Agenda – August 26

- Data Exploration
- Descriptive Statistics

Data Exploration

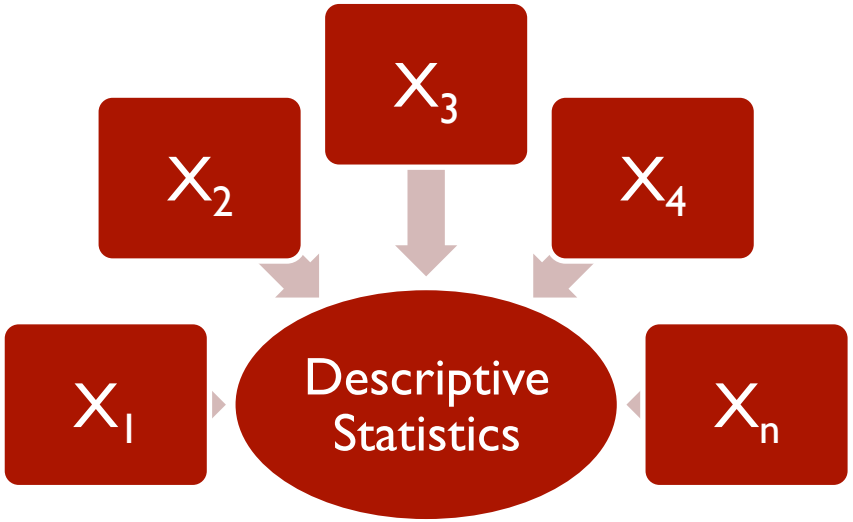
- An important first step in using data to make better decisions is data exploration
  - A large proportion of successful data mining efforts is devoted to data exploration!
- In order to understand data, we need to:
  - Characterize the distributions of the variables
  - Identify relationships among the data
  - Identify anomalies
- There are challenges to this process
  - Messy/dirty data: Missing, unstructured, or wrongly formatted data – Need to process or clean
  - Big data: Volume, variety, and velocity

Data Exploration: From this…

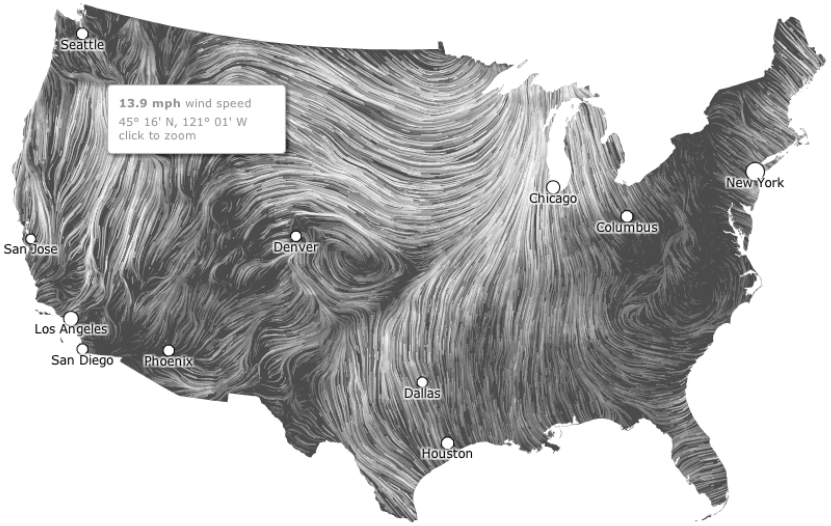
…to this!!

