Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_

8. Missing Values

Types of Missing Observations

Blank fields come in 4 flavours:

- **nonresponse**
  an observation was expected but none had been entered
- **data entry issue**
  an observation was recorded but was not entered in the dataset
- **invalid entry**
  an observation was recorded but was considered invalid and has been removed
- **expected blank**
  a field has been left blank, but expectedly so

Types of Missing Observations

Too many missing values (of the first three type) can be indicative of **issues with the data collection process** (more on this later).

Too many missing values (of the fourth type) can be indicative of **poor questionnaire design**.

Finding missing values can help you deal with other data science problems.

The Case for Imputation

Not all analytical methods can easily accommodate missing observations:

- **discard** the missing observation
  - not recommended, unless the data is MCAR in the dataset as a whole
  - acceptable in certain situations (e.g., small number of missing values in a large dataset)
- come up with a **replacement (imputation) value**
  - main drawback: we never know what the true value would have been
  - often the best available option

Missing Values Mechanism

Missing Completely at Random (MCAR)

- item absence is independent of its value or of auxiliary variables
- **example:** an electrical surge randomly deletes an observation in the dataset

Missing at Random (MAR)

- item absence is not completely random; can be accounted by auxiliary variables with complete info
- **example:** if women are less likely to tell you their age than men for societal reasons, but not because of the age values themselves)

Missing Values Mechanism

Not Missing at Random (NMAR)

- reason for nonresponse is related to item value (also called **non-ignorable non-response**)
- **example:** if illicit drug users are less likely to admit to drug use than teetotallers

In general, the missing mechanism **cannot be determined** with any certainty; we may need to make assumptions (domain expertise can help).

Imputation Methods

- list-wise deletion
- mean or most frequent imputation
- regression or correlation imputation
- stochastic regression imputation
- last observation carried forward
- next observation carried backward
- k-nearest neighbours imputation
- multiple imputation
- etc.

Imputation Methods

List-wise deletion: remove units with at least one missing values

- **assumption:** MCAR
- **cons:** can introduce bias (if not MCAR), reduction in sample size, increase in standard error

Mean/most frequent imputation: substitute missing values by average/most frequent value

- **assumption:** MCAR
- **cons:** distortions of distribution (spike at mean) and relationships among variables

Imputation Methods

Regression/correlation imputation: substitute missing values using fitted values based on other variables with complete information

- **assumption:** MAR
- **cons:** artificial reduction in variability, over-estimation of correlation

Stochastic regression imputation: regression/correlation imputation with a random error term added

- **assumption:** MAR
- **cons:** increased risk of type I error (false positives) due to small std error

