DATA SCIENCE
PPGIa/PUCPR

Prof. Jean Paul Barddal

PUCPR
GRUPO MARISTA

1

MISSING DATA

2

Missing data

- Missing data is very common
- Main reasons behind missing data:
  - Person did not want to share his/her personal data
  - Data loss during data transfer
  - Lack of information, etc
- Important: some algorithms will ignore missing values, while others will not even run if missing data exists

3

Approaches for handling missing data

We will analyze a couple of ideas to handle missing data:

1. Removing rows with missing values
2. Removing rows where most part of the values are missing
3. Ignoring columns with too many missing values
4. Data imputation
5. Data imputation using machine learning

4

#1: Removing rows with missing values

- Sometimes, the number of rows with missing values is not too big (< 5%)
- If this is the case, we can think about removing the rows with missing values
- **Important**: we should try to avoid data removal as much as possible

5

#2: Removing rows where most of the values are missing

- An often better approach is to remove instances in which most of the values are missing
- The threshold may change here, e.g., 70%, 75%, 80%, etc
- How do we pick this threshold up?
  - Feeling
  - We check how much data we will lose

6

#3: Ignoring columns with too many missing values

- If the number of missing values is too high for a column, perhaps we should get rid of the column
- Again: what is the threshold for flagging a column removal?

7

#4: Data imputation

- If you don't have too many missing values, we can replace these values with a "generic" value
- This is not necessarily the best approach, as we are not sure whether this "generic" value is a common behavior in data

8

More on imputation

- Static value
  – Replace missing values with a 0, +99999999 or -99999999
  – Do these values make sense? What is their impact when using machine learning?
- Statistic
  – Replace missing values with the mean/**median**/mode of the remainder of the values
  – Mean: what if the data has outliers?
- Predictive model
  – We can infer the missing values based on the present values for the remainder of the columns

9

Summing up

- If you don't have many missing values, we may ignore them
- If a column has too many missing values, we may remove it
- Imputation is a viable option, but requires testing

10

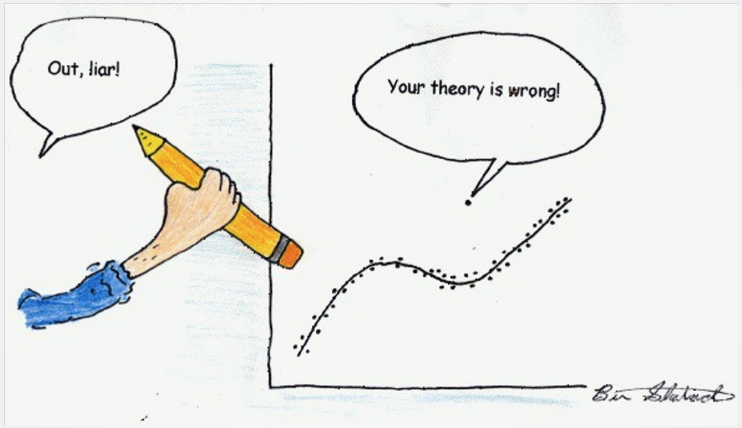
Activity

- Let's identify and handle missing values using the Titanic dataset

11

OUTLIERS

12



13

Outlier

- An outlier is a value that is very "far" from the others
- It can be much smaller or bigger than the rest
- Does not exhibit the same behavior as the other data
- It is hard to identify, as what is an outlier to a dataset might be different in other datasets

14

Reasons for Outliers

- Mistakes in data acquisition
- Data may be corrupted
- Or maybe, the outlier is true

15

Example

Do you see any outliers in this table?

Gender	Age	Height
M	20	159
F	21	191
M	24	173
M	24	181
F	28	156
M	26	192
F	19	280
F	22	162
M	26	190

16

Another example

What about in this table? Note that all of the data is correct, yet, Michael Phelps is still an outlier

Athlete	# of medals
Michael Phelps	28
Larisa Latiynina	18
Marit Bjorgen	15
Nikolai Andrianov	15
Ole Eimar Bjorndalen	13
Boris Shakhlin	13
Edoardo Mangiarotti	13

17

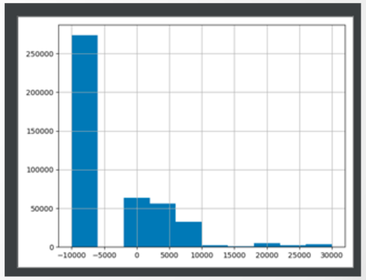
FINDING OUTLIERS WITH HISTOGRAMS

18

Finding outliers with histograms

Histogram

- Univariate plot that allows us to analyze the data distribution of a single variable
- Let's plot the "MAIORRENDACASA" variable histogram and see what happens :)



19

USING BOX-PLOTS FOR OUTLIER ANALYSIS

20

Agenda

- We will use Tukey's method to identify outliers
- Tukey's method is based on quartiles, more specifically on the Inter-quartile range, or IQR
- Overview:
  - Apply Tukey's method in each variable individually
  - If an instance is flagged as an outlier n times, then it should be removed
  - How do we define n?

