



Data Cleaning

Programming for Data Science

Hand in your code – A Tale of Instruction

Please send your code to claudio.hartmann@tu-dresden.de

- Regardless of programming language
- Accepted file types: .py, .R, .zip (with one file in it)
- No links! No repos! No data!
- Include your name, first name and task number in the file name.
- Deadline for all tasks: 07.02.2020 (We recommend a weekly hand-in cycle.)

Warm Up

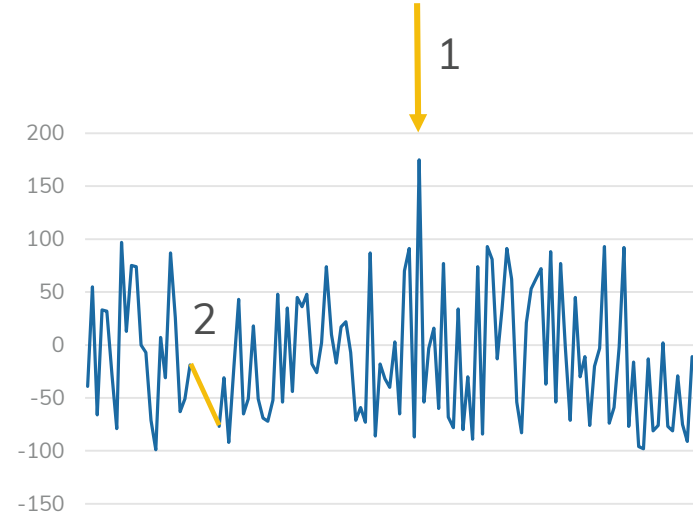
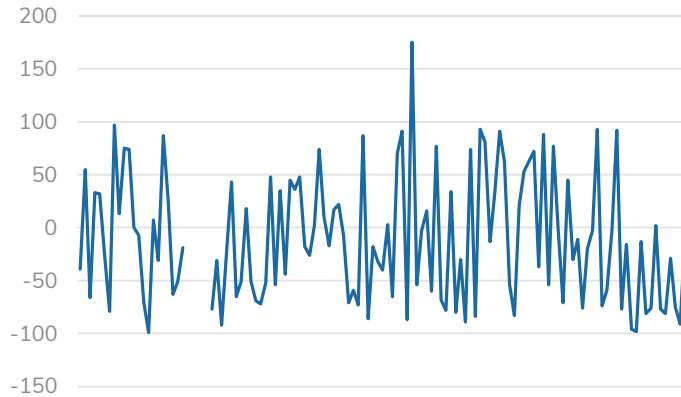


Answer the following statements! Give reason for your answers.

1. What are dimensions of data quality?
2. What are possible types of data impurities?
3. How can they effect data analyses?
4. What are methods to resolve data impurities?
5. What are the advantages of linear over step interpolation?

Task

Clean the given data.



Task

Step 0

- You will get a csv file from us. Load it in your language/environment.
- Explore the data in it.

Step 1

- Find outliers in the data.
- Implement a Interquartile range filter (IRQ)* and a z-score filter*.
- Replace outliers with NA values.

Step 2

- Fill all missing data points with NA.
- Implement a step interpolation* and a linear interpolation*.
- Replace all NA values with the interpolated values.

*use your own implementation

IQR

- Based on 25th and 75th percentile (quantile)
- Outliers are values which are larger than $Q_{75} + 1.5 \cdot IQR$ or smaller than $Q_{25} - 1.5 \cdot IQR$.
- $IQR = Q_{75} - Q_{25}$

z-score (3σ method)

- Outliers are values which are larger than 3 or smaller than -3 in the z-standardized data.
- z-standardized: $z(data) = \frac{data - \mu(data)}{\sigma(data)}$

step interpolation

- From y_{low} to y_{high} in the middle of the gap



linear interpolation

- Continuously from y_{low} to y_{high} with slope m



Package suggestions

R

- data.table

python3

- pandas
- numpy
- (matplotlib.pyplot)

Exercise Appointment

We compare and discuss the results

- Tuesday, 29.10.2019,
- Consultation: Thursday, 24.10.2019, APB/E008,
- Please prepare your solutions! Send us your code!

If you have questions, please mail us:

claudio.hartmann@tu-dresden.de Orga + Code

lucas.woltmann@tu-dresden.de Tasks + Python

lars.kegel@tu-dresden.de Tasks + R

