A CLOUD GURU

Numeric Feature Engineering

Brock Tubre
INSTRUCTOR

Numeric Feature Engineering

Numeric Feature Engineering

Transforming numeric values within our data so Machine Learning algorithms can better analyze them.

Changing numeric values in our datasets so they are easier to work with.

Common Techniques

- **Feature Scaling**

  Changes numeric values so all values are on the same scale.

  - Normalization
  - Standardization

- **Binning**

  Changes numeric values into groups or buckets of similar values.

  - Quantile Binning aims to assign the same number of features to each bin.

Scaling

scaling = feature scaling = normalization

Problem

Will a home loan be approved?

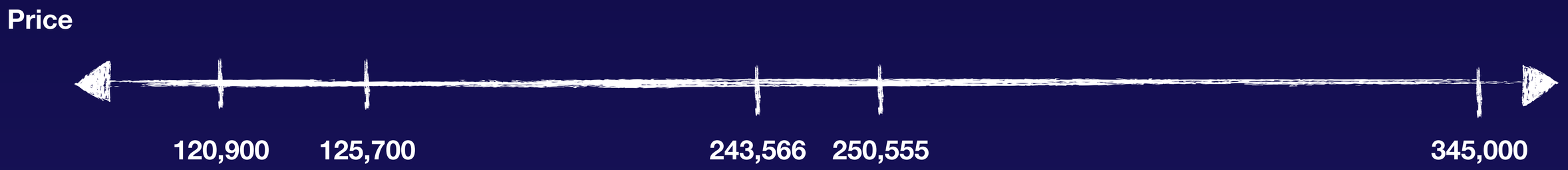
ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	N
2	house	4	2988	L	243566	Y
3	house	3	1877	N	125700	N
4	condo	5	3876	M	345000	Y
5	apartment	2	1250	N	120900	Y

Problem

Will a home loan be approved?

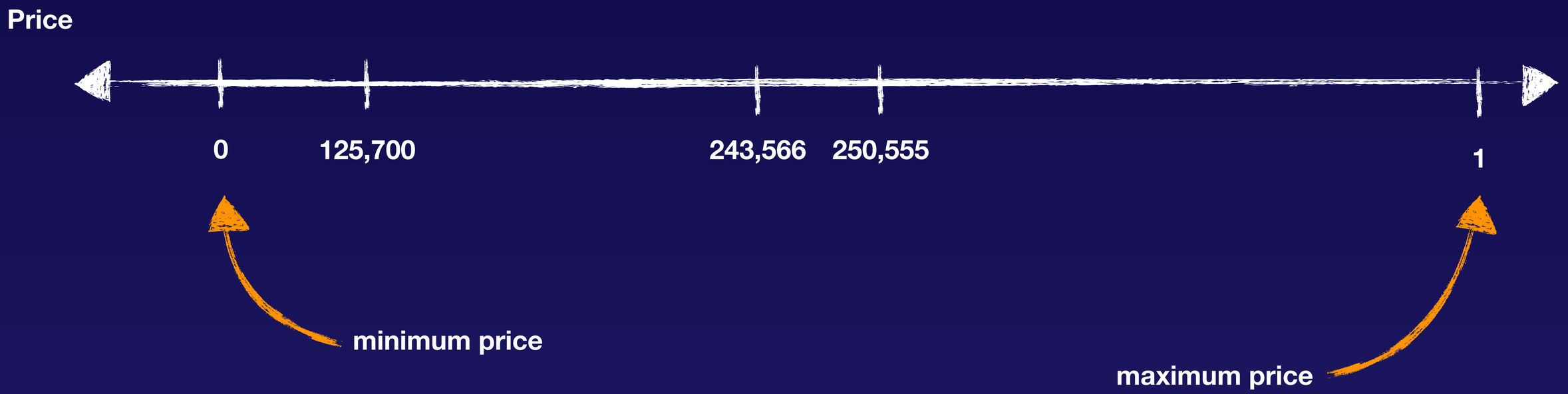
ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	N
2	house	4	2988	L	243566	Y
3	house	3	1877	N	125700	N
4	condo	5	3876	M	345000	Y
5	apartment	2	1250	N	120900	Y

