# DATA ANALYTICS

DR. BRENDA MULLALLY

# EXAMPLE 2.3: BASEBALL SALARIES 2011.XLSX

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Measures of central tendency** | | | | **Measures of variability** | |
| 2 | Mean | $3,305,055 | | | Range | $31,586,000 |
| 3 | Median | $1,175,000 | | | Interquartile range | $3,875,925 |
| 4 | Mode | $414,000 | 57 | | Variance | 20,563,887,478,833 |
| 5 | | | | | Standard deviation | $4,534,742 |
| 6 | **Min, max, percentiles, quartiles** | | | | Mean absolute deviation | $3,249,917 |
| 7 | Min | $414,000 | | | | |
| 8 | Max | $32,000,000 | | | **Measures of shape** | |
| 9 | P01 | $414,000 | 0.01 | | Skewness | 2.2568 |
| 10 | P05 | $414,000 | 0.05 | | Kurtosis | 5.7233 |
| 11 | P10 | $416,520 | 0.10 | | | |
| 12 | P20 | $424,460 | 0.20 | | **Percentages of values less than given values** | |
| 13 | P50 | $1,175,000 | 0.50 | | Value | Percentage less than |
| 14 | P80 | $5,500,000 | 0.80 | | $1,000,000 | 46.38% |
| 15 | P90 | $9,800,000 | 0.90 | | $1,500,000 | 54.69% |
| 16 | P95 | $13,590,000 | 0.95 | | $2,000,000 | 58.36% |
| 17 | P99 | $20,000,000 | 0.99 | | $2,500,000 | 63.23% |
| 18 | Q1 | $430,325 | 1 | | $3,000,000 | 66.55% |
| 19 | Q2 | $1,175,000 | 2 | | | |
| 20 | Q3 | $4,306,250 | 3 | | | |

# MEASURES OF VARIABILITY

- The **range** is the maximum value minus the minimum value.
- The **interquartile range** (**IQR**) is the third quartile minus the first quartile.
  - Thus, it is the range of the middle 50% of the data.
  - It is less sensitive to extreme values than the range.
- The **variance** is essentially the average of the squared deviations from the mean.
  - If $X_i$ is a typical observation, its squared deviation from the mean is $(X_i - mean)^2$.

# MEASURES OF VARIABILITY

- The **sample variance** is denoted by $s^2$, and the **population variance** by $\sigma^2$.

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n - 1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n}$$

  - If all observations are close to the mean, their squared deviations from the mean—and the variance—will be relatively small.
  - If at least a few of the observations are far from the mean, their squared deviations from the mean—and the variance—will be large.
  - In Excel, use the *VAR* function to obtain the sample variance and the *VARP* function to obtain the population variance.

# MEASURES OF VARIABILITY

- A fundamental problem with variance is that it is in squared units (e.g., \$ → $\$^2$).
- A more natural measure is the **standard deviation**, which is the square root of variance.
  - The **sample standard deviation**, denoted by *s*, is the square root of the sample variance.
  - The **population standard deviation**, denoted by $\sigma$, is the square root of the population variance.
  - In Excel, use the *STDEV* function to find the sample standard deviation or the *STDEVP* function to find the population standard deviation.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Low variability supplier** | | | | **High variability supplier** | |
| 2 | | | | | | |
| 3 | Diameter1 | Sq dev from mean | | | Diameter2 | Sq dev from mean |
| 4 | 102.61 | 6.610041 | | | 103.21 | 9.834496 |
| 5 | 103.25 | 10.310521 | | | 93.66 | 41.139396 |
| 6 | 96.34 | 13.682601 | | | 120.87 | 432.473616 |
| 7 | 96.27 | 14.205361 | | | 110.26 | 103.754596 |
| 8 | 103.77 | 13.920361 | | | 117.31 | 297.079696 |
| 9 | 97.45 | 6.702921 | | | 110.23 | 103.144336 |
| 10 | 98.22 | 3.308761 | | | 70.54 | 872.257156 |
| 11 | 102.76 | 7.403841 | | | 39.53 | 3665.575936 |
| 12 | 101.56 | 2.313441 | | | 133.22 | 1098.657316 |
| 13 | 98.16 | 3.530641 | | | 101.91 | 3.370896 |
| 14 | | | | | | |
| 15 | Mean | | | | Mean | |
| 16 | 100.039 | | | | 100.074 | |
| 17 | | | | | | |
| 18 | Sample variance | | | | Sample variance | |
| 19 | 9.1098 | 9.1098 | | | 736.3653 | 736.3653 |
| 20 | | | | | | |
| 21 | Population variance | | | | Population variance | |
| 22 | 8.1988 | 8.1988 | | | 662.7287 | 662.7287 |
| 23 | | | | | | |
| 24 | Sample standard deviation | | | | Sample standard deviation | |
| 25 | 3.0182 | 3.0182 | | | 27.1361 | 27.1361 |
| 26 | | | | | | |
| 27 | Population standard deviation | | | | Population standard deviation | |
| 28 | 2.8634 | 2.8634 | | | 25.7435 | 25.7435 |

