Data Preprocessing

Data Proprocesing

a. This labor-intensive phase covers all aspects of preparing the final data set, which shall be used for subsequent phases, from the initial, raw, dirty data.

b. Select the cases and variables you want to analyse, and that are appropriate for your analysis.

c. Perform transformations on certain variables, if needed.

d. Clean the raw data so that it is ready for the modeling tools.

Data Preprocessing

Data preparation

- *Evaluate* the quality of the data,
- Clean the raw data,
- Deal with missing data, and
- Perform transformations on certain variables

Data Understanding

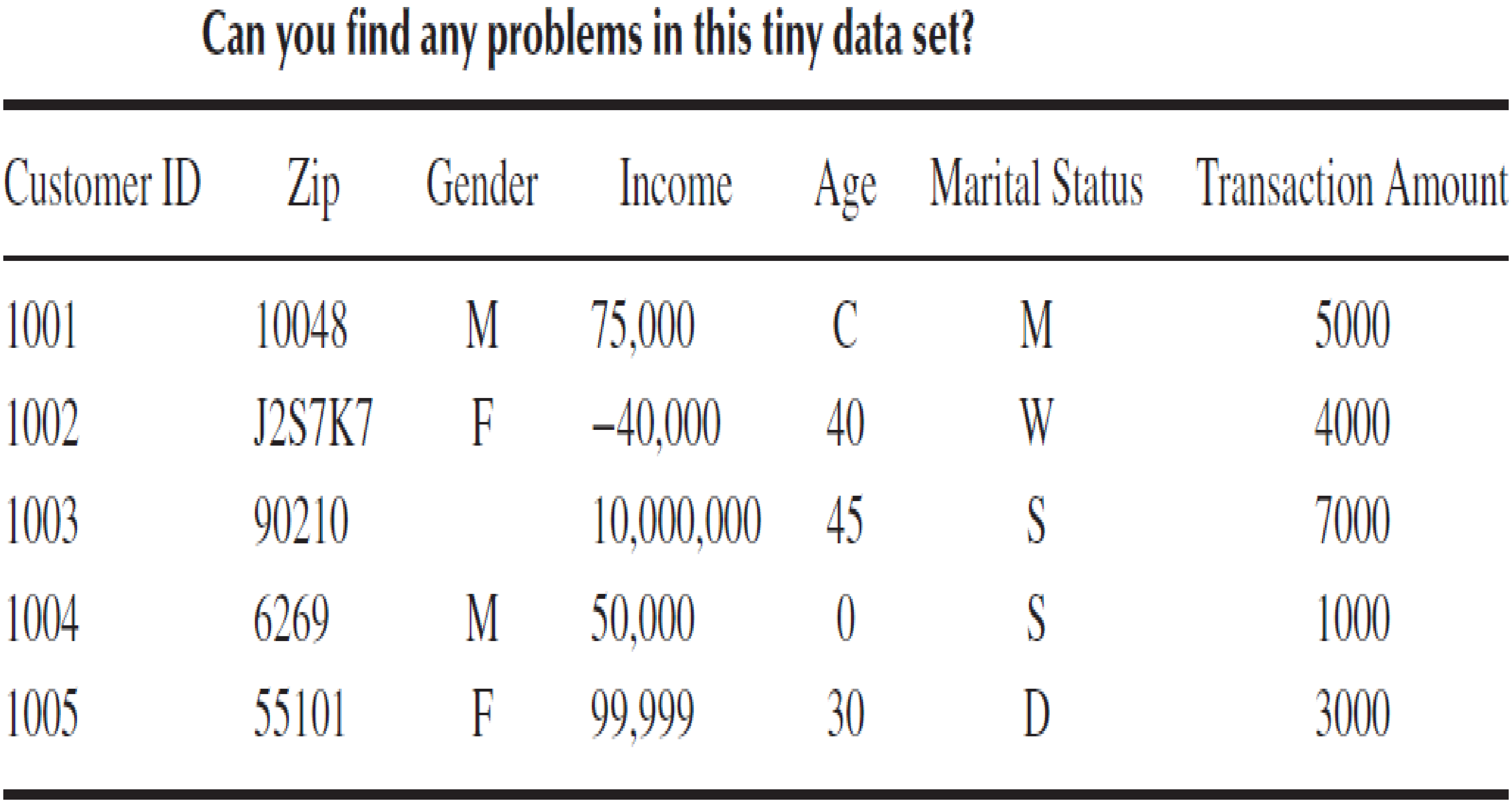
- *Exploratory Data Analysis (EDA)*

WHY DO WE NEED TO PREPROCESS THE DATA?

- fields that are obsolete or redundant;
- missing values;
- outliers;
- data in a form not suitable for the data mining models;
- values not consistent with policy or common sense.

Depending on the data set, data preprocessing alone can account for **10–60%** of all the time and effort for the entire data science process.

DATA CLEANING

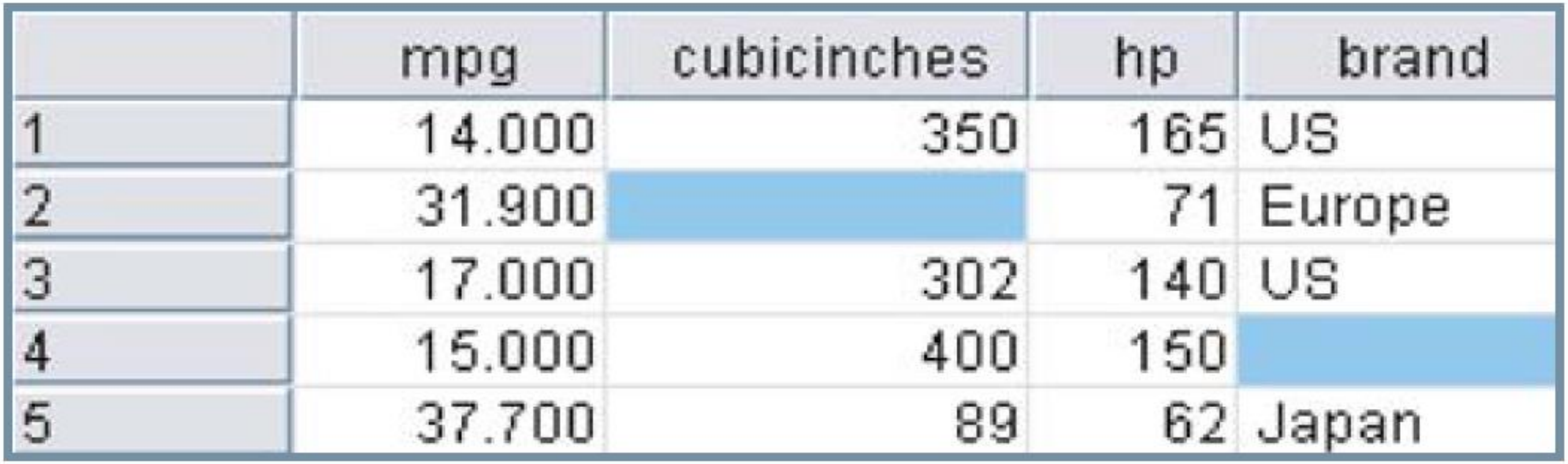


Can you find any problems in this tiny data set?

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75,000	C	M	5000
1002	J2S7K7	F	−40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

HANDLING MISSING DATA

- ❑ Missing data is a problem that continues to plague data analysis methods.
- ❑ Even as our analysis methods gain sophistication, but still encounter missing values in fields



	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900		71	Europe
3	17.000	302	140	US
4	15.000	400	150	
5	37.700	89	62	Japan

Common method of “handling” missing values is simply to omit the records from analysis. May be dangerous … lead to a biased subset of the data.

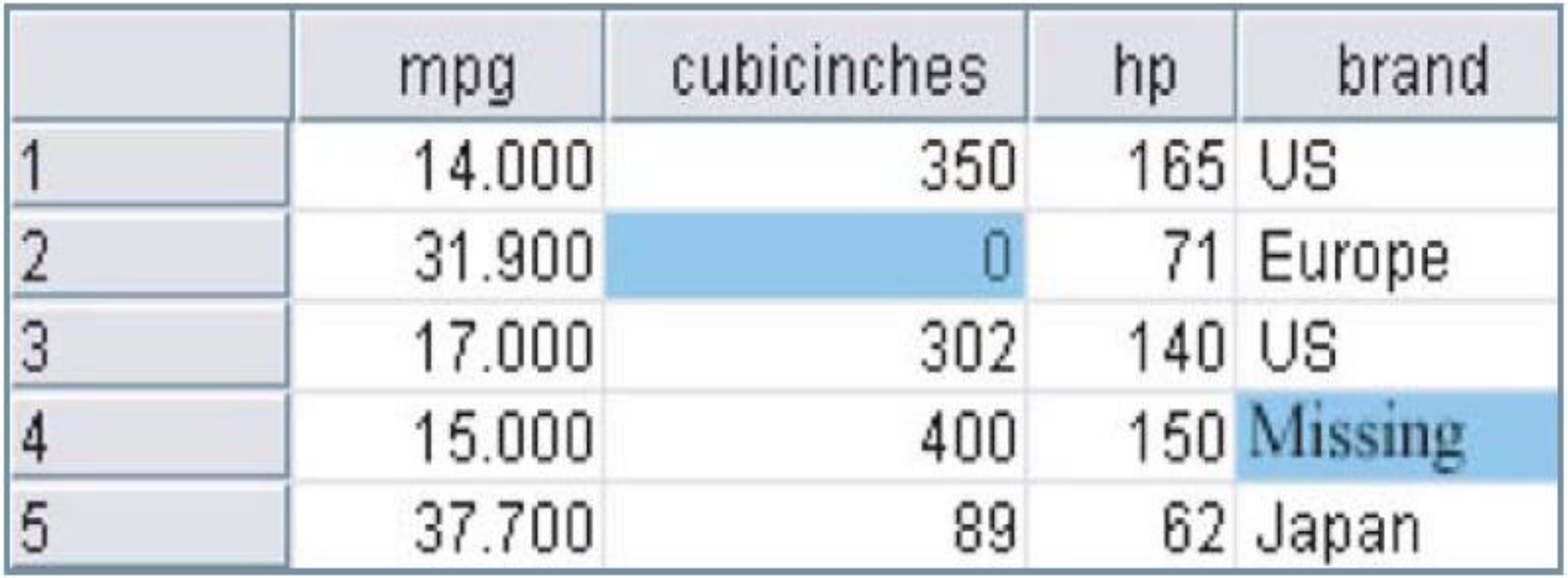
HANDLING MISSING DATA Cont…

Common criteria for choosing replacement values for missing data:

1. Replace the missing value with some constant, specified by the analyst.
2. Replace the missing value with the field mean (for numeric variables) or the mode (for categorical variables).
3. Replace the missing values with a value generated at random from the observed distribution of the variable.
4. Replace the missing values with **imputed** values based on the other characteristics of the record

HANDLING MISSING DATA Cont…

❑ Result of replacing the missing values with the **constant 0** for the numerical variable *cubicinches* and the label ***missing*** for the categorical variable *brand*.

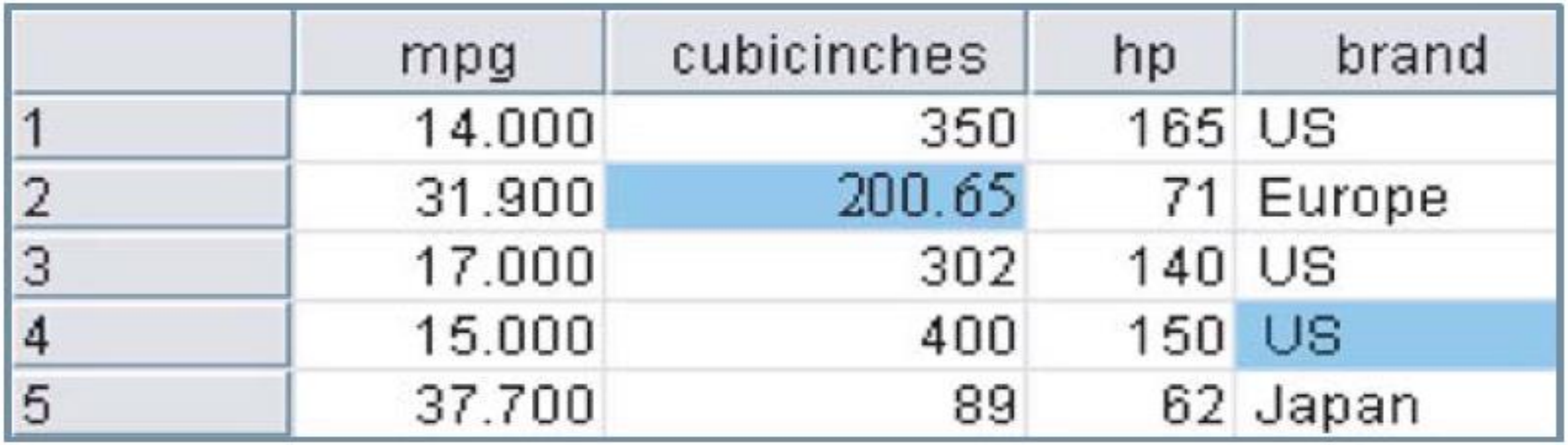


	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	0	71	Europe
3	17.000	302	140	US
4	15.000	400	150	Missing
5	37.700	89	62	Japan

Replacing missing field values with user-defined constants.