Two Key Data Exploration Steps

Data Aggregation



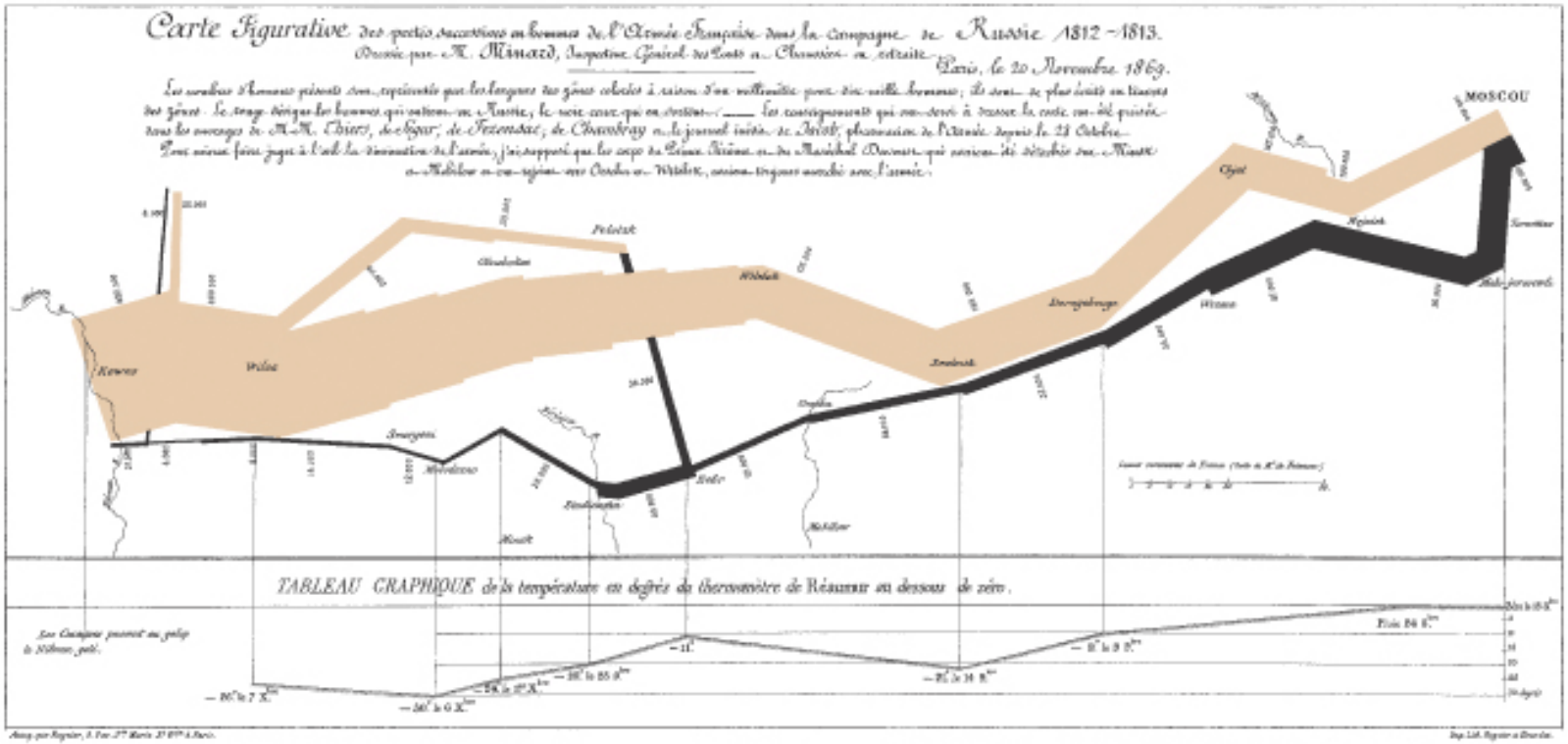
Data Visualization



Data Visualization

- Lots of great tools out there!
- Commercial software
  - Excel + Decision Tools, Tableau, MATLAB
- Open-source programming languages
  - R + ggplot2, Python + matplotlib + Seaborn
- Web-based
  - Many Eyes, Wolfram Alpha, D3.js, Google Charts
- Also, an entire field of theoretical and applied knowledge
  - Effective Data Visualization

A picture is worth a thousand words…



Napoleon's March to Moscow The War of 1812

Charles Joseph Minard

This classic of Charles Joseph Minard (1781–1870), the French engineer, shows the terrible fate of Napoleon's army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence, this combination of data map and time-series, drawn in 1869, portrays the devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Neman River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally straggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn.

Edward R. Tufte, *The Visual Display of Quantitative Information* Graphics Press Box 430 Cheshire, Connecticut 06410

From Data to Knowledge

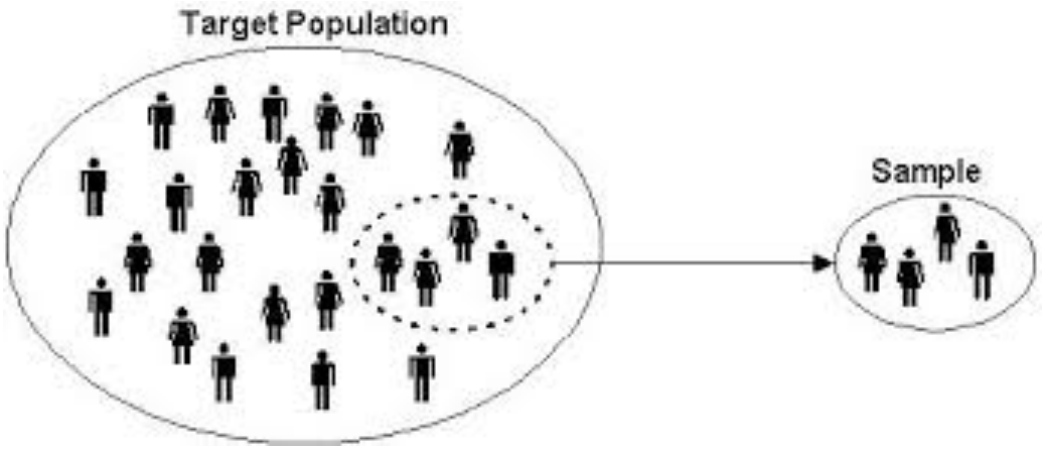
Data Exploration Outline

- Today: Descriptive Statistics
  - Comparing Averages
  - Summarizing and Visualizing Categorical Data
  - Summarizing and Visualizing Numerical Data

- Next Time: Finding Relationships among Variables

Populations and Samples

- A population includes all of the entities of interest
  - E.g., all potential voters in a presidential election or all invoices submitted for Medicare reimbursement
- A sample is a subset of the population, often randomly chosen and preferably representative of the population as a whole
  - Polling agencies (e.g., Gallup, U.S. Census)



Summary Measures: Population vs. Sample

- A summary measure of an entire population is called a ***parameter***

- A summary measure of a sample is called a ***statistic***

Data Sets, Variables, and Observations

- A data set is usually presented in tabular form, with observations in rows and variables in columns
- An observation (or record) is a list of all variable values for a single member of a population (or sample)
- A variable is a characteristic of members of a population, such as height, gender, or salary
  - Also called fields, attributes, features, predictors

Types of Data

- Numerical
  - Discrete vs. continuous
- Categorical
  - Ordinal vs. nominal
  - Encoding: Translate categories to discrete numbers
    - Special case: Dummy variables (0-1)
  - Binning/discretization: Translate numerical data into discrete bins
- Time series vs. cross-sectional
  - Time-dependent vs. time invariant

Example: Questionnaire Data

- Observations: Samples of people
- Variables: age, gender, state, children, salary, opinion
  - Numerical: age and salary (continuous), children and opinion (discrete)
  - Categorical: gender and state (nominal)
  - Index of the observation is often included in first column
  - Variable names should be concise but meaningful
- Time series or cross-sectional?