# EMPIRICAL RULES FOR INTERPRETING STANDARD DEVIATION

- The interpretation of the standard deviation can be stated as three **empirical rules**.
  - If the values of a variable are approximately *normally* distributed (symmetric and bell-shaped), then the following rules hold:
    - Approximately 68% of the observations are within one standard deviation of the mean.
    - Approximately 95% of the observations are within two standard deviations of the mean.
    - Approximately 99.7% of the observations are within three standard deviations of the mean.

# EMPIRICAL RULES FOR BASEBALL SALARIES
(SLIDE 2 OF 3)

- The empirical rules should be applied with caution, especially when the data are clearly skewed, as illustrated by the calculations for baseball salaries below.

| | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|
| 1 | **Do empirical rules apply?** | | | | | | | |
| 2 | | Lower endpoint | Upper endpoint | # below lower | # above upper | % below lower | % above upper | % between |
| 3 | Rule 1 | -$1,229,688 | $7,839,797 | 0 | 108 | 0% | 13.20% | 86.80% |
| 4 | Rule 2 | -$5,764,430 | $12,374,539 | 0 | 54 | 0% | 6.60% | 93.40% |
| 5 | Rule 3 | -$10,299,172 | $16,909,281 | 0 | 19 | 0% | 2.32% | 97.68% |

# MEASURES OF SHAPE
## (SLIDE 1 OF 2)

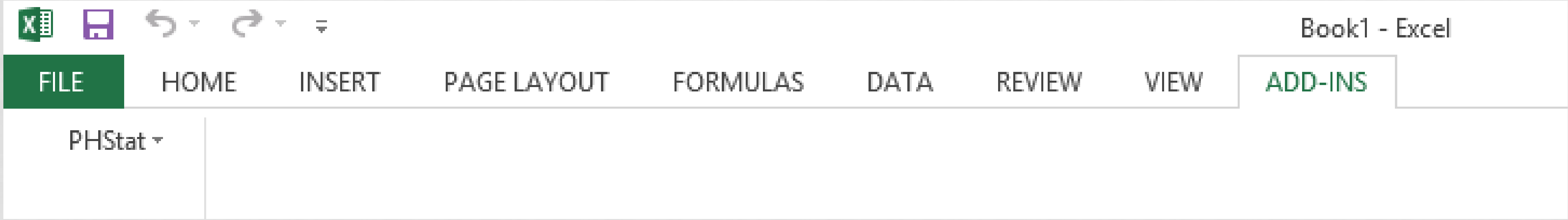
- **Skewness** occurs when there is a lack of symmetry.
  - A variable can be **skewed to the right** (or **positively skewed**) because of some really *large* values (e.g., really large baseball salaries).
  - Or it can be **skewed to the left** (or **negatively skewed**) because of some really *small* values (e.g., temperature lows in Antarctica).
- In Excel, a measure of skewness can be calculated with the *SKEW* function.

# MEASURES OF SHAPE
## (SLIDE 2 OF 2)

- **Kurtosis** has to do with the “fatness” of the tails of the distribution relative to the tails of a normal distribution.
- A distribution with high kurtosis has many more extreme observations.
- In Excel, kurtosis can be calculated with the *KURT* function.

# EXCEL TOOLS & ADD-INS

- Excel's built in functions (average, stdev and others) were used to calculate a number of summary measures. You can generate the same results using an add-in. In the next lab there is a download for PhStat add-in.
- Once downloaded and you run the PHStat.xlam file the  it appears in the Add-ins tab in excel.