HANDLING MISSING DATA Cont…

- Missing values may be replaced with the respective field means and modes



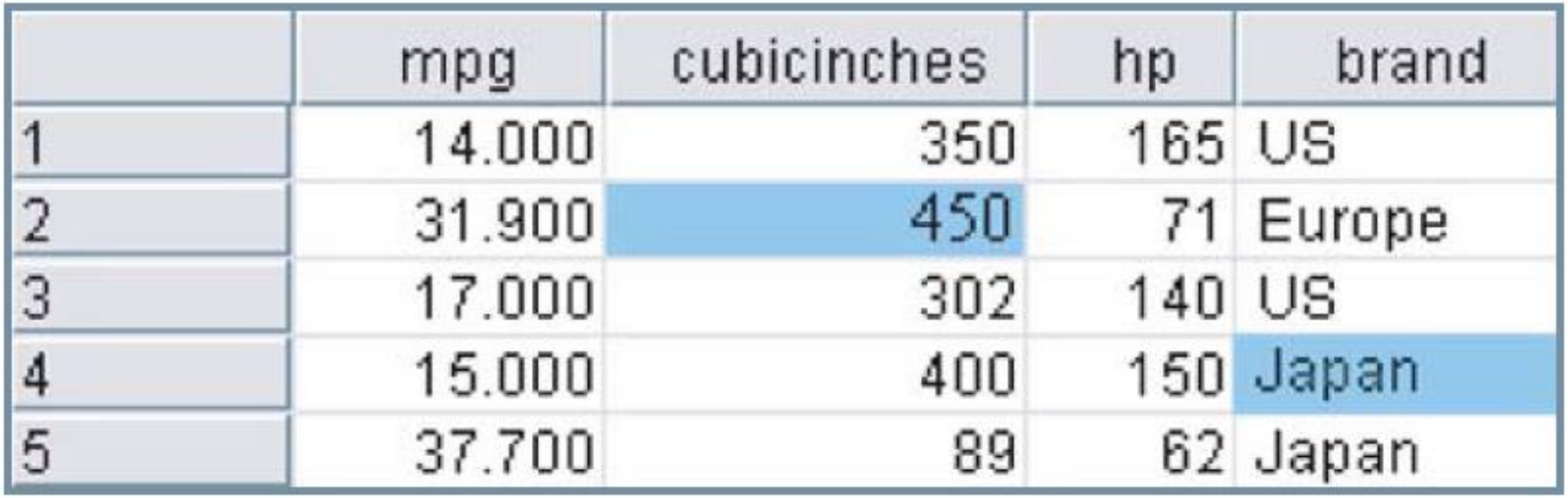
	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	200.65	71	Europe
3	17.000	302	140	US
4	15.000	400	150	US
5	37.700	89	62	Japan

Replacing missing field values with means or modes.

- May not always be the best choice
- Observed that mean is greater than the 81st percentile
- Measures of spread will be artificially reduced
- replacing missing values is a gamble

HANDLING MISSING DATA Cont…

- ❑ Missing values replaced with values generated at random from the observed distribution of the variable



	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	450	71	Europe
3	17.000	302	140	US
4	15.000	400	150	Japan
5	37.700	89	62	Japan

Replacing missing field values with random draws from the distribution of the variable.

- ❑ benefit is measures of center and spread should remain closer to the original
- ❑ no guarantee that the resulting records would make sense

HANDLING MISSING DATA Cont…

Data imputation methods

- In data imputation, we ask “What would be the most likely value for this missing value, given all the other attributes for a particular record?”
- For instance, an American car with 300 cubic inches and 150 horsepower would probably be expected to have more cylinders than a Japanese car with 100 cubic inches and 90 horsepower.
- This is called imputation of missing data.
- Tools needed, such as multiple regression or classification and regression trees.

HANDLING MISSING DATA (more methods) Cont…

KNN (K Nearest Neighbors)

- Machine Learning algorithms can be used to handle missing data/for data imputation
- ML techniques like KNN, XGBoost and Random Forest can be used
- k neighbors are chosen based on some distance measure and their average is used as an imputation estimate.
- **Method requires**
  - ✓ Selection of the number of nearest neighbors, and
  - ✓ Distance metric
- KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes

Distance metric varies according to type of data:

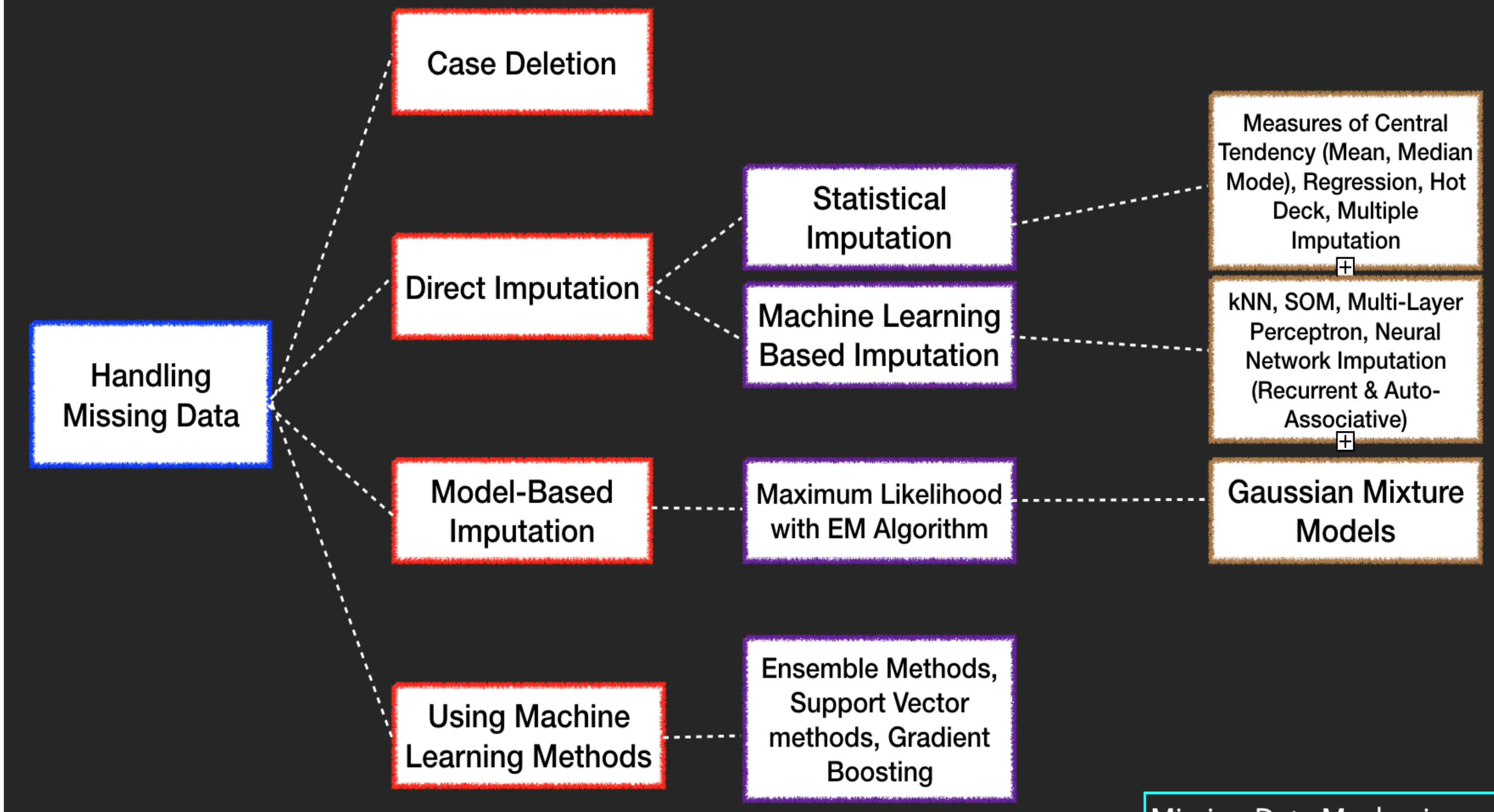
1. Continuous Data: The commonly used distance metrics for continuous data are **Euclidean, Manhattan and Cosine**
2. Categorical Data: **Hamming distance** is generally used in this case.

HANDLING MISSING DATA (more methods) Cont…

Xgboost learns for missing values

- Once a tree structure has been trained it isn’t too hard to also consider the presence of missing values in the test set: it’s enough to attach a default direction to each decision node.
- Optimum default direction is determined and the missing values will go in that direction

Handling Missing Data



Handling Missing Data

Case Deletion

Direct Imputation

Statistical Imputation

Measures of Central Tendency (Mean, Median Mode), Regression, Hot Deck, Multiple Imputation

Machine Learning Based Imputation

kNN, SOM, Multi-Layer Perceptron, Neural Network Imputation (Recurrent & Auto-Associative)

Model-Based Imputation

Maximum Likelihood with EM Algorithm

Gaussian Mixture Models

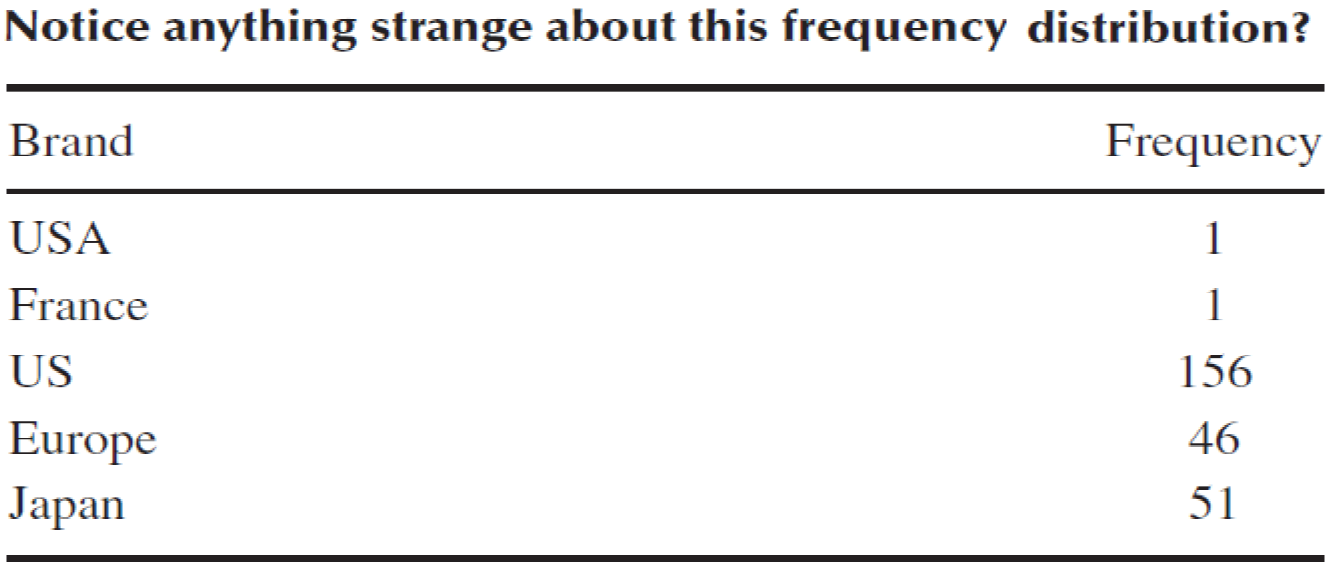
Using Machine Learning Methods

Ensemble Methods, Support Vector methods, Gradient Boosting

Missing Data Mechanism

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing Not at Random (MNAR)

IDENTIFYING MISCLASSIFICATIONS



Notice anything strange about this frequency distribution?

Brand	Frequency
USA	1
France	1
US	156
Europe	46
Japan	51

- Frequency distribution shows five classes
- Two of the classes, USA and France, have a count of only one automobile each.
- Two of the records have been inconsistently classified with respect to the origin of manufacture
- To maintain consistency, the record with origin USA should have been labelled US, and France should have been labelled Europe.

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS

- **Outliers** are extreme values that go against the trend of the remaining data.