	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1

Excel Tables

- Facilitates filtering, sorting, summarizing, and formatting data
- To convert a basic range of data into a table:
  - Select the Tables tab
  - Highlight data (use Shift+Ctrl+Arrows)
  - New > Insert Table with Headers (if applicable)
- Other features
  - Name the table
  - Easily generate a Pivot Table (more later)
  - Convert back to basic data range
- See Example 2-7 for details

Descriptive Statistics for Categorical Variables

Mostly based on counts and proportions

- Counts: number of observations in each category
  - $x_1, x_2, \ldots, x_n$
- Proportions: proportion of observations in each category, relative to total number of observations
  - $x_1 / n, x_2 / n, \ldots, x_n / n$
  - Can also convert to percentages (multiply by 100%)

Example: Supermarket Transactions

- In Excel, use the COUNTIF function to count the number of observations in each category
  - =COUNTIF(data_range, criterion)
  - =COUNTIF(A1:A10, "M") or =COUNTIF(A1:A10, ">10")
- Divide each count by the total number of observations to generate the proportions

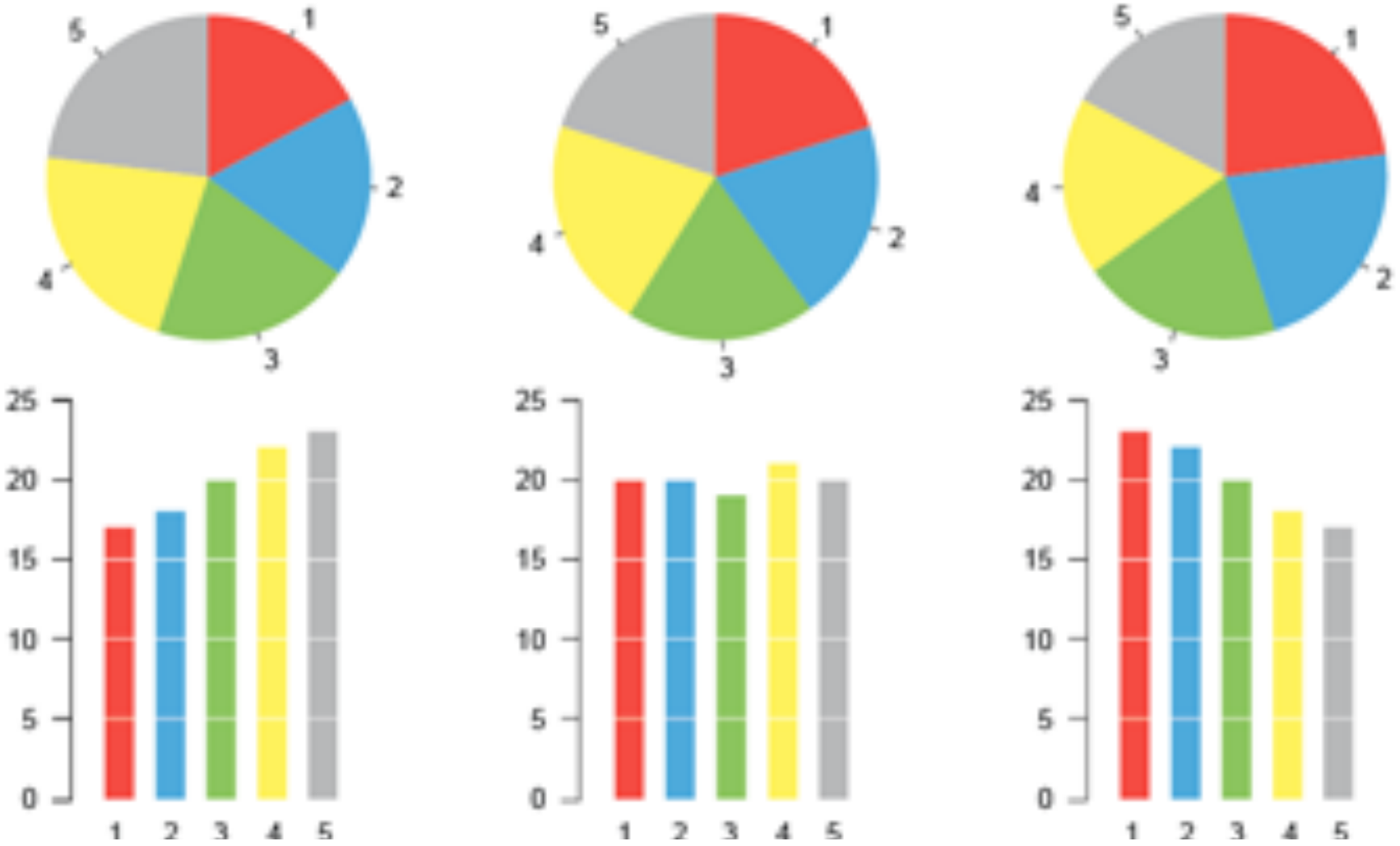
	R	S	T
1	**Categorical summaries**		
2	Gender	Count	Percent
3	M	6889	49.0%
4	F	7170	51.0%
5			100.0%

Alternative Method for Counting

Another efficient way to find the counts and proportions for a categorical variable is to use dummy variables

- For each categorical value, create a new column that encodes the observations as either 1 (in the category) or 0 (not in the category)
  - The IFELSE function comes in handy for this
- Then,
  - Count the frequency for each value by summing the 0s and 1s in each column
  - Calculate the proportions by dividing the sums by the total number of observations (COUNT function)

Visualizing Categorical Data: Bar/Column vs. Pie Charts



Descriptive Statistics for Numerical Variables

- Many ways to summarize numerical variables
  - Aggregate statistical measures
  - Visualization
- We can ask many questions to learn how the values of a numerical variable are distributed:
  - What are the most typical values?
  - How spread out are the values?
  - What are the extreme values?
  - Are the data symmetric or skewed in some direction?