# CHARTS FOR NUMERICAL VARIABLES

- There are many graphical ways to indicate the distribution of a numerical variable.
  - For cross-sectional variables:
    - Histograms
    - Box plots
  - For time series variables:
    - Time series graphs

# HISTOGRAMS

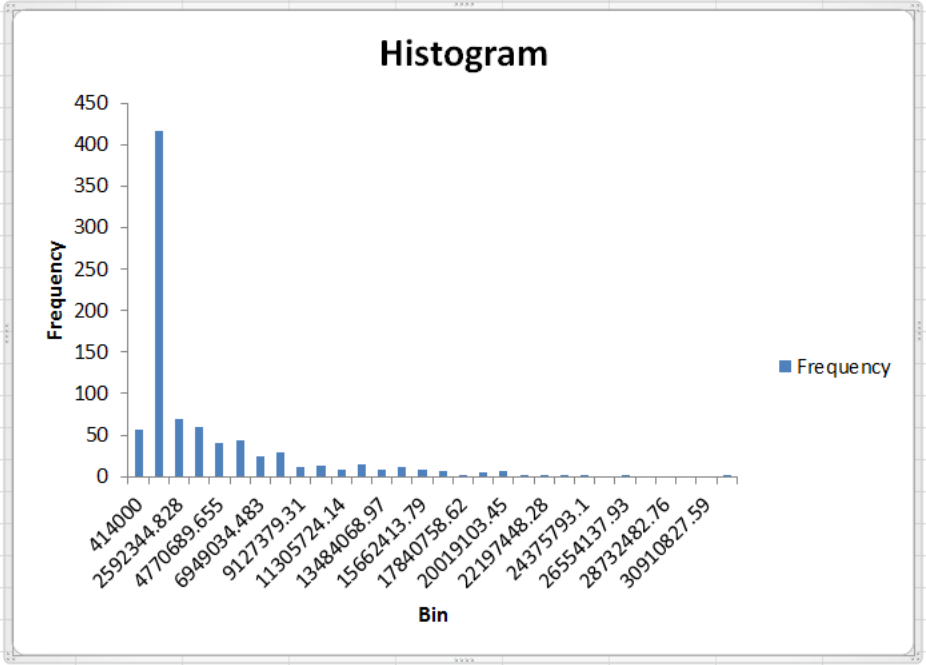
- A **histogram** is the most common type of chart for showing the distribution of a numerical variable.
  - It is based on binning the variable—that is, dividing it up into discrete categories.
  - It is a column chart of the counts in the various categories (with no gaps between the vertical bars).
- A histogram is great for showing the shape of a distribution—whether the distribution is symmetric or skewed in one direction.

# BASEBALL SALARIES 2011.XLSX

- **Objective**: To see the shape of the salary distribution through a histogram.
- **Solution**: It is possible to create a histogram with Excel tools only—but it can be a tedious process.
  - The resulting table of counts is usually called a **frequency table**.
  - The counts are called **frequencies**.
- It is easier to create a histogram with some add-ins but many of these are at a cost.