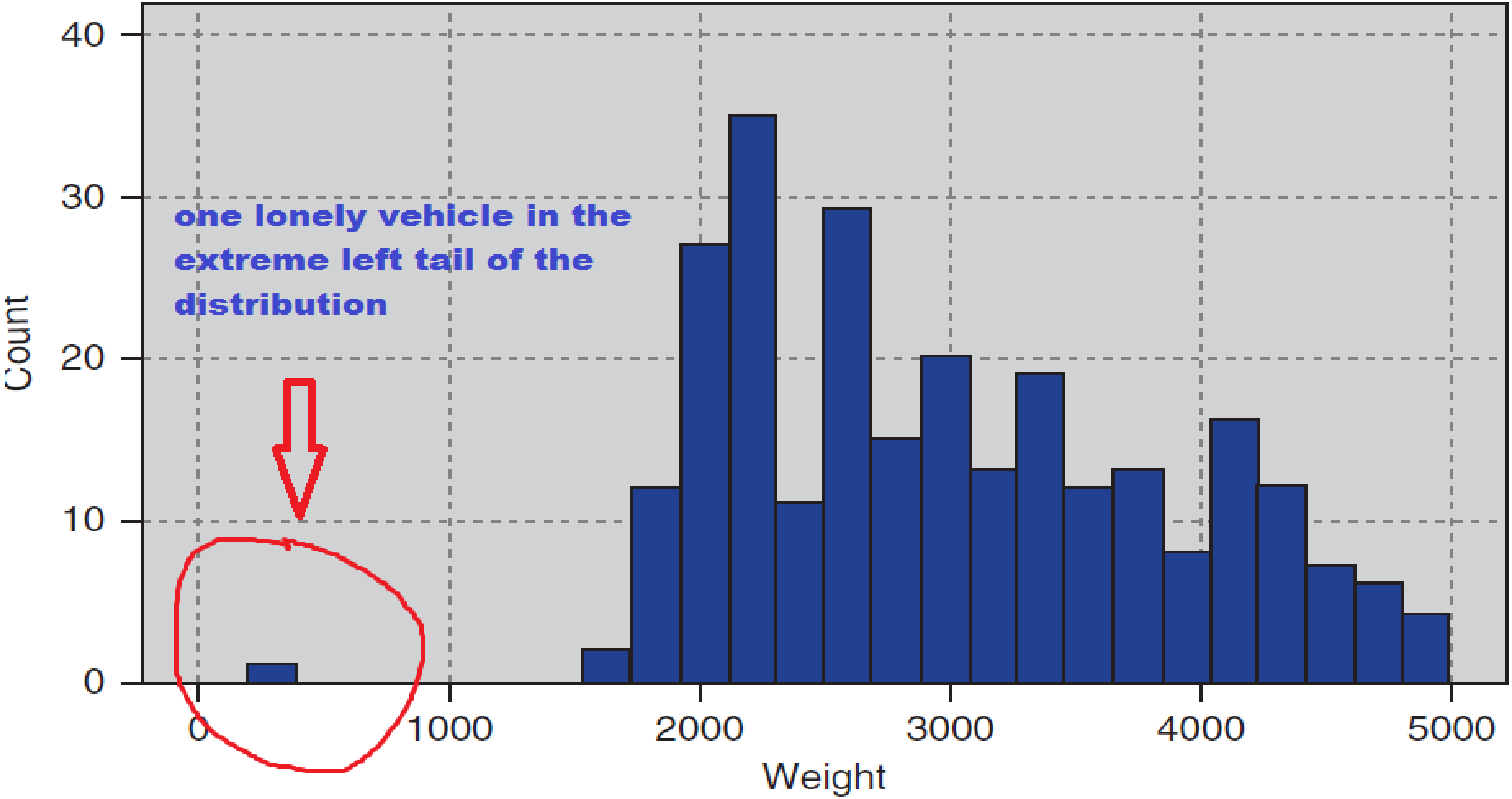
- **Wikipedia Definition** -In statistics, an **outlier** is an observation point that is distant from other observations.

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS

- In data mining, **outlier detection** is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- Identifying outliers is important because they may represent errors in data entry.
- Statistical methods are sensitive to the presence of outliers, and may deliver unreliable results
- Method for identifying outliers for numeric variables is to examine a *histogram* of the variable.

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS

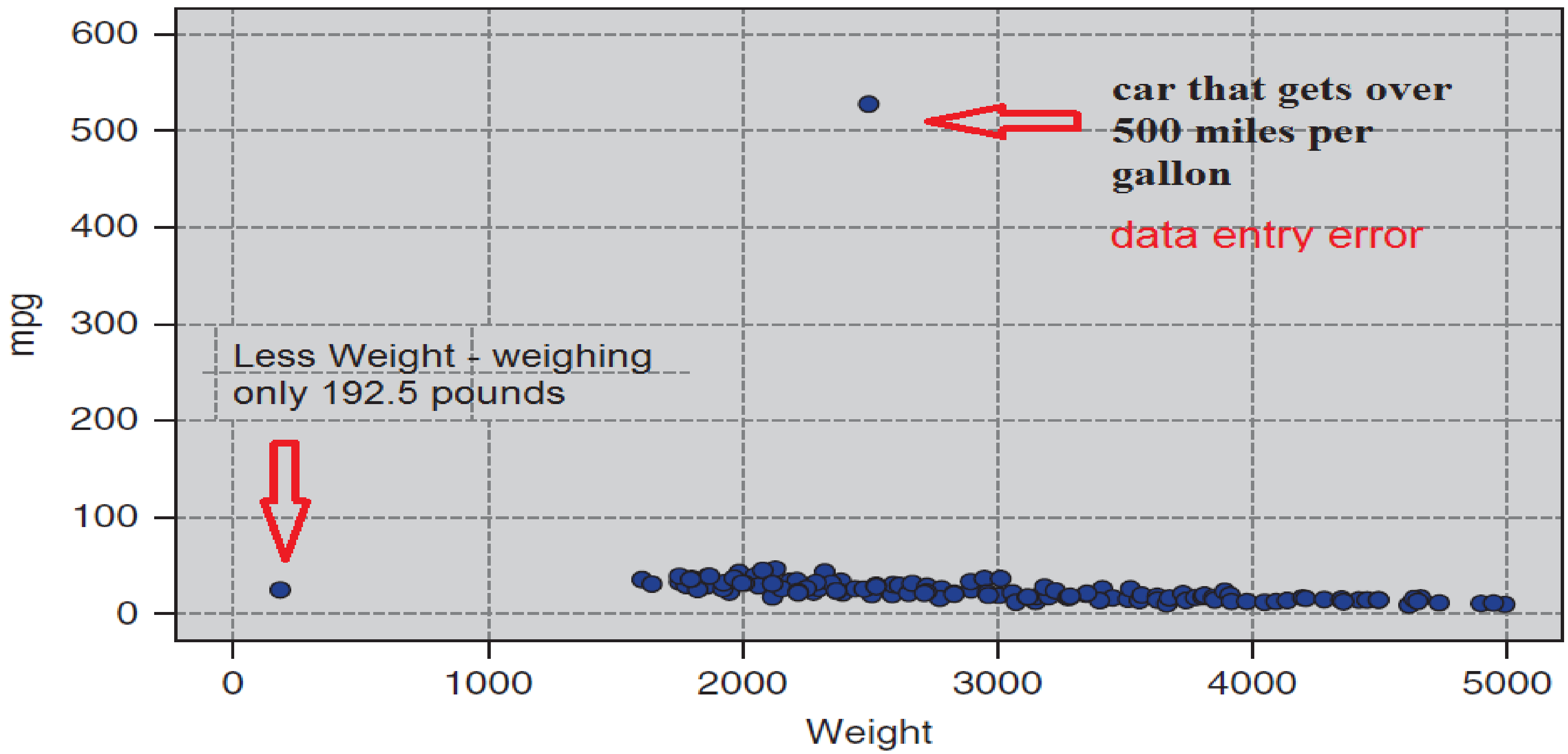
cars data set histogram



one lonely vehicle in the extreme left tail of the distribution

GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS

Sometimes two-dimensional scatter plots5 can help to reveal outliers in more than one variable



car that gets over 500 miles per gallon

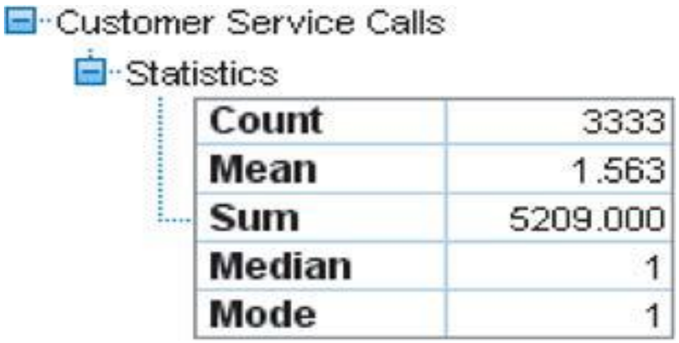
data entry error

Less Weight - weighing only 192.5 pounds

- ➢ **Record may be outlier in a particular dimension but not in another**

MEASURES OF CENTER AND SPREAD

- If interested in estimating where the center of a particular variable lies
- Measures of center are a special case of ***measures of location***
- Numerical *measures of center, commonly used are* mean, median, and mode
  - Examples of the measures of location are percentiles and quantiles
- *Mean of a variable is simply the average of the valid values taken by the* variable.
- Statistical summaries of *churn data set*



Customer Service Calls

Statistics

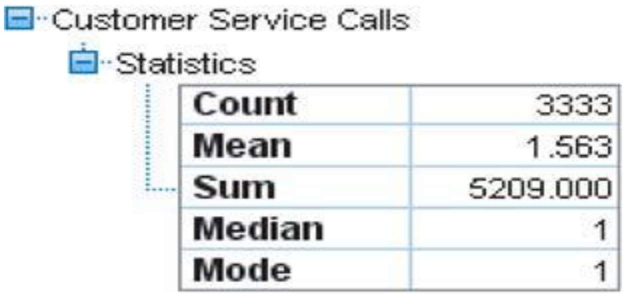
Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

MEASURES OF CENTER AND SPREAD

- For variables that are not extremely skewed, the mean is usually not too far from the variable center
- For extremely skewed data sets, the mean becomes less representative of the variable center
- Mean is sensitive to the presence of outliers
- Analysts sometimes prefer median as alternative measures of center
- Median is resistant to the presence of outliers
- Other analysts may prefer to use the mode

MEASURES OF CENTER AND SPREAD

- Note that the measures of center do not always concur as to where the center of the data set lies.
- In Figure, the median is 1, which means that half of the customers made at least one customer service call;
- Mode is also 1, which means that most frequent number of customer service calls was 1.
- The median and mode agree.
- However, mean is 1.563, which is 56.3% higher than other measures.
- This is due to the mean’s sensitivity to the right-skewness of the data.



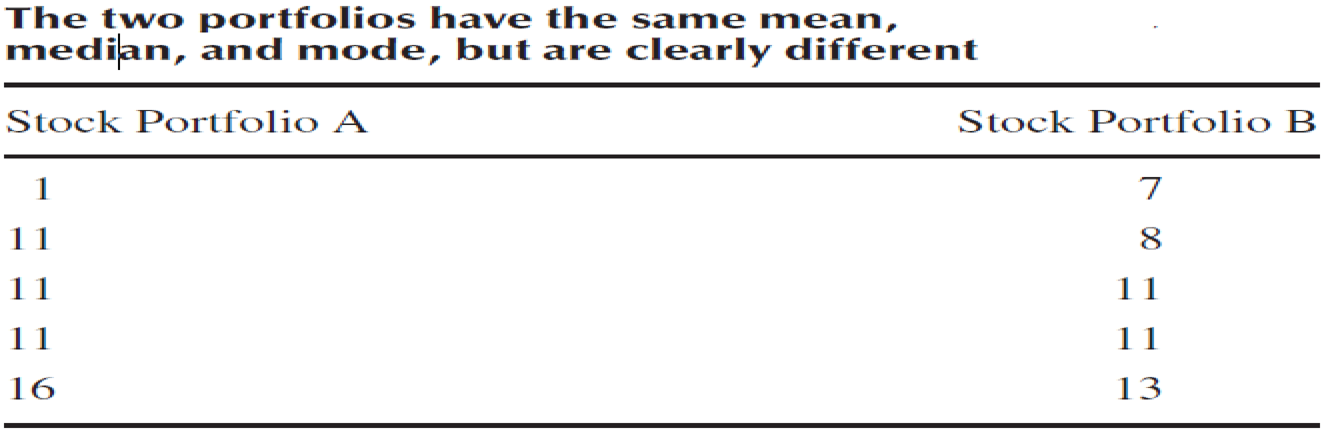
Customer Service Calls

Statistics

Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

MEASURES OF CENTER AND SPREAD

- Measures of location are not sufficient to summarize a variable effectively
- Two variables may have the very same values for the mean, median, and mode, and yet have different natures



The two portfolios have the same mean, median, and mode, but are clearly different

Stock Portfolio A	Stock Portfolio B
1	7
11	8
11	11
11	11
16	13

- The mean P/E ratio is 10, the median is 11, and the mode is 11 for each portfolio
- These measures of center do not provide us with a complete picture
- Missing are the ***measures of spread*** *or the* ***measures of variability***
- Describes how spread out the data values are
- Portfolio A's P/E ratios are more spread out than those of portfolio B
- So the measures of variability for portfolio A should be larger than those of B.

MEASURES OF CENTER AND SPREAD

- Typical **measures of variability** include the ***range*** *(maximum−minimum), the* **deviation** (SD), the **mean absolute deviation**, and the **interquartile range (IQR).**
- The sample *SD is defined by*

$$s = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n - 1}}$$

- SD is sensitive to the presence of outliers
- SD is distance between a field value and the mean
- Analysts to prefer other measures of spread, **mean absolute deviation** in situations involving extreme values.
- Formula for Mean Absolute Deviation (MAD) is

$$MAD = \sum_{i-1}^{n} \frac{|x_i - \overline{x}|}{n}$$

DATA TRANSFORMATION

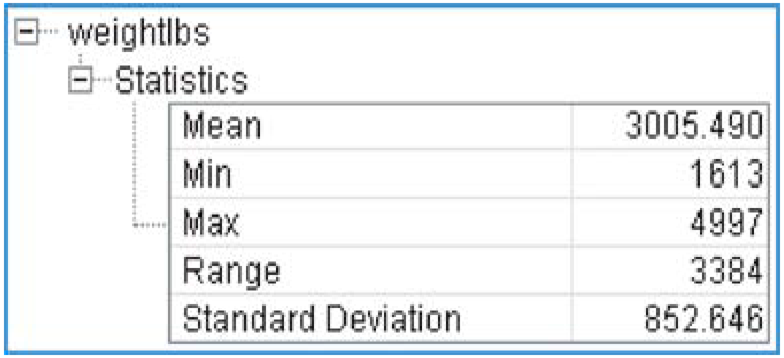
- Variables tend to have ranges that vary greatly from each other
- major league baseball, players’ batting averages will range from zero to less than 0.400
- Number of home runs hit in a season will range from zero to around 70
- Such differences in the ranges will lead to a tendency for the variable with greater range to have undue influence on the results
- That is, greater variability in home runs will dominate the lesser variability in batting averages
- Data miners/Data Scientist should *normalize their numeric variables.*
- Neural networks benefit from *normalization.*
- Benefit to algorithms that make use of distance measures, such as the *k nearest neighbors algorithm.*

MIN–MAX NORMALIZATION

- Let X refer to our original field value, and X* refer to the normalized field value.

