SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Applied Data Science
Data Preprocessing

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

Somaiya TRUST

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*=“42”, *Birthday*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

Missing Values

Col1	Col2	Col3	Col4	Col5
12456	0.99	Male	43	Small
98567	1.23		55	Medium
34567	9999	Female	NA	Large
67231	0.72	Male	35	?

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Treating Missing Values

- Remove rows
- Substitute a specific values
- Interpolate values
- Forward fill
- Backward fill
- Impute

Note

- Many ML algorithm don’t deal with missing values

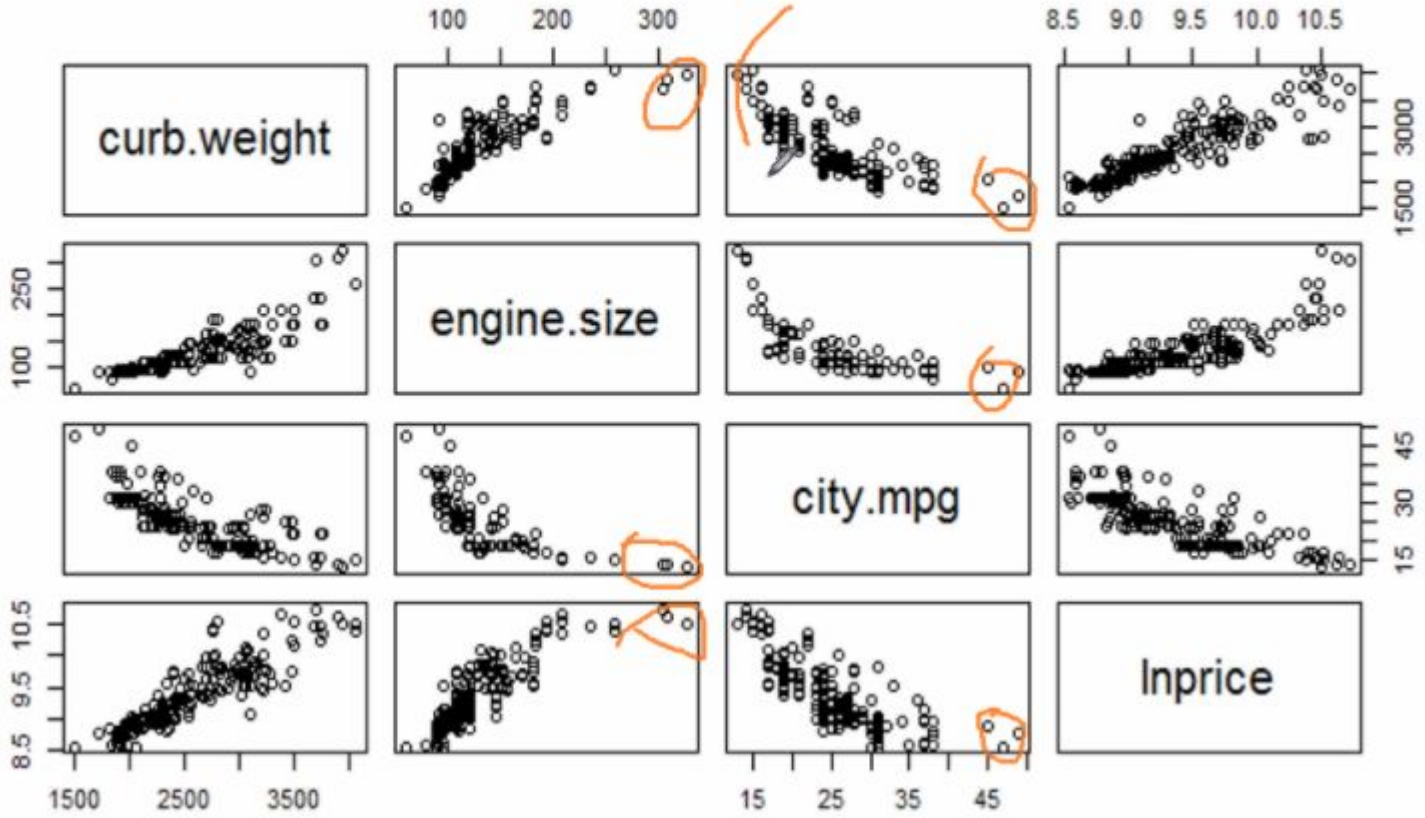
Note

- Repeated Values bias results

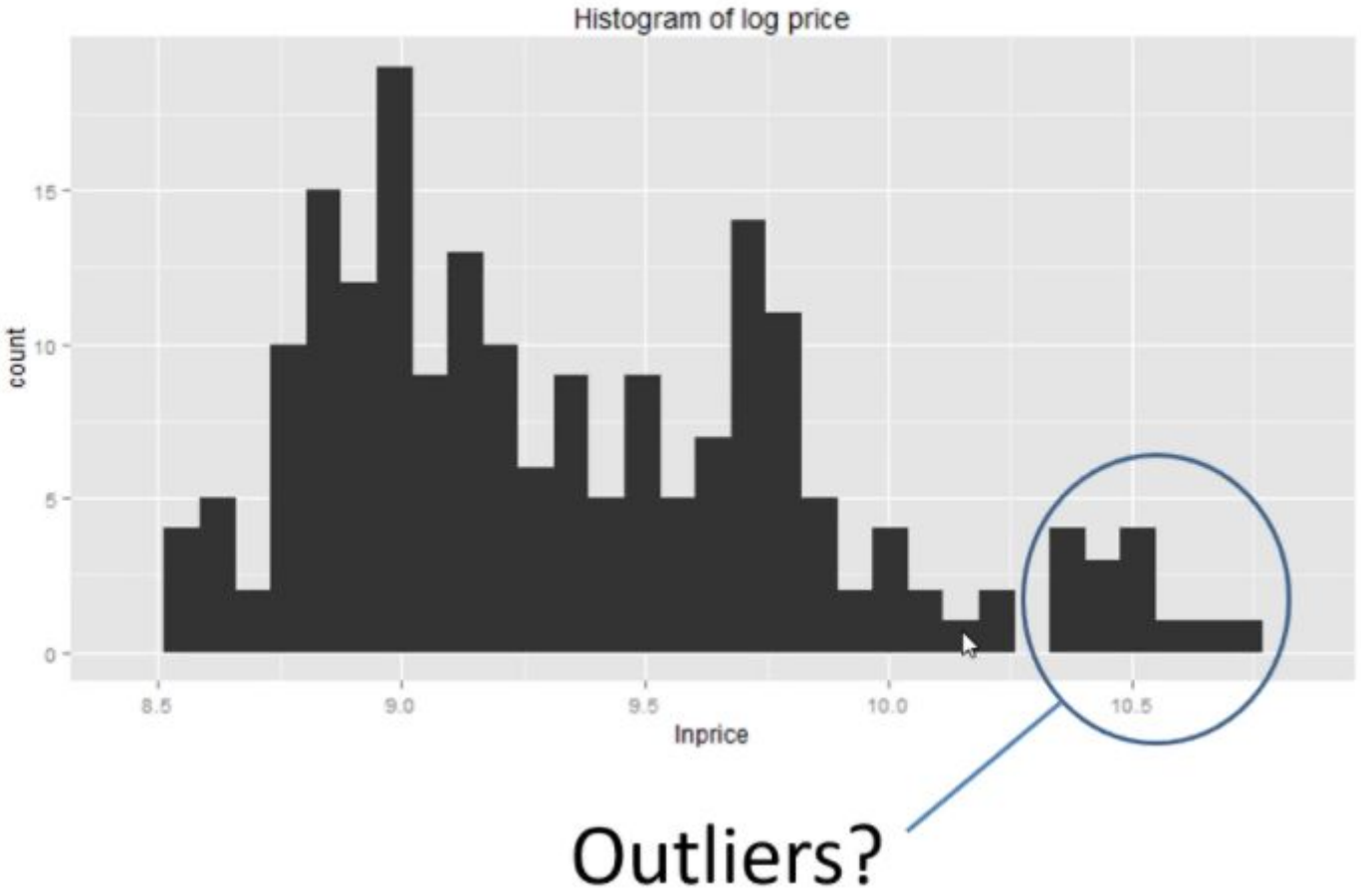
Treating Outlier and Error

- Error and Outliers can bias model training
- Many Possible Sources of errors
  - Erroneous measurement
  - Entry errors
  - Transposed values in table

