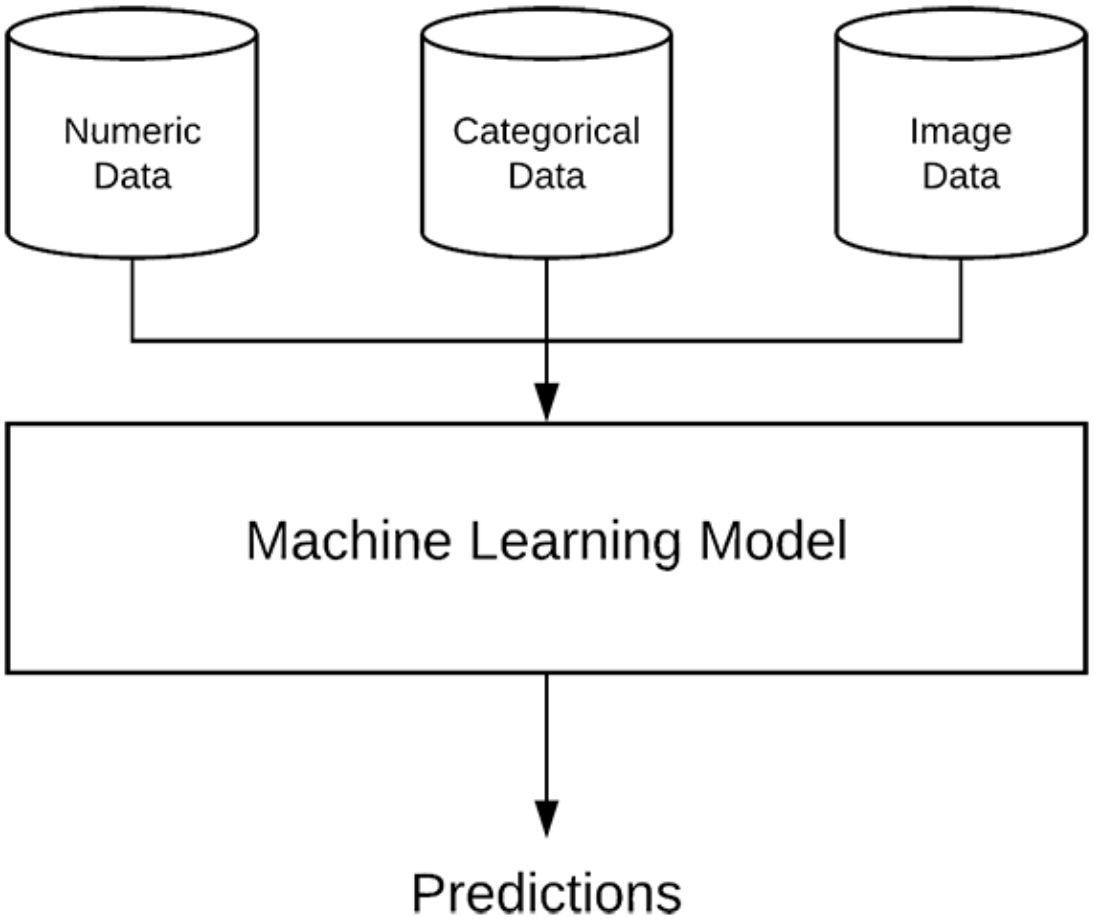
Preprocessing, visualisation and quality of data

Paul van Leeuwen (paul2.vanleeuwen@wur.nl), Wageningen Economic Research

26 October 2020

Introduction

Introduction



Handling Missing Data

Presence Of Missing Data

- Supervised Machine Learning problems are missing data problems.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.7	1.5	0.4	NA
6.3	2.8	5.1	1.5	virginica
6.6	2.9	4.6	1.3	versicolor
5.1	3.8	1.9	0.4	setosa
5.8	2.6	4.0	1.2	versicolor
4.4	3.0	1.3	0.2	NA
5.8	2.8	5.1	2.4	virginica
4.4	2.9	1.4	0.2	setosa
5.3	3.7	1.5	0.2	NA
6.3	2.5	5.0	1.9	NA

Handling Missing Data

- Most algorithms require no missing data points.
- Not advised: drop rows or columns that contain missing values.
- There are several ways to deal with missing data points.
  - Impute the correct value (which is often just zero).
  - Estimate what the correct value could be.

Estimation Procedures

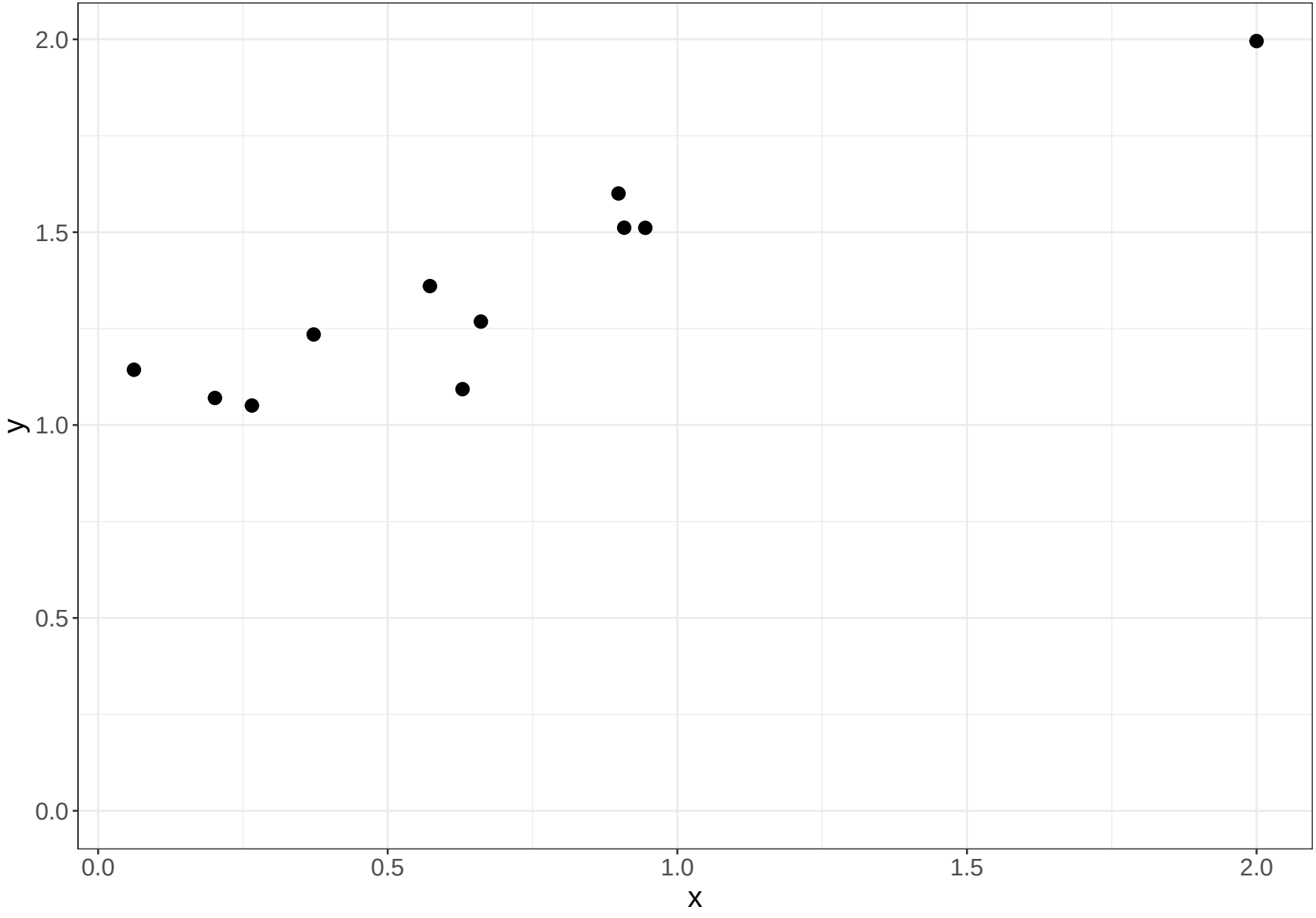
- Instead, try to determine what caused a value to be missing.
- A missing value can be informative as well.
- See Van Buuren [2018] for an introduction.