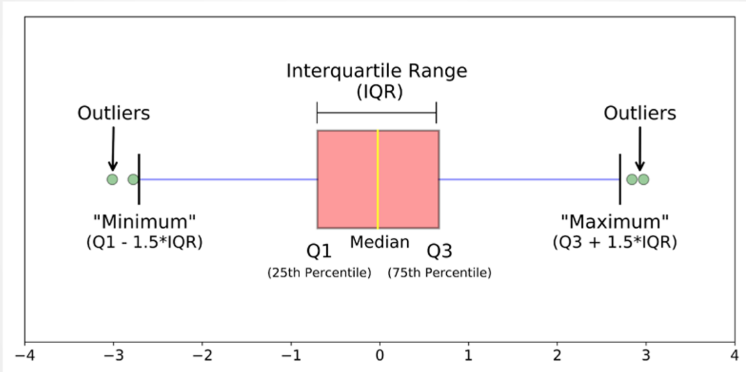
21

Inter-quartile Range

- The interval between the 1st and 3rd quartiles
- IQR = Q3 - Q1
- If a value is greater than 1.5 * IQR + Q3 or smaller than Q1 - 1.5 * IQR, then it is flagged as an outlier
- 1.5 is a "rule of thumb" proposed by Tukey
- 3 is a threshold for determining "far out" values
- Let's analyze this using boxplots

22

Box-plot



23

Using box-plots for outlier analysis

- Be cautious
- There is no guarantee that data outside the IQR are, indeed, outliers
- Approaches:
  - Check the data that has been flagged
  - Verify the IQR results for multiple variables, and if it is flagged multiple times, then it is very likely that this data point is an outlier

24

Activity

Let's use box-plots to analyze the variables available in the **OMMLBD_FAMILIAR.csv** file

25

ISOLATION FORESTS

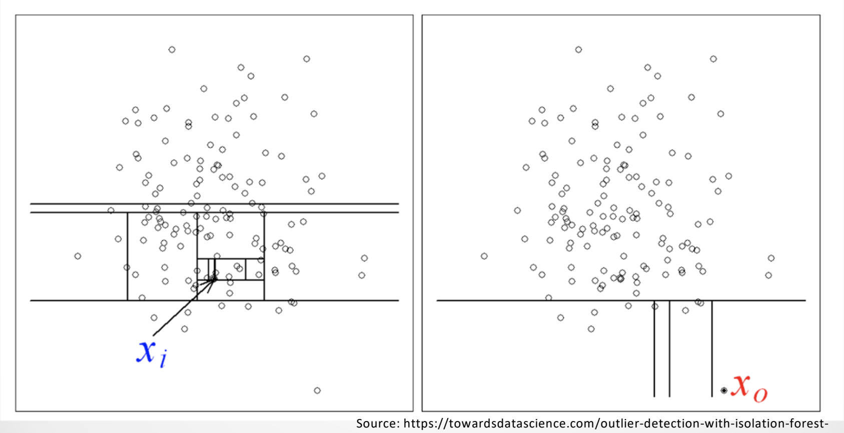
26

Rationale

- Outliers are less frequent than regular observations
- If the data is randomly partitioned, outliers will require, in average, less partitions to be **isolated** from the rest of the data points

27

Rationale



Source: https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

28

Isolation Forest

- A set of isolation trees are created, and each tree randomly partitions the data points
- An "anomaly score" is obtained for a data point as follows:
- A score close to 1 stands for an outlier
- A score smaller than 0.5 indicates normal observations
- If all scores are close to 0.5, the entire sample does not contain distinct anomalies

29

Example

Let's use isolation forests in a toy example and understand it visually

30

ACTIVITY

31

Enron

- Back in the 2000s, Enron was one of the biggest companies in USA
- In 2002, it went bankrupt due to frauds
- Most of the data has been made public
- And today we will work with the data on salaries and emails

32

Outlier analysis

- Using the Enron dataset, try and identify individuals with weird (outlier) behavior
- The rationale here is that if someone has an outlier behavior, it is likely to be a fraudster
- Use any tools you wish, but we expect you to find at least **3 major outliers**
- **Data is available at: http://www.ppgia.pucpr.br/~jean.barddal/datascience/enron.csv**

33