BUSINESS ANALYTICS: CHARTS

DR. BRENDA MULLALLY

CHARTS FOR NUMERICAL VARIABLES

- THERE ARE MANY GRAPHICAL WAYS TO INDICATE THE DISTRIBUTION OF A NUMERICAL VARIABLE.
  - FOR CROSS-SECTIONAL DATA:
    - HISTOGRAMS
    - BOX PLOTS
  - FOR TIME SERIES DATA:
    - TIME SERIES GRAPHS

CROSS SECTIONAL DATA: 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

Histogram
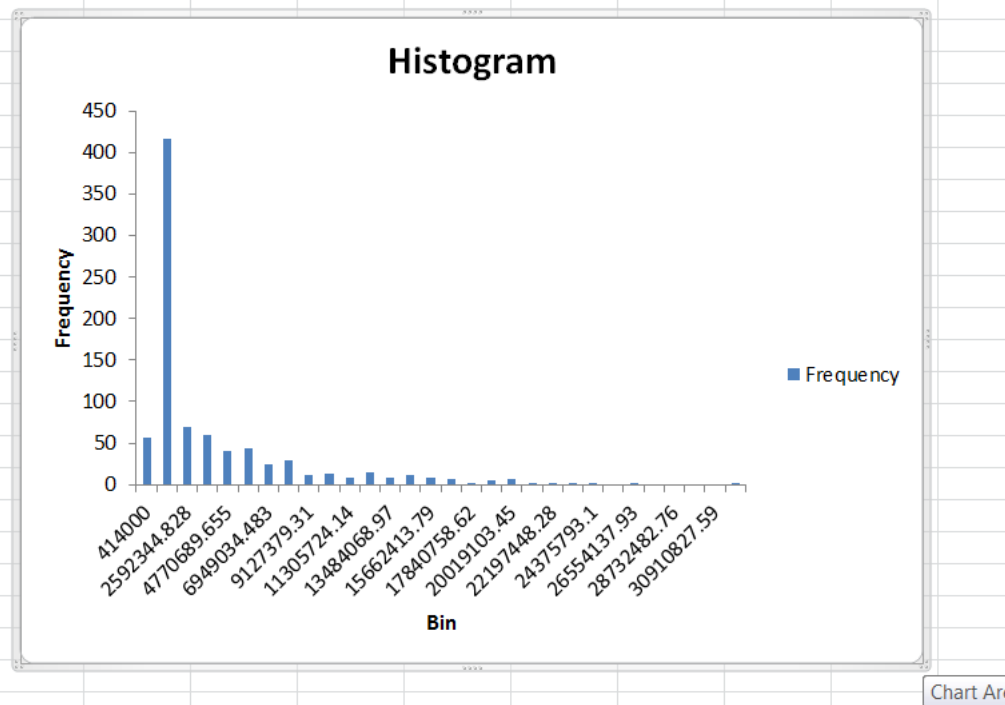
CROSS SECTIONAL DATA:
BOX PLOTS

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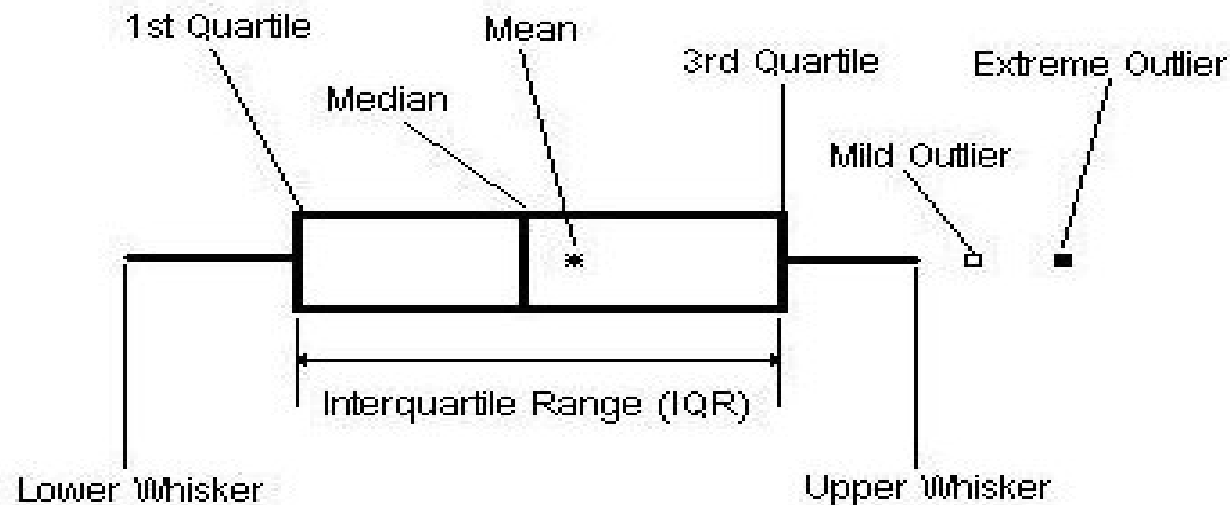
Whiskers extend to the furthest observations that are no more than 1.5 IQR from the edges of the box. Mild outliers are observations between 1.5 IQR and 3 IQR from the edges of the box. Extreme outliers are greater than 3 IQR from the edges of the box.