Descriptive Statistics for Numerical Variables

Numerical summary measures can be categorized into several groups:

- Measures of central tendency
- Minimum, maximum, percentiles, and quartiles
- Measures of variability
- Measures of shape

Central Tendency

- Why does it matter?
  - It helps to know what's typical or most common

- Measures
  - Mean
  - Median
  - Mode

Measures of Central Tendency: The Mean

- The mean is the average of all values of a variable
- If we have sample data, we call this measure the sample mean and denote it by $\overline{X}$
- If we have population data, we call it the population mean and denote it by μ
- Formally, we compute

the mean by:

$$\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- In Excel, calculate the mean with the AVERAGE function

Measures of Central Tendency: The Median

- The median is the middle observation when the data is sorted from smallest to largest, i.e., in *ascending order*
  - If the number of observations is odd, the median is the middle observation
  - If the number of observations is even, the median is the average of the two middle observations
- One advantage of the median over the mean is that it is not sensitive to outliers
- In Excel, calculate the median with the MEDIAN function

Measures of Central Tendency: The Mode

- The mode is the value that appears most often
  - Not very interesting for continuous numerical data, but can be useful for discrete numerical or categorical data
- In Excel, calculate the mode with the MODE function

Minimum, Maximum, Percentiles, and Quartiles

- For any percentage p, the p^{th} percentile is the value such that p% of all values are less than it
  - The median is a special case, i.e., the 50th percentile
- Quartiles divide the data into four approximately equal-sized groups
  - The 1st, 2nd, and 3rd correspond to the 25th, 50th, and and 75th percentiles
  - Similarly, deciles are divided into 10% percentiles
- In Excel, use MIN, MAX, PERCENTILE, and QUARTILE functions

Variability

- Why does it matter?
  - In operations and supply chain management, variability could mean less efficient processes or poor quality
  - In finance, variability could mean volatility and risk
  - In marketing, variability means heterogeneity, i.e., need to market to different types of consumers
- Less clear on how to calculate ➔ More measures!
  - Range and interquartile range
  - Variance and standard deviation
  - Mean absolute deviation

Measures of Variability: Range and Interquartile Range

- The range is the difference between the maximum and minimum values
  - Fairly crude measure of variability, very sensitive to outliers
- The interquartile range (IQR) is the difference between the 1st and 3rd quartiles (Q3 – Q1)
  - In other words, the range of the middle 50% of the data
  - Less sensitive to extreme values

Measures of Variability: Variance and Standard Deviation

- The variance is approximately the average of the squared deviations from the mean

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

  - Sample variance is denoted by S^2, population variance by σ^2
  - Difficult to interpret because of squared units (e.g. \$ → \$^{2})
- A more interpretable measure is the standard deviation, which is the square root of variance (S, σ)
- In Excel, use the VAR and STDEV functions

Interpreting Sample Variance

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

- If the observations are all close to the mean, variance will be relatively small
- If at least a few observations are far from the mean, the variance will be large
- Because deviations from the mean are squared, observations below the mean contribute the same amount to variance observations equally above the mean

Interpreting Sample Standard Deviation

- The interpretation of the standard deviation can be stated as three empirical rules
- If the variable is approx. normally distributed (symmetric and bell-shaped), then:
  - Approx. 68% of the observations are within one standard deviation of the mean $\overline{X} \pm s$
  - Approx. 95% of the observations are within two standard deviations of the mean $\overline{X} \pm 2s$
  - Approx. 99.7% of the observations are within three standard deviations of the mean $\overline{X} \pm 3s$
- Fortunately, many variables in real-world data are indeed approximately normally distributed

Measures of Variability: Mean Absolute Deviation

- The mean absolute deviation (MAD) is another measure of variability

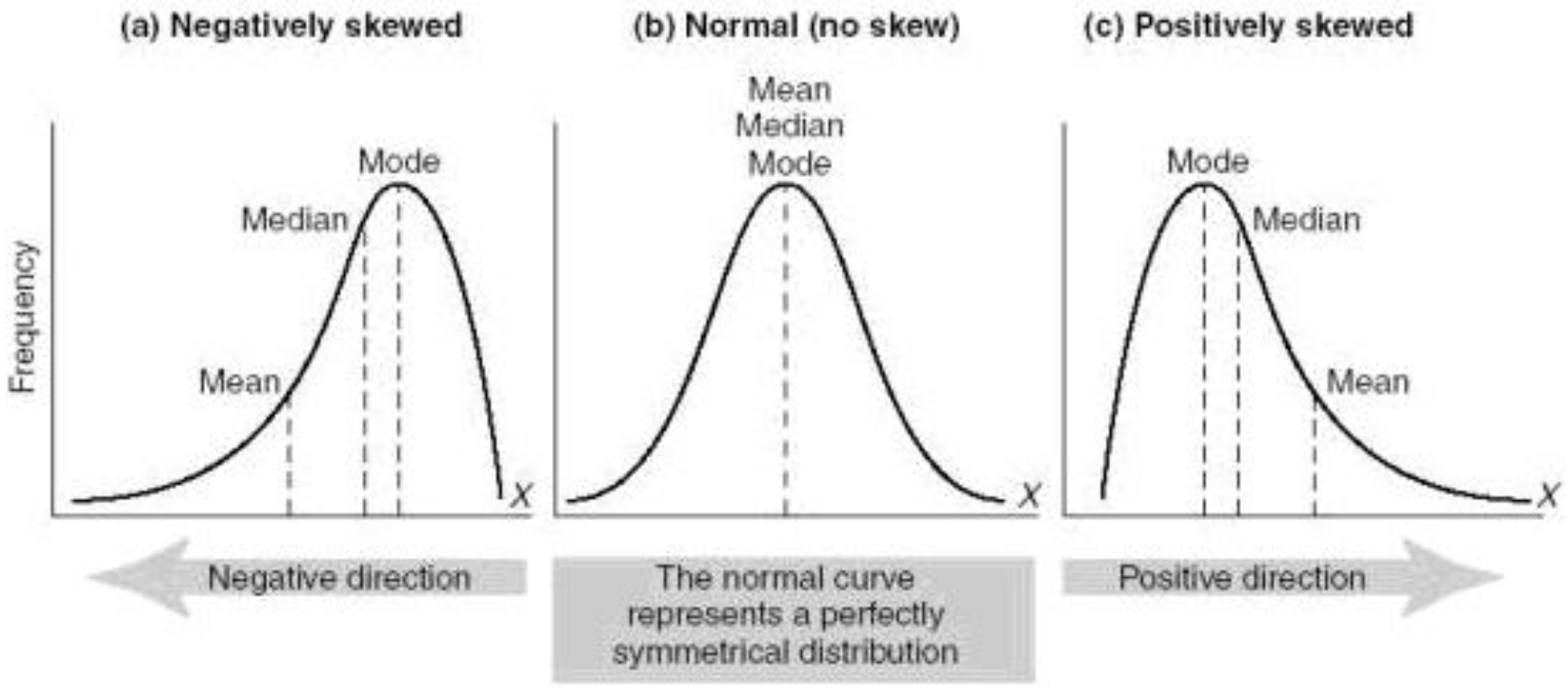
$$\text{MAD} = \frac{\sum_{i=1}^{n} \left(\left| X_i - \overline{X} \right| \right)}{n}$$

