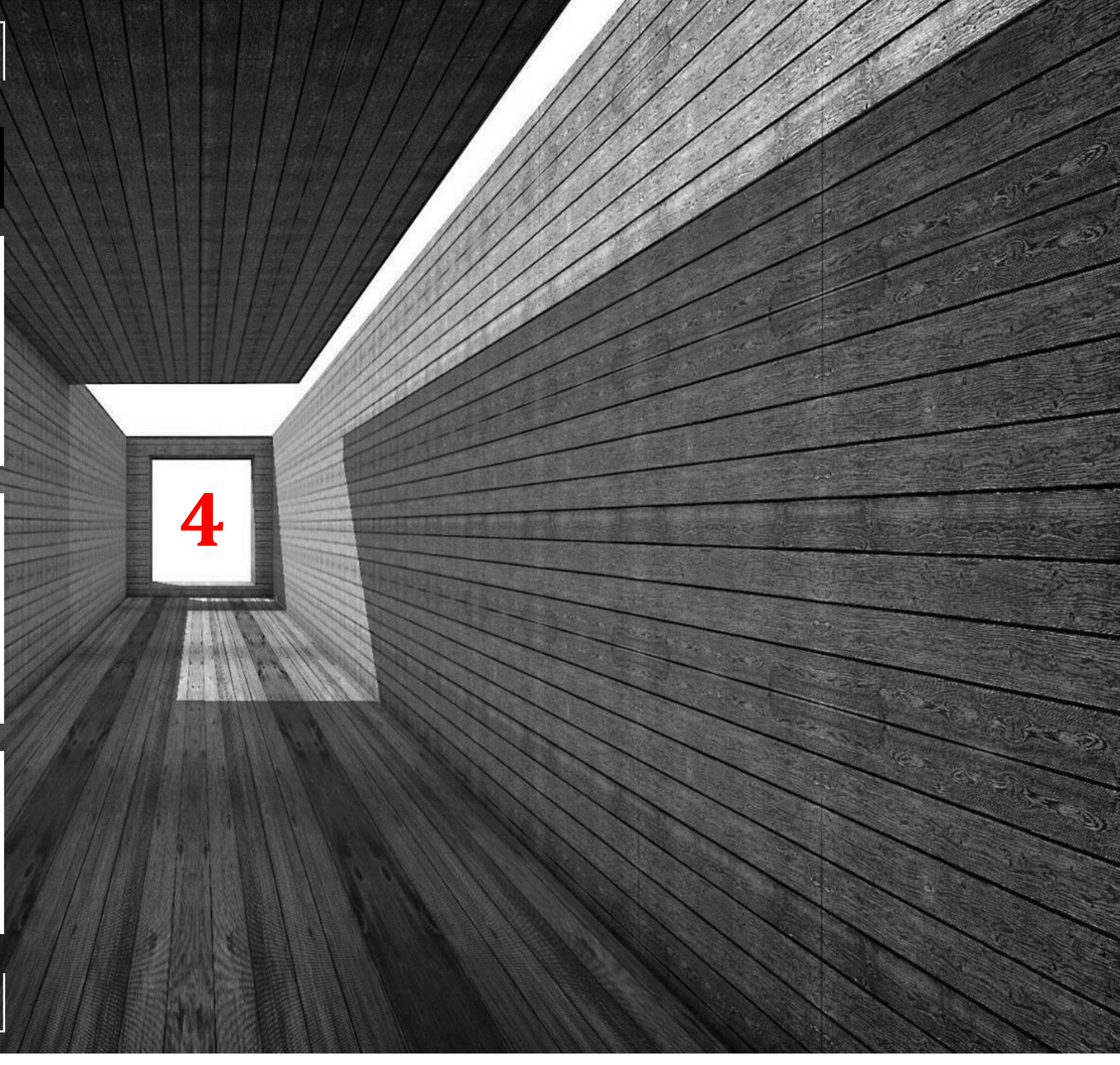
The German University in Cairo

CSEN1095 Data Engineering

Lecture 4
Data Preprocessing II

Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg



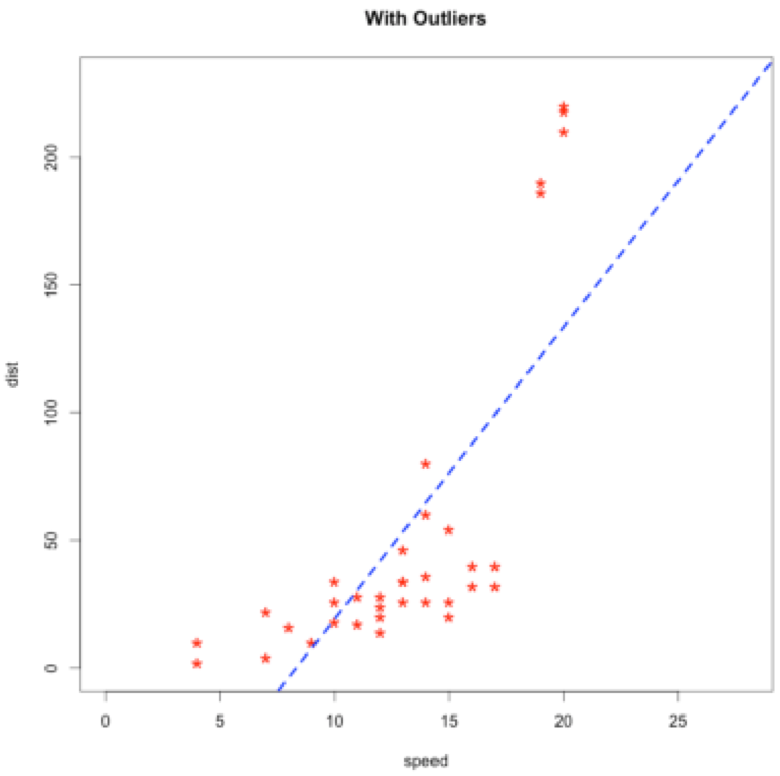
Data Cleaning (Cont.)