Imputation Methods

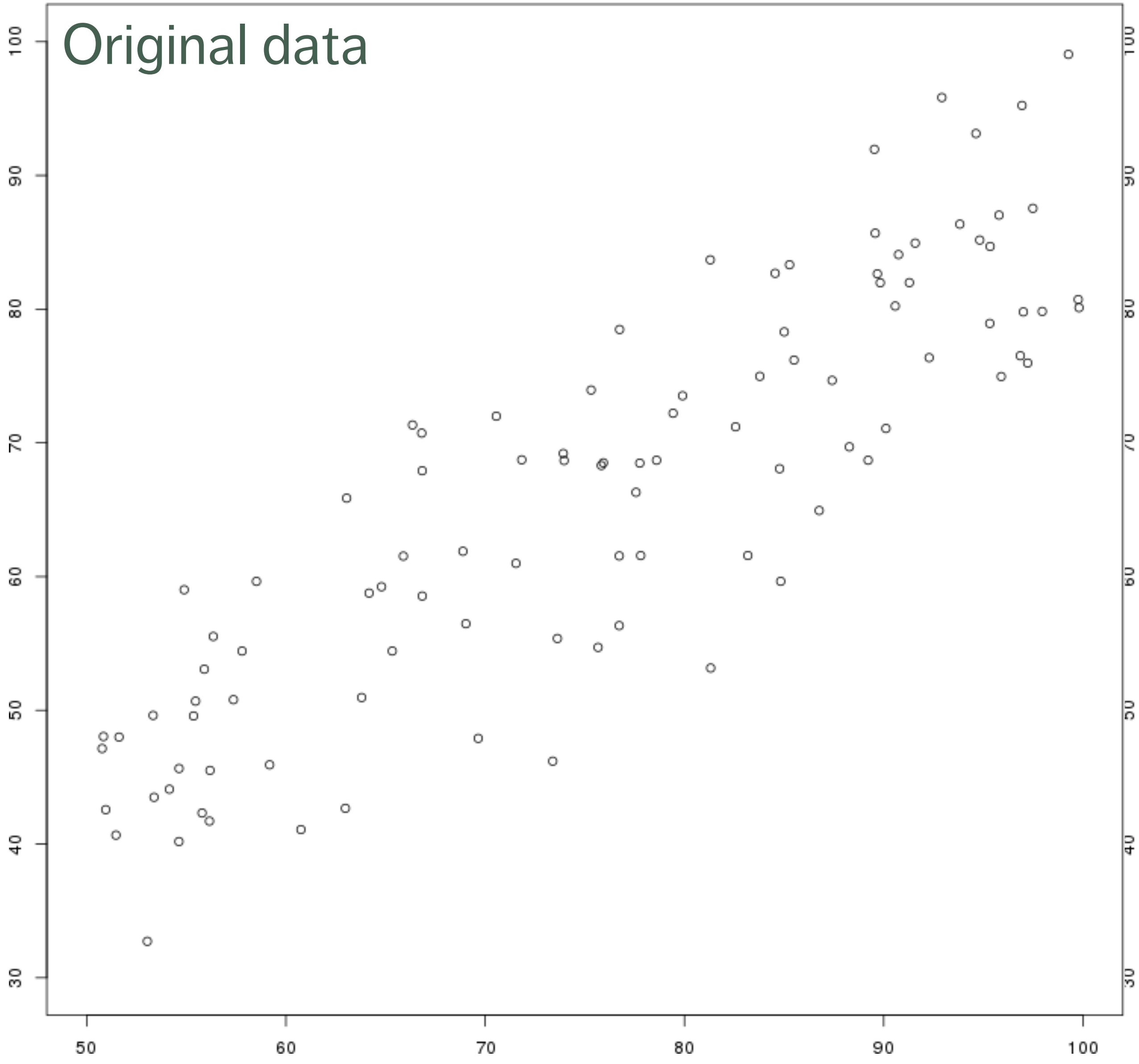
Last observation carried forward: substitute the missing values with latest previous values (in a longitudinal study)

- **assumption:** MCAR, values do not vary greatly over time
- **cons:** may be too "generous", depending on the nature of study