- In Excel, use the AVEDEV function

Measures of Shape: Skewness

- Skewness occurs because of a lack of symmetry
  - A variable can be skewed to the right (positively skewed) because of some really large values (e.g. professional athletes' salaries)
  - Or it can be skewed to the left (negatively skewed) because of some really small values (e.g. temperature lows in Antarctica)
- Skewness is easily seen through visualization
- In Excel, calculate skewness with the SKEW function

Central Tendency and Skewness



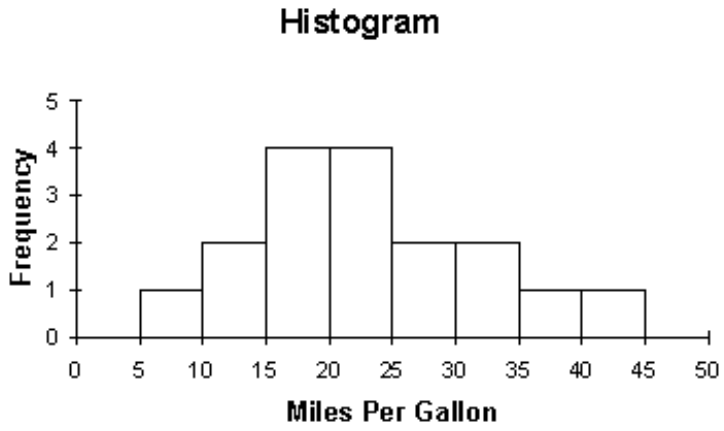
Measures of Shape: Kurtosis

- Kurtosis relates to the "peakedness" of the distribution or the "fatness" of the tails of the distribution relative to the Normal distribution
  - A distribution with high kurtosis has many extreme observations
- In Excel, calculate Kurtosis with the KURT function

Visualizing Numerical Data

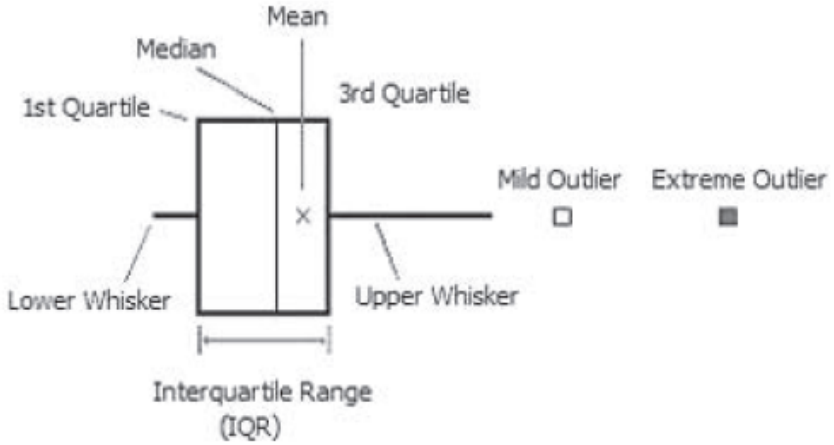
Histogram

- Most common type
- Based on binning the variable and plotting the frequency or proportion of each bin
- Great for showing the shape of a distribution



Box Plot (Box-and-whisker)

- Percentile-based plot for visualizing the distribution of a variable
- Side-by-side box plots are often used for comparing distributions
  - E.g., Gross sales for movie genres



Visualizing Numerical Data

What to look for:

- Where is the center?
  - Mean, median, mode
- What is the variability?
  - Range?
  - Variance/standard deviation?
  - Min, max?
  - Is it bounded?
- What is the shape?
  - Number of peaks? Is it skewed?
- Are there outliers?