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.
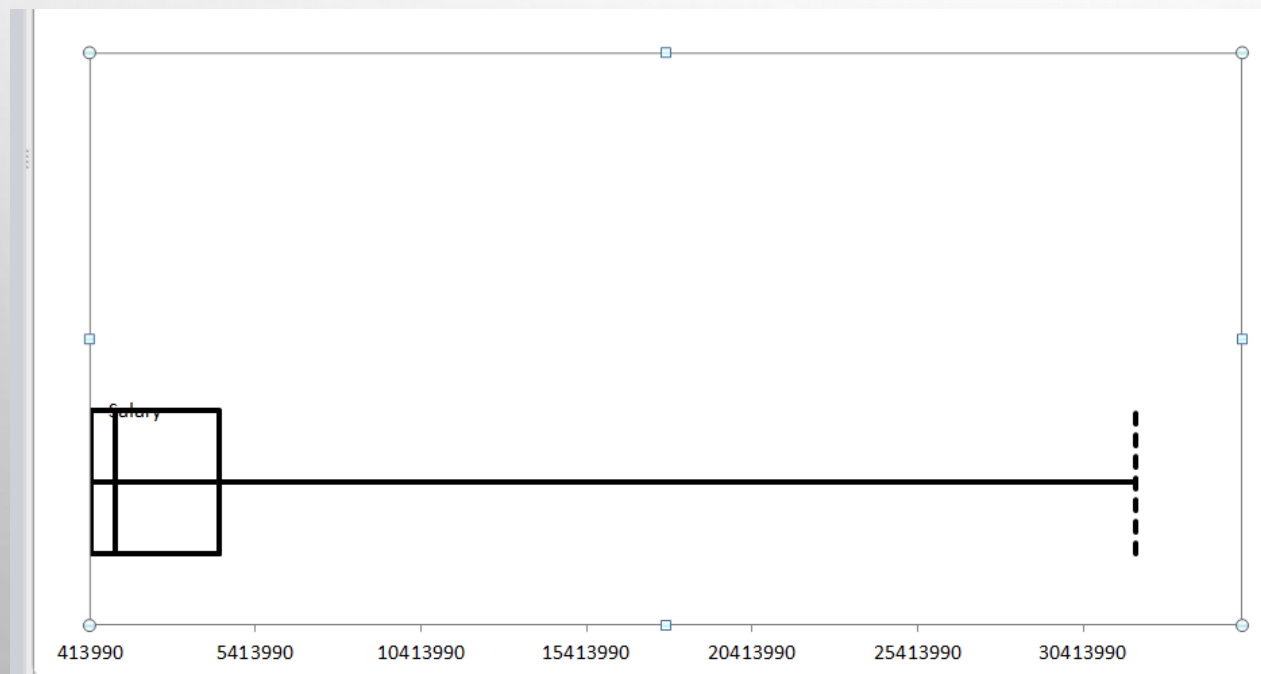
TIME SERIES DATA:
LINE GRAPH

