Data Preprocessing

Data Cleaning: Handling Missing Values, Noisy Data and Outliers

Data Cleaning (Data Cleansing)

- Real world data are tend to be incomplete, noisy and inconsistent
- Data cleaning routines attempt to identify missing values, fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data

> *80 percent of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis...*
>
> IBM Data Analytics

- One of the biggest data cleaning task is handling missing values

Data Cleaning: Missing Values

- Many tuple (records) have no recorded value for several attributes
- Identifying missing values:
  - When Pandas library for python is used, it detect the missing values as "NaN" [1]
  - It automatically consider "blank" in the attribute value, "NaN/nan/NAN" in the attribute value , "NA" in the attribute value, "n/a" in the attribute value, "NULL/null" in the attribute value as NaN

[1] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

Methods to Handle Missing Values

- Ignore the tuples:
  - This method is effective only when the tuples contain several attributes (> 50% of attributes) with missing value

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018			83.14912	
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	24.29851	87.68657	963
08-07-2018	t11			
09-07-2018	t11	26.8494	61.10241	15
10-07-2018	t11	27.88806	75.07463	13583.25
11-07-2018	t11	27.35915	76.02113	19768.5
23-07-2018	t12	24.39024	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.75
25-07-2018				
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	24.29851	87.68657	963
09-07-2018	t11	26.8494	61.10241	15
10-07-2018	t11	27.88806	75.07463	13583.25
11-07-2018	t11	27.35915	76.02113	19768.5
23-07-2018	t12	24.39024	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.75
26-07-2018	t12	22.19718	99	864

Tuples contain several attributes (> 50% of attributes) with missing value

Methods to Handle Missing Values

- Ignore the tuples:
  - This method is effective only when the tuples contain several attributes (> 50% of attributes) with missing value
  - This method is also used when the target variable (class label) is missing

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	83.14912	
10-07-2018	NaN	25.17021	85.34043	652.5
11-07-2018	t10	24.29851	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494		15
10-07-2018	t11	27.88806	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12		94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	NaN	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	83.14912	
11-07-2018	t10	24.29851	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494		15
10-07-2018	t11	27.88806	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12		94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
26-07-2018	t12	22.19718	99	864

Target attribute (StationID) with missing value

Methods to Handle Missing Values

- Fill in the missing values (imputing values) manually:
  - Time consuming
  - Not feasible given a large data set with many missing values
- Use a global constant to fill in missing value (Imputing global constant):
  - Replace all missing attribute values by a same constant
  - Imputed value may not be correct

Methods to Handle Missing Values

- Use attribute mean/median/mode to fill in the missing value (mean/median/mode imputation):
  - Applicable to numeric data
  - Centre of the data won't change

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	NaN	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	NaN	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Methods to Handle Missing Values

- Use attribute mean/median/mode to fill in the missing value (mean/median/mode imputation):
  - Applicable to numeric data
  - Centre of the data won't change

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	NaN	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	NaN	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.1368	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	25.1368	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	25.1368	94.4065	1071
24-07-2018	t12	24.16197	(Ctrl) 01	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Methods to Handle Missing Values

- Use attribute mean/median/mode to fill in the missing value (mean/median/mode imputation):
  - Applicable to numeric data
  - Centre of the data won't change
  - However, it does not preserve the relationship with other variables

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.1368	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	25.1368	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	25.1368	94.4065	1071
24-07-2018	t12	24.16197	01	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	85.42	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.1368	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	85.42	15
10-07-2018	t11	25.1368	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	25.1368	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Methods to Handle Missing Values

- Filling with local mean/median/mode:
  - Use attribute mean/median/mode of all samples belonging to a group (class) to fill in the missing value
    - Applicable to numeric data
    - Centre of the data of a group won't change

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	NaN	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	NaN	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.612	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	NaN	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Methods to Handle Missing Values

- Filling with local mean/median/mode:
  - Use attribute mean/median/mode of all samples belonging to a group (class) to fill in the missing value
    - Applicable to numeric data
    - Centre of the data of a group won't change