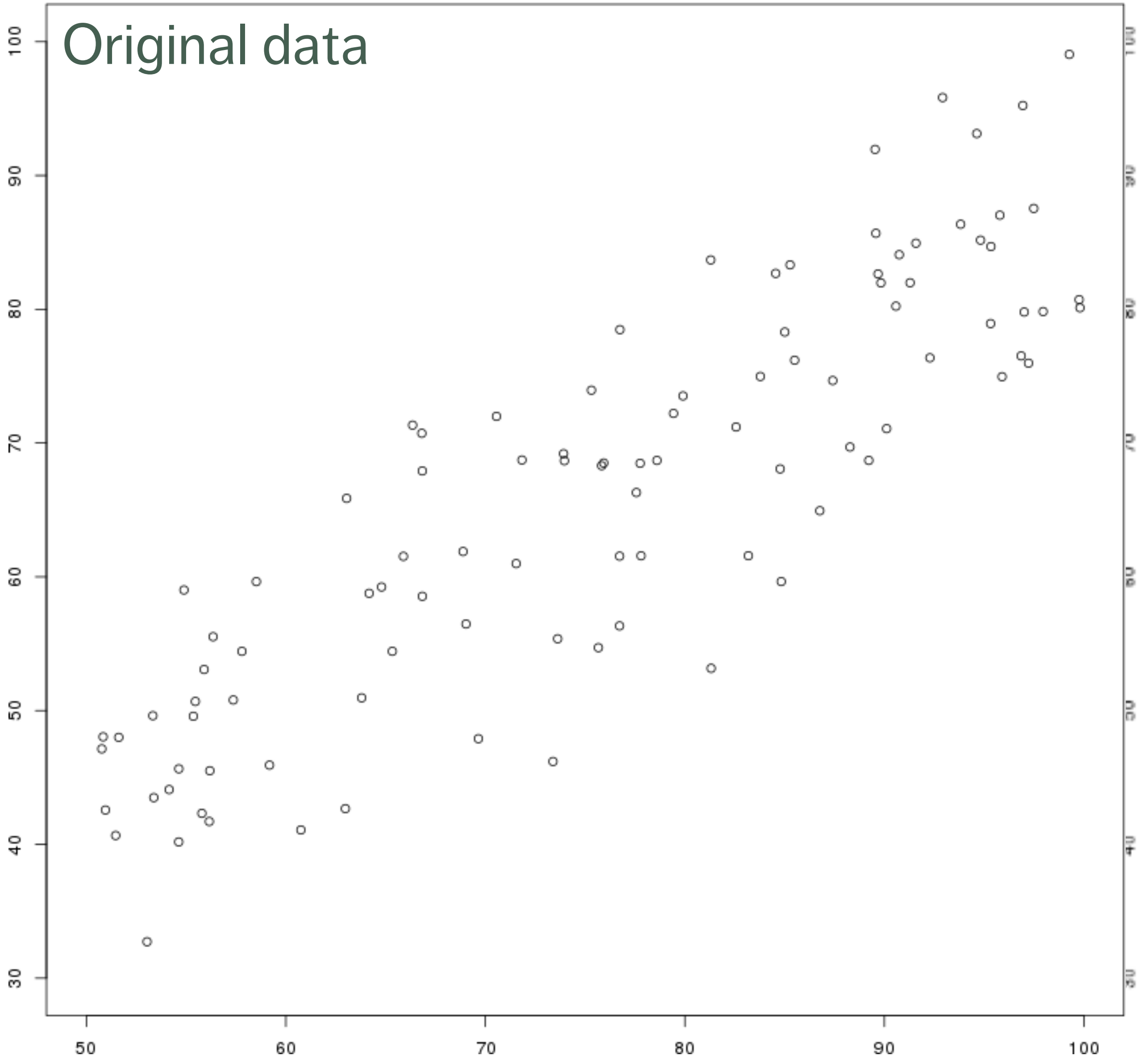
k nearest neighbour imputation (kNN): substitute the missing entry with the average from the group of the k most similar complete cases