Handling Outliers

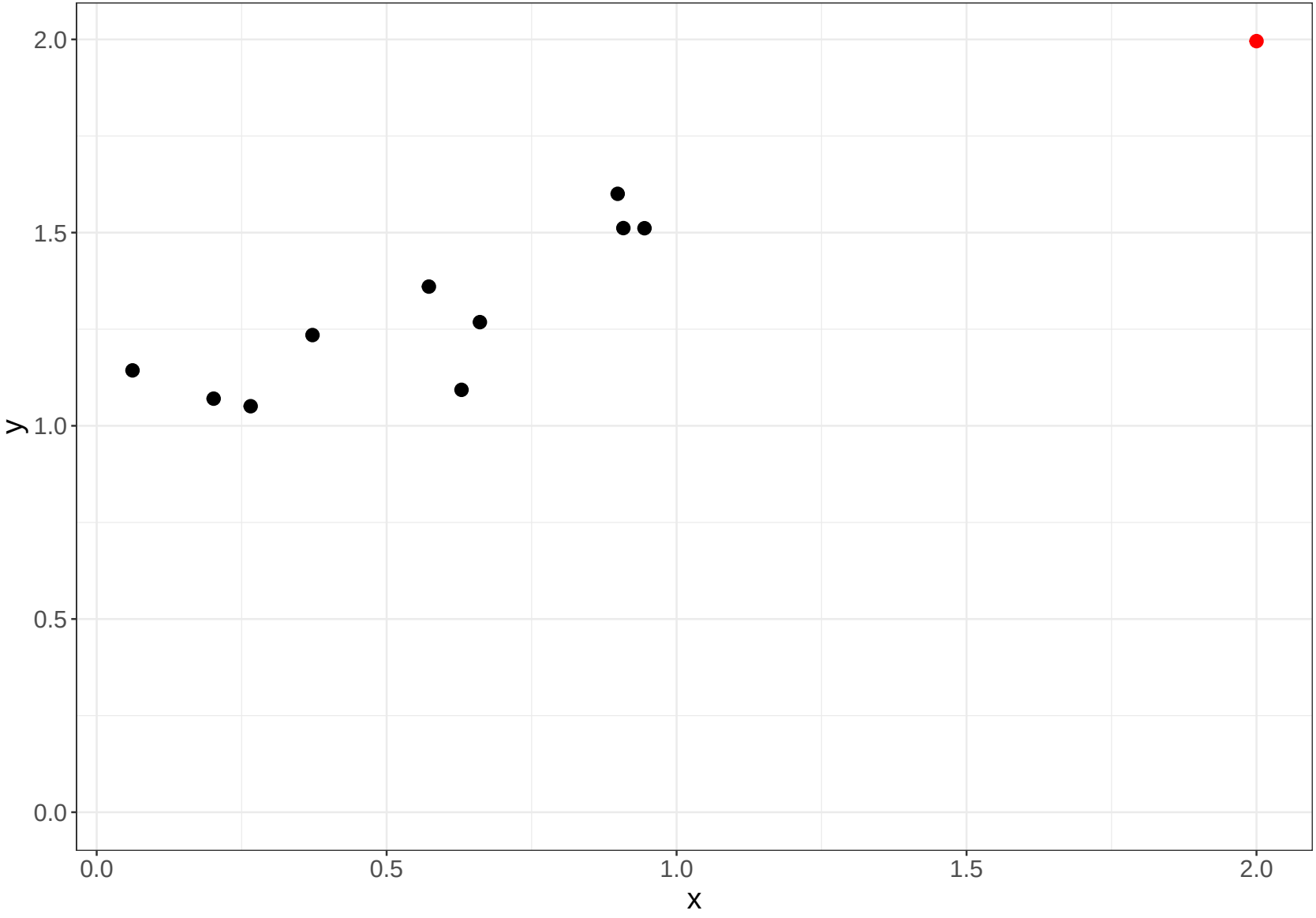
What Is An Outlier?

- 'An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.' (Hawkins [1980])
- Outliers *might* distort your model results.