FIGURE 2.6
A Symmetric Data Pattern

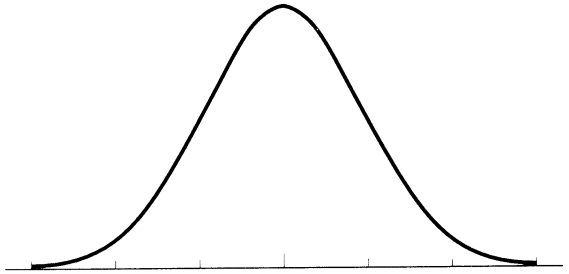


FIGURE 2.7
A Right-Tailed or Right Skewed Data Pattern

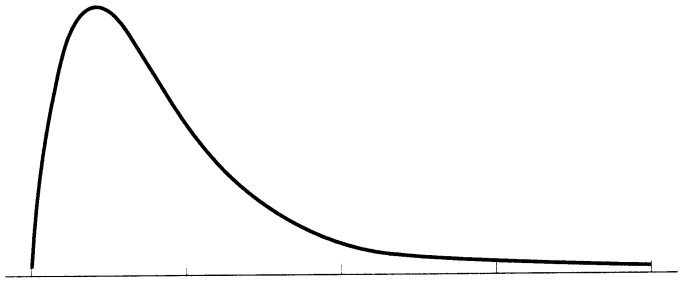
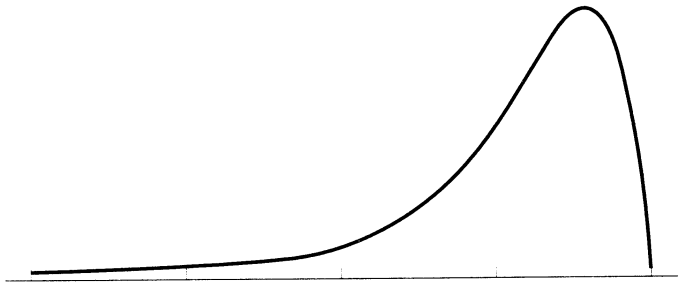
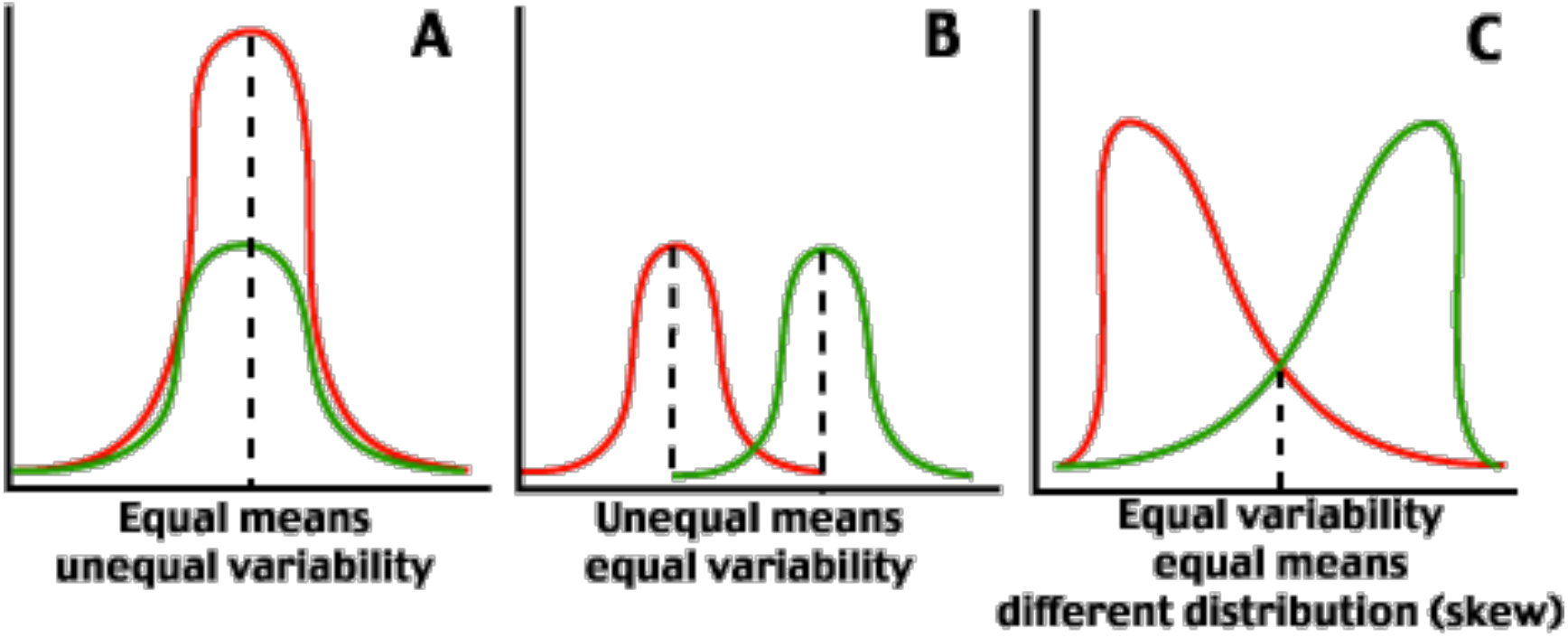


FIGURE 2.8
A Left-Tailed or Left Skewed Data Pattern



Central Tendency, Variability, and Shape



Example: 2009 MLB Player Baseball Salaries

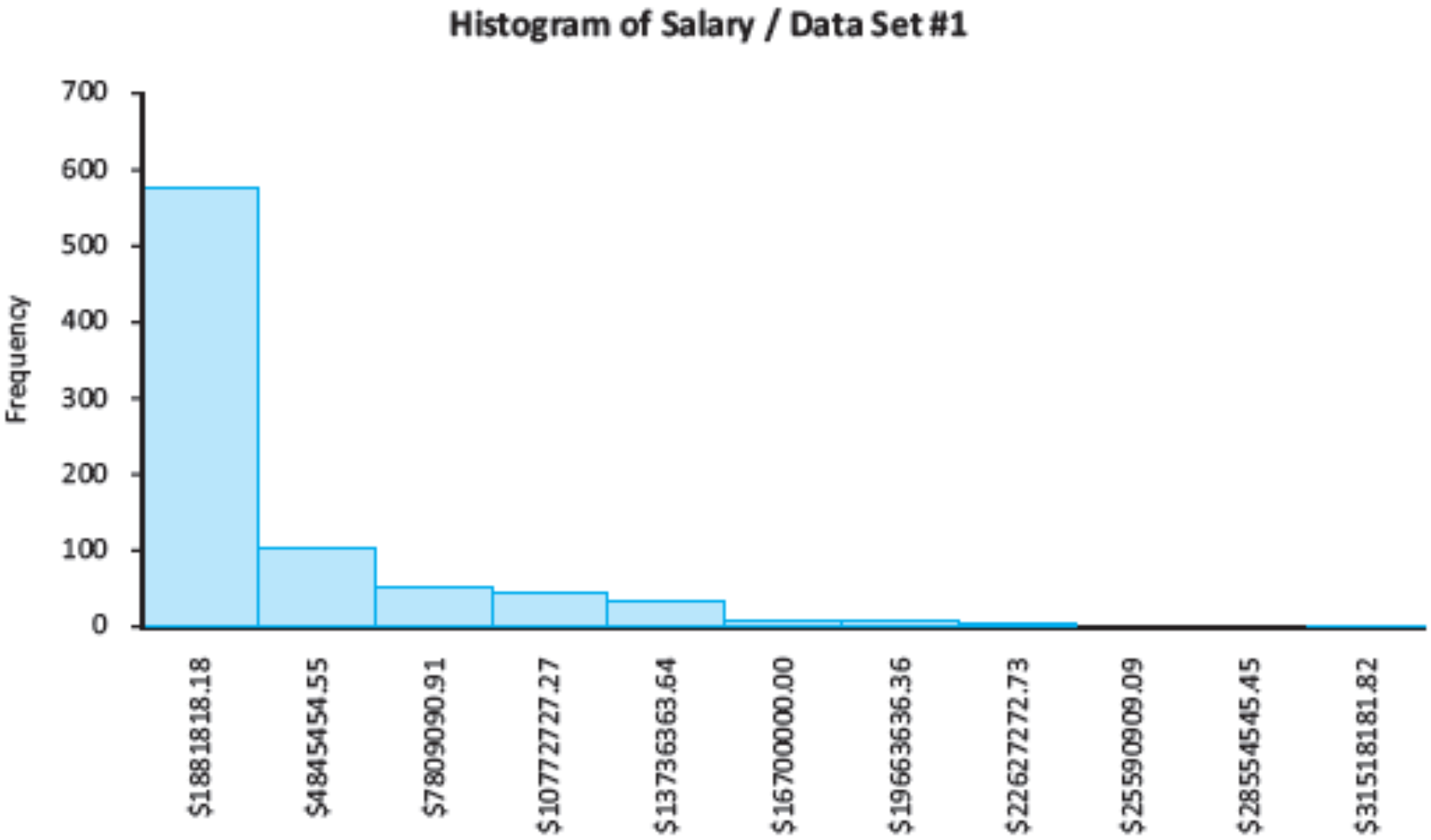
Player	Team	Position	Salary
Aardsma, Dave	Seattle Mariners	Pitcher	$419,000
Abreu, Bobby	Los Angeles Angels	Outfielder	$5,000,000
Adams, Mike	San Diego Padres	Pitcher	$414,800
Adenhart, Nick	Los Angeles Angels	Pitcher	$400,000
Affeldt, Jeremy	San Francisco Giants	Pitcher	$3,500,000
Albaladejo, Jon	New York Yankees	Pitcher	$403,075
Albers, Matt	Baltimore Orioles	Pitcher	$410,000
Amezaga, Alfredo	Florida Marlins	Shortstop	$1,300,000
Anderson, Brett	Oakland Athletics	Pitcher	$400,000
Anderson, Brian Nikoli	Chicago White Sox	Outfielder	$440,000
Anderson, Garret	Atlanta Braves	Outfielder	$2,500,000
Anderson, Josh	Detroit Tigers	Outfielder	$400,000
Anderson, Marlon	New York Mets	Second Baseman	$1,150,000
Andino, Robert	Baltimore Orioles	Infielder	$400,000