- Min–max normalization works by seeing how much greater the field value is than the minimum value min(X), and scaling this difference by the range. That is,7

$$X^*_{\mathrm{mm}} = \frac{X - \min(X)}{\mathrm{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$



weightlbs

Statistics

Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

- The minimum weight is 1613 pounds, and the range = max(*X*) − *min*(*X*) = *4997* − *1613* = 3384 pounds.

MIN–MAX NORMALIZATION

➢ Let us find the min–max normalization for three automobiles weighing 1613, 3384, and 4997 pounds, respectively.

- For an ultralight vehicle, weighing only 1613 pounds (the field minimum), the min–max normalization is

$$X^*_{\mathrm{mm}} = \frac{X - \min(X)}{\mathrm{range}(X)} = \frac{1613 - 1613}{3384} = 0$$

- The *midrange* equals the average of the maximum and minimum values in a data set. That is,

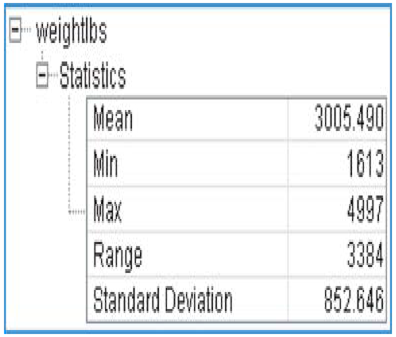
$$\mathrm{Midrange}(X) = \frac{\max(X) + \min(X)}{2} = \frac{4997 + 1613}{2} = 3305 \ \text{pounds}$$

- For a “midrange” vehicle (if any), which weighs exactly halfway between the minimum weight and the maximum weight, the min–max normalization is

$$X^*_{\mathrm{mm}} = \frac{X - \min(X)}{\mathrm{range}(X)} = \frac{3305 - 1613}{3384} = 0.5$$

- The heaviest vehicle has a min–max normalization value of

$$X^*_{\mathrm{mm}} = \frac{X - \min(X)}{\mathrm{range}(X)} = \frac{4497 - 1613}{3384} = 1$$



weightlbs

Statistics

Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

Z-SCORE STANDARDIZATION

- ***Z-score standardization***, which is very widespread in the world of statistical analysis
- Works by taking the difference between the field value and the field mean value, and scaling this difference by the SD

$$Z\text{-score} = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

- For the vehicle weighing only 1613 pounds, the Z-score standardization is

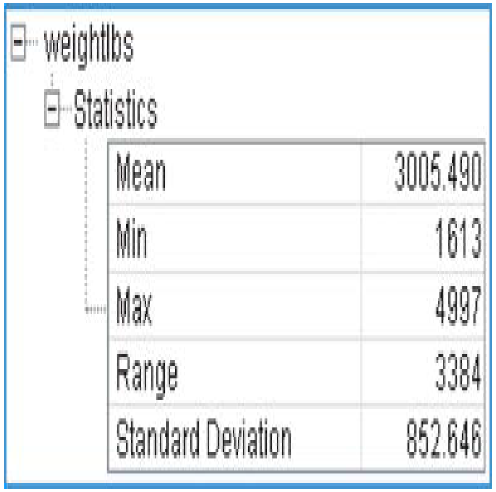
$$Z\text{-score} = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{1613 - 3005.49}{852.49} \approx -1.63$$

- For an "average" vehicle (if any), with a weight equal to mean$(X) = 3005.49$ pounds, the Z-score standardization is

$$Z\text{-score} = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{3005.49 - 3005.49}{852.49} = 0$$

- For the heaviest car, the Z-score standardization is

$$Z\text{-score} = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{4997 - 3005.49}{852.49} \approx 2.34$$



weightlbs

Statistics

Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

DECIMAL SCALING

- ***Decimal scaling* ensures** that every normalized value lies between −1 and 1.

$$X^*_{\text{decimal}} = \frac{X}{10^d}$$

where *d* represents the number of digits in the data value with the largest absolute value

- For the weight data, the largest absolute value is |4997| = 4997, which has *d*=4 digits.
- The decimal scaling for the minimum and maximum weight are

$$\text{Min}: X^*_{\text{decimal}} = \frac{1613}{10^4} = 0.1613 \quad \text{Max}: X^*_{\text{decimal}} = \frac{4997}{10^4} = 0.4997$$

NORMALIZATION TECHNIQUES COMMENTS

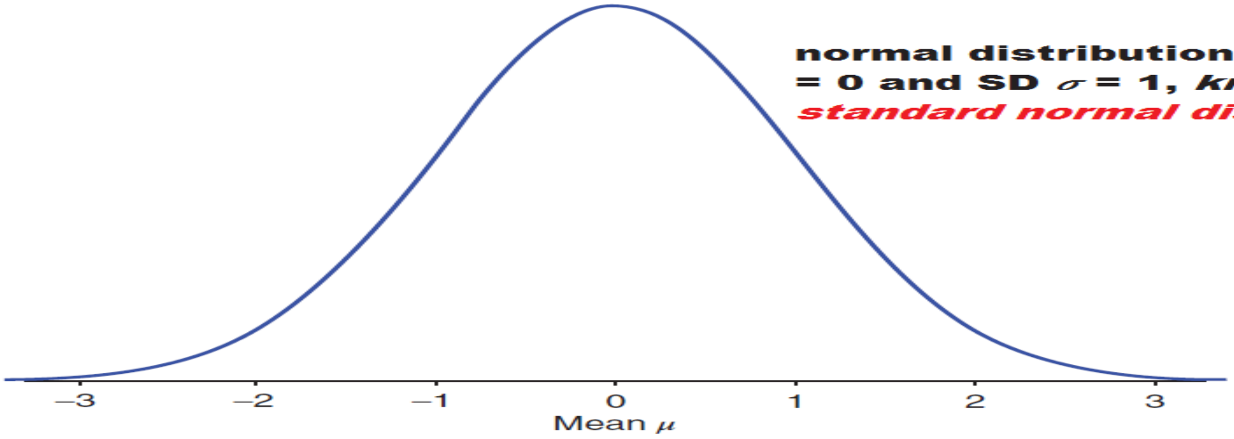
What are the best normalization methods (Z-Score, Min-Max, etc.)?

- **Z-score**
  - preserve range (maximum and minimum)
  - If your data follow a **Gaussian distribution (normal)**, they are converted into a N(0,1) distribution and probabilities calculation will be easier.
- **More Researcher Comments**
  - Depend on the data to be normalized. Normally Z-score is very common for data normalization.
  - Min-Max and Z-Score not suitable for sparse
  - It depends on the aims of the study and the nature of the data.

TRANSFORMATIONS TO ACHIEVE NORMALITY

- Variables must be normally distributed for some data mining/ML algorithms
- The normal distribution is a continuous probability distribution commonly *known as* the bell curve, which is symmetric.
- **A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.**
- It is centered at mean μ and has its spread determined by SD σ (sigma)

- Normal distribution has mean $\mu = 0$ and SD $\sigma = 1$

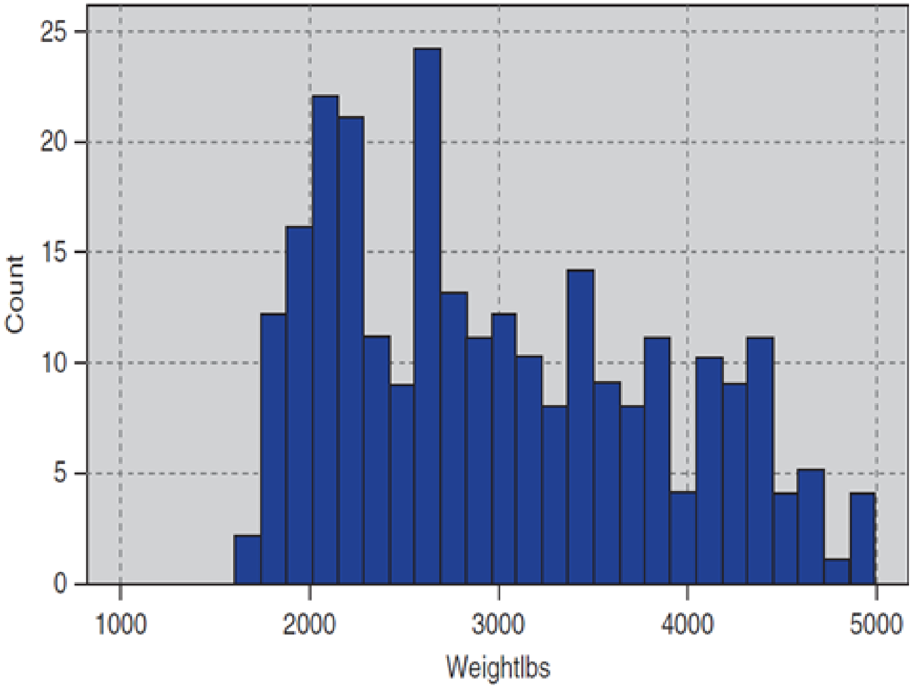


normal distribution that has mean μ = 0 and SD σ = 1, *known as* the *standard normal distribution Z.*

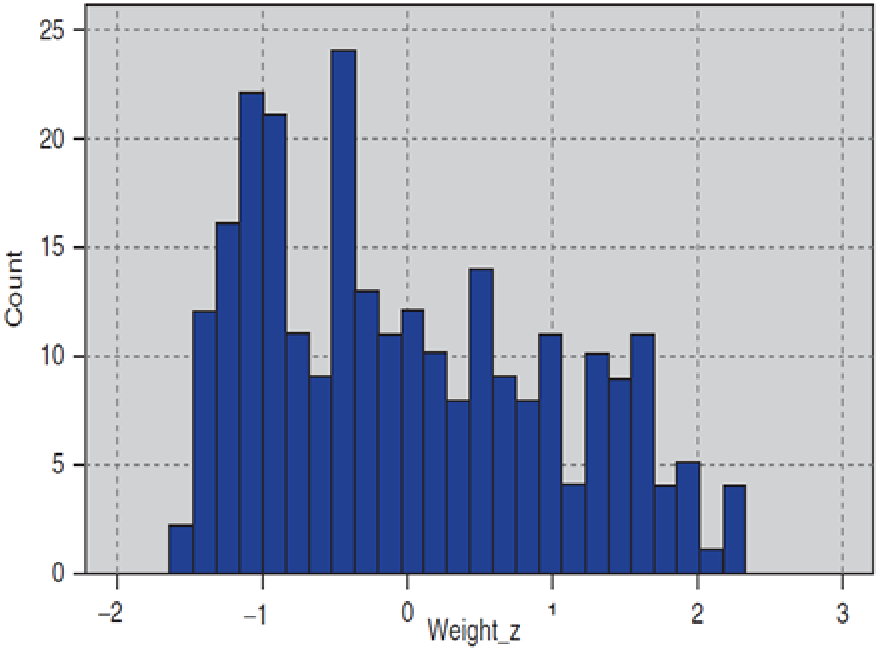
Standard normal Z distribution.

TRANSFORMATIONS TO ACHIEVE NORMALITY

- ❑ **Common misconception** - *Z*-score standardization applied to them follow the standard **normal *Z* distribution**
- ❑ It is true that the Z-standardized data will have mean 0 and SD=1, but the distribution may still be skewed