Impact Of Outliers

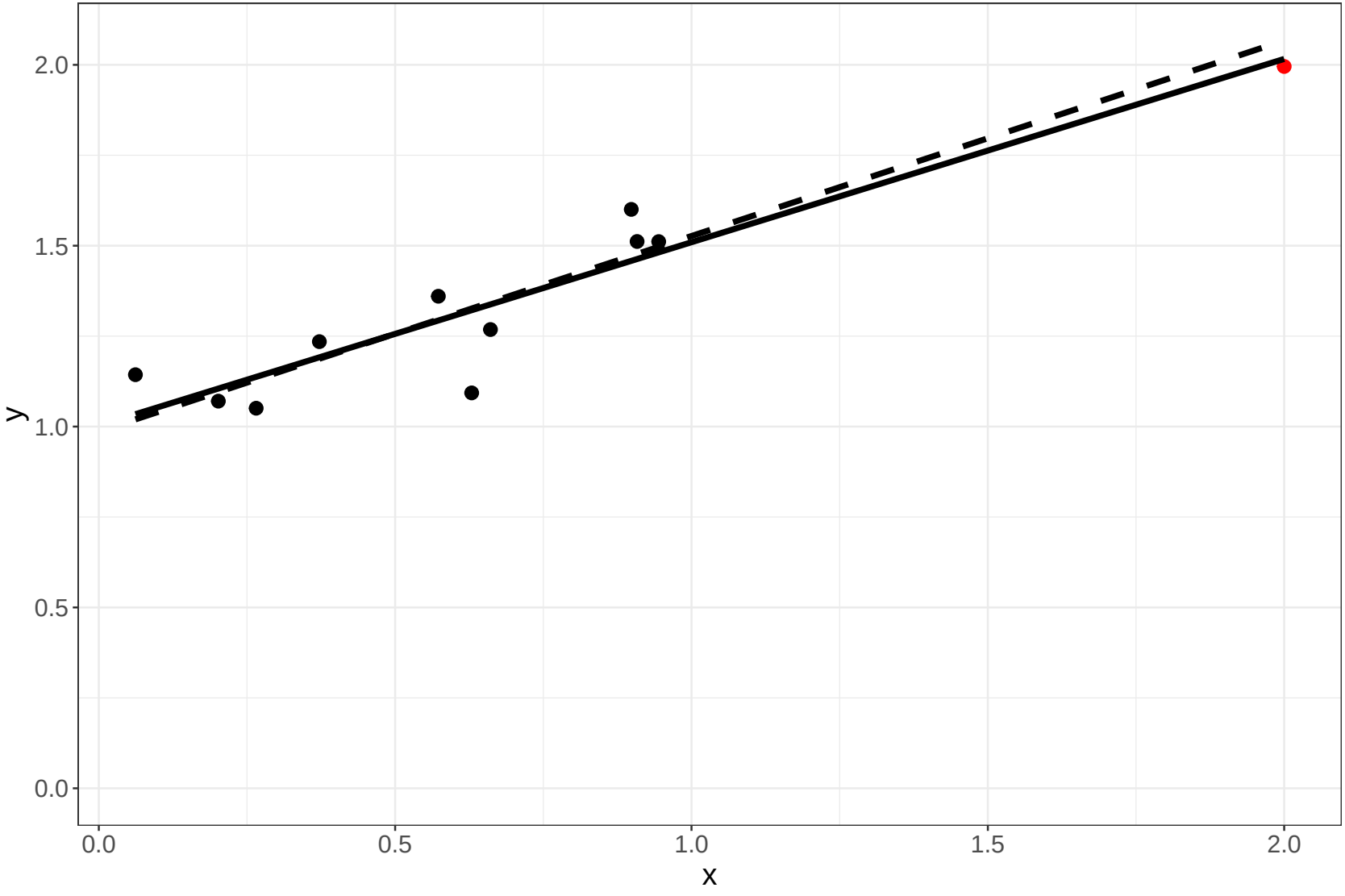


Impact Of Outliers



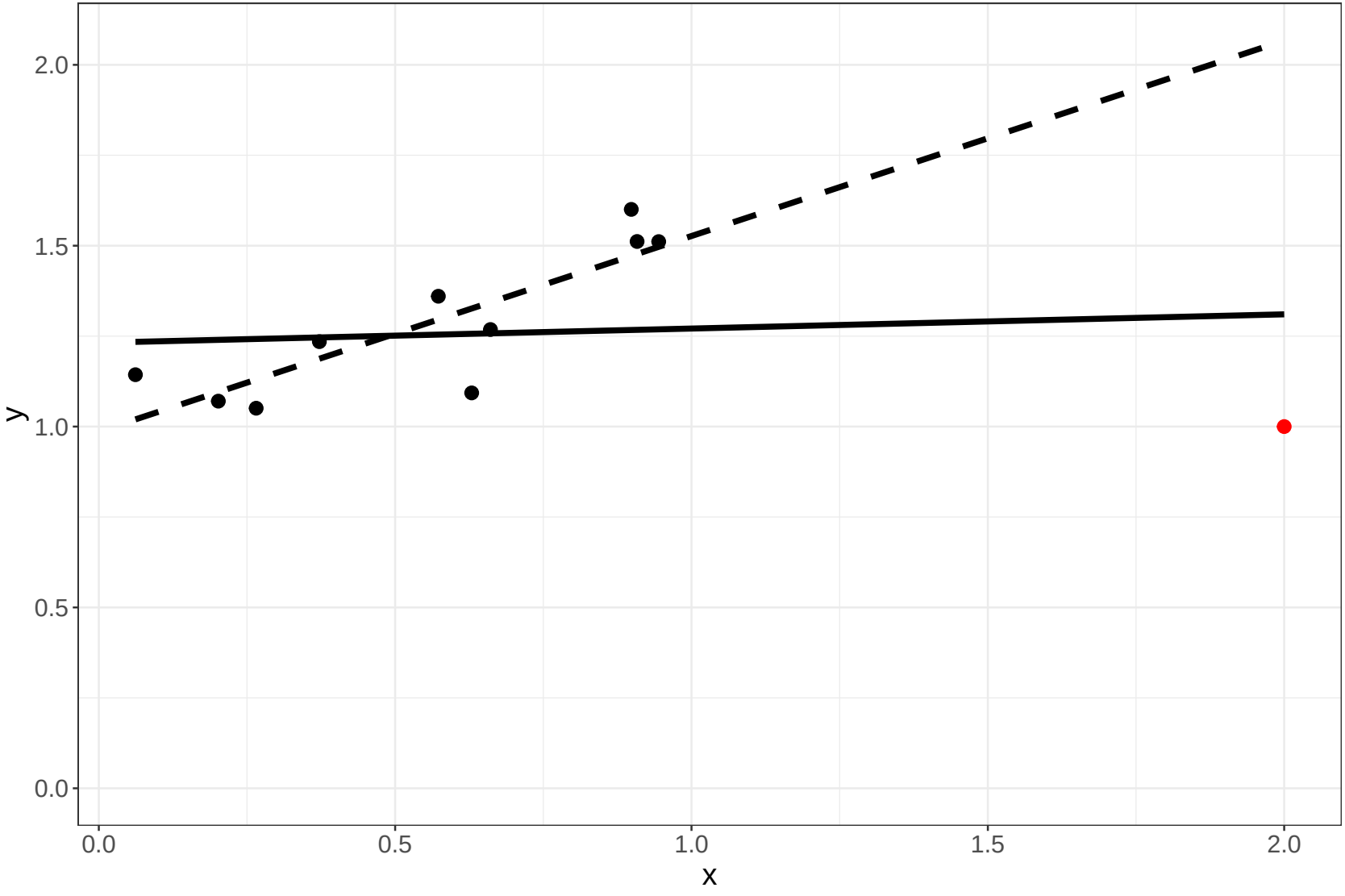
Impact Of Outliers

Linear regression with (solid) and without (dashed) the outlying data point.