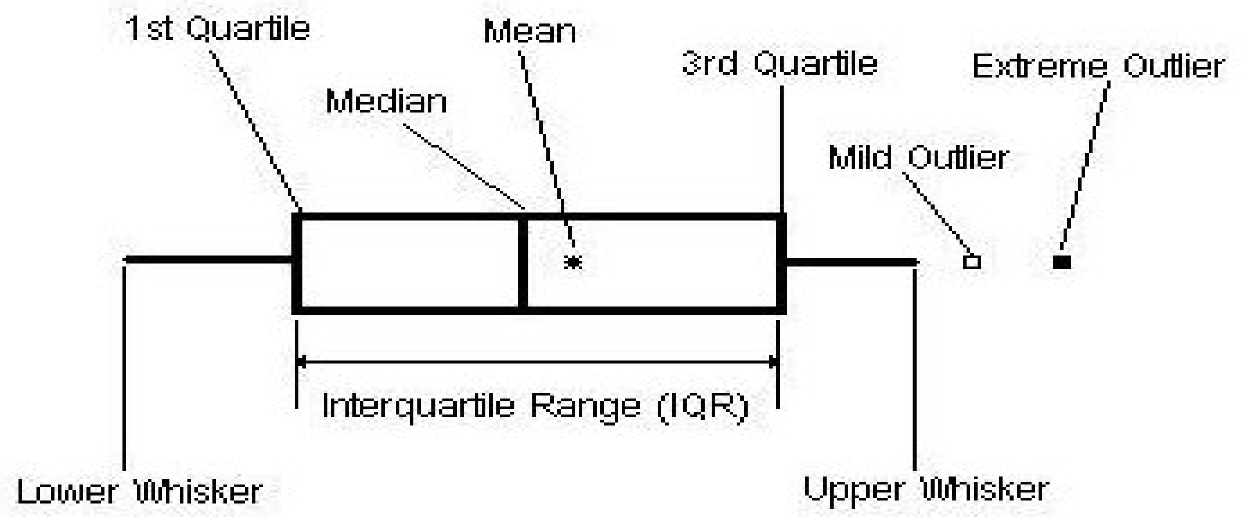
# BASEBALL SALARIES 2011.XLSX

| Bin | Frequency |
|---|---|
| 414000 | 57 |
| 1503172 | 417 |
| 2592345 | 70 |
| 3681517 | 60 |
| 4770690 | 41 |
| 5859862 | 43 |
| 6949034 | 25 |
| 8038207 | 29 |
| 9127379 | 11 |
| 10216552 | 13 |
| 11305724 | 9 |
| 12394897 | 14 |
| 13484069 | 9 |
| 14573241 | 12 |
| 15662414 | 8 |
| 16751586 | 6 |
| 17840759 | 1 |
| 18929931 | 5 |
| 20019103 | 6 |