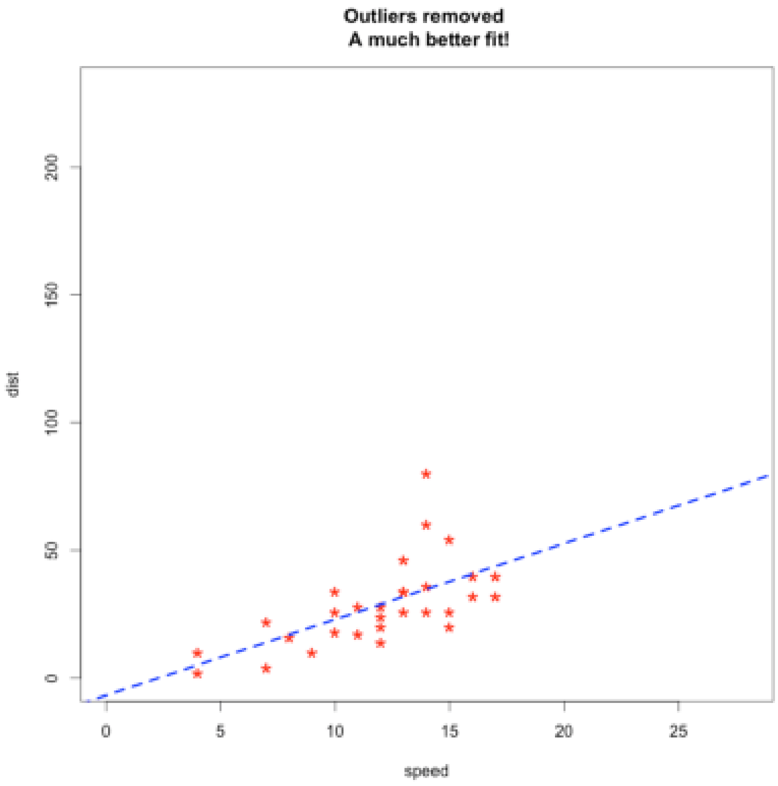
Handling Noisy Data – Outliers

- **Outlier**: an object that ***deviates*** significantly from the rest of the objects
  - *e.g. a student with exceptionally high grades?*
- Normal versus anomalous data objects → *how to define* ***normalcy?***
- Outliers versus noise →***randomness, repetition patterns***
  - Noise should be removed or fixed before outlier detection?
- **Outliers are interesting**: they violate the mechanism that generates normal data, and they may be the most interesting objects to analyze
- **Outlier detection** vs. ***novelty detection***: early stage outlier; but later merged into the model

Why Do We Need to Handle Outliers?

- They increase error variance
- If non-randomly distributed, they can decrease dataset normality
- They can impact assumptions of some ML techniques (e.g. regression)





Handling Noisy Data – Outlier Types

Global
(point anomaly)

Deviate significantly from the rest of the dataset

- Ex: *abnormally large age value*

How to measure deviation?

Contextual
(*conditional outlier*)

Deviate *with respect to context* (e.g. time, location)

- Ex: *Temperature values - 40^o in Cairo, when is it an outlier?*
- ***Contextual*** attributes used to evaluate context
- ***Behavioral*** attributes used to evaluate outlier behavior

How to define context?

Collective

A *subset* of objects collectively deviates significantly from the dataset

- Ex: *Multiple order delays*

How to define group behavior?

