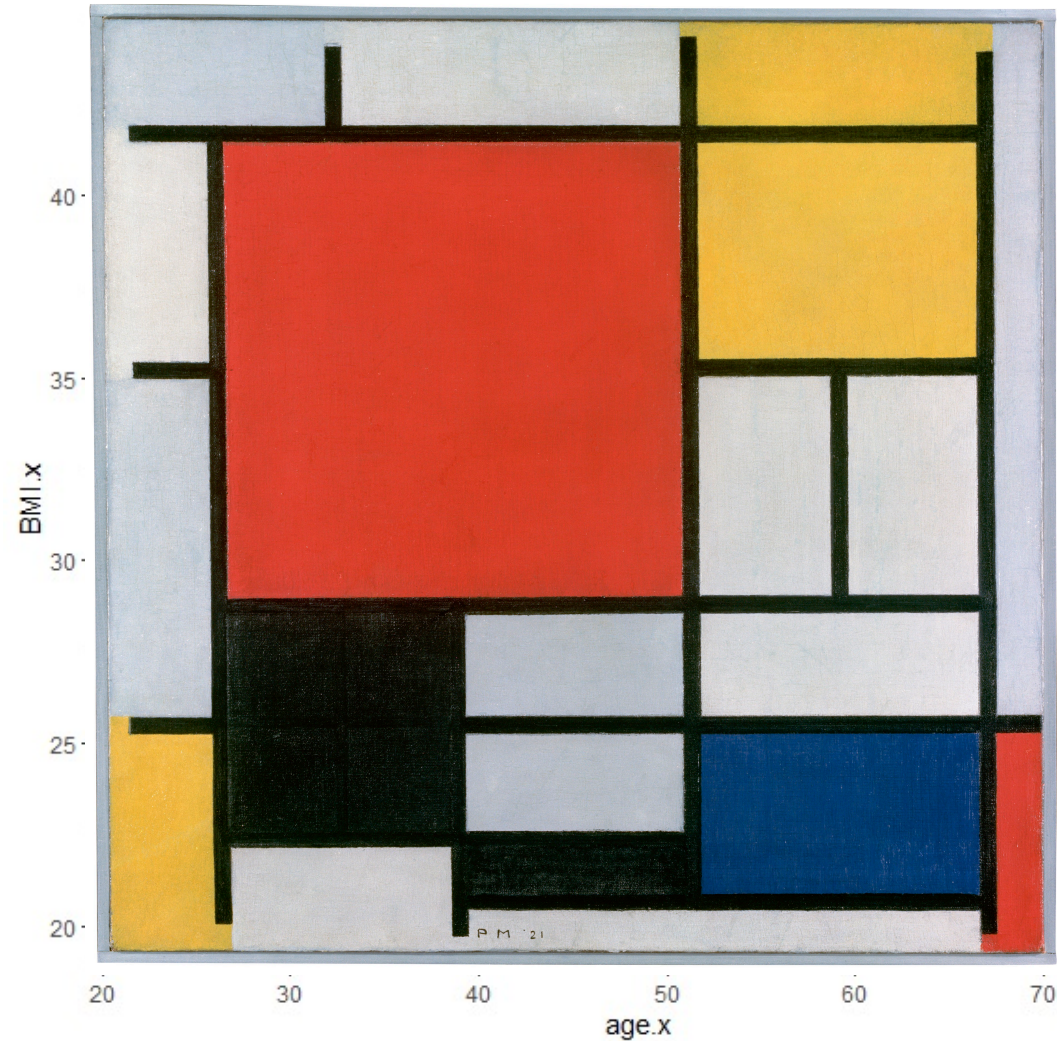


Non-categorical variables

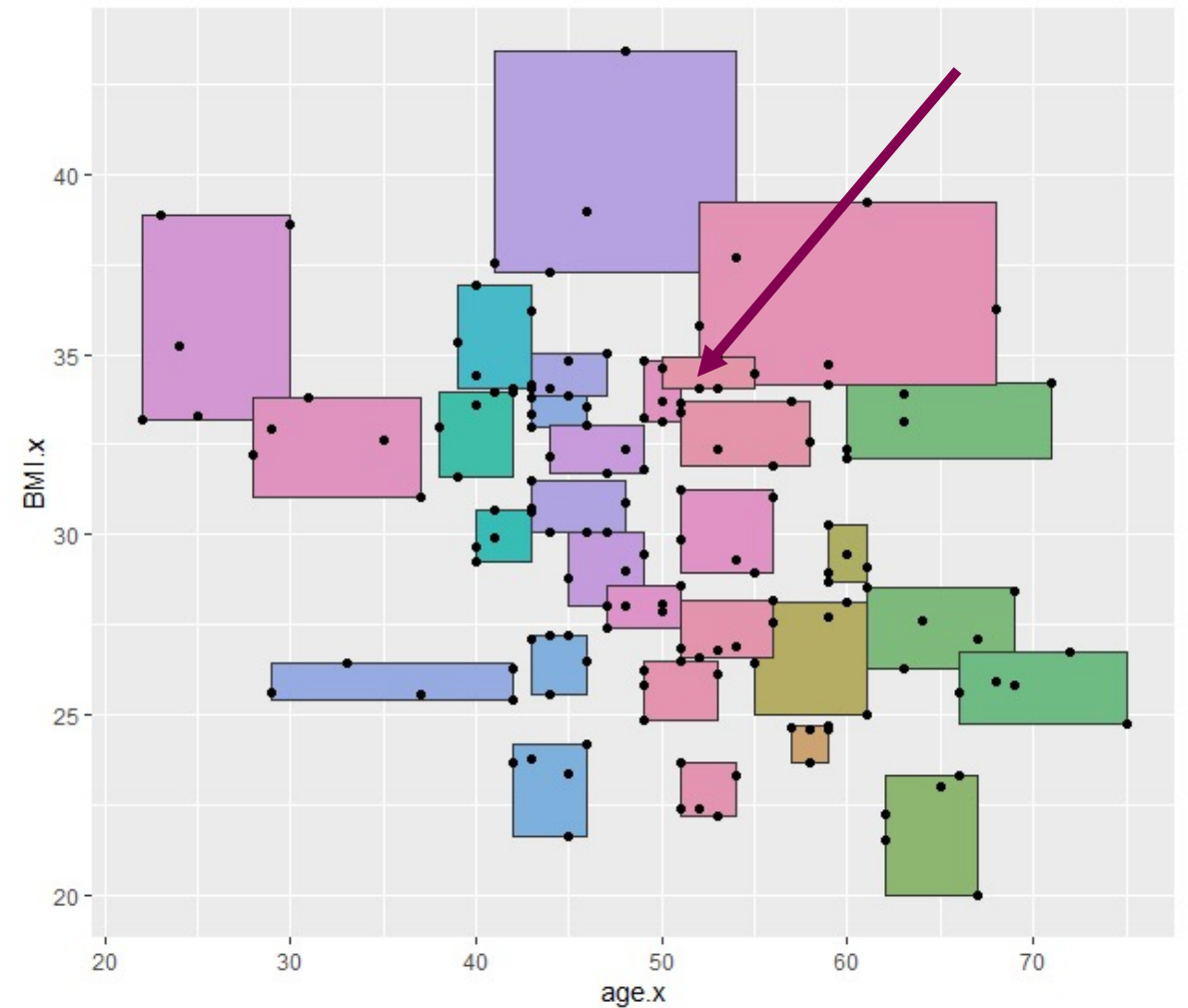


The multi in multi-dimensional

Partitioning



Clustering



Considerations

- Requesters of the data might not prefer this if they have a standard workflow.
- Clinically significant cut-offs might be more useful
- Many different techniques available, not all might be suitable for your specific data
- When using “random/mean within band” using these banding techniques might be preferable.



Recommendations



Thank you

- Please reach out with comments, discussion and questions to:
jasper.enderman@astrazeneca.com