| 21108276 | 1 |
| 22197448 | 1 |
| 23286621 | 2 |
| 24375793 | 1 |
| 25464966 | 0 |
| 26554138 | 1 |
| 27643310 | 0 |
| 28732483 | 0 |
| 29821655 | 0 |
| 30910828 | 0 |
| More | 1 |



# BOX PLOTS

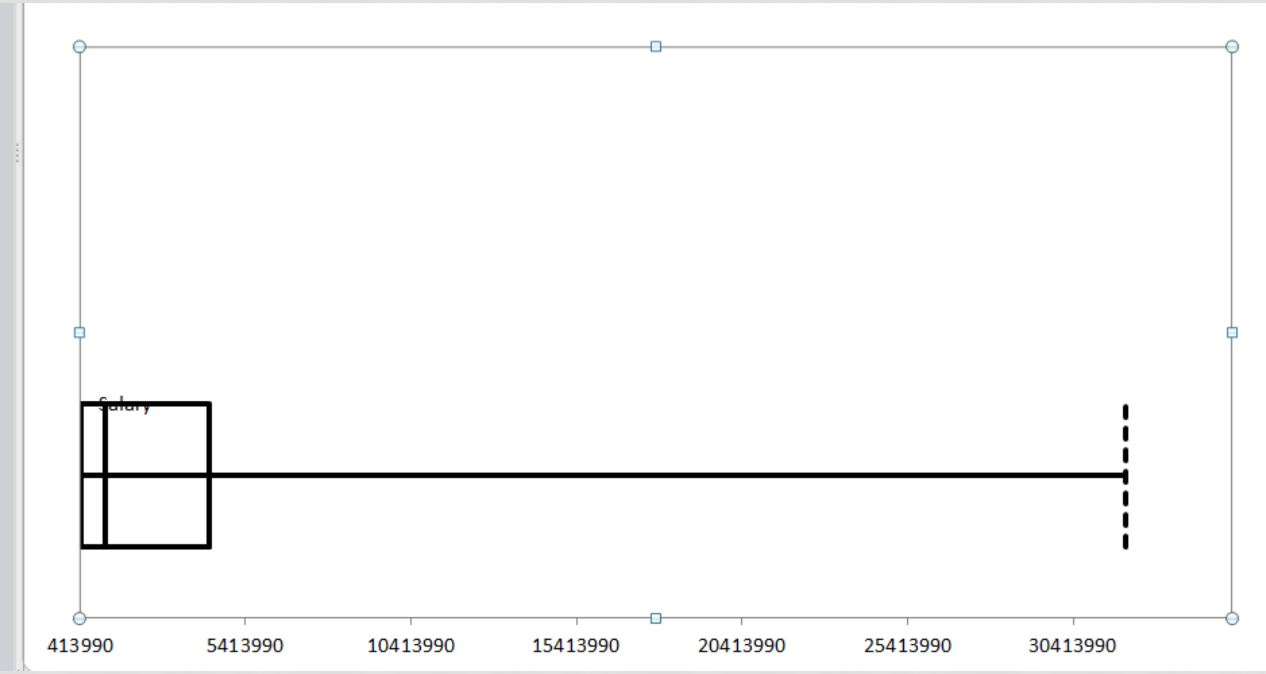
- A **box plot** (or **box-whisker plot**) is an alternative type of chart for showing the distribution of a variable.
  - The elements of a generic box plot are shown below:




Whiskers extend to the furthest observations that are no more than 1.5 IQR from the edges of the box. Mild outliers are observations between 1.5 IQR and 3 IQR from the edges of the box. Extreme outliers are greater than 3 IQR from the edges of the box.

# BASEBALL SALARIES 2011.XLSX

- **Objective**: To illustrate the features of a box plot, particularly how it indicates skewness.
- **Solution**: In PhStat, select Box-Plot from the descriptive statistics dropdown list and fill in the dialog box.



# TIME SERIES DATA

- Our main interest in time series variables is how they change over time, and this information is lost in traditional summary measures and in histograms or box plots.
- For time series data, a **time series graph** is used. This is a graph of the values of one or more time series, using time on the horizontal axis.
  - This is always the place to start a time series analysis.

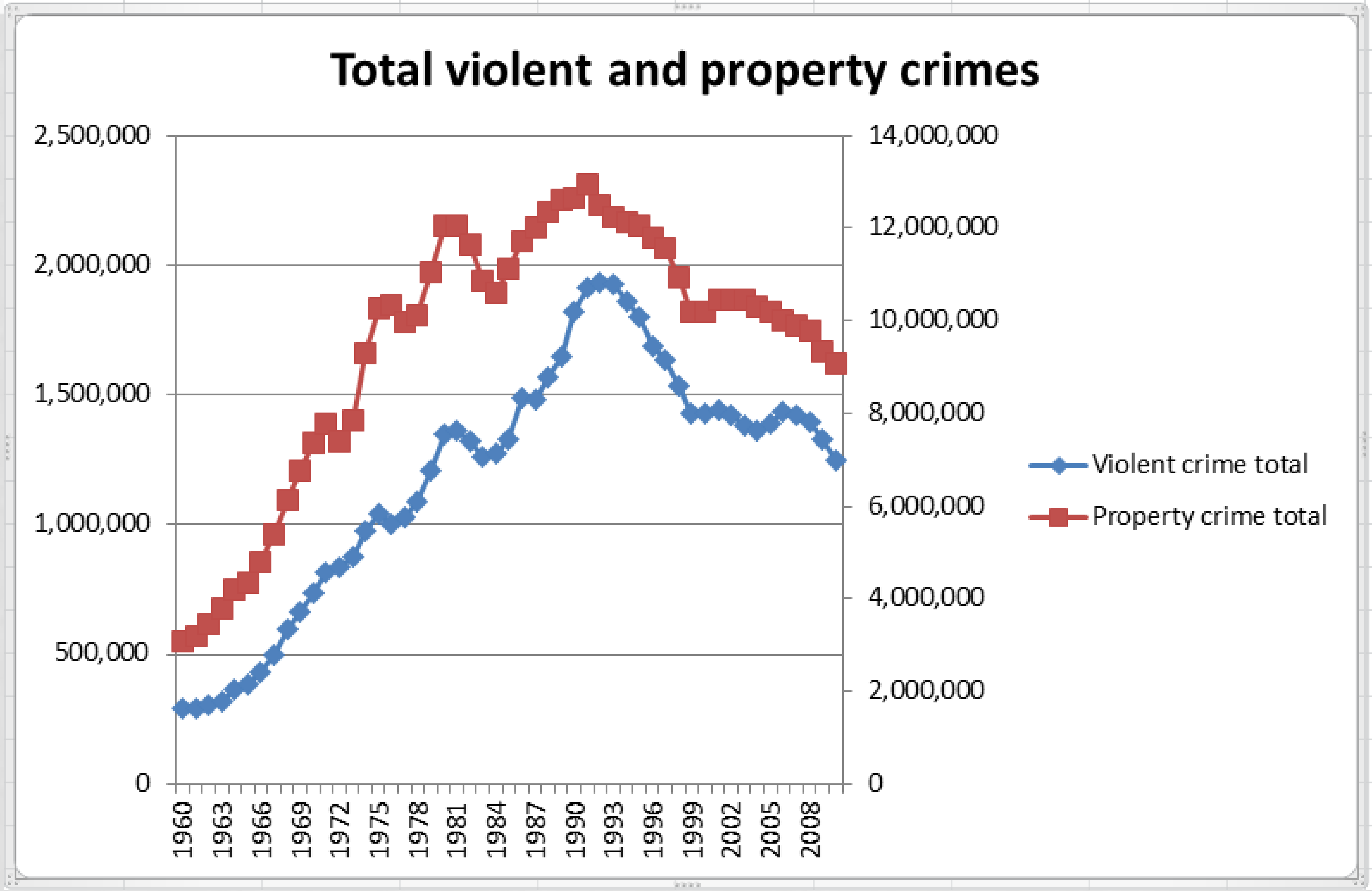
# CRIME IN US.XLSX

- **Objective**: To see how time series graphs help to detect trends in crime data.
- **Solution**: Data set contains annual data on violent and property crimes for the years 1960 to 2010.