Handling Noisy Data – Outlier Detection Methods

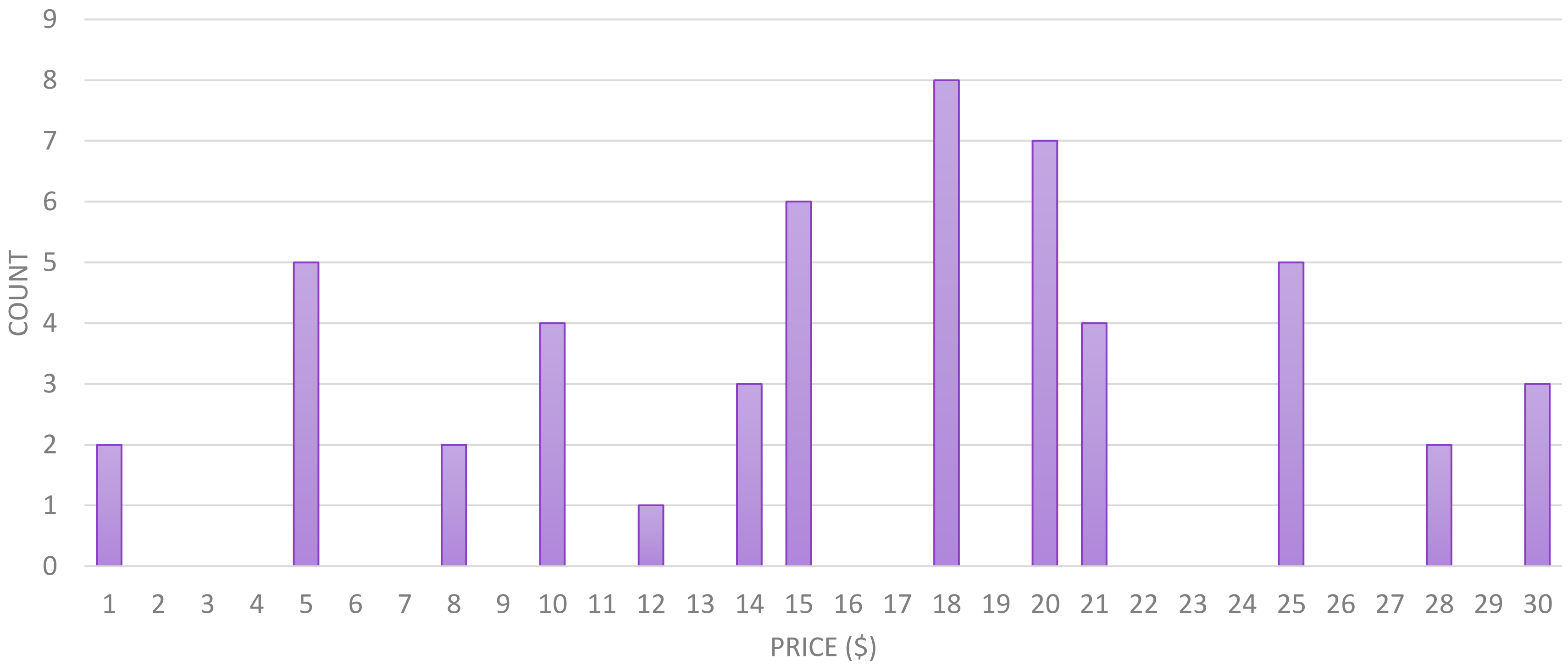
- Statistical → e.g. boxplots, histograms
- Model-based → e.g. regression
- Distributional → e.g. clustering

Detecting Noisy Data – **Statistical Outlier Detection**

- **Parametric:** assume normal data is generated by a distribution with parameter θ
  - *PDF* of distribution $f(x, \theta)$ yields probability that x is generated by distribution → ***smaller means outlier***
  - For **univariate outliers** → *boxplots* → parameters are *median* and *IQR*
  - For multivariate outliers → χ^2*-statistic* → parameter is *mean*
- **Nonparametric:** learn normal model from input data
  - *histograms*