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.612	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	NaN	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.612	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	25.916	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Methods to Handle Missing Values

- Filling with local mean/median/mode:
  - Use attribute mean/median/mode of all samples belonging to a group (class) to fill in the missing value
    - Applicable to numeric data
    - Centre of the data of a group won't change
    - However, it does not preserve the relationship with other variables

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.612	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	25.916	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	NaN	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Dates	Station Id	Temperature	Humidity	Rain
08-07-2018	t10	25.46875	82.1875	6.75
09-07-2018	t10	26.19298	NaN	1762
10-07-2018	t10	25.17021	85.34043	652.5
11-07-2018	t10	25.612	87.68657	963
08-07-2018	t11	23.53846	61.92308	3
09-07-2018	t11	26.8494	NaN	15
10-07-2018	t11	25.916	75.07463	13583
11-07-2018	t11	27.35915	76.02113	19769
23-07-2018	t12	23.884	94.4065	1071
24-07-2018	t12	24.16197	97.66901	438.8
25-07-2018	t12	25.29323	94.84211	13667
26-07-2018	t12	22.19718	99	864

Methods to Handle Missing Values

- Use the values from the previous/next record (with in a group) to fill in missing value (Padding)

	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018	t10	26.19298	NaN	1762
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	NaN	87.68657	963
6	08-07-2018	t11	23.53846	61.92308	3
7	09-07-2018	t11	26.8494	NaN	15
8	10-07-2018	t11	NaN	75.07463	13583
9	11-07-2018	t11	27.35915	76.02113	19769
10	23-07-2018	t12	NaN	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.8
12	25-07-2018	t12	25.29323	94.84211	13667
13	26-07-2018	t12	22.19718	99	864

	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018	t10	26.19298	82.1875	1762
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	25.17021	87.68657	963
6	08-07-2018	t11	23.53846	61.92308	3
7	09-07-2018	t11	26.8494	61.92308	15
8	10-07-2018	t11	26.8494	75.07463	13583
9	11-07-2018	t11	27.35915	76.02113	19769
10	23-07-2018	t12	24.16197	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.8
12	25-07-2018	t12	25.29323	94.84211	13667
13	26-07-2018	t12	22.19718	99	864

- If the data is categorical or text, one can replace the missing values by most frequent observations

Methods to Handle Missing Values

- Use most probable value to fill the missing value:
  - Use interpolation technique to predict the missing value
    - **Linear interpolation** is achieved by geometrically rendering a straight line between two adjacent points on a graph or plane
    - Interpolation happens column wise
    - Popular strategy
    - It does not preserves the relationship with other variables

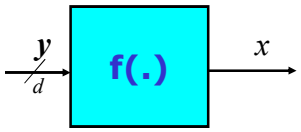
	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	NaN	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	NaN	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	NaN	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	NaN	99	513
16	25-07-2018	23.18182	98.9697	195.75

	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	24.2	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	21.5	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	23.7	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	21.6	99	513
16	25-07-2018	23.18182	98.9697	195.75

Methods to Handle Missing Values

- Use most probable value to fill the missing value:
  - Use regression techniques to predict the missing value (regression imputation)
    - Let $y_1, y_2, \ldots, y_d$ be a set of d attributes
    - Regression (multivariate): The n^{th} value is predicted as

$$x_n = f(y_{n1}, y_{n2}, \ldots, y_{nd})$$



    - Linear regression (multivariate): $x_n = w_1 y_{n1} + w_2 y_{n2} + \ldots + w_d y_{nd}$