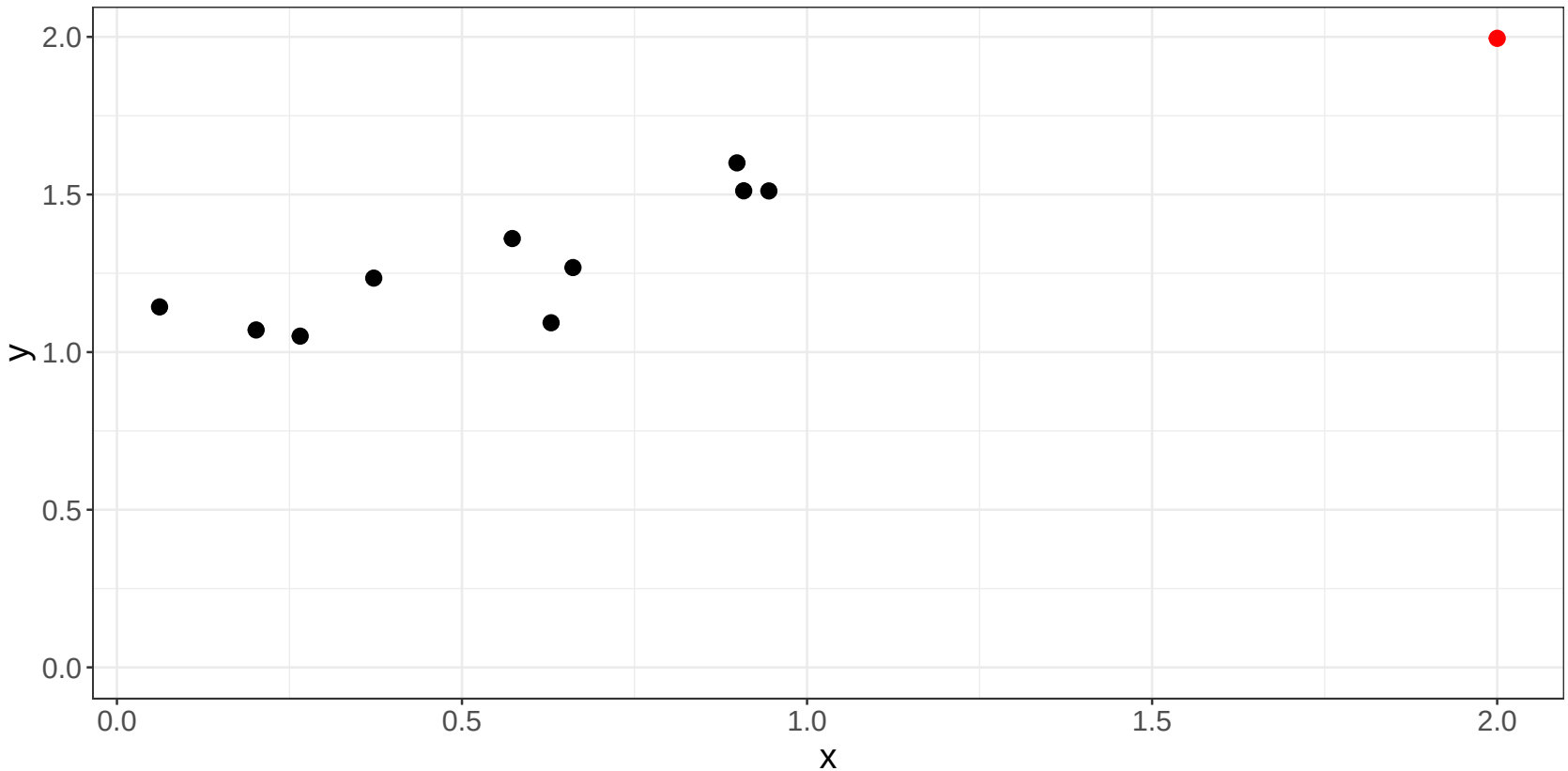
Impact Of Outliers

Linear regression with (solid) and without (dashed) the outlying data point.



Identification Of Outliers

- Lot of different techniques.
- One interesting approach is a *local density-based method*.
- How lonely is a data point relative to the loneliness of its closest neighbours?



Treatment of Outliers

- Depends whether the probability distribution is heavy-tailed.
- Before discarding outliers, determine their cause.
- Robust Linear Regression automatically removes (some) outliers.
- See Aggarwal [2015] for an introduction.
  - LOCI (Papadimitriou et al. [2003]) is appealing.

Improve Your Feature Space

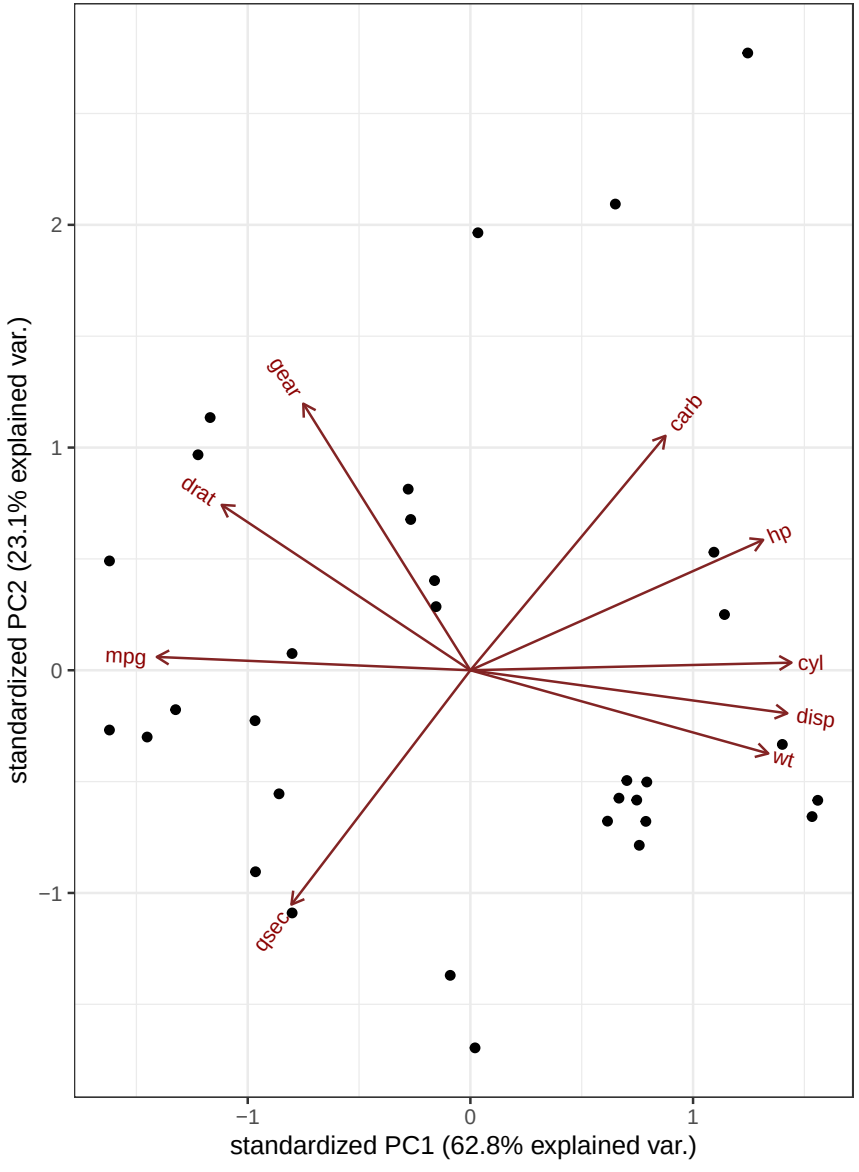
Improve Your Feature Space

- Nowadays datasets are huge.
- Effectively adjusting your feature space can improve your model.
- Approaches to decrease your feature space:
  - Principal Component Analysis (PCA);
  - stacking.
- Approaches to increase your feature space:
  - add higher order terms;
  - add relevant external data.