Summarizing Salary Data

Statistic	Value
Mean	$3,260,059
Variance	19,045,050,733,784
Std. Dev.	$4,364,064
Skewness	2.10
Median	$1,151,000
Mode	$400,000
Minimum	$400,000
Maximum	$33,000,000
Range	$32,600,000
Count	818
Sum	$2,666,728,494
1st Quartile	$419,550
3rd Quartile	$4,237,500

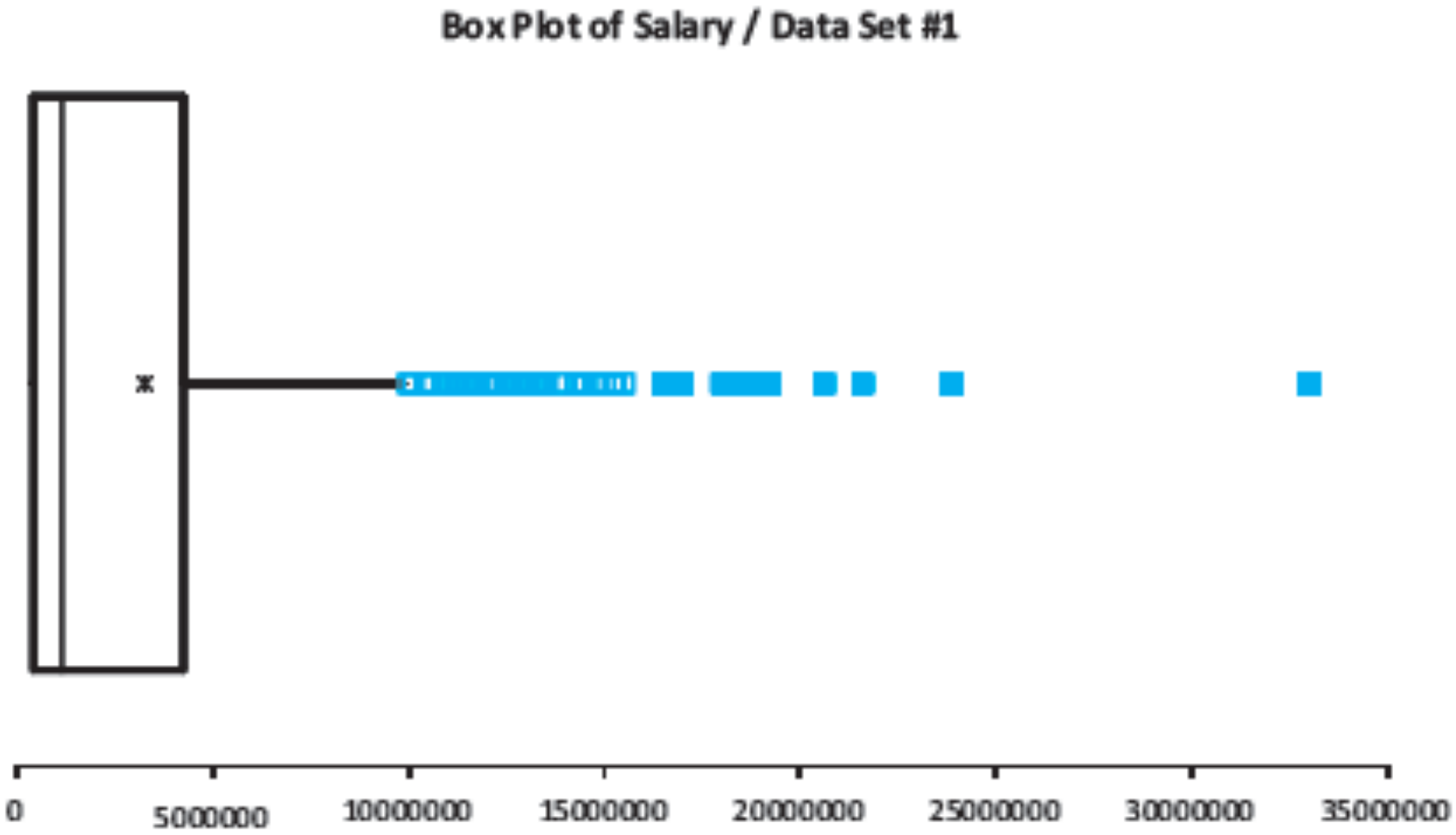
MLB Player Salaries -- Histogram Plot

- Create with StatTools or standard Analysis ToolPack add-in



MLB Player Salaries -- Boxplot

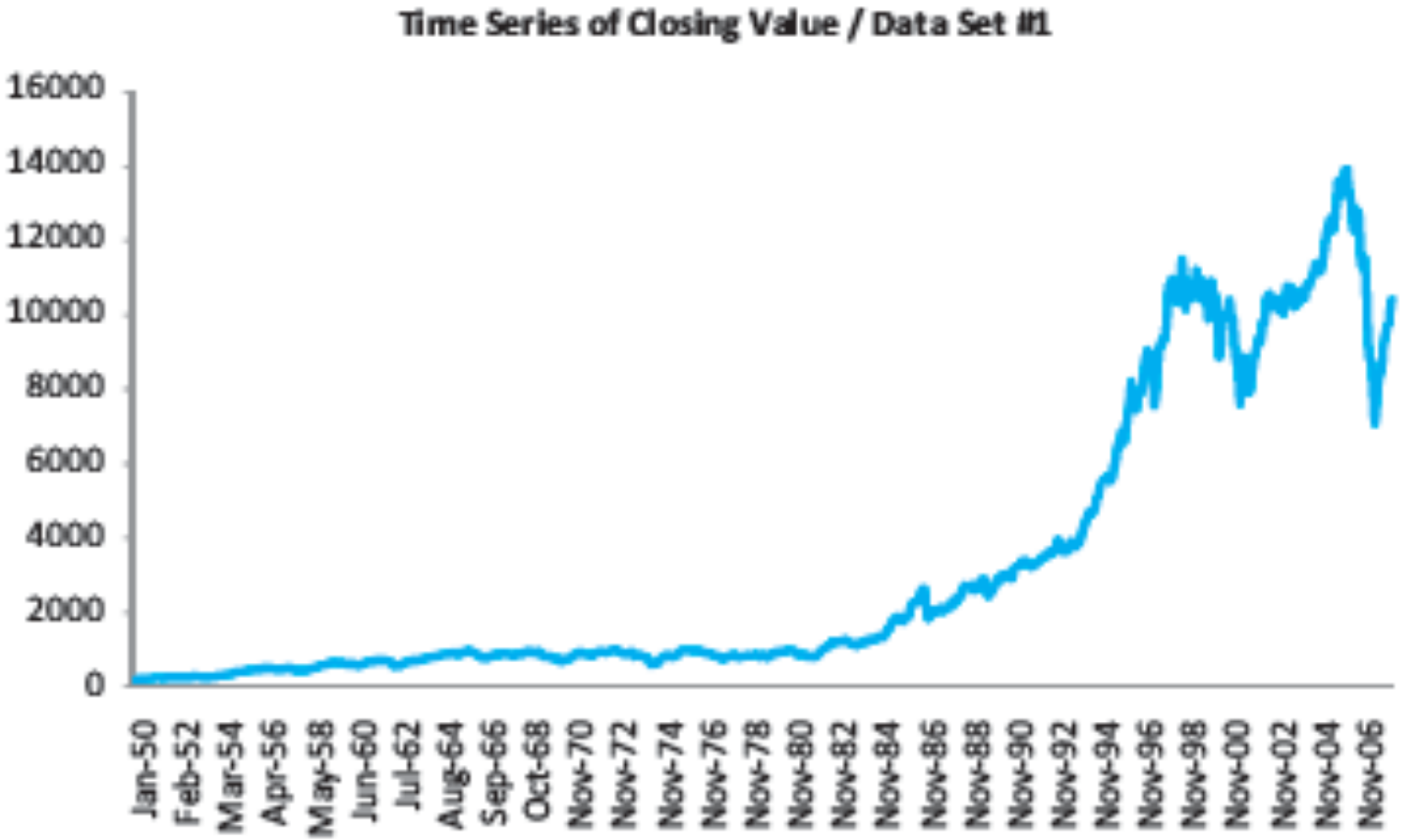
- Create with StatTools



Time Series Data

- Summary measures and visualizations used for cross-sectional data do not make much sense
- Our main interest is how variables change over time
  - This information is lost in traditional summary measures and in histograms or boxplots
- Instead, we use **line charts**, which plot one or more time series variables with time on the x-axis

Example: DJIA Monthly Close



Visualizing Time Series Data -- Sparklines

- Miniature data visualizations used for time series
- Available as a special Chart type in Excel 2010 and newer, select data range and insert into single cell

Daily price history for Dow Jones and S&P 500 indices
Line chart

Run differential for first 26 games of 2013 Chicago Cubs baseball season
Win-Loss chart

Outliers

- An outlier is an observation that lies well outside of the norm, with respect to one variable or a combination of variables
- General rule of thumb
  - An outlier is any value more than three standard deviations from the mean
- Best practice
  - Run analysis two ways: With outliers and without
- Applications – Outlier/anomaly detection
  - Fraud detection, diagnostic medicine, (structural) fault detection, superstar athletes

Missing Values

- As with outliers, we need to know how to detect missing values and what to do about them
- Missing values are coded in many ways (e.g., NA, blank)
  - In Excel, do a Find/Replace to standardize missing values
- More importantly, what to do with missing values?
  1. Ignore them, but you need to know how the software deals with them
  2. Fill in missing values with central measure of existing values
  3. Examine the existing values in the row of a missing value; they may provide information on what a missing value should be

Next Time…

Finding Relationships among Variables