Normalization

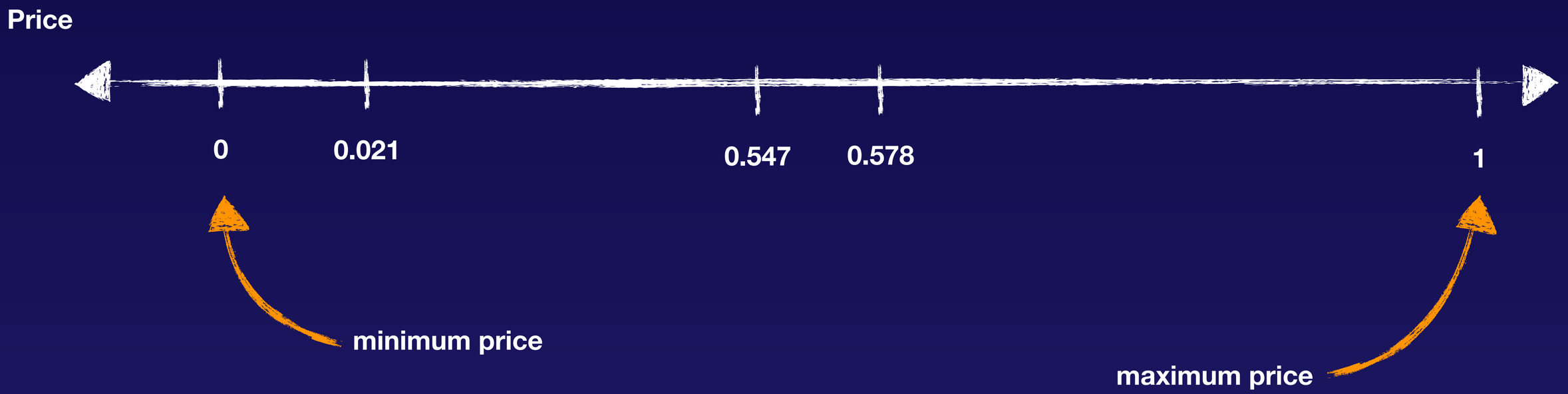


Normalization

Normalization



Normalization



Normalization

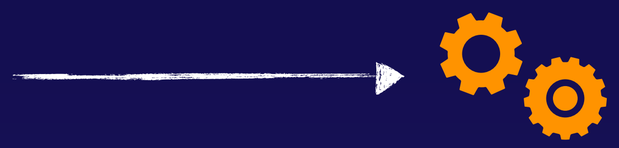
Price
250555
243566
125700
345000
120900

Normalization

Price_Scaled
0.578559
0.547372
0.021419
1.0000
0.0000

Normalization

Price
250555
243566
125700
345000
120900



Normalization

Price_Scaled
0.578559
0.547372
0.021419
1.0000
0.0000

Normalization

Price
250555
243566
125700
345000
120900

x

$max(x)$

$min(x)$

Normalization

Price_Scaled
0.578559
0.547372
0.021419
1.0000
0.0000

x'

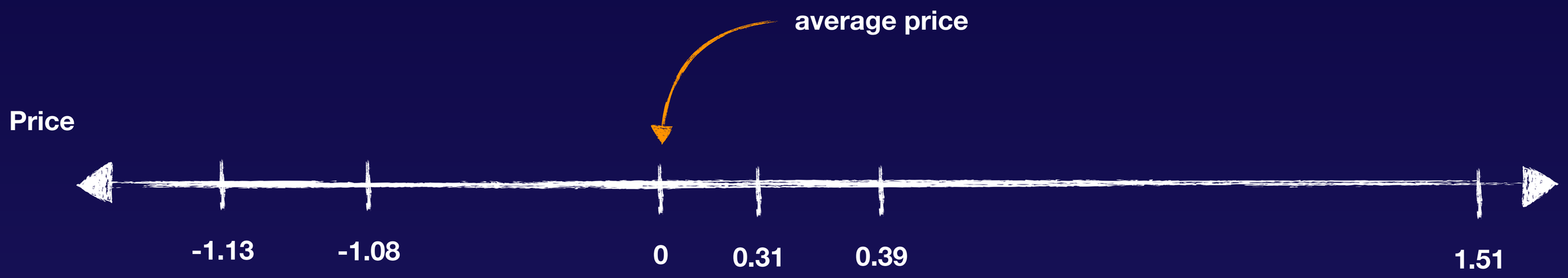
$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