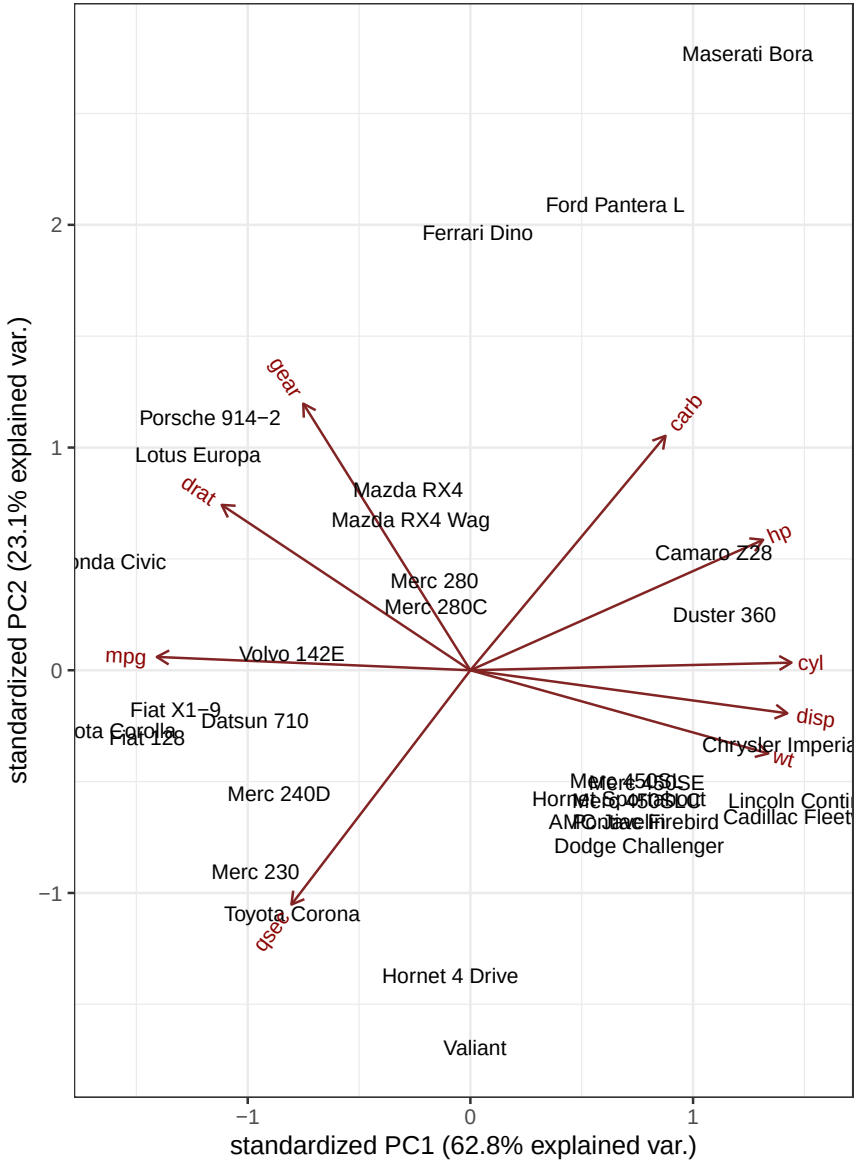
Principal Component Analysis

- The principal components summarise the data.
  - None of the information is thrown away.
  - The dataset is transformed to see it from another angle.
- PCA allows you to verify your intuition.
- The coefficients are in general statistically more significant.

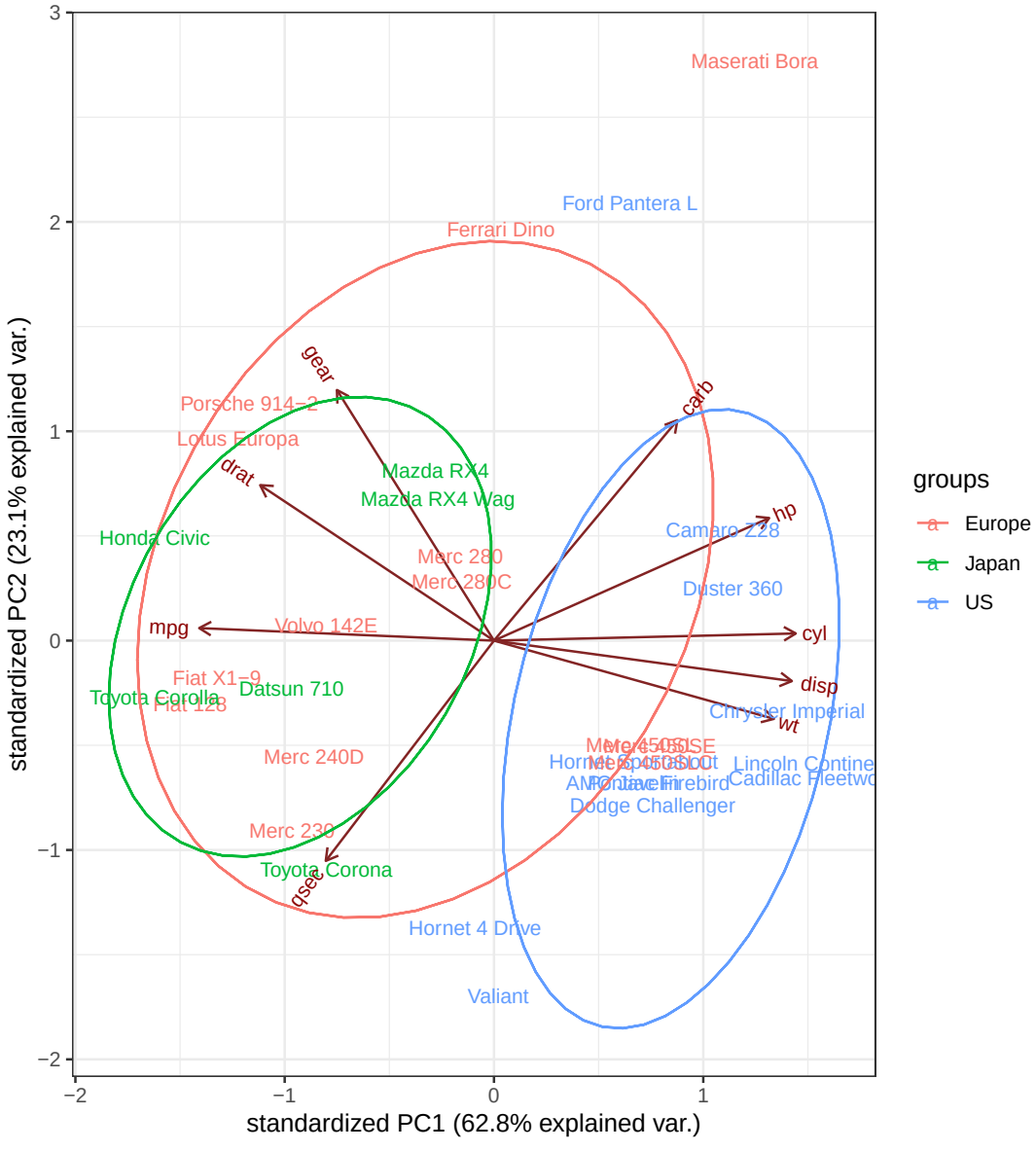
An Example


gear
carb
drat
hp
mpg
cyl
disp
wt
qsec
2
1
0
−1
−1
0
1
standardized PC2 (23.1% explained var.)
standardized PC1 (62.8% explained var.)

An Example


Maserati Bora
Ford Pantera L
Ferrari Dino
gear
carb
Porsche 914−2
Lotus Europa
drat
Mazda RX4
Mazda RX4 Wag
hp
Camaro Z28
Merc 280
Merc 280C
Duster 360
mpg
Volvo 142E
cyl
disp
Fiat X1−9
Datsun 710
Fiat 128
wt
Chrysler Imperia
Merc 240D
Lincoln Conti
Cadillac Fleet
Dodge Challenger
Merc 230
Toyota Corona
qsec
Hornet 4 Drive
Valiant
standardized PC2 (23.1% explained var.)
standardized PC1 (62.8% explained var.)
2
1
0
−1
−1
0
1