- 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
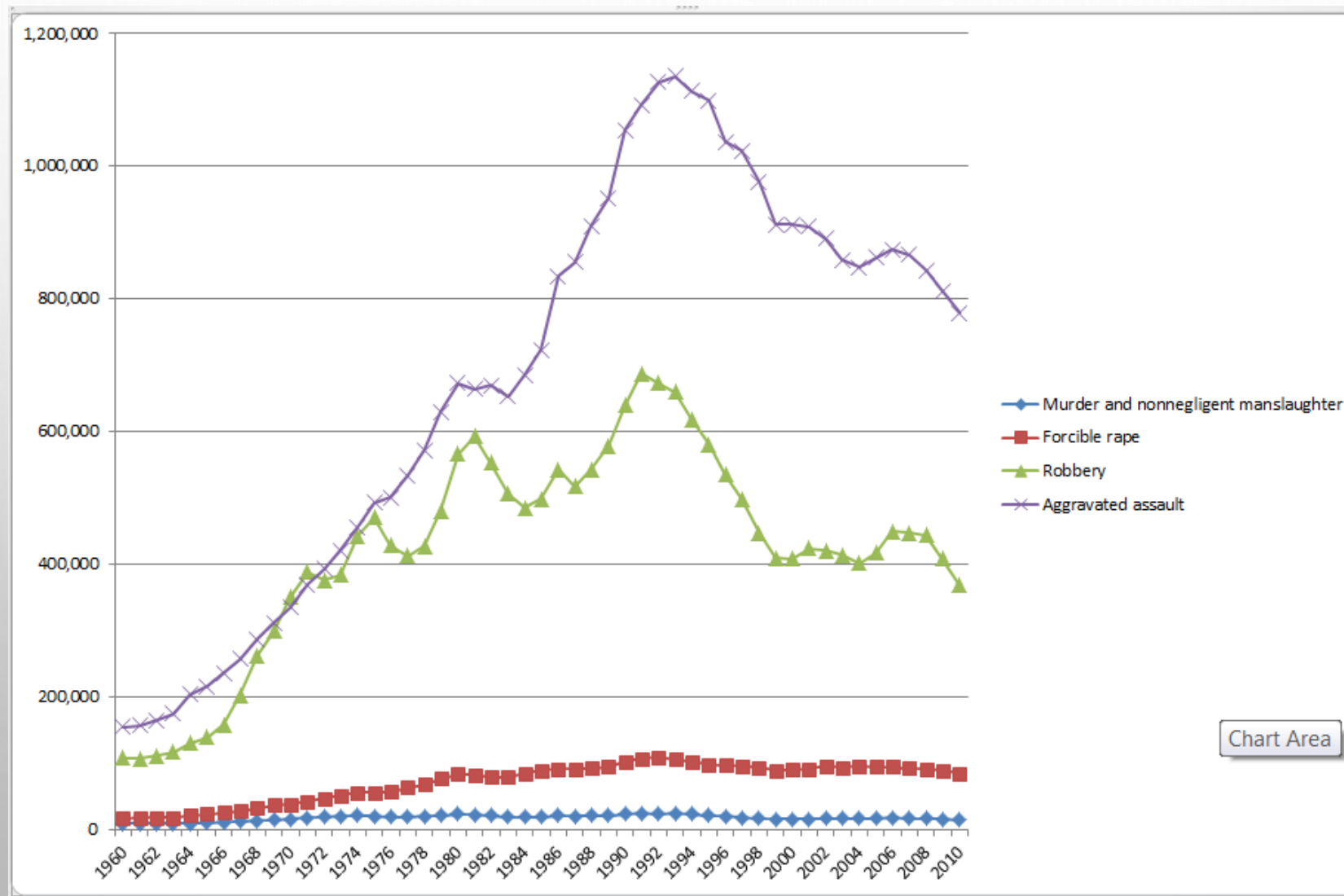