$$\frac{125700 - 120900}{345000 - 120900} = 0.021419$$

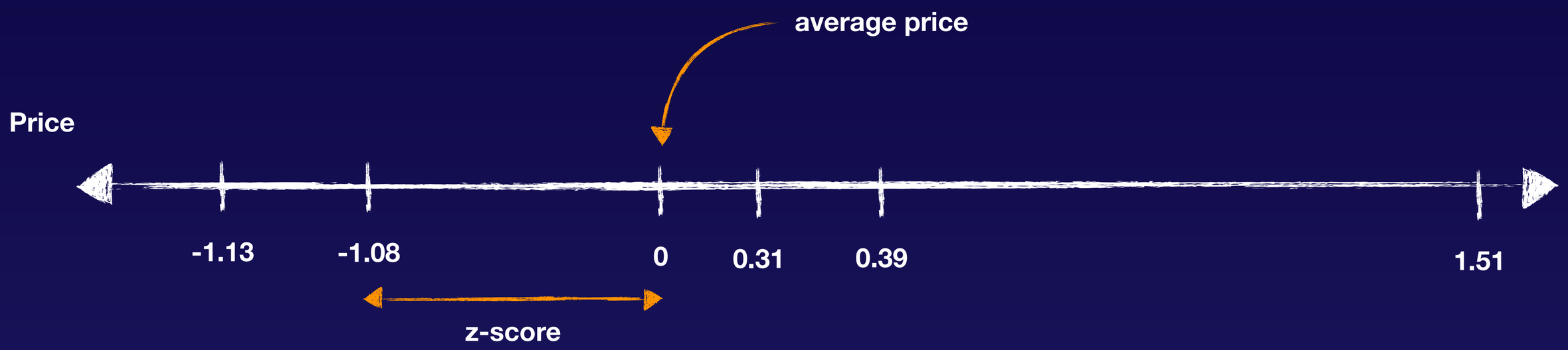
Normalization

Outliers can throw off normalization!

Standardization



Standardization



Standardization

Price
250555
243566
125700
345000
120900

Standardization

Price_Scaled
0.394923
0.312311
-1.080890
1.511283
-1.137627

Standardization

Price
250555
243566
125700
345000
120900

x

Standardization

$\bar{x}$ = **Price mean: 217144.2**

σ = **Price std: 84600.81**

Price_Scaled
0.394923
0.312311
-1.080890
1.511283
-1.137627

z

$$z = \frac{x - \bar{x}}{\sigma}$$

$$\frac{125700 - 217144.2}{84600.81} = -1.080890$$

Scaling Summary

1 Scaling Features

Is required for many algorithms like linear/non-linear regression, clustering, neural networks, and more. Scaling features depends on the algorithm you use.

2 Normalization

Rescales values from 0 to 1 but doesn't handle outliers very well.

3 Standardization

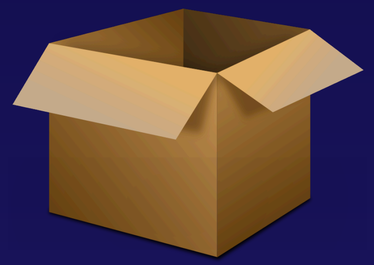
Rescales values by making the values of each feature in the data have zero mean and is much less affected by outliers.

4 Translating Back

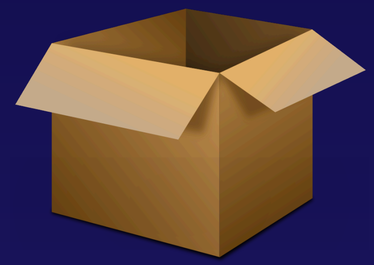
When you're done you can always scale back the data to the original representation.