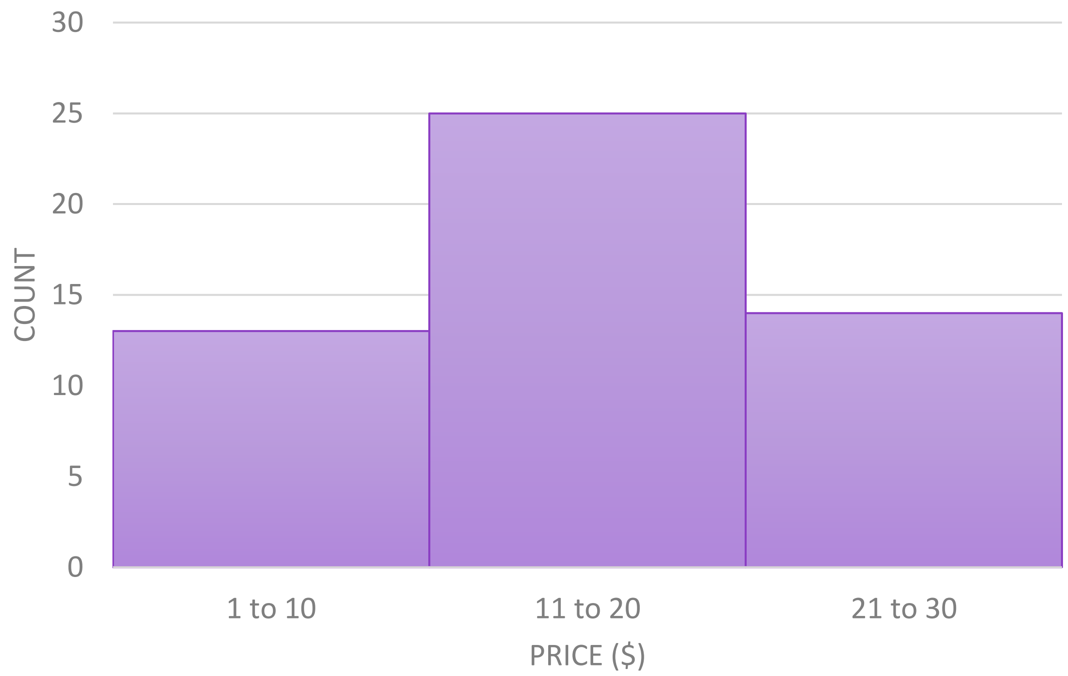
Histograms

- An approximate representation of the distribution of numerical data
  - e.g. original observations of number of products for each price value in a store
- Divide the entire range of values into a series of intervals – then count how many values fall into each interval



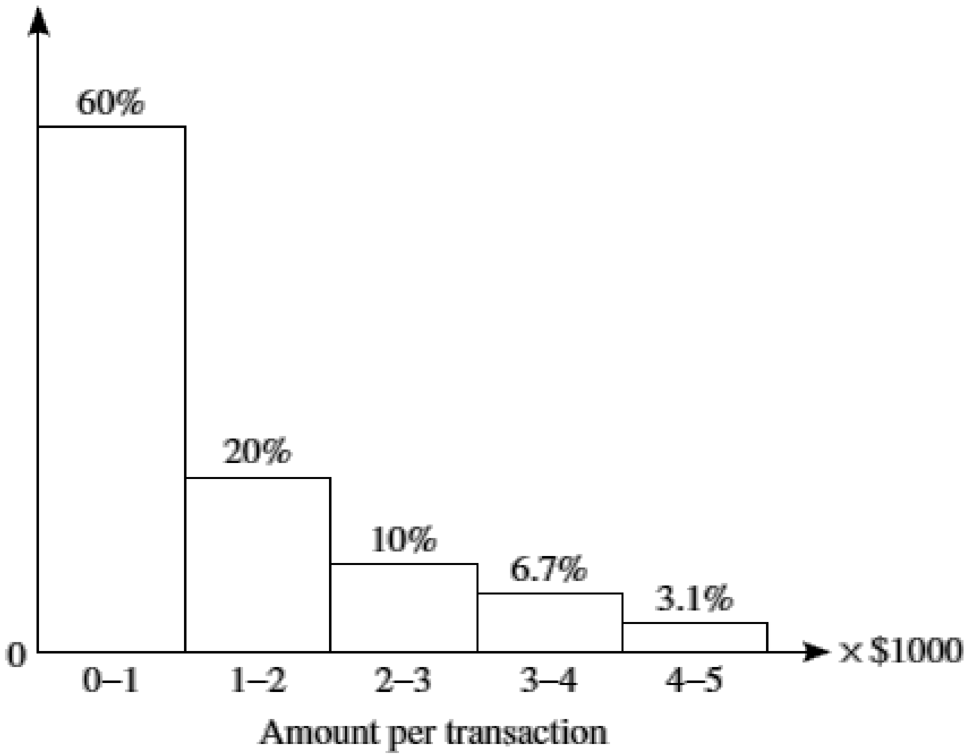
Binning

- You can "bin/bucket" the range of values and further merge observations
  - Bins are consecutive, non-overlapping intervals
- To reduce further → change width of bins/buckets (e.g. from example $10 range)



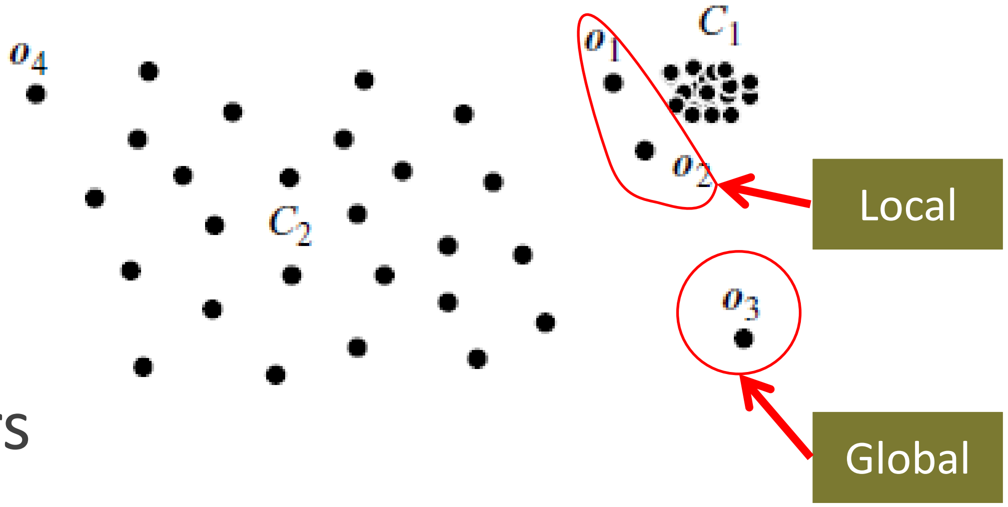
Detecting Noisy Data – Statistical Outlier Detection