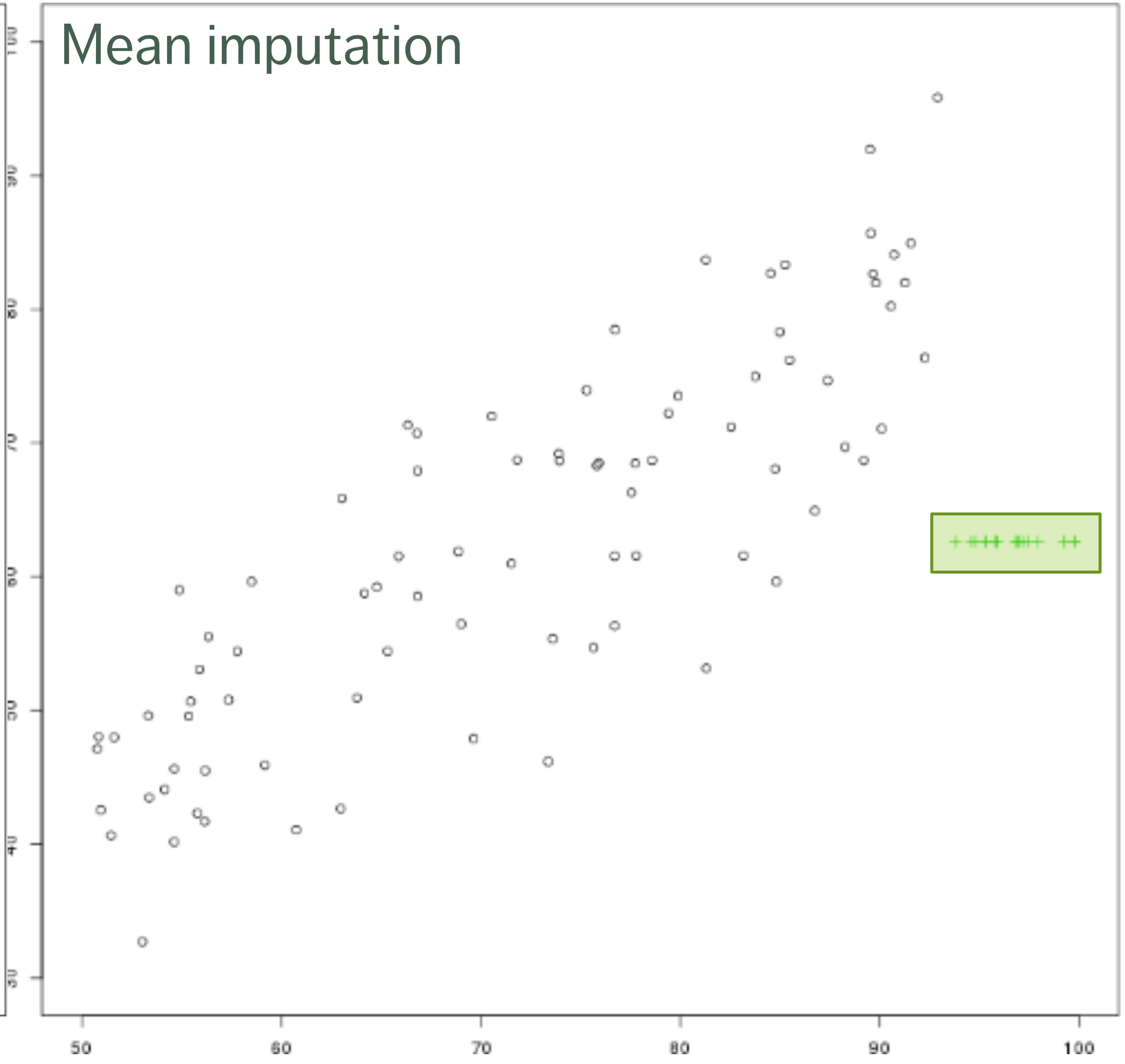
- **assumption:** MAR
- **cons:** difficult to choose appropriate value for k; possible distortion in data structure

Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



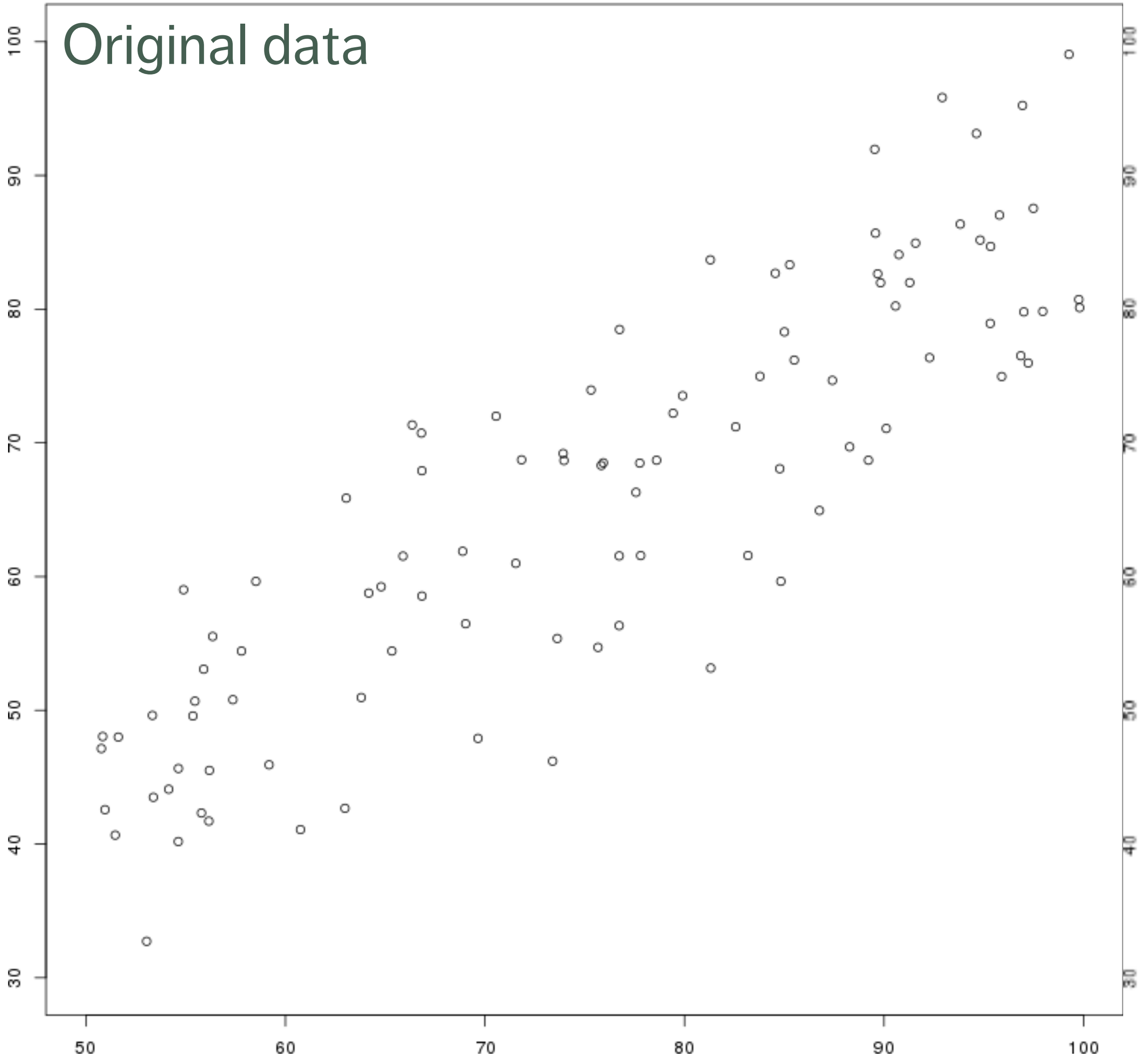
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.