Binning

Binning



30s and below

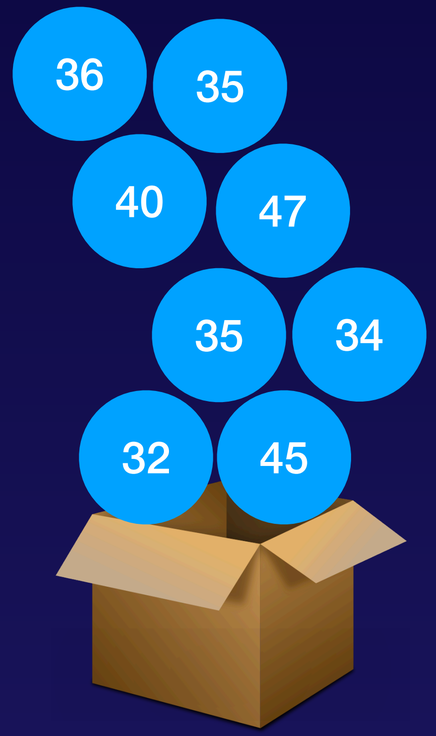


30s-50s

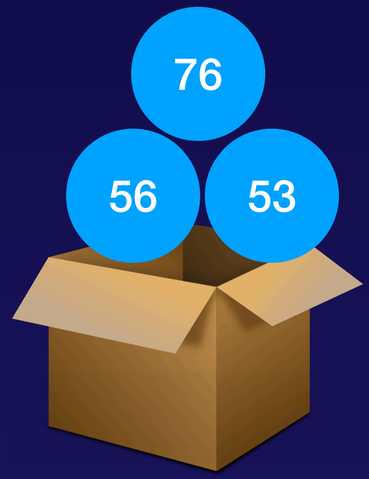
50 and above

Binning

30s and below



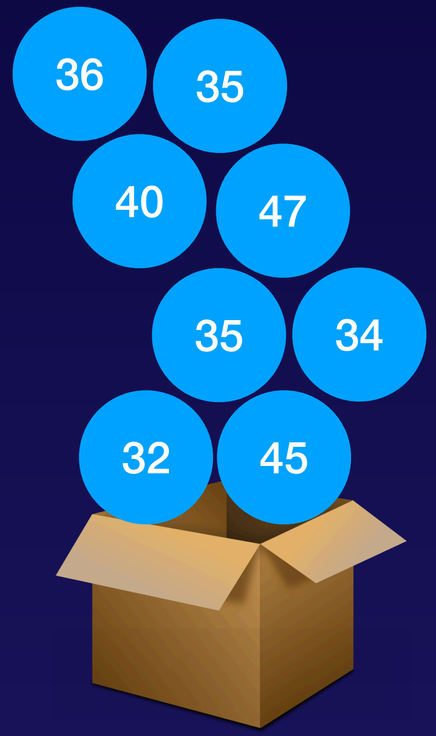
30s-50s



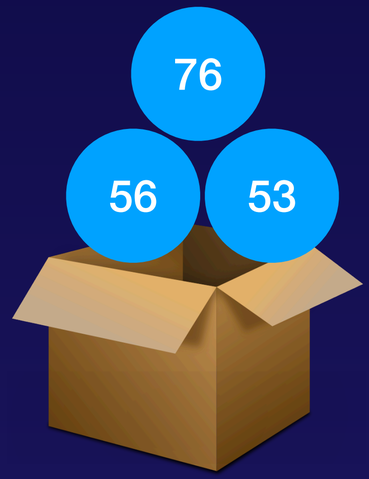
50 and above

Binning

30s and below



30s-50s



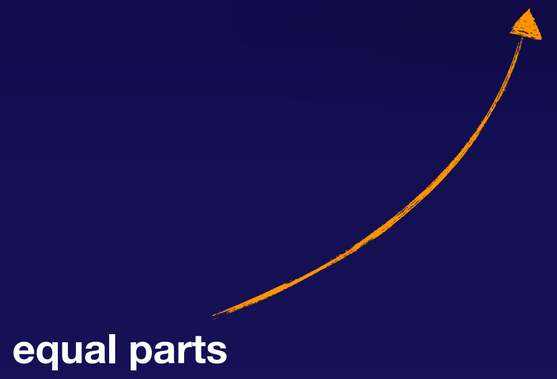
50 and above

We can end up with irregular bins which are not uniform…

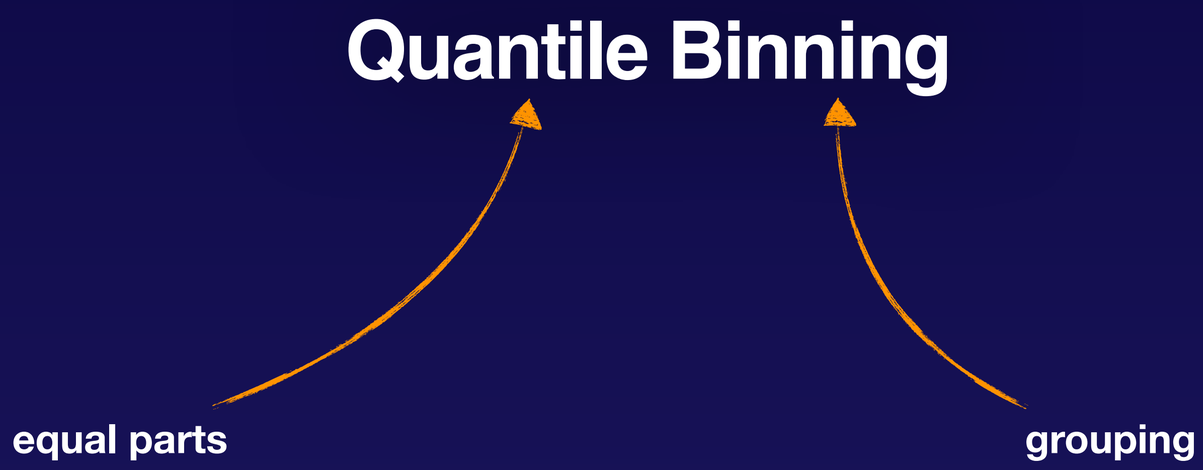
Quantile Binning

Quantile Binning

Quantile Binning



Quantile Binning



Quantile Binning

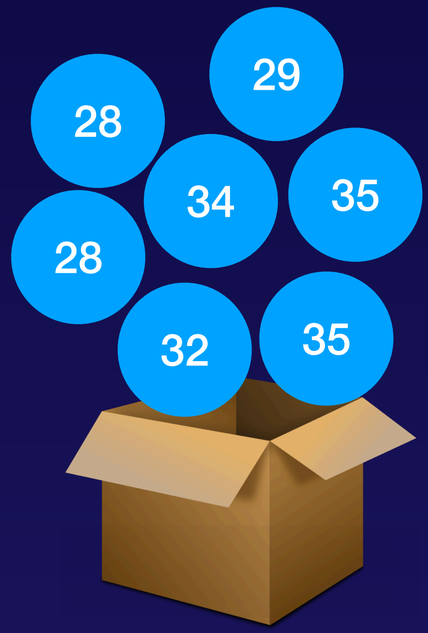
25 and below

26-35

36 and above

Quantile Binning

Youth



Young Adult

Adult

Quantile Binning

age
24
45
18
32
40
76

age_bin
Youth
Young Adult
Youth
Young Adult
Adult
Adult

Quantile Binning

age
24
45
18
32
40
76

age_bin
Youth
Young Adult
Youth
Young Adult
Adult
Adult

one-hot encoding

Youth	Young Adult	Adult
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	0	1

Quantile Binning Summary

1 Binning

Is used to group together values to reduce the effects of minor observation errors.

2 Quantile Binning

Bins values into equal number of bins.

3 Optimum Number of Bins

Depends on the characteristics of the variables and its relationship to the target. This is best determined through experimentation.

Numeric Feature Engineering Summary

Technique	Function
Normalization	From 0 to 1 0 - minimum value 1 - maximum value
Standardization	0 is the average value is the z-score
Quantile Binning	creates equal number of bins