- **Histograms** → e.g. a transaction with the amount of $7500 is considered an outlier
  - Does not belong to any of the bins (0.2% of transactions > $5000)
- **Problem** → hard to choose an appropriate bin size for histogram
  - Too small bin size → normal objects in empty/rare bins, *false positives*
  - Too big bin size → outliers in some frequent bins, *false negatives*



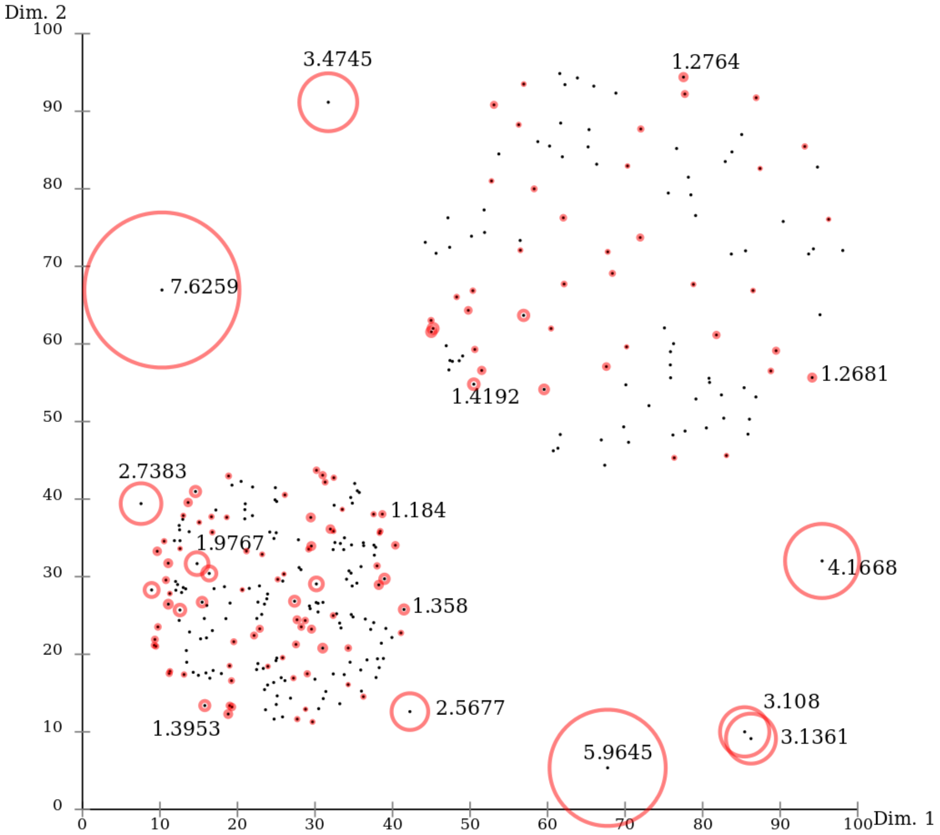
Detecting Noisy Data – **Distributional Outlier Detection**

- **Distance-based** → for object $\boldsymbol{o}$, examine the number of other objects in its ***r-neighborhood***
  - if < fraction threshold π then flag $\boldsymbol{o}$ as outlier
  - $\boldsymbol{r}$ is a distance threshold, π is a fraction threshold (min # objects needed in neighborhood)
- **Density-based** → for object $\boldsymbol{o}$, examine its density *relative to* density of its local neighbors
  - A local outlier factor (LOF) is computed in terms of the *K-Nearest Neighbors* of an object in comparison to its neighbors



KNN as an Outlier Detector

- For each object, identify distance to its kth nearest neighbor, use distance as an outlier score
- Define a distance threshold and flag as outliers the objects whose outlier score (i.e. kth NN distance) is larger than the distance threshold
- Alternatively, use average distance of the k nearest neighbors as object's outlier score, and compare to threshold



LOF scores visualized

How to Handle Outliers?

- Delete outlier observations/objects
  - If # outlier observations is small
  - If outliers are random and not interesting phenomena
- Transform entire attribute to smooth out outliers
  - Binning
  - Log transformation
- Impute outlier values with mean or estimated value
- Keep and use outlier analysis methods
  - Clustering

Handling Noisy Data – **Binning/Bucketing**

Binning → smooth a sorted data value by consulting its "neighborhood"