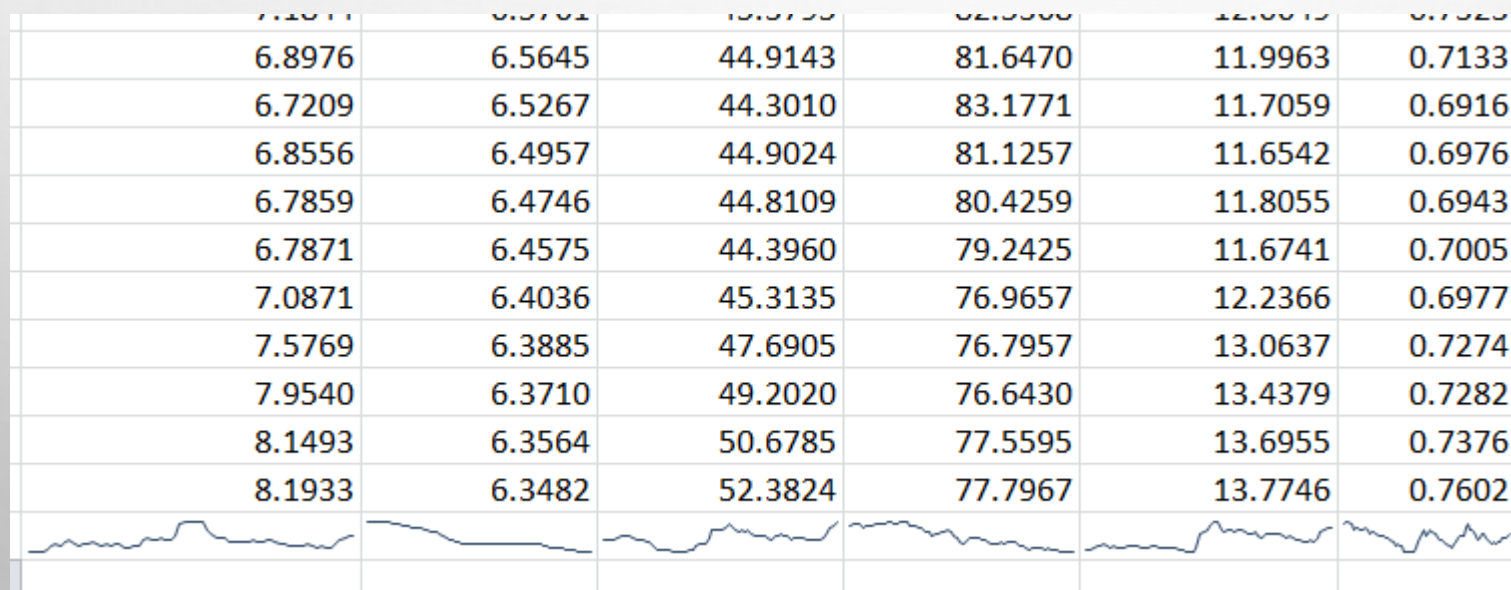
TOTAL VIOLENT AND