Original Data

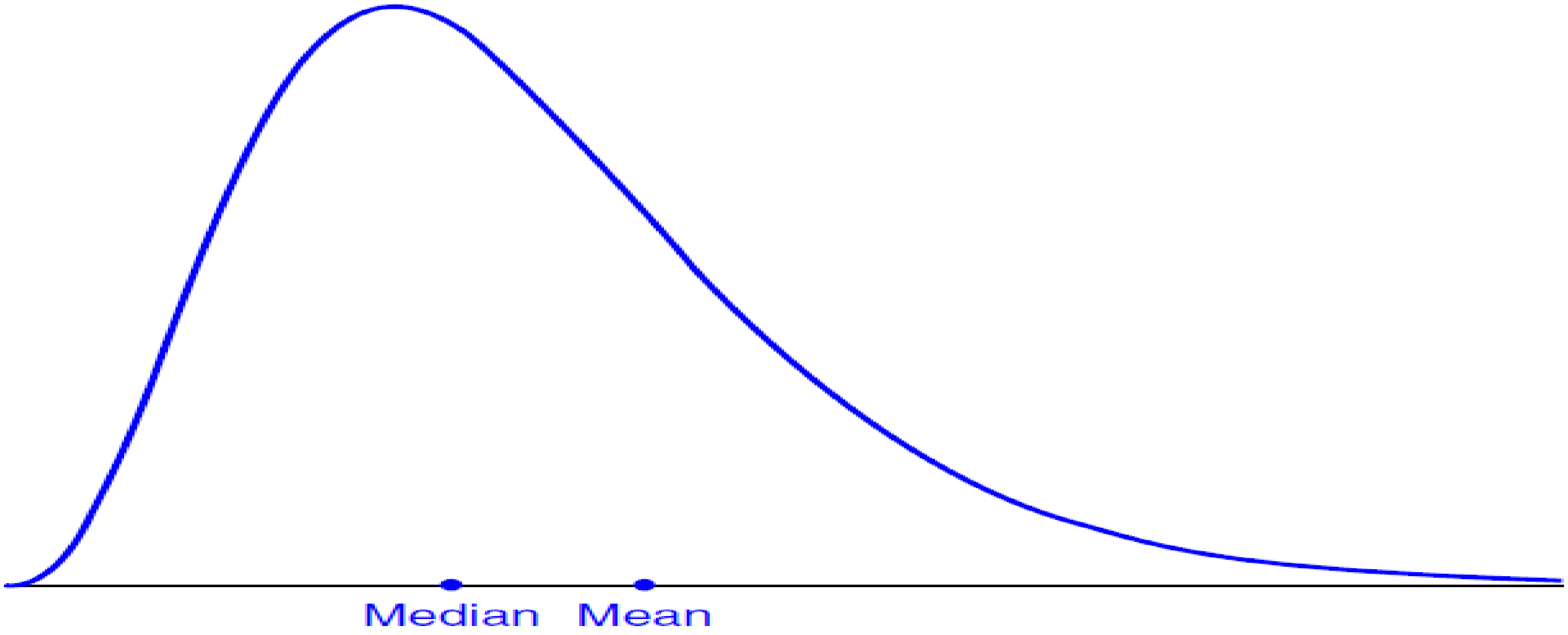


Z-Standardized Data is still right-skewed, not normally distributed

TRANSFORMATIONS TO ACHIEVE NORMALITY

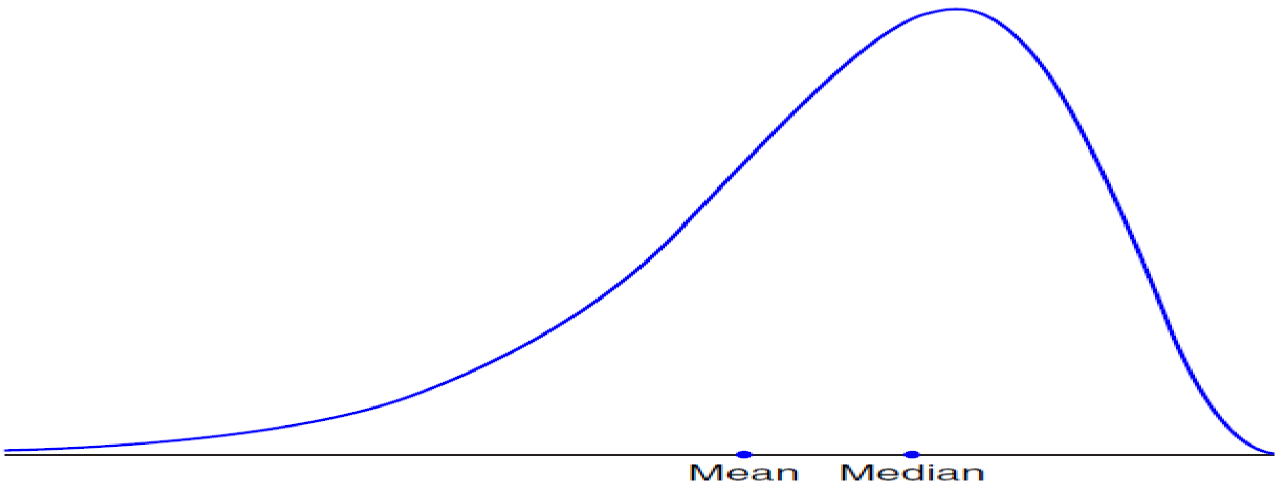
- Statistic to measure the ***skewness*** of a distribution

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$



- Right-skewed data has positive skewness
- Mean is greater than the median

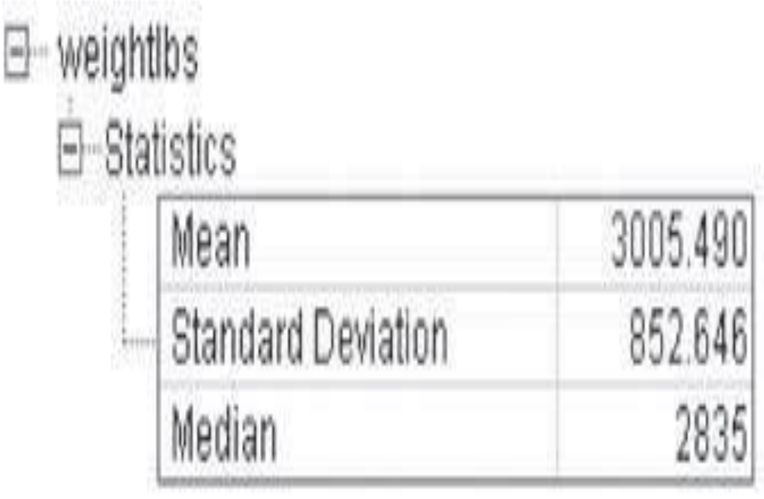
TRANSFORMATIONS TO ACHIEVE NORMALITY



- Left-skewed data, the mean is smaller than the median,
- Negative values for skewness
- For perfectly **symmetric (and unimodal) data**, **the mean, median, and mode are all equal**, and so the skewness equals zero.

TRANSFORMATIONS TO ACHIEVE NORMALITY

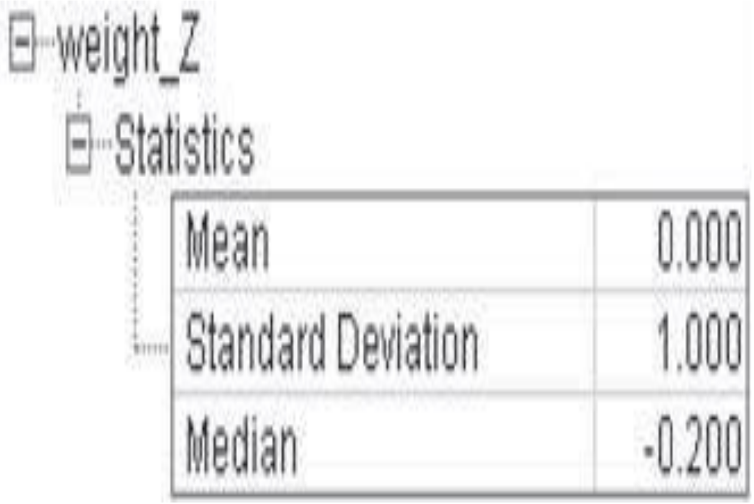
- **Z-score standardization has no effect on skewness**



weightlbs

Statistics

Mean	3005.490
Standard Deviation	852.646
Median	2835



weight_Z

Statistics

Mean	0.000
Standard Deviation	1.000
Median	-0.200

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(3005.490 - 2835)}{852.646} = 0.6$$

For *weight_Z* we have

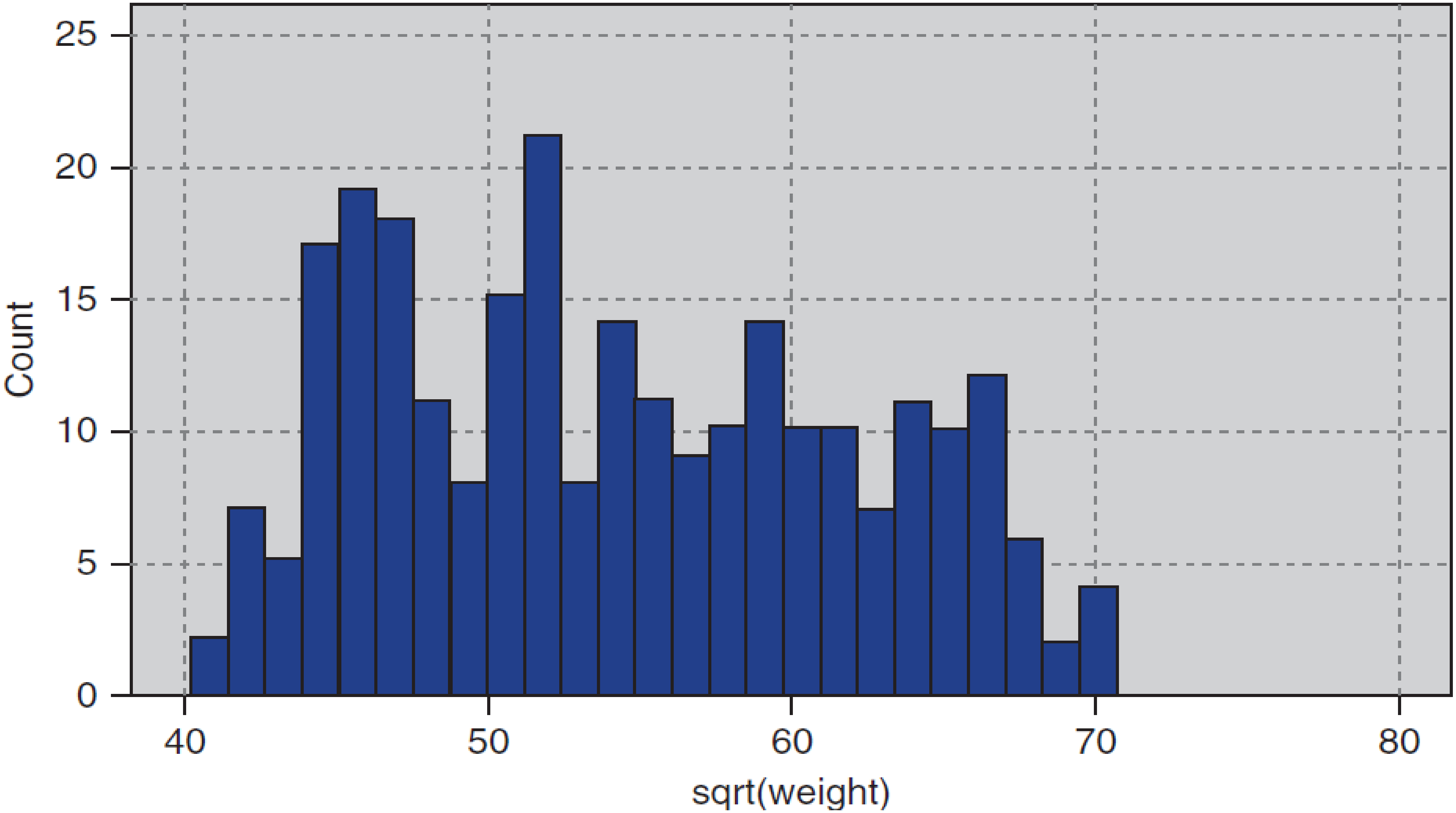
$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(0 - (-0.2))}{1} = 0.6$$

TRANSFORMATIONS TO ACHIEVE NORMALITY

- To make data “more normally distributed,” make it symmetric - eliminate skewness.
- **Common transformations are**
  - Natural log transformation
  - Square root transformation, and
  - Inverse square root transformation