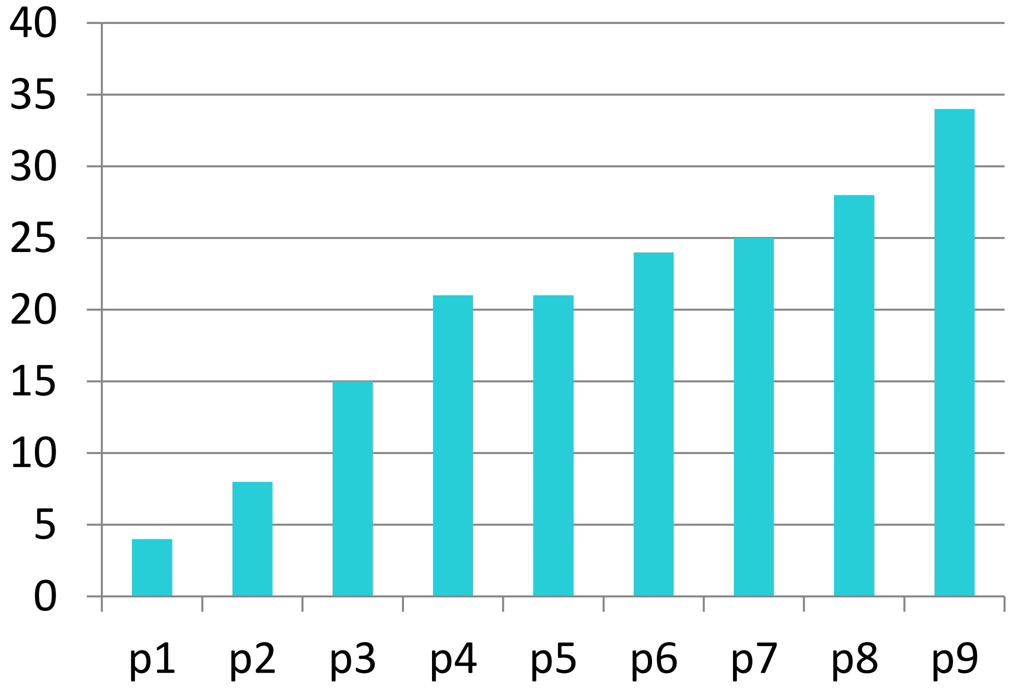
- sorted values are partitioned into a number of "buckets," or *bins* → *local* smoothing
- *equal-depth* bins → each bin has same frequency of values
- *equal-width* bins → interval range per bin is constant
- Either method produces uniform bins
  - Smoothing by *bin means* → each bin value is replaced by the bin *mean*
  - Smoothing by *bin medians* → each bin value is replaced by the bin *median*

➢ Cluster-based binning technique will be discussed in transformation

Handling Noisy Data – **Binning**

Example: Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

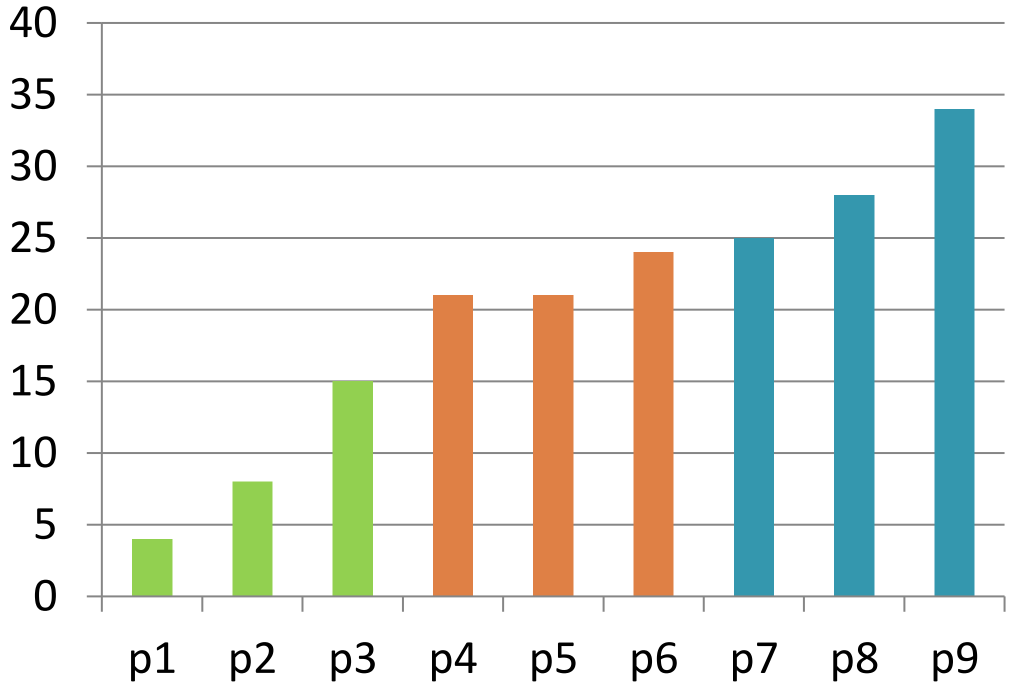
Partition into (equal-frequency) bins
Bin 1: 4, 8, 15 Bin 2: 21, 21, 24 Bin 3: 25, 28, 34
Smoothing by bin means
Bin 1: 9, 9, 9 Bin 2: 22, 22, 22 Bin 3: 29, 29, 29