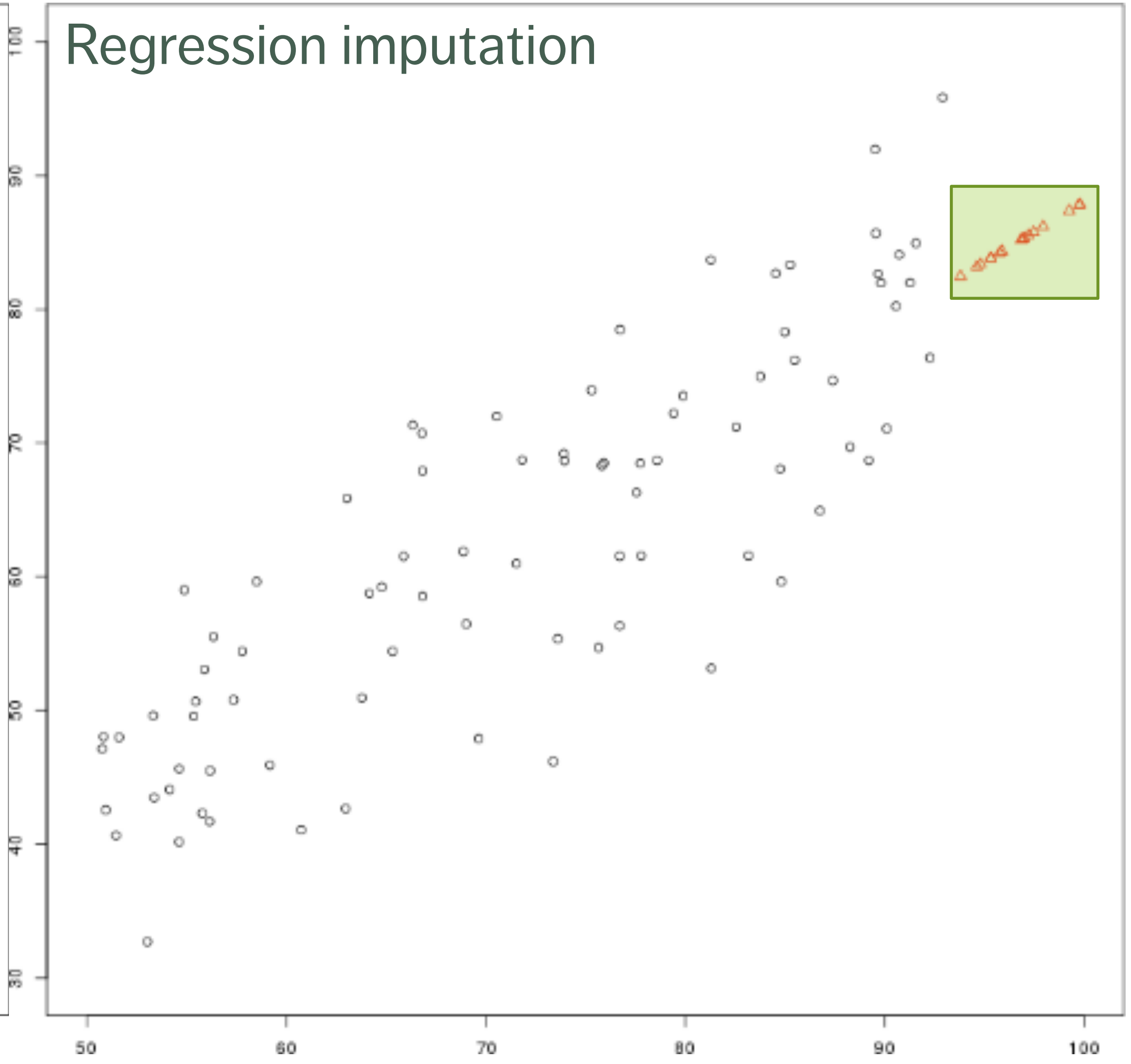




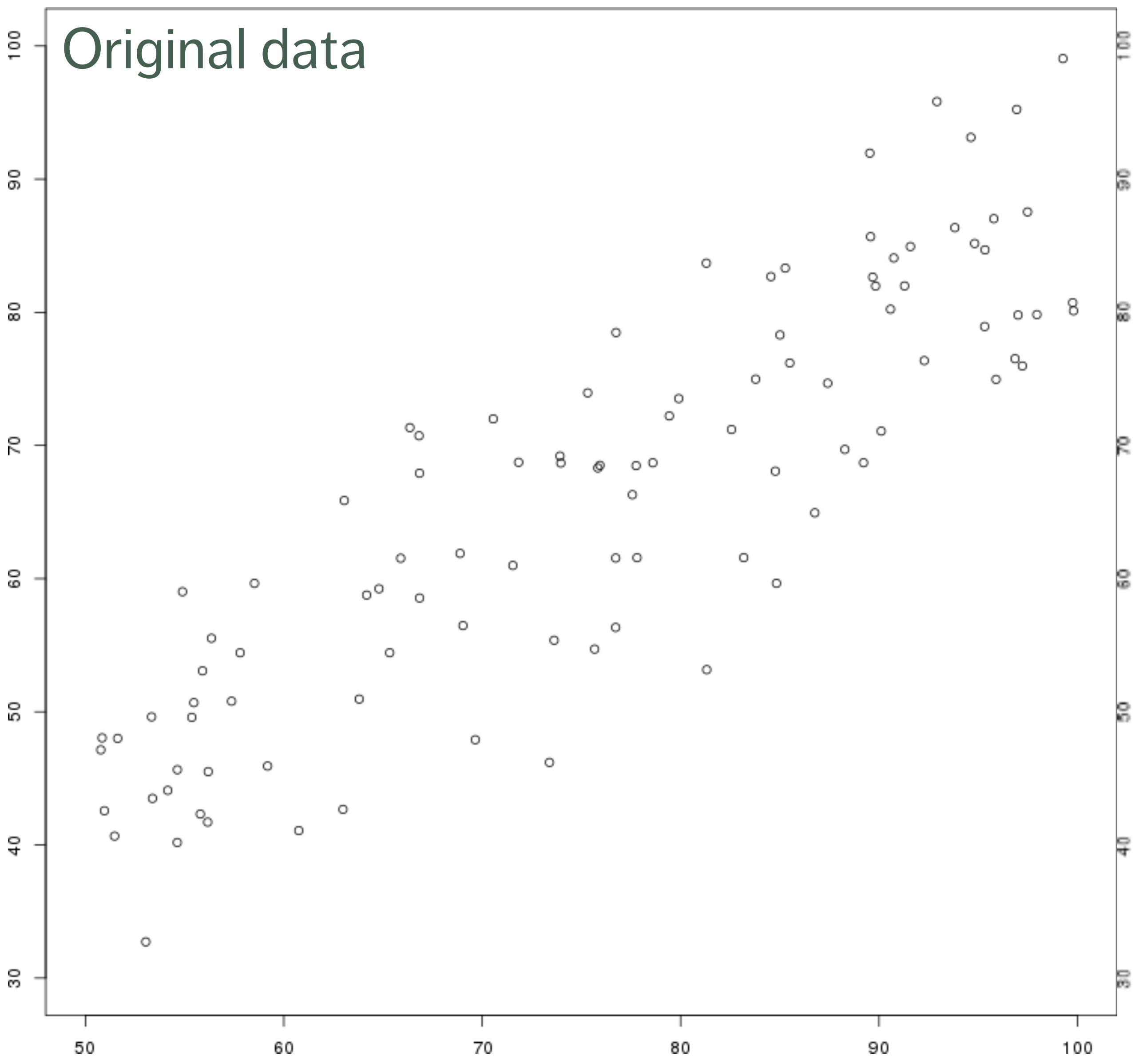
Session 3

Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



Multiple Imputation

Imputations increase the noise in the data.

In **multiple imputation**, the effect of that noise can be measured by consolidating the analysis outcome from multiple imputed datasets

Steps:

1. repeated imputation creates m versions of the dataset
2. each of these datasets is analyzed, yielding m outcomes
3. the m outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known

Multiple Imputation

Advantages

- **flexible**; can be used in a various situations (MCAR, MAR, even NMAR in certain cases)
- accounts for **uncertainty** in imputed values
- fairly easy to implement

Disadvantages

- m may need to be fairly **large** when there are many missing values in numerous features, which slows down the analyses
- if the analysis output is not a single value but some complicated mathematical object, this approach is unlikely to be useful

Take-Aways

Missing values **cannot simply be ignored**.

The missing mechanism **cannot typically be determined** with any certainty.

Imputation methods work best when values are **MCAR** or **MAR**, but imputation methods tend to produce biased estimates.

In single imputation, imputed data is treated as the actual data; multiple imputation can help reduce the noise.

Is stochastic imputation best? In our example, yes – but … **No-Free Lunch theorem**!

Suggested Reading

Missing Values

Data Understanding, Data Analysis, Data Science
Data Preparation

Missing Values

- Missing Value Mechanisms
- Imputation Methods
- Multiple Imputation

Exercises

Missing Values

1. Recreate the examples of Imputation Methods.
2. Recreate the missing value imputation process (data cleaning) used in Example: Algae Bloom.
3. Conduct kNN imputation on the grades dataset with various values of k.
4. Conduct multiple imputation on the grades dataset using stochastic regression in order to estimate the slope and intercept for the line of best fit.