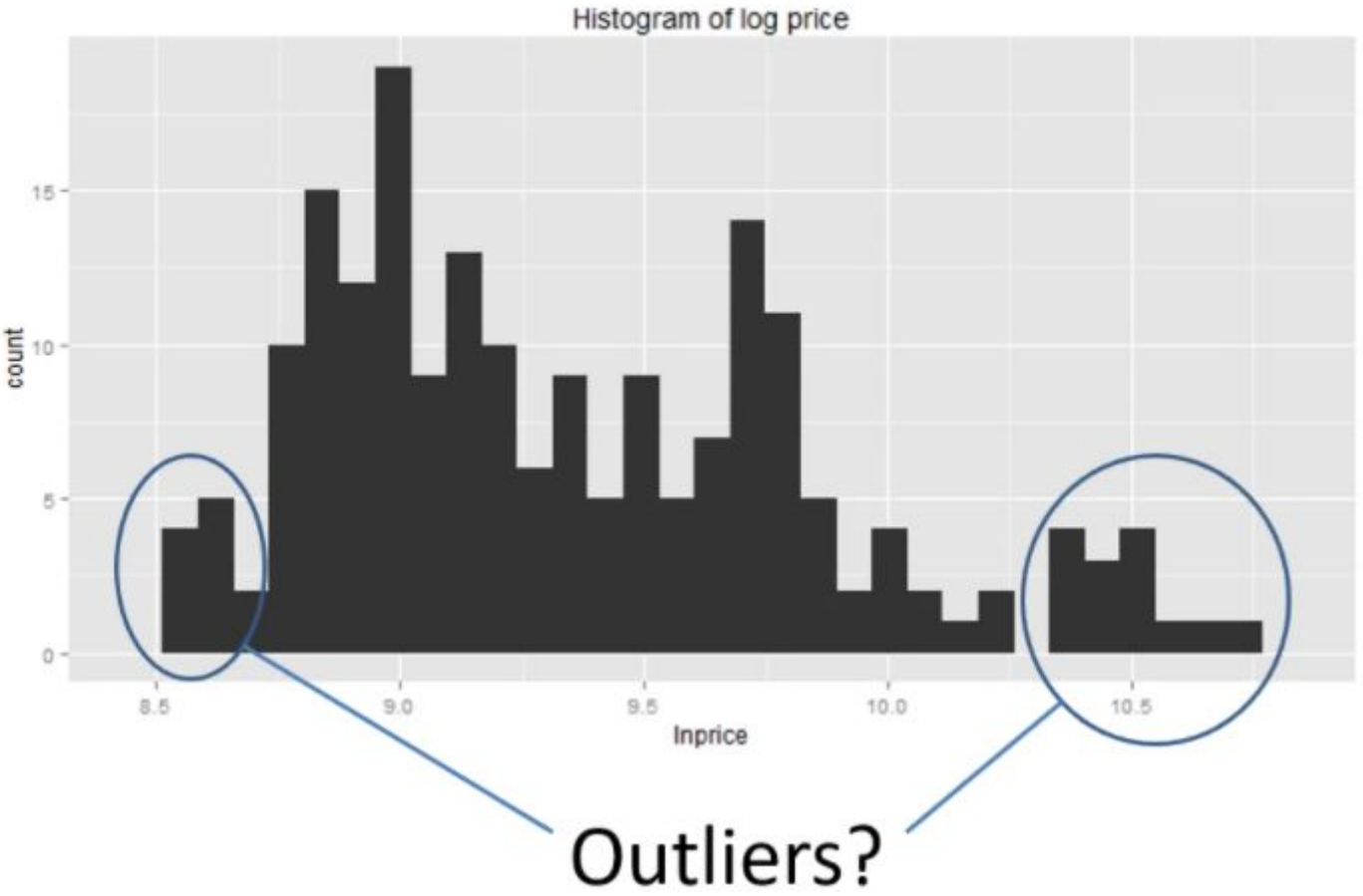
Visualizing Outliers



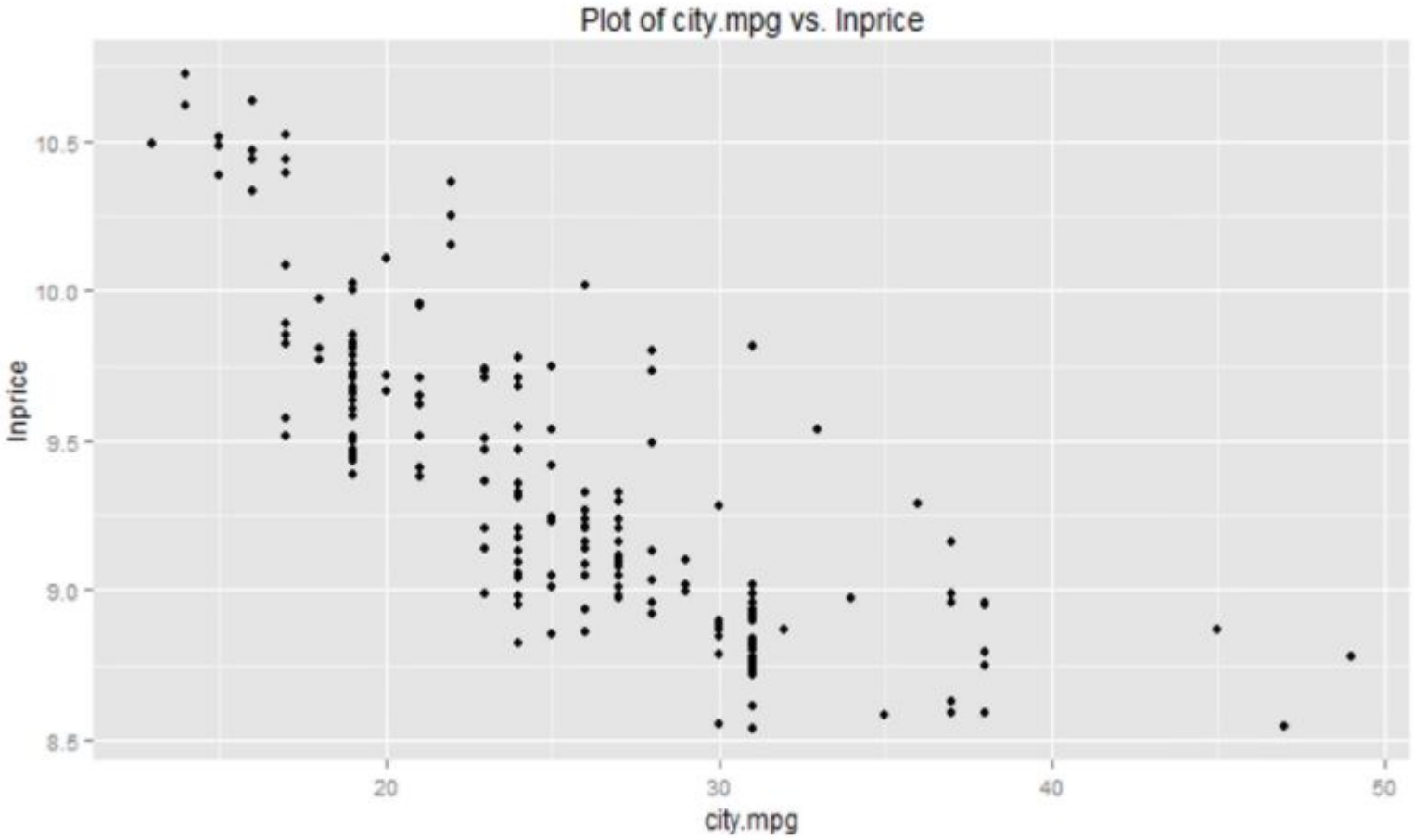
Identify Outliers and Errors



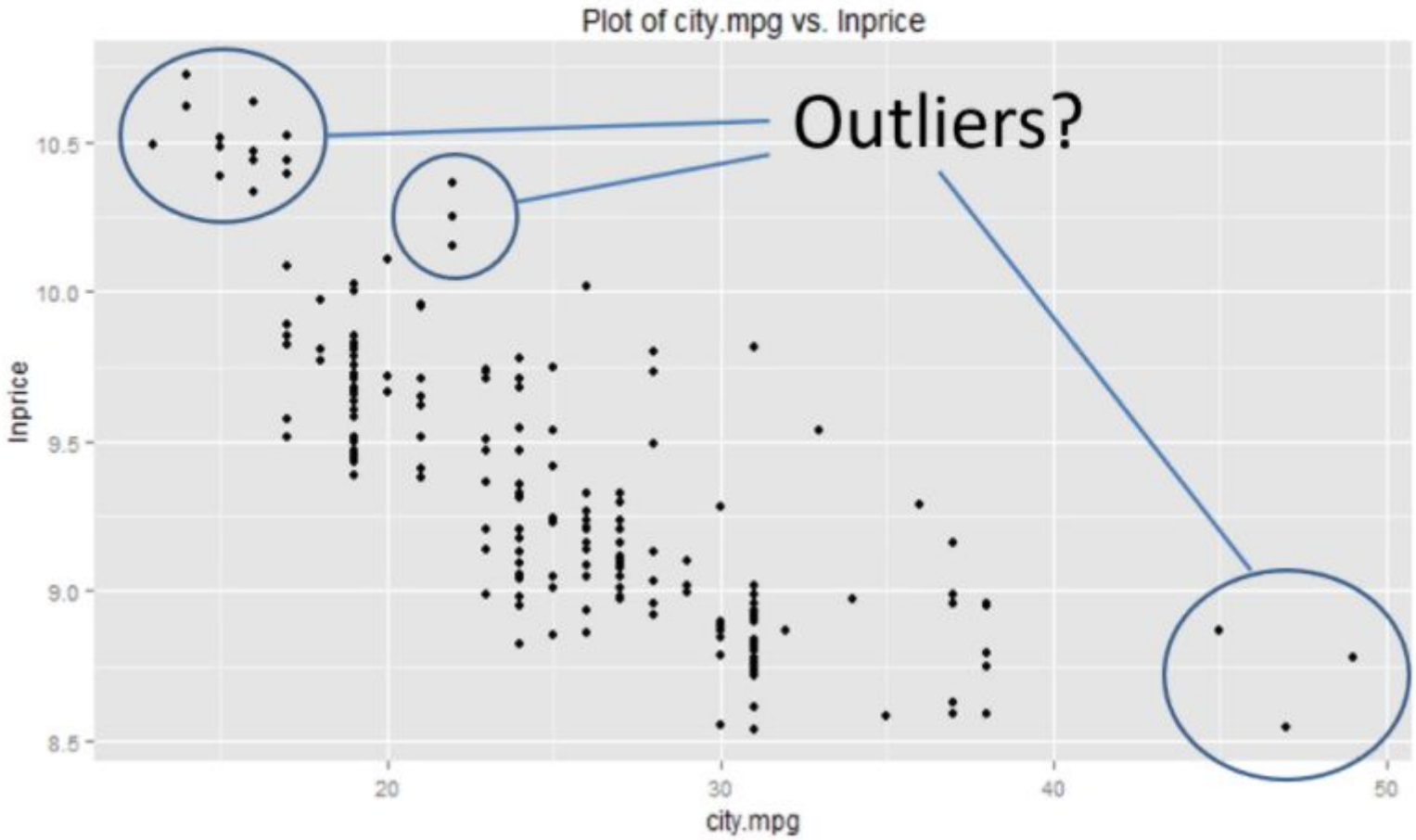
Identify Outliers and Errors



Identify Outliers and Errors



Identify Outliers and Errors



Clean outliers and errors

- Error treatments
  - Censor
  - Trim
  - Interpolate
  - Substitute
- Clip values module

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Smoothing (noisy data)

Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- equal-frequency partitioning

Smoothing (noisy data)

Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- equal-frequency partitioning

bin 1	5,10,11,13
bin 2	15,35,50,55
bin 3	72,92,204,215

Smoothing (noisy data)

Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- equal-frequency partitioning

bin 1	5,10,11,13
bin 2	15,35,50,55
bin 3	72,92,204,215

What is Smoothing by bin mean/median/boundary?