Handling Noisy Data – **Binning**

Example: Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

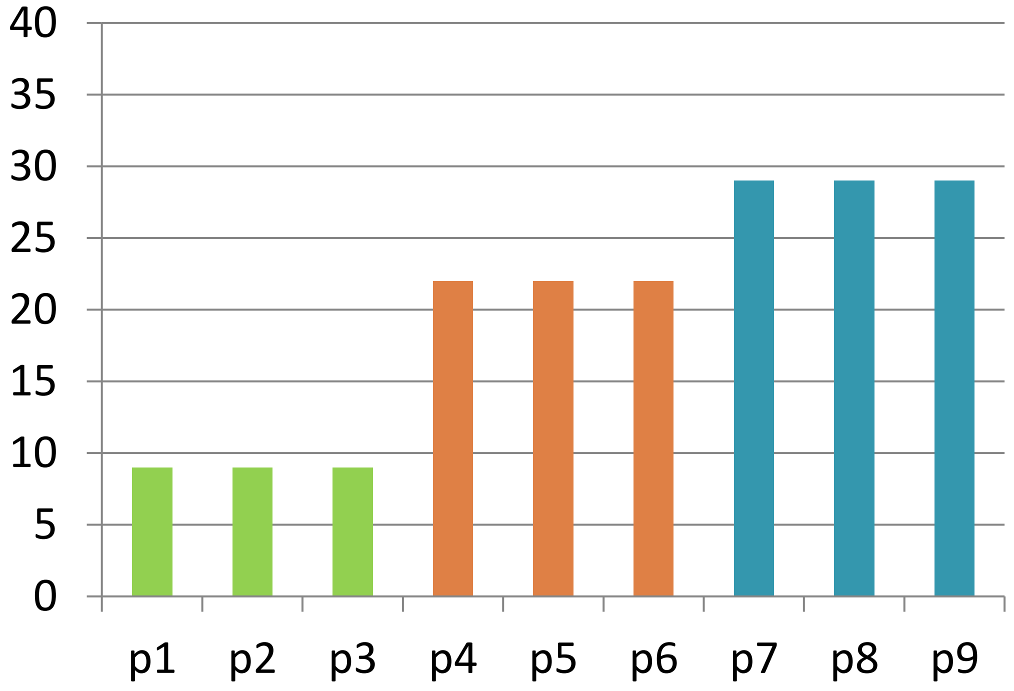
Partition into (equal-frequency) bins
Bin 1: 4, 8, 15 Bin 2: 21, 21, 24 Bin 3: 25, 28, 34
Smoothing by bin means
Bin 1: 9, 9, 9 Bin 2: 22, 22, 22 Bin 3: 29, 29, 29



Handling Noisy Data – **Binning**

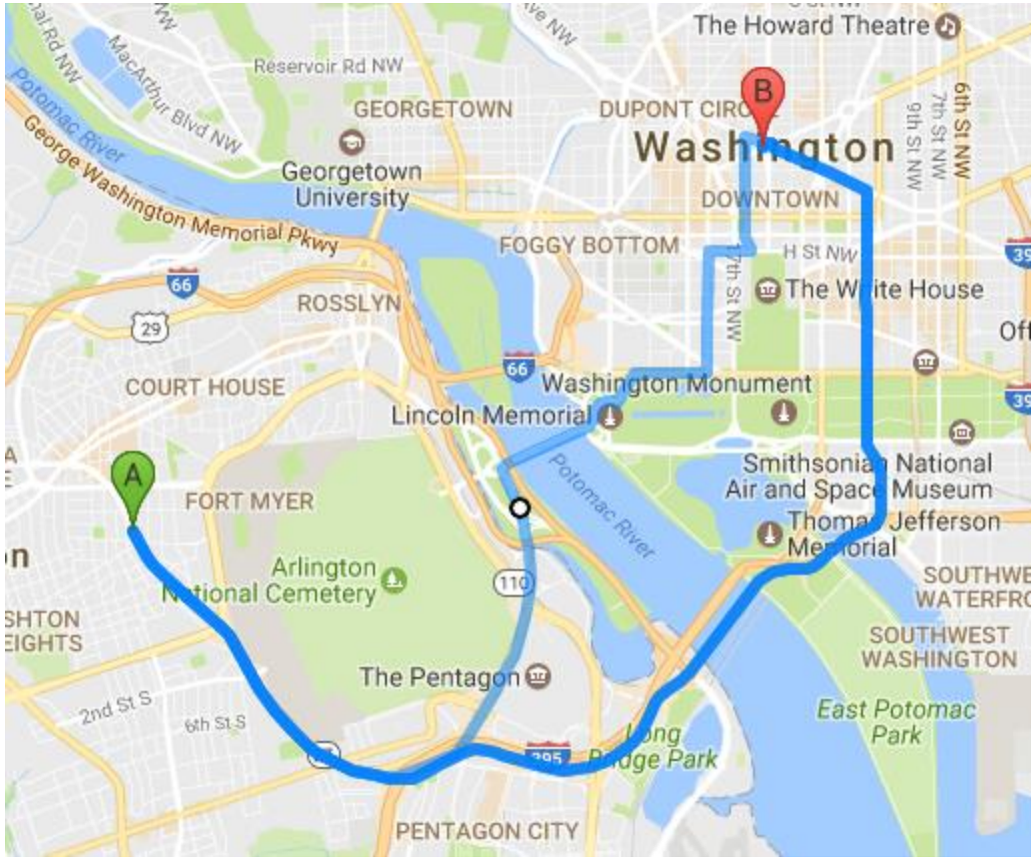
Example: Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins
Bin 1: 4, 8, 15 Bin 2: 21, 21, 24 Bin 3: 25, 28, 34
Smoothing by bin means
Bin 1: 9, 9, 9 Bin 2: 22, 22, 22 Bin 3: 29, 29, 29