PROPERTY CRIMES
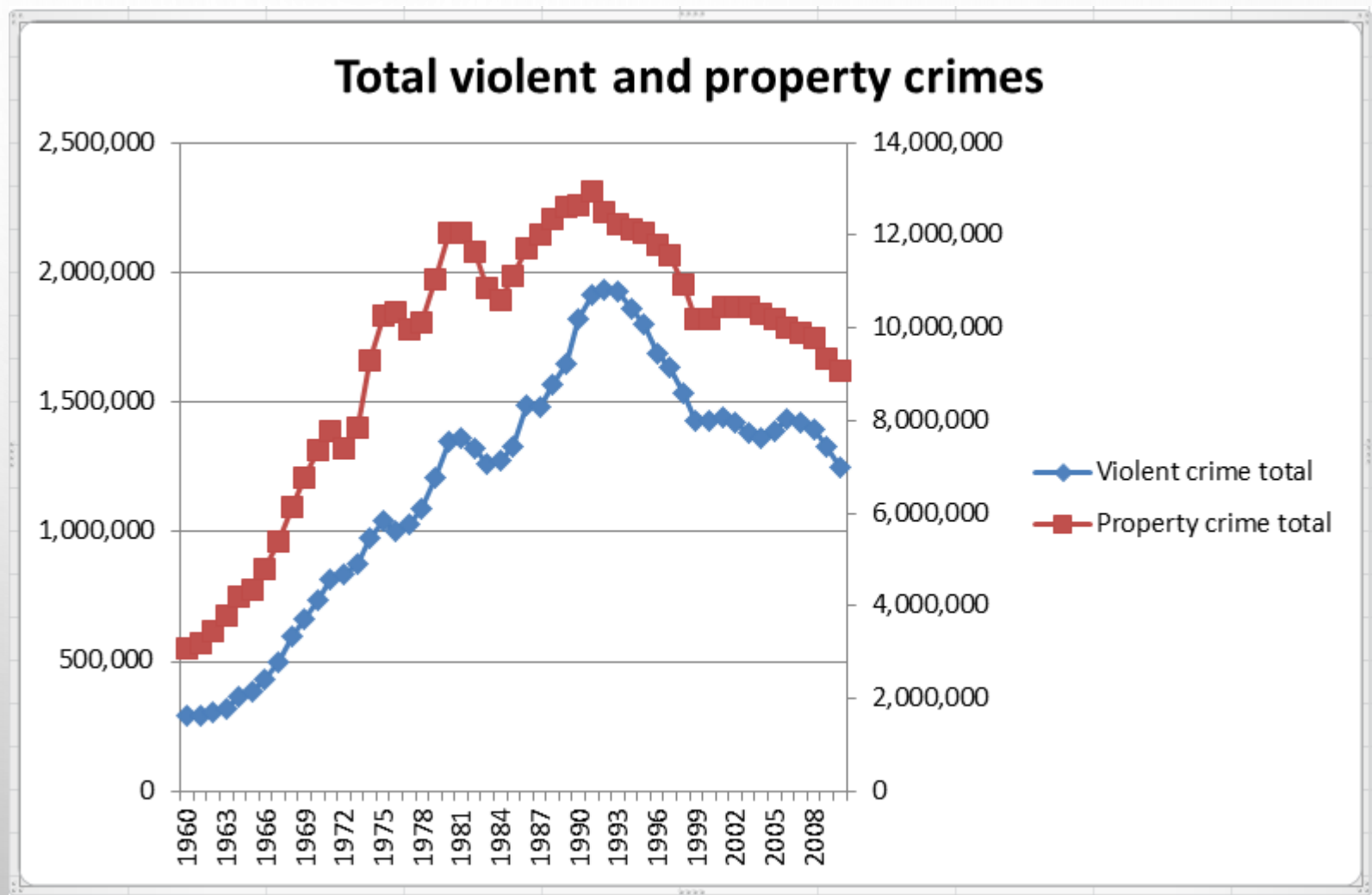
CRIME IN US.XLSX

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

7.1011	6.5701	45.5795	82.5500	12.0015	0.7525
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.

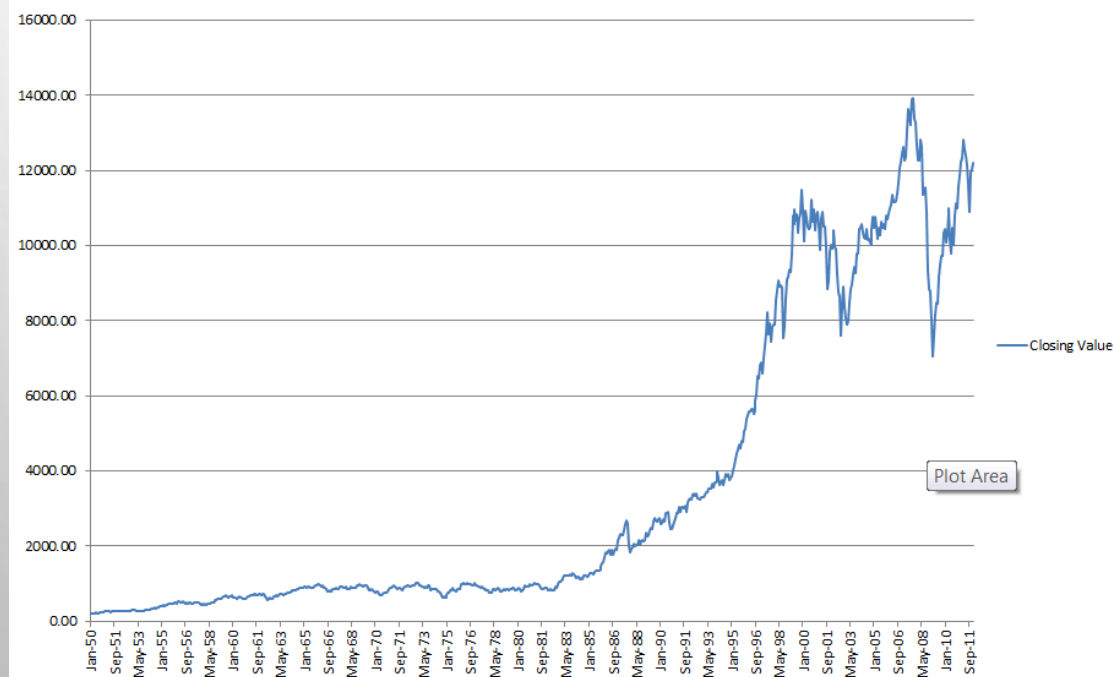|	Closing Value
One Variable Summary	**DJIA Data**
Mean	3484.13
Std. Dev.	4044.57
Median	969.26
1st Quartile	764.58
3rd Quartile	5616.21

