TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Data Cleaning

Programming for Data Science

Warm Up

Dresden Database Systems Group

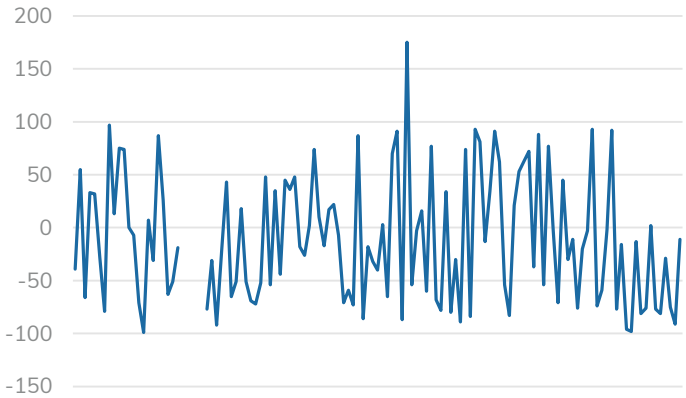
Answer the following statements! Give reason for your answers.

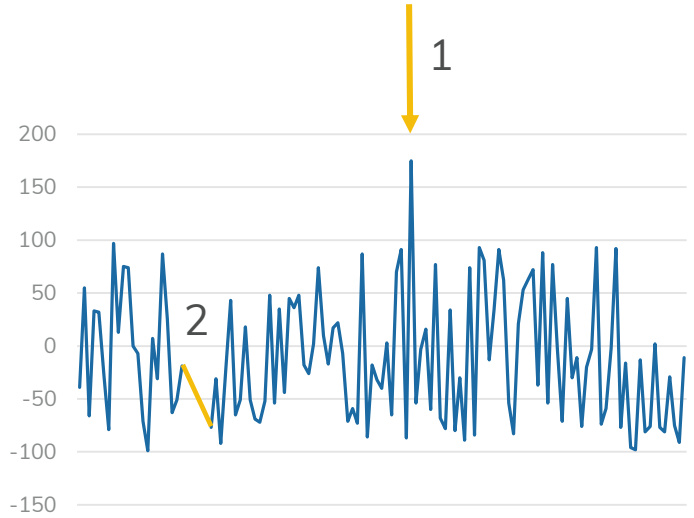
1. What are dimensions of data quality?
2. What are possible types of data impurities?
3. How can they effect data analyses?
4. What are methods to resolve data impurities?
5. What are the advantages of linear over step interpolation?

https://amcs.website

Task

Clean the given data.





Task

Dresden Database Systems Group

IQR

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

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

Task

Dresden Database Systems Group

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

Package suggestions

Dresden Database Systems Group

R

- data.table

python3

- pandas
- numpy
- (matplotlib.pyplot)

Exercise Appointment

Dresden Database Systems Group

We compare and discuss the results

- Tuesday, 10.11.2020,
- Consultation: Please use the forum in Opal.
- Please prepare your solutions! Send us your code!

If you have questions, please mail us:

claudio.hartmann@tu-dresden.de Orga + Code + R
lucas.woltmann@tu-dresden.de Tasks + Python