Food for Thought

NaNs and Noise in location data

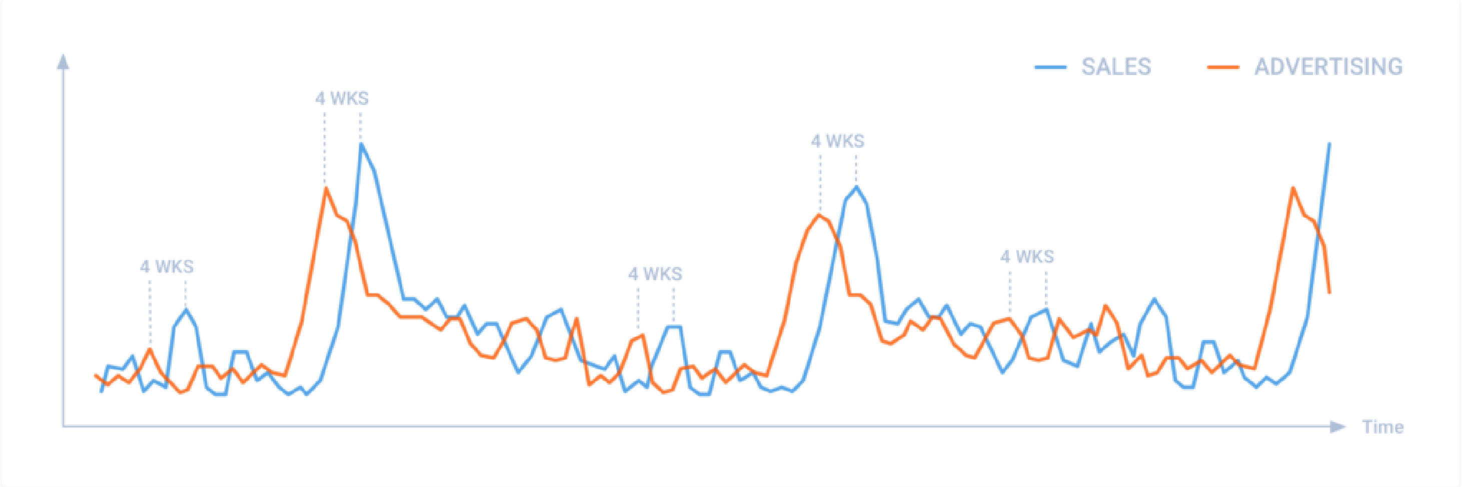


2/2/2008 13:38	116.48088	39.89025
2/2/2008 13:43	116.48087	39.89023
2/2/2008 13:53	116.48087	39.8902
2/2/2008 13:58	116.48088	39.89019
2/2/2008 14:03	116.47666	39.88997
2/2/2008 14:08	116.46695	39.89574
2/2/2008 14:13	116.46694	39.89579
2/2/2008 14:18	116.46695	39.8958
2/2/2008 14:23	116.46708	39.8956
2/2/2008 14:33	116.46701	39.89598
2/2/2008 14:38	116.46699	39.89597
2/2/2008 14:43	116.46698	39.89596
2/2/2008 14:48	116.46697	39.89595
2/2/2008 14:48	116.46697	39.89595
2/2/2008 14:53	116.46711	39.8954
2/2/2008 14:58	116.46685	39.89582
2/2/2008 15:03	116.467	39.89588
2/2/2008 15:08	116.4669	39.89587
2/2/2008 15:13	116.46689	39.89584

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF)

Food for Thought

NaNs and Noise in time series data



	date	open	high	low	close	volume	Name
1	2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
2	2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
3	2013-02-12	14.45	14.51	14.1	14.27	8126000	AAL
4	2013-02-13	14.3	14.94	14.25	14.66	10259500	AAL
5	2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL
6	2013-02-15	13.93	14.61	13.93	14.5	15628000	AAL
7	2013-02-19	14.33	14.56	14.08	14.26	11354400	AAL
8	2013-02-20	14.17	14.26	13.15	13.33	14725200	AAL
9	2013-02-21	13.62	13.95	12.9	13.37	11922100	AAL
10	2013-02-22	13.57	13.6	13.21	13.57	6071400	AAL
11	2013-02-25	13.6	13.76	13.0	13.02	7186400	AAL
12	2013-02-26	13.14	13.42	12.7	13.26	9419000	AAL

- More on imputation can be found in the book: Flexible Imputation of Missing Data (available online at https://stefvanbuuren.name/fimd/)
- Or try the Kaggle project here
- Additional read

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF)

GUC

Thank You