An Example

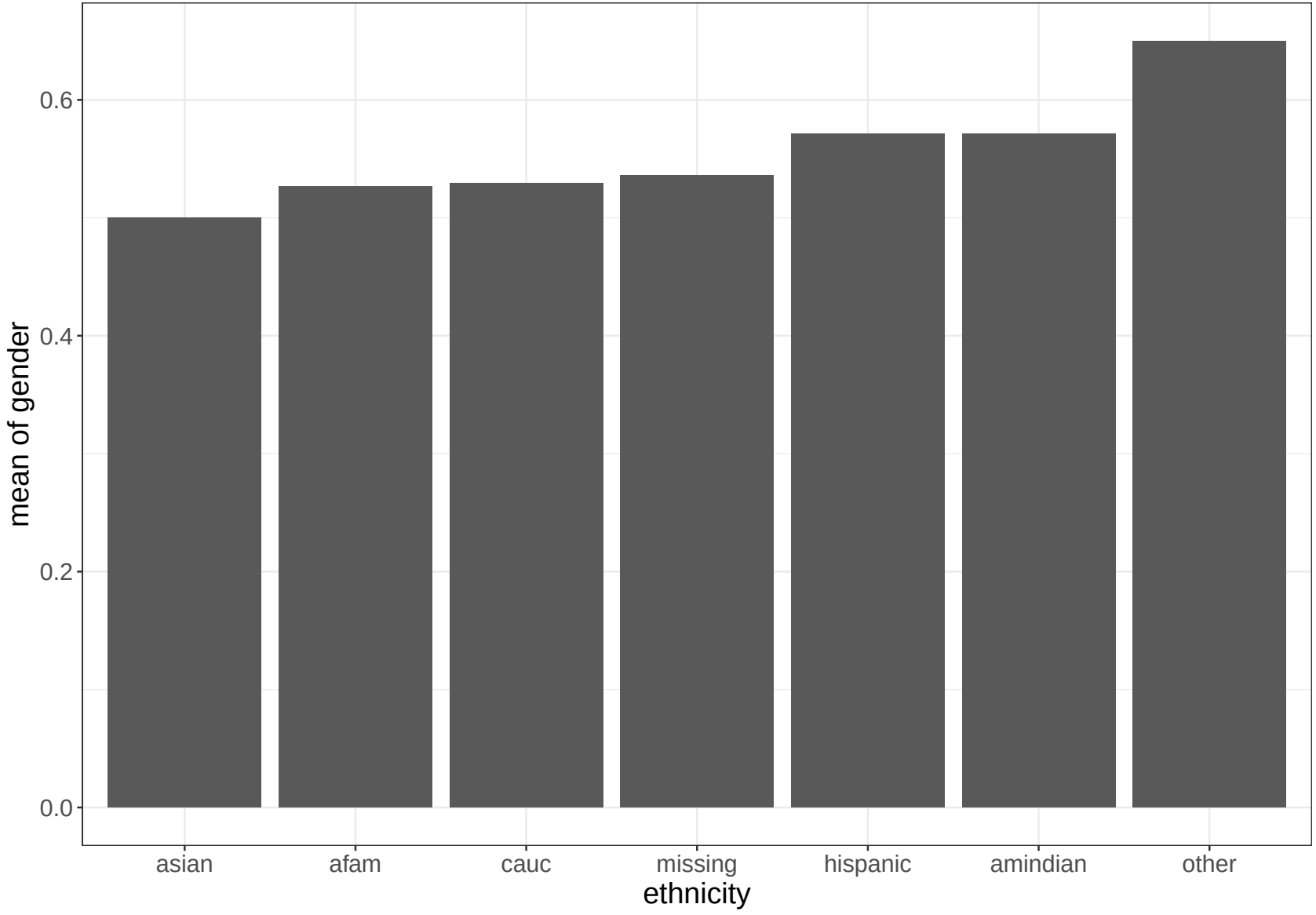

standardized PC2 (23.1% explained var.)
standardized PC1 (62.8% explained var.)
3
2
1
0
−1
−2
−2
−1
0
1
Maserati Bora
Ford Pantera L
Ferrari Dino
Porsche 914−2
Lotus Europa
Mazda RX4
Mazda RX4 Wag
Honda Civic
Merc 280
Merc 280C
Camaro Z28
Duster 360
Volvo 142E
Fiat X1−9
Datsun 710
Toyota Corolla
Fiat 128
Chrysler Imperial
Merc 240D
Lincoln Continental
Cadillac Fleetwood
Dodge Challenger
Merc 230
Toyota Corona
Hornet 4 Drive
Valiant
gear
carb
drat
hp
mpg
cyl
disp
wt
qsec
groups
Europe
Japan
US

Stacking

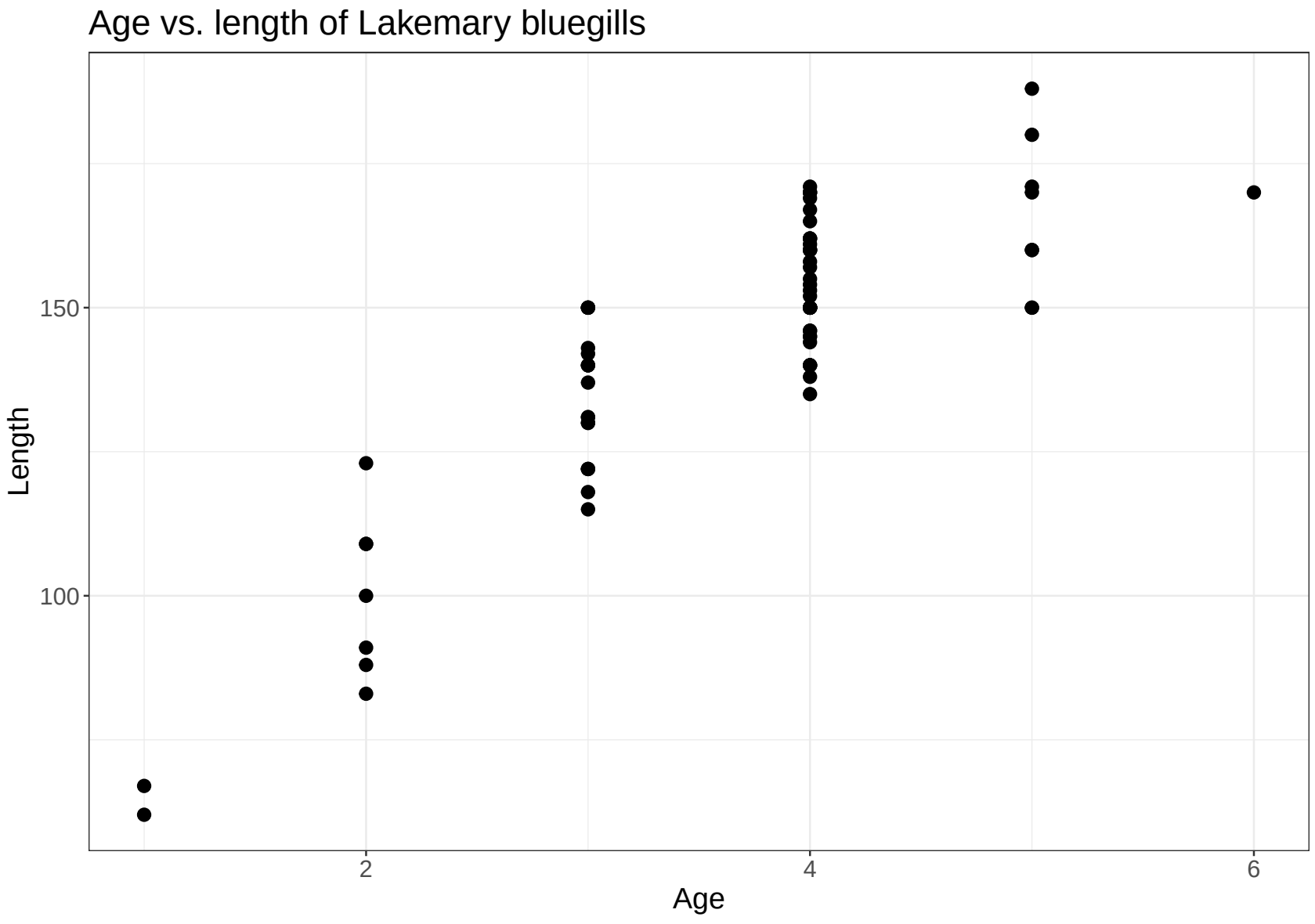
- A categorical variable can be dummified (a.k.a one-hot encoding).
- When there are a lot categories (e.g. municipality), we could run into issues.
- Alternative is to consider the mean of the target variable for each category.
- As an example, take the STAR dataset.
  - Explanatory variable is ethnicity, the target variable the ratio females vs. males.