Smoothing(noisy data)

Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- equal-frequency partitioning

bin 1	5,10,11,13
bin 2	15,35,50,55
bin 3	72,92,204,215

What is Smoothing by bin mean/median/boundary?
Replace each bin value is replaced by mean/median/nearest boundary

Smoothing(noisy data)

Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- equal-width partitioning

Smoothing(noisy data)

Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- equal-width partitioning

The width of each interval is (215 – 5)/3 = 70.

bin 1	5,10,11,13,15,35,50,55,72
bin 2	92
bin 3	204,215

Perform Smoothing by bin mean/median/boundary.

- 4,8, 15, 21, 21, 24, 25, 28, 34
- #no of bins=3
- Equal frequency bins
- Bin1:4,8,15
- Bin2:21,21, 24
- Bin3:25,28,34
- Smooting by bin bin boundaries
- Bin1: 4,4,15
- Bin 2:21,21,24
- **Bin 3: 25,25,34**

Encoding categorical data

- dataset$Country = factor(dataset$Country, levels = c('France', 'Spain', 'Germany'),labels = c(1, 2, 3))
- dataset$Purchased = factor(dataset$Purchased, levels = c('No', 'Yes'), labels = c(0, 1))

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

- **Min-Max Normalization:** This transforms the original data linearly.
- Suppose that: min_A is the minima and max_A is the maxima of an attribute, P
- We Have the Formula:

$$v' = \frac{v - min_P}{max_P - min_P}(new\_max_P - new\_min_P) + new\_min_P$$

- Where v is the value you want to plot in the new range.
- v' is the new value you get after normalizing the old value.
- Solved **example**:
  Suppose the minimum and maximum value for an attribute profit(P) are Rs. 10, 000 and Rs. 100, 000. We want to plot the profit in the range [0, 1]. Using min-max normalization the value of Rs. 20, 000 for attribute profit can be plotted to:

$$\frac{20000 - 10000}{100000 - 10000}(1 - 0) + 0 = 0.11$$

And hence, we get the value of v' as 0.11

- **Z-Score Normalization:**In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation
- A value, v, of attribute A is normalized to v' by computing

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

- For **example**:
  Let mean of an attribute P = 60, 000, Standard Deviation = 10, 000, for the attribute P. Using z-score normalization, a value of 85000 for P can be transformed to:

$$\frac{85000 - 60000}{10000} = 2.50$$

And hence we get the value of v' to be 2.5

- **Decimal Scaling:**
- It normalizes the values of an attribute by changing the position of their decimal points
- The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
- A value, v, of attribute A is normalized to v' by computing

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that Max(|v'|) < 1.

- For **example**:
- Suppose: Values of an attribute P varies from -99 to 99.
- The maximum absolute value of P is 99.
- For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on.

Data Discretization

- Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy.
- Data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.
- There are two forms
  - Supervised discretization, and
  - Unsupervised discretization.
- Supervised discretization refers to a method in which the class data is used.
- Unsupervised discretization refers to a method depending upon the way which operation proceeds.
- It means it works on the top-down splitting strategy and bottom-up merging strategy.

- Suppose we have an attribute of Age with the given values

Age	1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13, 18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization can be done by :
  - Interval labels can then be used to replace actual data values
  - Divide the range of a continuous attribute into intervals
  - Reduce data size by discretization(hierarchy)
  - Supervised vs. unsupervised(
  - Split (top-down) vs. merge (bottom-up)(specialization and generalization)
  - Discretization can be performed recursively on an attribute(profit per month from each day)
  - Prepare for further analysis, e.g., classification

Question ?