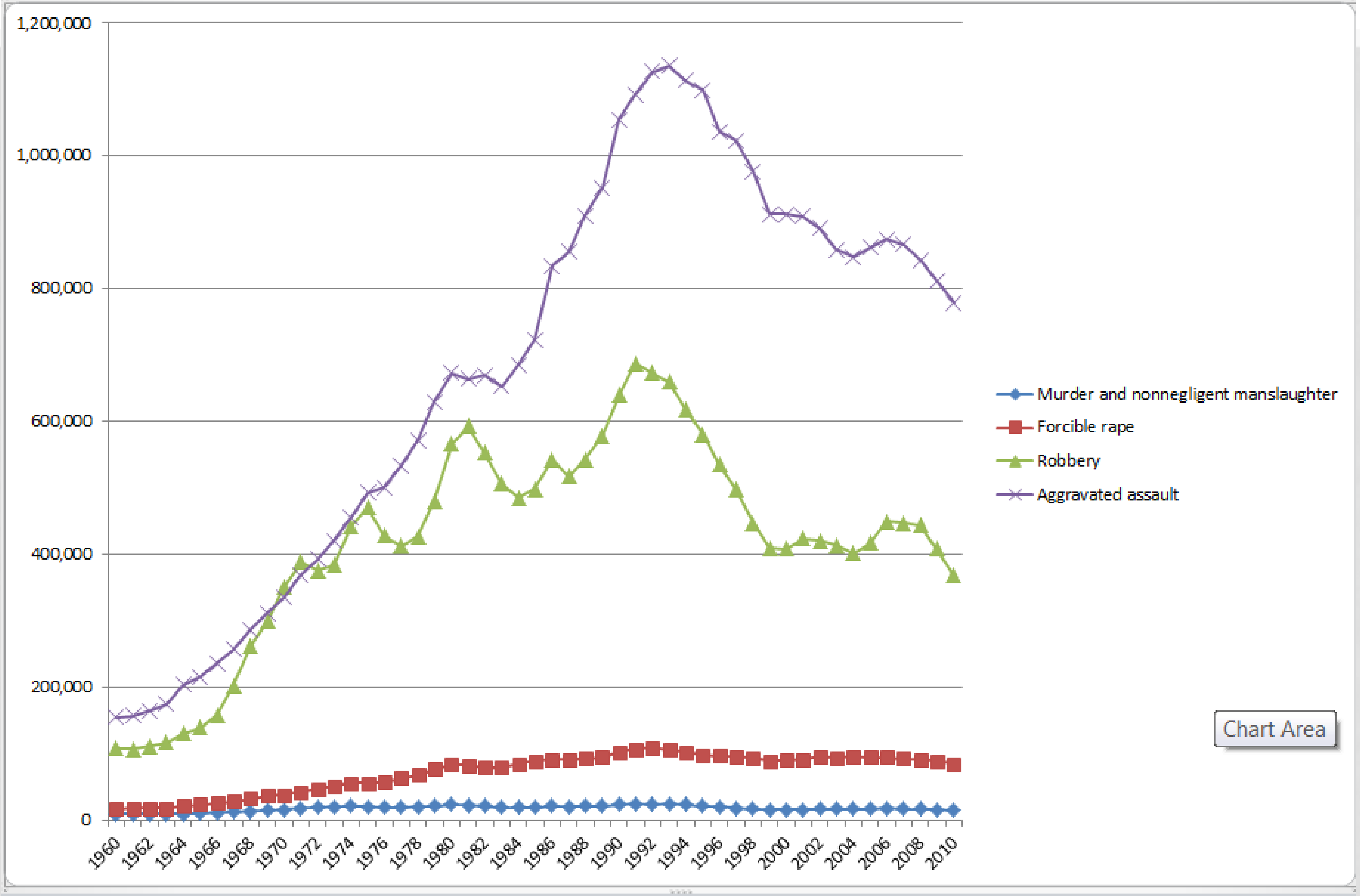
| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Year | Population | Violent crime total | Murder and nonnegligent manslaughter | Forcible rape | Robbery | Aggravated assault | Property crime total | Burglary | Larceny-theft | Motor vehicle theft |
| 2 | 1960 | 179,323,175 | 288,460 | 9,110 | 17,190 | 107,840 | 154,320 | 3,095,700 | 912,100 | 1,855,400 | 328,200 |
| 3 | 1961 | 182,992,000 | 289,390 | 8,740 | 17,220 | 106,670 | 156,760 | 3,198,600 | 949,600 | 1,913,000 | 336,000 |
| 4 | 1962 | 185,771,000 | 301,510 | 8,530 | 17,550 | 110,860 | 164,570 | 3,450,700 | 994,300 | 2,089,600 | 366,800 |
| 5 | 1963 | 188,483,000 | 316,970 | 8,640 | 17,650 | 116,470 | 174,210 | 3,792,500 | 1,086,400 | 2,297,800 | 408,300 |
| 6 | 1964 | 191,141,000 | 364,220 | 9,360 | 21,420 | 130,390 | 203,050 | 4,200,400 | 1,213,200 | 2,514,400 | 472,800 |
| 7 | 1965 | 193,526,000 | 387,390 | 9,960 | 23,410 | 138,690 | 215,330 | 4,352,000 | 1,282,500 | 2,572,600 | 496,900 |
| 8 | 1966 | 195,576,000 | 430,180 | 11,040 | 25,820 | 157,990 | 235,330 | 4,793,300 | 1,410,100 | 2,822,000 | 561,200 |
| 9 | 1967 | 197,457,000 | 499,930 | 12,240 | 27,620 | 202,910 | 257,160 | 5,403,500 | 1,632,100 | 3,111,600 | 659,800 |
| 10 | 1968 | 199,399,000 | 595,010 | 13,800 | 31,670 | 262,840 | 286,700 | 6,125,200 | 1,858,900 | 3,482,700 | 783,600 |