Column	Type	Domain	# of uniques	% of uniques	# of missings	% of missings
gender	binary	{male; female}	2	0.02	20	0.17
ethnicity	categorical unordered	{cauc; afam; asian; hispanic; other; amindian}	6	0.05	145	1.25
stark	categorical unordered	{regular+aide; regular; small}	3	0.03	5,273	45.46

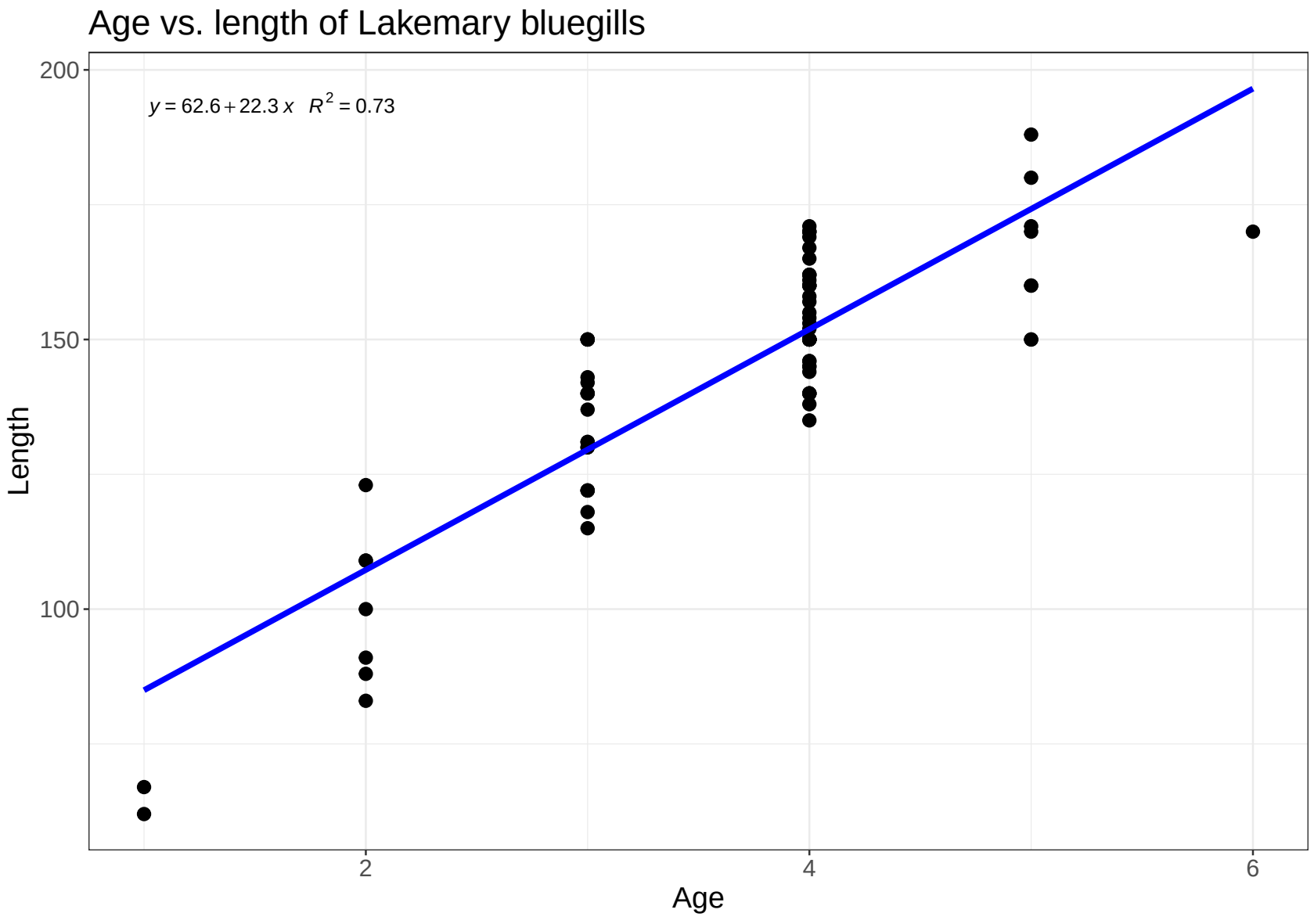
Stacking


0.6
0.4
0.2
0.0
mean of gender
asian
afam
cauc
missing
hispanic
amindian
other
ethnicity