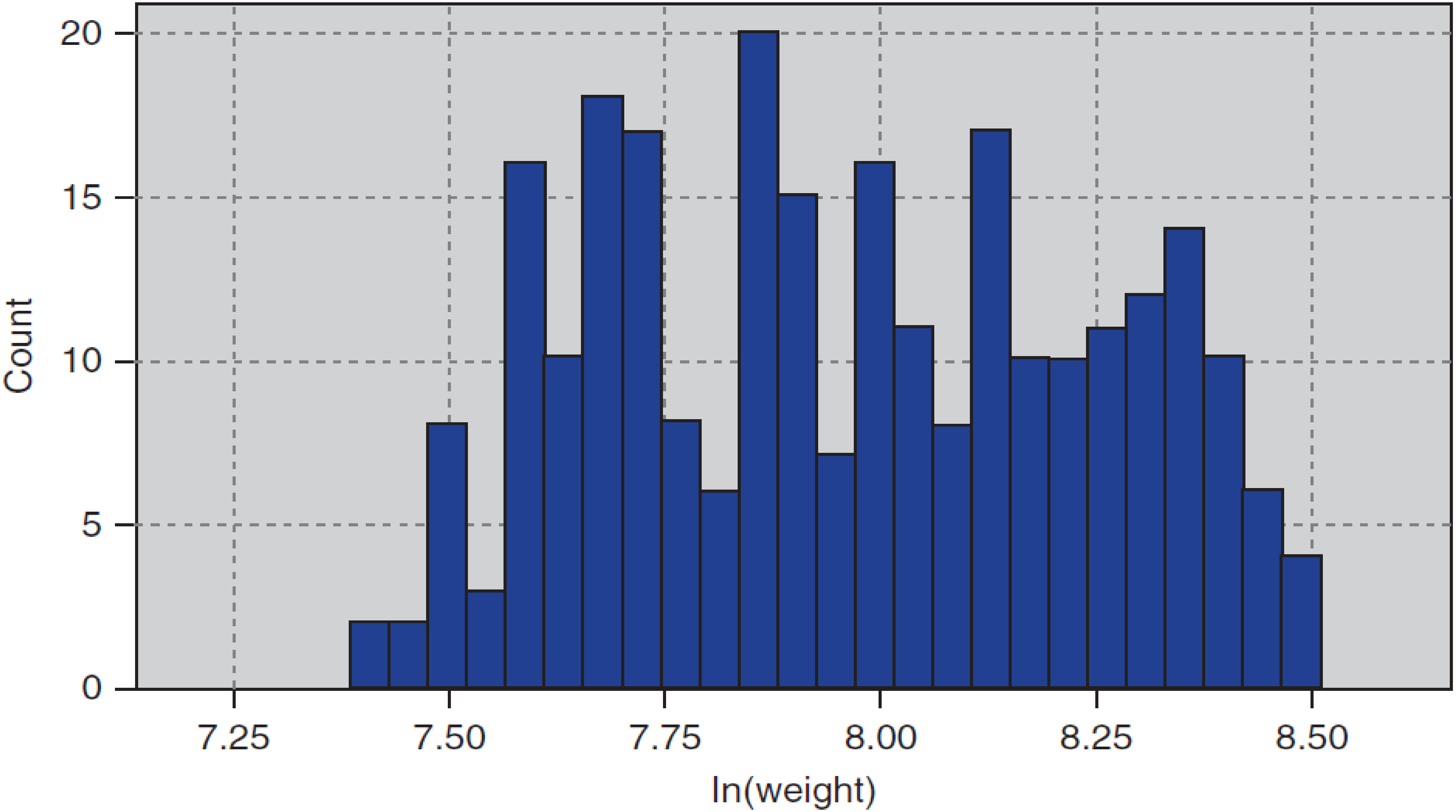
TRANSFORMATIONS TO ACHIEVE NORMALITY

Application of the square root transformation

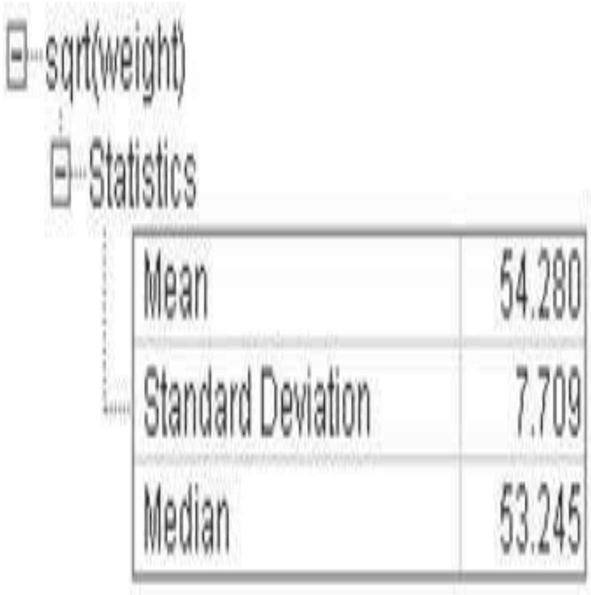


TRANSFORMATIONS TO ACHIEVE NORMALITY

Application of the Natural log transformation



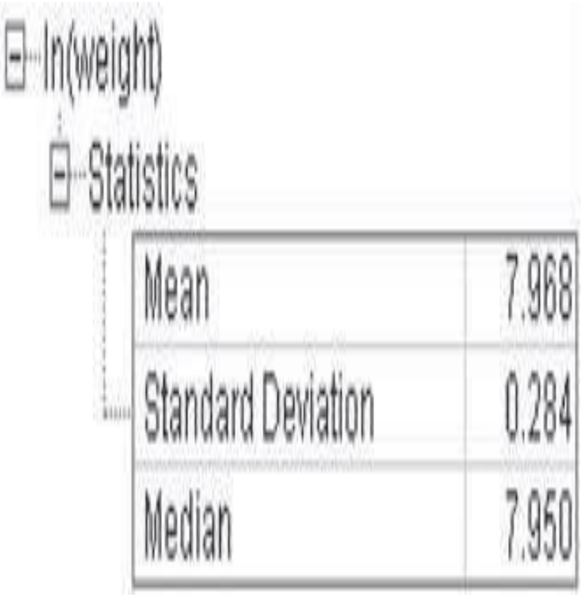
TRANSFORMATIONS TO ACHIEVE NORMALITY



sqrt(weight)

Statistics

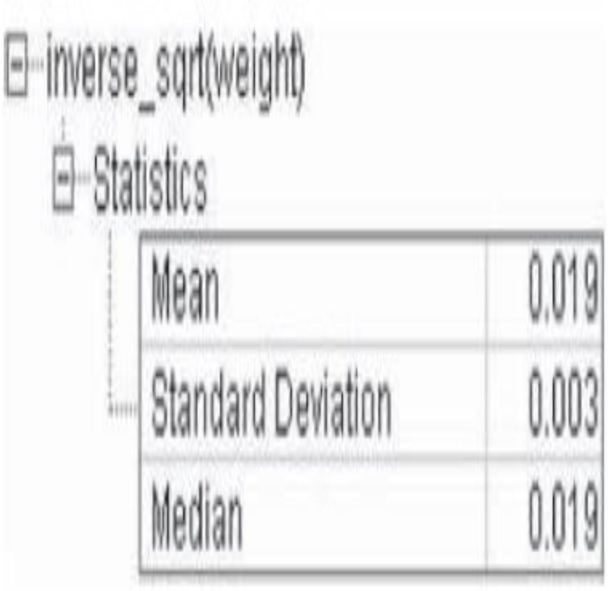
Mean	54.280
Standard Deviation	7.709
Median	53.245



ln(weight)

Statistics

Mean	7.968
Standard Deviation	0.284
Median	7.950



inverse_sqrt(weight)

Statistics

Mean	0.019
Standard Deviation	0.003
Median	0.019

$$\text{Skewness}\ \ (\text{sqrt}(\text{weight})) = \frac{3(54.280 - 53.245)}{7.709} \approx 0.40$$

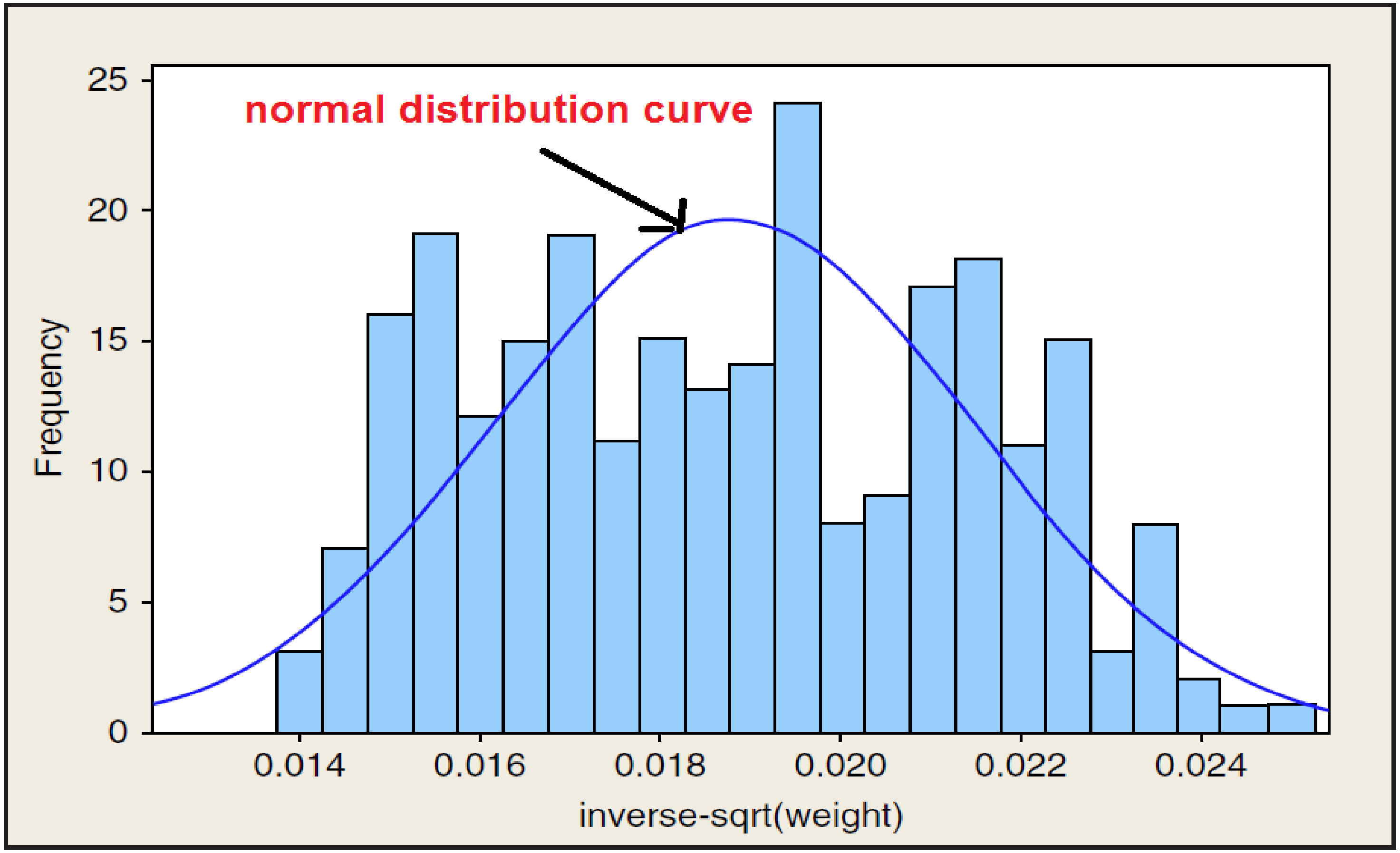
$$\text{Skewness}\ \ (\ln(\text{weight})) = \frac{3(7.968 - 7.950)}{0.284} \approx 0.19$$

$$\text{Skewness}\ \ (\text{inverse\_sqrt}(\text{weight})) = \frac{3(0.019 - 0.019)}{0.003} = 0$$

TRANSFORMATIONS TO ACHIEVE NORMALITY

- Achieved symmetry, not at normality.
- Normality can be checked by ***normal probability plot.***
- Plots the quantiles of a particular distribution against the quantiles of the standard normal distribution.
- Similar to a ***percentile***, the ***pth quantile*** of a distribution is the value x_p such that p% of the distribution values are ***less than or equal to*** x_p.

TRANSFORMATIONS TO ACHIEVE NORMALITY



The transformation *inverse_sqrt*(*weight*) has eliminated the skewness, but is still not normal.

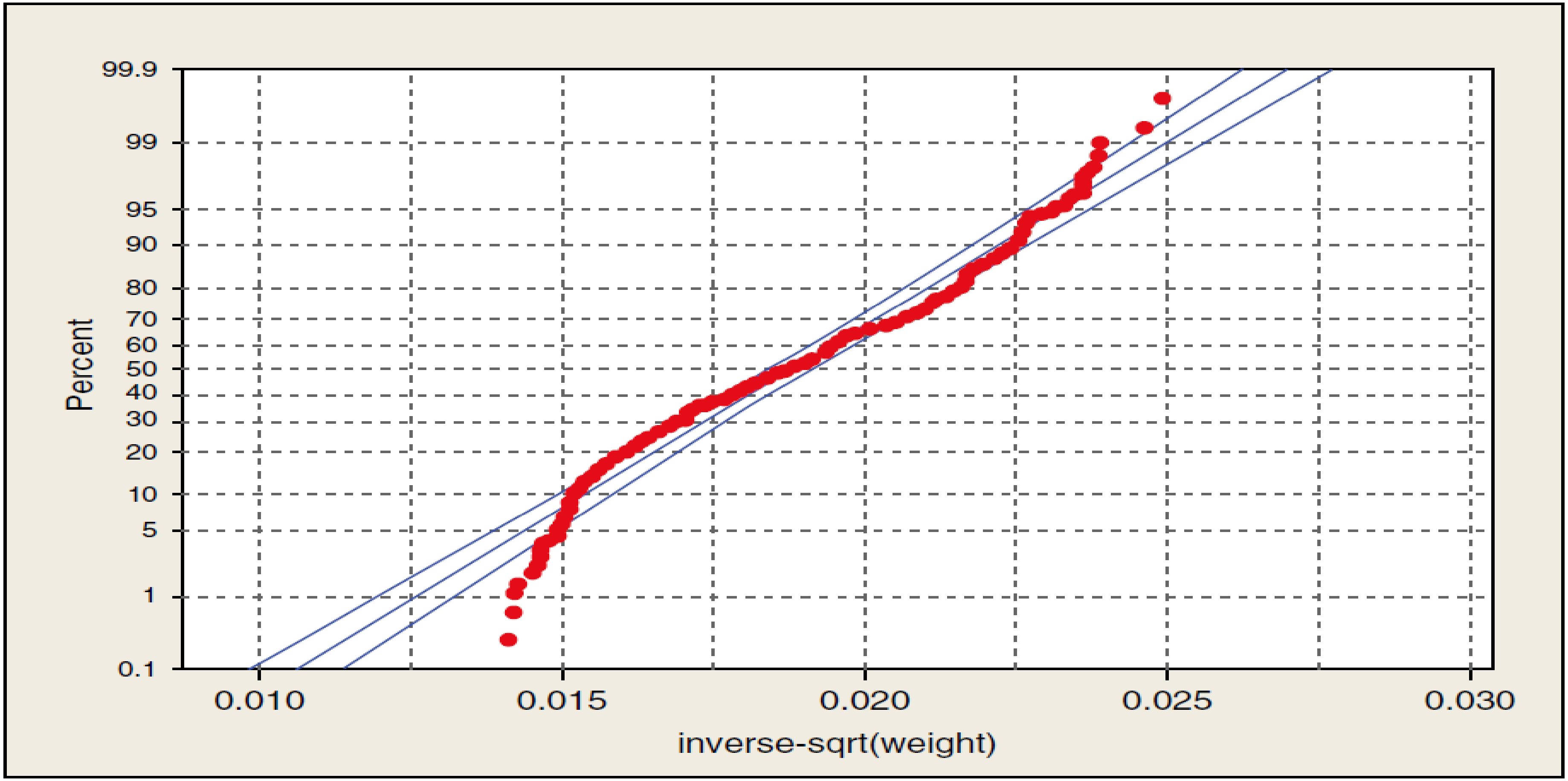
TRANSFORMATIONS TO ACHIEVE NORMALITY

How to Draw a Normal Probability Plot By Hand

1. Arrange your x-values in ascending order.
2. Calculate f_i = (i-0.375)/(n+0.25), where i is the position of the data value in the ordered list and n is the number of observations.
3. Find the z-score for each f_i
4. Plot your x-values on the horizontal axis and the corresponding z-score on the vertical axis.