Methods to Handle Missing Values

- Use most probable value to fill the missing value:
  - Use regression techniques to predict the missing value (regression imputation)
    - Let $y_1, y_2, \ldots, y_d$ be a set of d attributes
    - Regression (multivariate): The n^{th} value is predicted as

$$x_n = f(y_{n1}, y_{n2}, \ldots, y_{nd})$$

	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018	t10	26.19298	NaN	1762
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	NaN	87.68657	963
6	08-07-2018	t11	23.53846	61.92308	3
7	09-07-2018	t11	26.8494	NaN	15
8	10-07-2018	t11	NaN	75.07463	13583
9	11-07-2018	t11	27.35915	76.02113	19769
10	23-07-2018	t12	NaN	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.8
12	25-07-2018	t12	25.29323	94.84211	13667
13	26-07-2018	t12	22.19718	99	864

$$Temperature = f(Humidity, Rain)$$

$$Temperature = w_{T1} Humidity + w_{T2} Rain$$

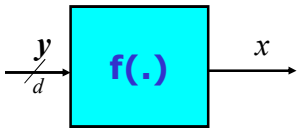
$$Humidity = f(Temperature, Rain)$$

$$Humidity = w_{H1} Temperature + w_{H2} Rain$$

Methods to Handle Missing Values

- Use most probable value to fill the missing value:
  - Use regression techniques to predict the missing value (regression imputation)
    - Let y_1, y_2, …, y_d be a set of d attributes
    - Regression (multivariate): The n^{th} value is predicted as

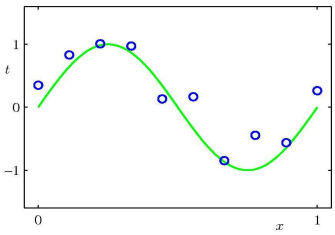
$$x_n = f(y_{n1}, y_{n2}, \ldots, y_{nd})$$



    - Linear regression (multivariate): $x_n = w_1 y_{n1} + w_2 y_{n2} + \ldots + w_d y_{nd}$
    - Popular strategy
    - It uses the most information from the present data to predict the missing values
    - *It preserves the relationship with other variables*

Data Cleaning: Smoothing the Noisy Data

- Noise is a random error or variance in a measured variable
- Due to noise, many tuple (records) have incorrect value for several attributes
- Mostly data is full of noise
- Smooth out the data to remove the effect of noise
- Data smoothing allows important patterns to stand out
- The idea is to sharpen the patterns (values) in the data and highlight trends the data is pointing to



- Methods for data smoothing:
  - Binning
  - Regression (function approximation)

Binning Methods for Data Smoothing

- Binning method smooth a sorted data value of a noisy attribute by consulting its neighbourhood i.e., the values around it
- It perform local smoothing as this method consult the neighbourhood of values
- The sorted values are partitioned into (almost) equal-frequency bins

Binning Methods for Data Smoothing

- ***Different approaches for smoothing by bin:***

1. Smoothing by bin means:
   - Each value in a bin is replaced by the mean value of the bin
2. Smoothing by bin medians:
   - Each value in a bin is replaced by the median value of the bin
3. Smoothing by bin boundaries:
   - The minimum and maximum values in a given bin are identified as bin boundaries
   - Each bin value is then replaced by the closest boundary value

- Larger the width, the greater the effect of the smoothing

Illustration of Binning Methods for Data Smoothing

- Example:
- Noisy data for price (in Rs) : 8, 15, 34, 24, 4, 21, 28, 21, 25
- Sorted data for price (in Rs) : 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into bins:
Bin1: 4, 8, 15
Bin2: 21, 21, 24
Bin3: 25, 28, 34

Smoothing by bin means:
Bin1: 9, 9, 9
Bin2: 22, 22, 22
Bin3: 29, 29, 29

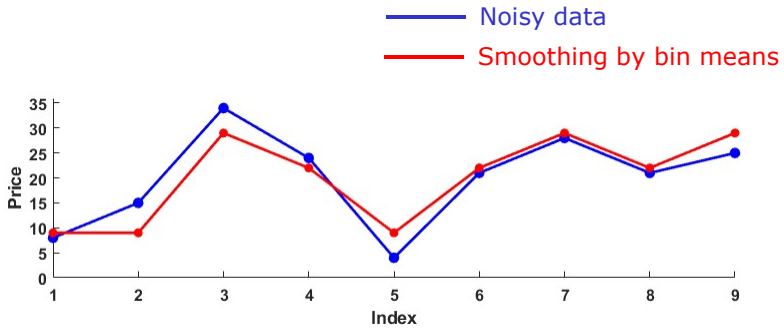


Illustration of Binning Methods for Data Smoothing

- Example:
- Noisy data for price (in Rs) : 8, 15, 34, 24, 4, 21, 28, 21, 25
- Sorted data for price (in Rs) : 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into bins:
Bin1: 4, 8, 15
Bin2: 21, 21, 24
Bin3: 25, 28, 34