# CRIME IN US.XLSX

Total Violent and Property Crimes

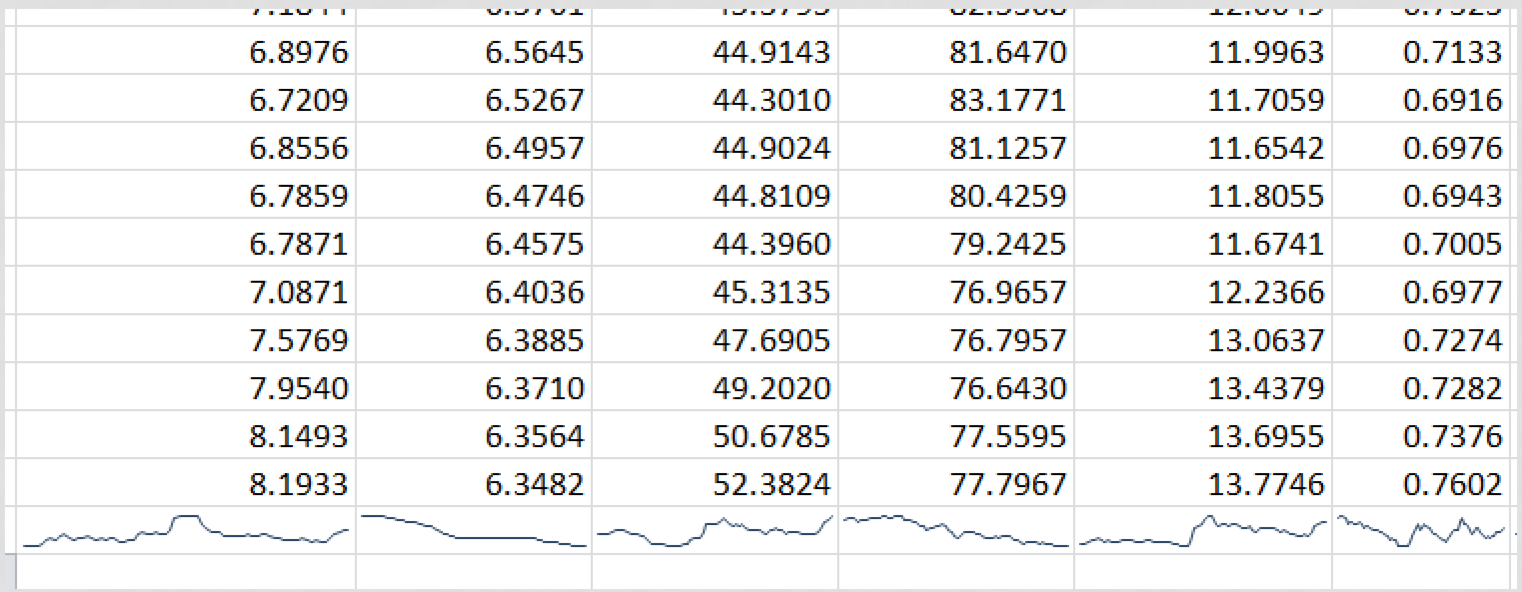


# CRIME IN US.XLSX



# SPARKLINE GRAPH

- New to Excel 2010 is the mini-chart embedded in a cell. It is especially useful for time series data.
- In the cell under a set of time series data include a sparkline



| | | | | | |
|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| 6.8976 | 6.5645 | 44.9143 | 81.6470 | 11.9963 | 0.7133 |
| 6.7209 | 6.5267 | 44.3010 | 83.1771 | 11.7059 | 0.6916 |
| 6.8556 | 6.4957 | 44.9024 | 81.1257 | 11.6542 | 0.6976 |
| 6.7859 | 6.4746 | 44.8109 | 80.4259 | 11.8055 | 0.6943 |
| 6.7871 | 6.4575 | 44.3960 | 79.2425 | 11.6741 | 0.7005 |
| 7.0871 | 6.4036 | 45.3135 | 76.9657 | 12.2366 | 0.6977 |
| 7.5769 | 6.3885 | 47.6905 | 76.7957 | 13.0637 | 0.7274 |
| 7.9540 | 6.3710 | 49.2020 | 76.6430 | 13.4379 | 0.7282 |
| 8.1493 | 6.3564 | 50.6785 | 77.5595 | 13.6955 | 0.7376 |
| 8.1933 | 6.3482 | 52.3824 | 77.7967 | 13.7746 | 0.7602 |

# DJIA MONTHLY CLOSE.XLSX

- **Objective**: To find useful ways to summarize the monthly Dow data.
- **Solution**: Data set contains monthly values of the Dow from 1950 through 2011.
- Create summary measures and time series graphs for monthly values and percentage changes of the Dow.

| | Closing Value |
|---|---|
| *One Variable Summary* | DJIA Data |
| Mean | 3484.13 |
| Std. Dev. | 4044.57 |
| Median | 969.26 |
| 1st Quartile | 764.58 |
| 3rd Quartile | 5616.21 |

| | Percentage Change |
|---|---|
| *One Variable Summary* | DJIA Data |
| Mean | 0.00642 |
| Std. Dev. | 0.04182 |
| Median | 0.00851 |
| 1st Quartile | -0.01721 |
| 3rd Quartile | 0.03289 |

# DJIA MONTHLY CLOSE.XLSX