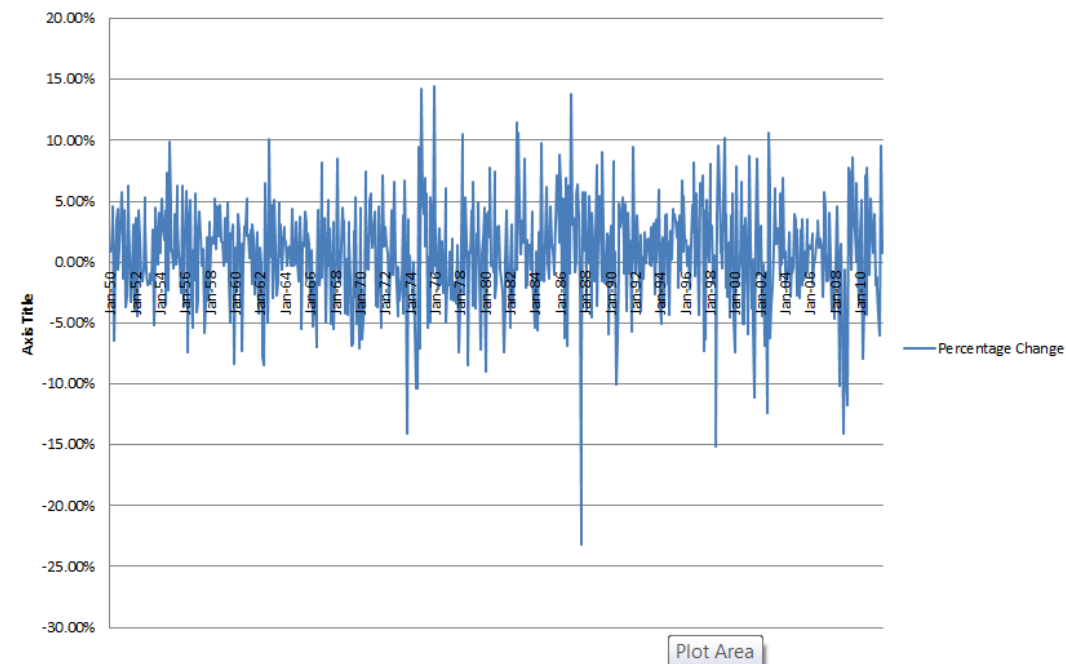|	Percentage Change
One Variable Summary	**DJIA Data**
Mean	0.00642
Std. Dev.	0.04182
Median	0.00851
1st Quartile	-0.01721
3rd Quartile	0.03289

DJIA MONTHLY CLOSE.XLSX

OUTLIERS

- AN **OUTLIER** IS A VALUE OR AN ENTIRE OBSERVATION (ROW) THAT LIES WELL OUTSIDE OF THE NORM.
  - SOME STATISTICIANS DEFINE AN OUTLIER AS ANY VALUE MORE THAN THREE STANDARD DEVIATIONS FROM THE MEAN, BUT THIS IS ONLY A RULE OF THUMB.
- EVEN IF VALUES ARE NOT UNUSUAL BY THEMSELVES, THERE STILL MIGHT BE UNUSUAL *COMBINATIONS* OF VALUES.
- WHEN DEALING WITH OUTLIERS, IT IS BEST TO RUN THE ANALYSES TWO WAYS: WITH THE OUTLIERS AND WITHOUT THEM.