Smoothing by bin Boundaries:
Bin1: 4, 4, 15
Bin2: 21, 21, 24
Bin3: 25, 25, 34

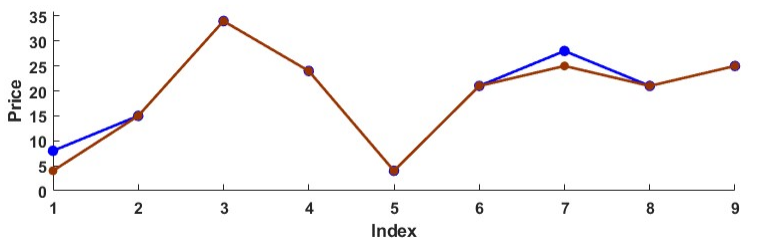
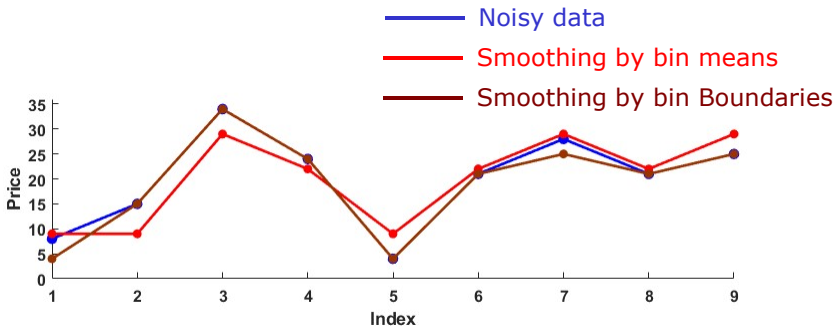


Illustration of Binning Methods for Data Smoothing

- Example:
- Noisy data for price (in Rs) : 8, 15, 34, 24, 4, 21, 28, 21, 25
- Sorted data for price (in Rs) : 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into bins:	Smoothing by bin means:	Smoothing by bin Boundaries:
Bin1: 4, 8, 15	Bin1: 9, 9, 9	Bin1: 4, 4, 15
Bin2: 21, 21, 24	Bin2: 22, 22, 22	Bin2: 21, 21, 24
Bin3: 25, 28, 34	Bin3: 29, 29, 29	Bin3: 25, 25, 34



Outlier Detection and Replacing with Centre of Tendency

- Compute first quartile (Q1) and third quartile (Q3) for an attribute
- Compute the interquartile range (IQR) as IQR=Q3-Q1 for that attribute
- Compute
  - *Bottom-whisker* = | Q1 – (1.5 x IQR) |
  - *Upper-whisker* = | Q3 + (1.5 x IQR) |
- Detect attribute value as outlier if
  - it is less than *Bottom-whisker* **OR**
  - it is larger than *Upper-whisker*
- Replace these outlier values with mean/median/mode of the attribute

24

Summary of Data Cleaning

- 80% of data analyst's time spent in cleaning that data
- Data cleaning routines attempt to identify missing values, fill in missing values, smooth out noise while identifying outliers
- One of the biggest data cleaning task is handling missing values
- Among the different methods for filling the missing values
  - Filling by central tendency (mean/median/mode)
  - Filling by interpolation
  - Filling by regression are popular methods
- When data is mostly full of noise, smooth out the data to remove the effect of noise (binning and regression)
- Outliers can be detected using quartiles and IQR
  - Detected outliers can be replaced by mean/median/mode

25