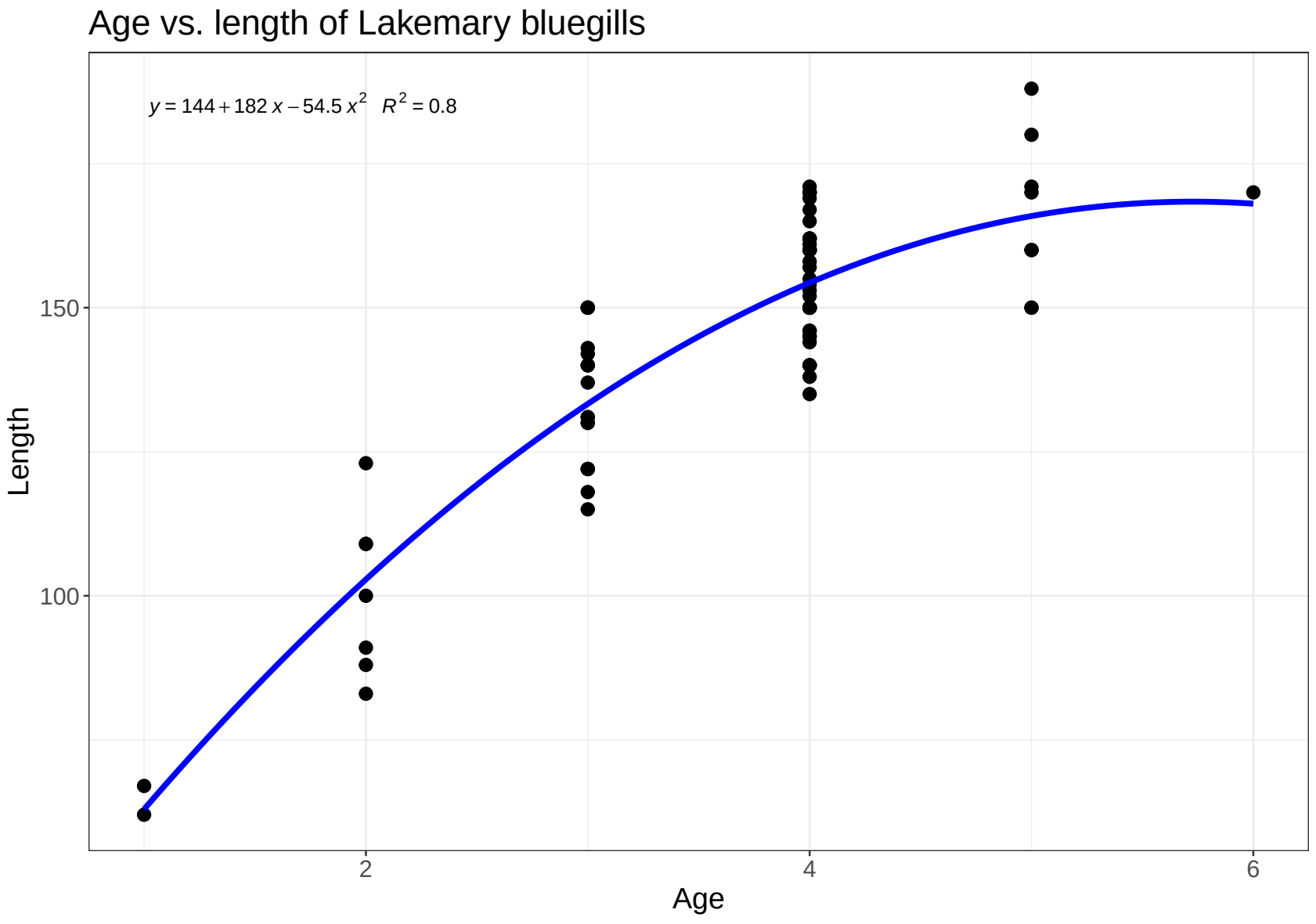
Add Higher Order Terms



Add Higher Order Terms



Add Higher Order Terms



Add Higher Order Terms

- Every deterministic function can be approximated with every desirable accuracy with higher order terms.
- When the target variable is likely to be quadratically correlated with a certain explanatory variable, add the quadratic variant of that variable.
- Take caution as towards the boundaries of the train data, predictions might explode!
  - For example, suppose you catch a rare one of 8 years, would it's length be 118?
- Interpretation of the model coefficients might become counter-intuitive.

Add External Data

- Even the best Machine Learning algorithms are only to squeeze relevant information out of the data that is given.
- Try to add external data conveying information that is not part of the current data.
- For example, in the case of the bluegills weight and different length types help to reduce the unexplained variance.
- Weather data could explain the variation in length.

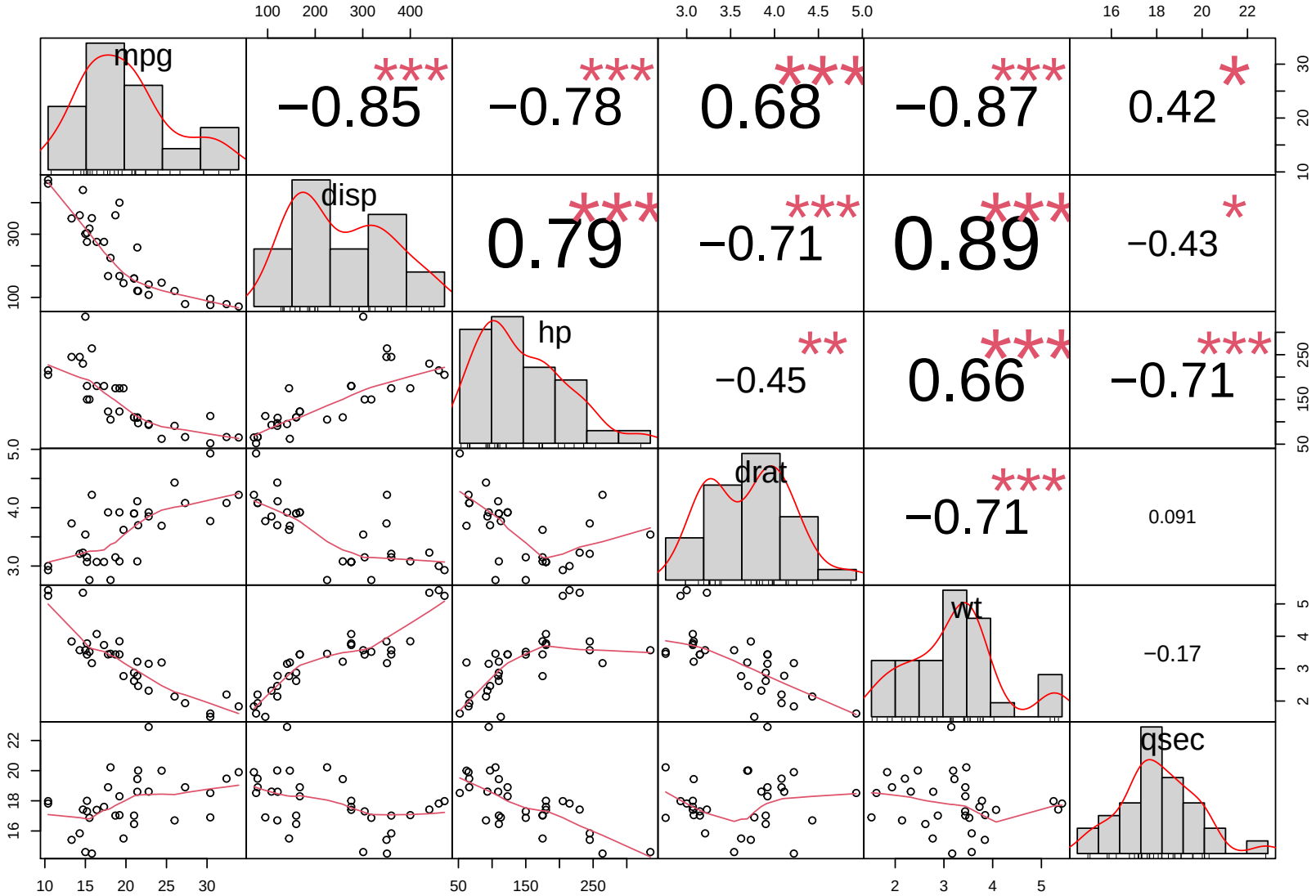
Visualisation

Use Your Intuition!

- Consider histograms and univariate analysis.
- Suppose a variable is numerical by origin but has become character.

Column	Type	Domain	# of uniques	% of uniques	# of missings	% of missings	Multiple of
Sepal.Length	categorical unordered	{5; 5.1; 6.3; 5.7; 6.7; 5.5; 5.8; 6.4; 4.9, ...}	35	23.33	0	0	-
Sepal.Width	decimal	[2.0; 4.4]	23	15.33	0	0	0.1
Petal.Length	decimal	[1.0; 6.9]	43	28.67	0	0	0.1
Petal.Width	decimal	[0.1; 2.5]	22	14.67	0	0	0.1
Species	categorical unordered	{setosa; versicolor; virginica}	3	2	0	0	-

Use Your Intuition!



References

Charu C Aggarwal. Outlier analysis. In *Data mining*, pages 237–263. Springer, 2015.

Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B Gibbons, and Christos Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Proceedings 19th international conference on data engineering (Cat. No. 03CH37405)*, pages 315–326. IEEE, 2003.

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.