TRANSFORMATIONS TO ACHIEVE NORMALITY

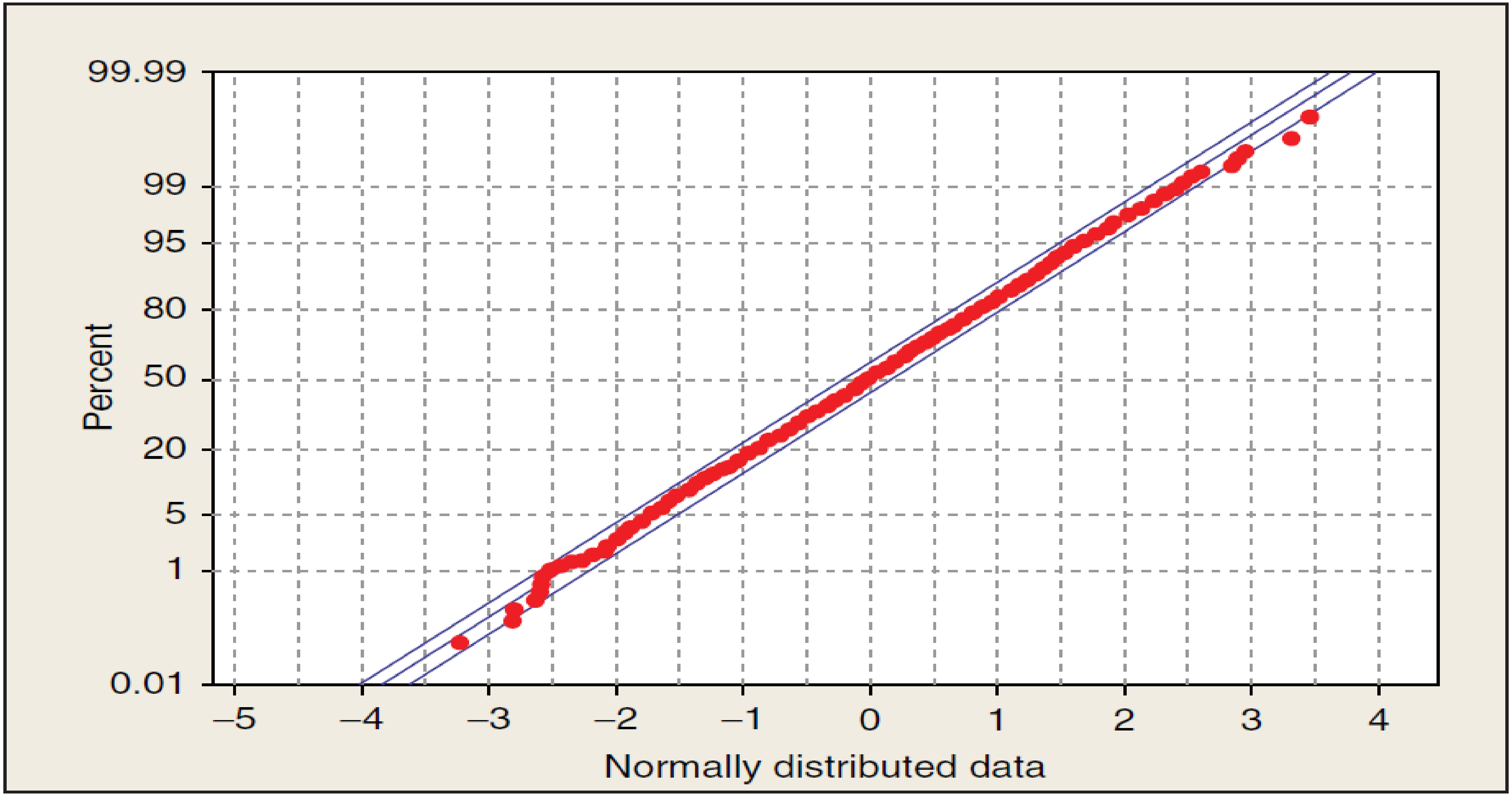
- shows systematic deviations from linearity, indicating non-normality



Normal probability plot of *inverse_sqrt(weight)* indicates nonnormality.

TRANSFORMATIONS TO ACHIEVE NORMALITY

- Shows no systematic deviations from linearity



Normal probability plot of normally distributed data.

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

- The Z-score method states that a data value is an outlier if it has a Z-score that is either less than –3 or greater than 3.
- Variable much beyond this range may bear further investigation (data entry errors or other issues)
- Should not automatically omit outliers from analysis.
- No outliers among the vehicle weights - Z-scores - for 1613 pounds - 1.63, 4997-pound vehicle-Z-score of 2.34.
- As neither Z-scores are either less than –3 or greater than 3, we conclude that there are no outliers among the vehicle weights.

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

- Mean and SD, part of the formula for the *Z*-score standardization, are ***sensitive*** to the presence of outliers
- Values of mean and SD will both be unduly affected by the presence or absence of this new data value.
- Not appropriate to use measures that are themselves sensitive to their presence.
- Data analysts have developed more *robust* statistical methods for outlier detection, which are less sensitive to the presence of the outliers.
- One elementary robust method is to use the Interquartile Range (IQR)

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

- *Quartiles* of a data set divide data set *four parts*
- Each containing 25% of the data:
  - *first quartile* (Q1) is the 25th percentile.
  - *second quartile (Q2)* is 50th percentile, (median).
  - *The third quartile (Q3)* is the 75th percentile.
- IQR is calculated as IQR=Q3−Q1
- Robust measure of outlier detection - A data value is an outlier if

  a. it is located 1.5(IQR) or more below Q1, or

  b. it is located 1.5(IQR) or more above Q3.

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

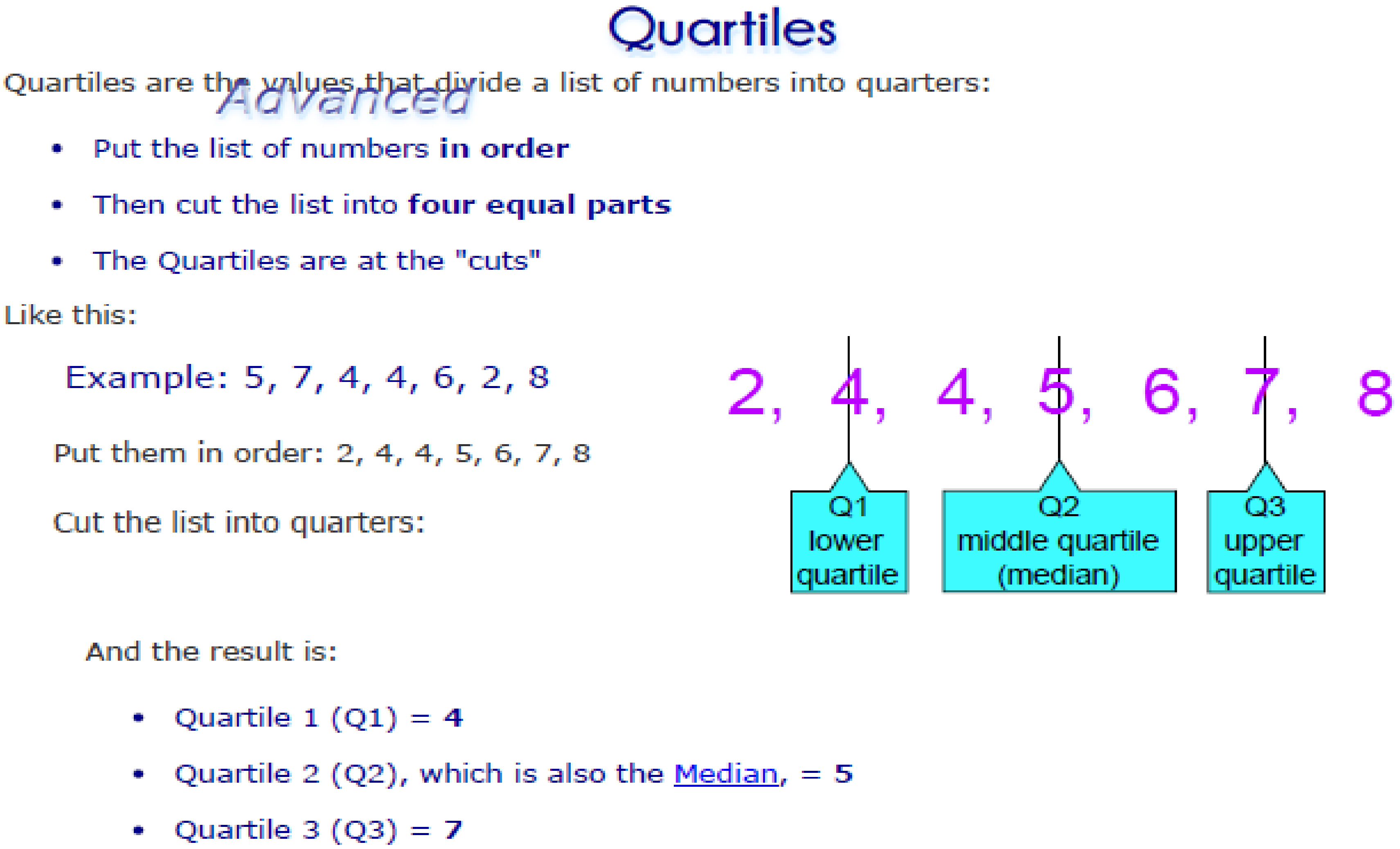
Example: Suppose for a set of test scores

- Q1=70
- Q3=80
- Interquartile Range IQR=80−70=10
- A test score would be robustly identified as an outlier if
  - it is lower than

    Q1−1.5(IQR)=70−1.5(10)=55, or
  - it is higher than

    Q3+1.5(IQR)=80+1.5(10)=95.

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS



Quartiles

Quartiles are the values that divide a list of numbers into quarters:

- Put the list of numbers **in order**
- Then cut the list into **four equal parts**
- The Quartiles are at the "cuts"

Like this:

Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

Cut the list into quarters:

And the result is:

- Quartile 1 (Q1) = **4**
- Quartile 2 (Q2), which is also the Median, = **5**
- Quartile 3 (Q3) = **7**

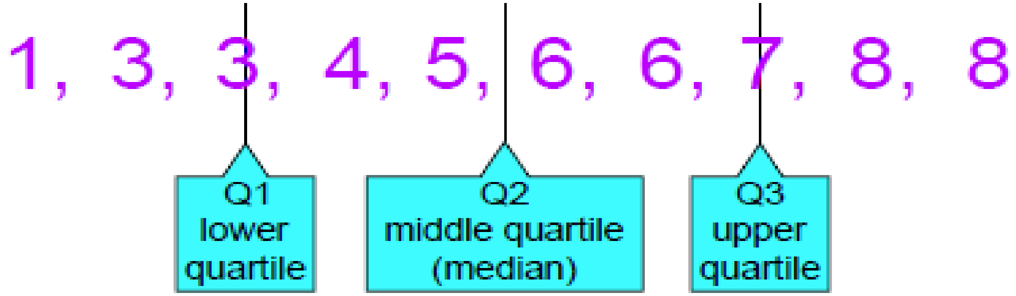
NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

Sometimes a "cut" is between two numbers ... the Quartile is the average of the two numbers.

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:



1, 3, 3, 4, 5, 6, 6, 7, 8, 8

Q1
lower
quartile

Q2
middle quartile
(median)

Q3
upper
quartile

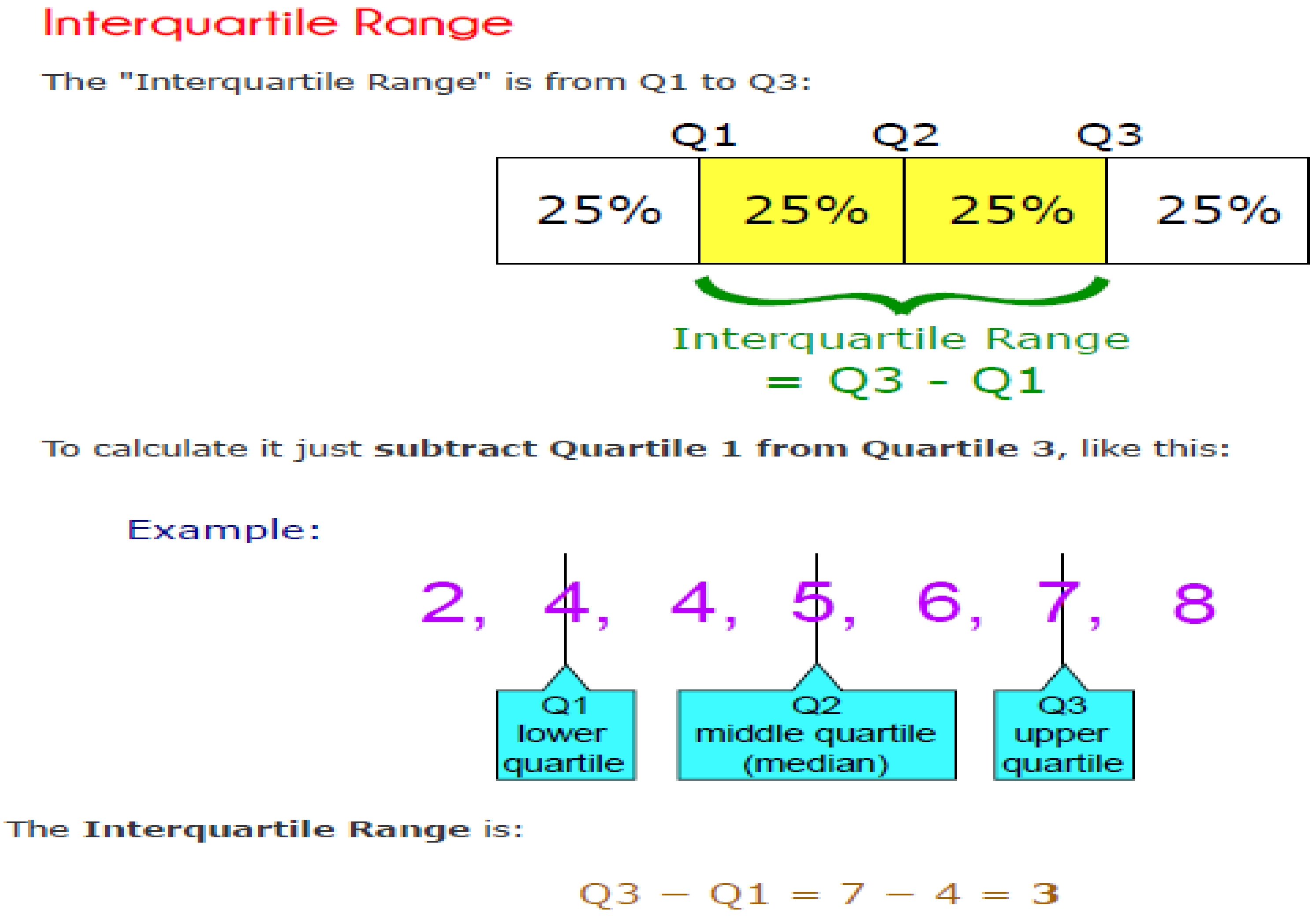
In this case Quartile 2 is half way between 5 and 6:

Q2 = (5+6)/2 = **5.5**

And the result is:

- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS



Interquartile Range

The "Interquartile Range" is from Q1 to Q3:

Q1 Q2 Q3

25% 25% 25% 25%

Interquartile Range
= Q3 - Q1

To calculate it just **subtract Quartile 1 from Quartile 3**, like this:

Example:

2, 4, 4, 5, 6, 7, 8

Q1
lower
quartile

Q2
middle quartile
(median)

Q3
upper
quartile

The **Interquartile Range** is:

$$Q3 - Q1 = 7 - 4 = 3$$

FLAG VARIABLES

- Some analytical methods, such as regression, require predictors to be numeric

- Need to recode the categorical variable into one or more flag variables

- A **flag variable** (or dummy variable, or indicator variable) is a categorical variable taking only two values, 0 and 1.

- For e.g., the categorical predictor **Gender**, taking values for female and male could be recoded into the flag variable gender_flag as follows:

  If Gender = female = then Gender_flag = 0;

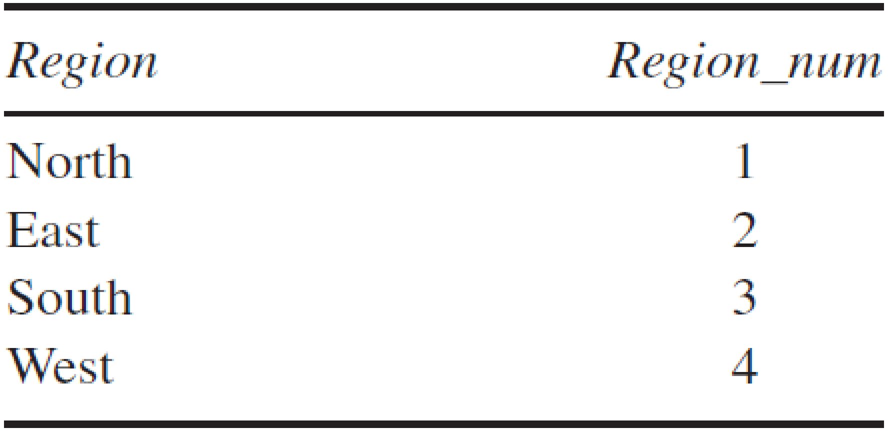
  if Gender = male then Gender _flag = 1.

FLAG VARIABLES

- When a categorical predictor takes k ≥ 3 possible values, then define k−1 dummy variables and use the unassigned category as the reference category
- For example, region has k=4 possible categories, {north, east, south, west}, then the analyst could define the following k−1=3 flag variables
  - north_flag: If region = north then north_flag = 1; otherwise north_flag = 0
  - east_flag: If region = east then east_flag = 1; otherwise east_flag = 0
  - south_flag: If region = south then south_flag = 1; otherwise south_flag = 0.
- Flag variable for west is not needed, as region=west is already uniquely identified by zero values for each of the three existing flag variables

TRANSFORMING CATEGORICAL VARIABLES INTO NUMERICAL VARIABLES

Would it not be easier to simply transform the categorical variable region into a single numerical variable rather than using several different flag variables?



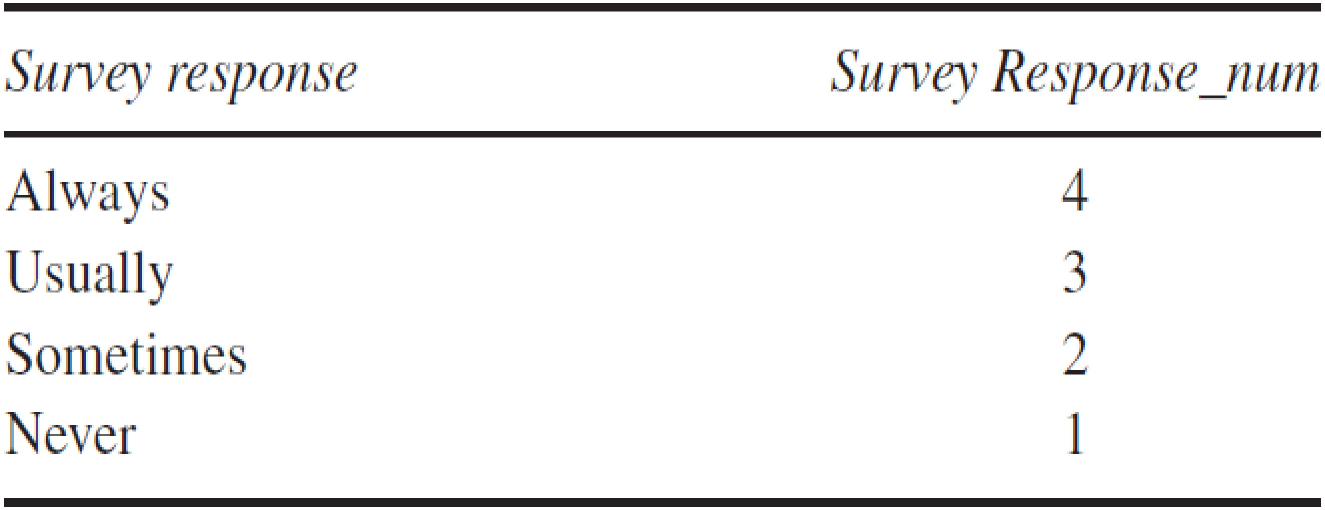
Region	*Region_num*
North	1
East	2
South	3
West	4

Unfortunately, this is a common and hazardous error.
The algorithm now erroneously thinks the following:

- The four regions are ordered.
- West>South>East>North.
- West is three times closer to South compared to North, and so on.

TRANSFORMING CATEGORICAL VARIABLES INTO NUMERICAL VARIABLES

- In most instances, data analyst should avoid transforming categorical variables to numerical variables
- Exception - for categorical variables that are clearly ordered



Survey response	*Survey Response_num*
Always	4
Usually	3
Sometimes	2
Never	1

BINNING NUMERICAL VARIABLES

- Some algorithms prefer categorical rather than continuous predictors
- Need to partition any numerical predictors into bins or bands
- Wish to partition the numerical predictor house value into low, medium, and high
- **Four common methods**
  1. *Equal width binning*
  2. *Equal frequency binning*
  3. *Binning by clustering*
  4. *Binning based on predictive value*
- Equal width binning is not recommended, as the width of the categories can be greatly affected by the presence **of outliers**

BINNING NUMERICAL VARIABLES

- Data set - X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44} and k=3
- **equal width binning**
  - *Low*: $0 \leq X < 15$, which contains all the data values except one
  - *Medium*: $15 \leq X < 30$, which contains no data values at all
  - *High*: $30 \leq X < 45$, which contains a single outlier
- **equal frequency binning** - we have n=12, k=3, and n/k=4
  - *Low*: Contains the first four data values, all X=1.
  - *Medium*: Contains the next four data values, {1, 2, 2, 11}.
  - *High*: Contains the last four data values, {11, 12, 12, 44}.
- **k-means - seems to be the correct partition**

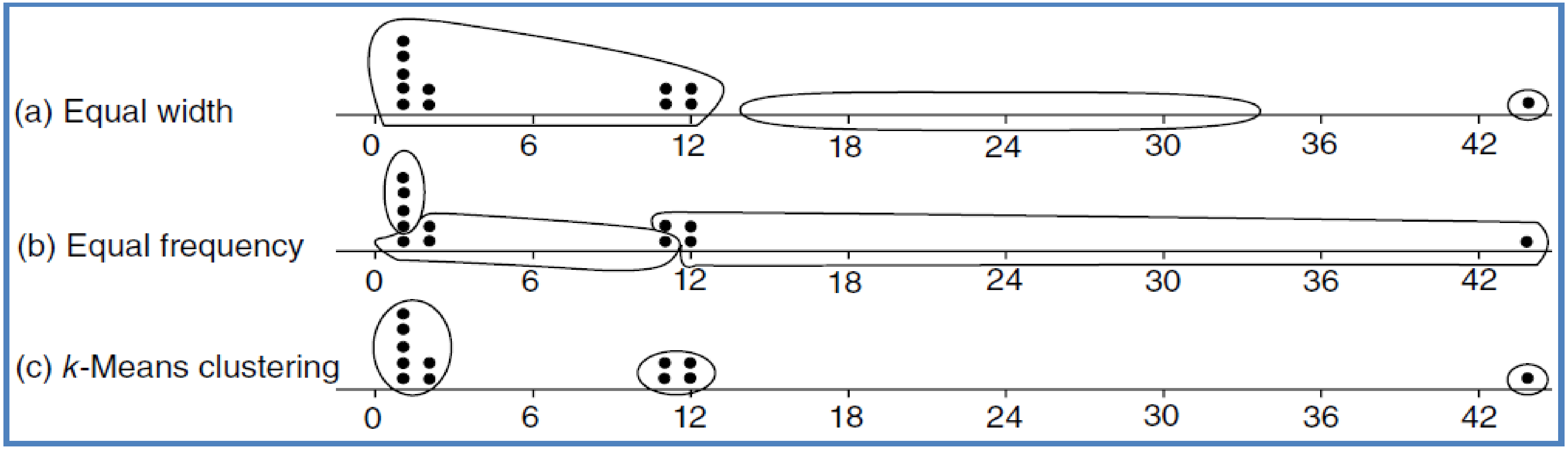


Illustration of binning methods.

RECLASSIFYING CATEGORICAL VARIABLES

- Reclassifying categorical variables is the categorical equivalent of binning numerical variables
- Categorical variable will contain too many easily analysable field values
- For example, the predictor state could contain 50 different field values
- Data analyst should reclassify the field values
- 50 states could each be reclassified as the variable region
- Alternatively, the 50 states could be reclassified as the variable economic_level (richer states, the midrange states, and the poorer states)
- Data analyst should choose a reclassification that supports the objectives of the business problem or research question.

ADDING AN INDEX FIELD

- Recommended that the data analyst create an index field,
- tracks the sort order of the records
- Data mining data gets partitioned at least once (and sometimes several times).
- It is helpful to have an index field so that the original sort order may be recreated.
- For example, using IBM/SPSS Modeler, you can use the *@Index* function in the *Derive* node to create an index field.

REMOVING VARIABLES THAT ARE NOT USEFUL

- ❑ Data analyst may remove variables that will not help analysis
- ❑ Such variables include
  - ➢ unary variables and
  - ➢ variables that are very nearly unary.
- ❑Unary variables take on only a single value
- ❑ Sample of students at an all-girls private school would gender as female.
- ❑ Sometimes a variable can be very nearly unary
- ❑ For example, suppose that 99.95% of the players in a field hockey league are female, with the remaining 0.05% male.

VARIABLES THAT SHOULD PROBABLY NOT BE REMOVED

- **Common –** practice to remove following variables
  1. Variables for which 90% or more of the values are missing
  2. Variables that are strongly correlated
- May be a pattern in the missingness, and therefore useful information
- Challenge to any strategy for imputation of missing data
- Donation Survey example – people may not donate, will skip the survey question
- preferable to construct a flag variable, donation_flag
- if 10% are representative, then choose to imputation of the missing 90%.
- imputation be based on the regression or decision tree methods
- **Bottom line:** avoid removing variables having missing values

VARIABLES THAT SHOULD PROBABLY NOT BE REMOVED

- An example of correlated variables may be precipitation and attendance at a state beach.
- As precipitation increases, attendance at the beach tends to decrease (**negatively correlated**)
- Inclusion of correlated variables may
  - at best double-count a particular aspect of the analysis,
  - and at worst lead to instability of the model results.
- Some data analysts may decide to simply remove one of the variables
- Avoid - important information may thereby be discarded
- Instead, principal components analysis may be applied, where the common variability in correlated predictors may be translated into a **set of uncorrelated principal components.**

REMOVAL OF DUPLICATE RECORDS

- Records may have been inadvertently copied, thus creating duplicate records.
- Duplicate records lead to an **overweighting** of the data values
- Only one set of them should be retained
- For example, if the ID field is duplicated, then definitely remove the duplicate records.
- Data analyst should apply common sense
- Suppose a data set contains three nominal fields, and each field takes only three values, then 3 × 3 × 3 = 27 possible different records
- If there are more than 27 records, at least one of them has to be a duplicate

REMOVAL OF DUPLICATE RECORDS

- Removing duplicate records is not particularly difficult.
- Most statistical packages and database systems have built-in commands that group records together.
- In fact, in the database language SQL, this command is called Group By.

Thank You !!!