BUSINESS ANALYTICS

DR. BRENDA MULLALLY

DESCRIPTIVE STATISTICS

- THE ABILITY TO DESCRIBE A DATASET IS VERY IMPORTANT.
- DESCRIBE EACH VARIABLE USING STATISTICAL TECHNIQUES (NUMERICAL/CATEGORICAL)

HOW TO DESCRIBE CATEGORICAL VARIABLES?

- THERE ARE ONLY A FEW POSSIBILITIES FOR DESCRIBING A CATEGORICAL VARIABLE, ALL BASED ON *COUNTING*:
  - COUNT THE NUMBER OF CATEGORIES.
  - GIVE THE CATEGORIES NAMES.
  - COUNT THE NUMBER OF OBSERVATIONS IN EACH CATEGORY (REFERRED TO AS THE **COUNT OF CATEGORIES**).
    - ONCE YOU HAVE THE COUNTS, YOU CAN DISPLAY THEM GRAPHICALLY, USUALLY IN A COLUMN CHART OR A PIE CHART.

EXAMPLE 2.2: SUPERMARKET TRANSACTIONS.XLSX

- **OBJECTIVE**: TO SUMMARIZE CATEGORICAL VARIABLES IN A LARGE DATA SET.
- **SOLUTION**: DATA SET CONTAINS TRANSACTIONS MADE BY SUPERMARKET CUSTOMERS OVER A TWO-YEAR PERIOD.
- CHILDREN, UNITS SOLD, AND REVENUE ARE NUMERICAL.
- PURCHASE DATE IS A DATE VARIABLE.
- TRANSACTION AND CUSTOMER ID ARE USED ONLY TO IDENTIFY.
- ALL OF THE OTHER VARIABLES ARE CATEGORICAL.

	A	B	C	D	E	F	G	H	I	J	K	O	P
1	Transaction	Purchase Date	Customer ID	Gender	Marital Status	Homeowner	Children	Annual Income	City	State or Province	Country	Units Sold	Revenue
2	1	12/18/2011	7223	F	S	Y	2	$30K - $50K	Los Angeles	CA	USA	5	$27.38
3	2	12/20/2011	7841	M	M	Y	5	$70K - $90K	Los Angeles	CA	USA	5	$14.90
4	3	12/21/2011	8374	F	M	N	2	$50K - $70K	Bremerton	WA	USA	3	$5.52
5	4	12/21/2011	9619	M	M	Y	3	$30K - $50K	Portland	OR	USA	4	$4.44
6	5	12/22/2011	1900	F	S	Y	3	$130K - $150K	Beverly Hills	CA	USA	4	$14.00
7	6	12/22/2011	6696	F	M	Y	3	$10K - $30K	Beverly Hills	CA	USA	3	$4.37
8	7	12/23/2011	9673	M	S	Y	2	$30K - $50K	Salem	OR	USA	4	$13.78
9	8	12/25/2011	354	F	M	Y	2	$150K +	Yakima	WA	USA	6	$7.34
10	9	12/25/2011	1293	M	M	Y	3	$10K - $30K	Bellingham	WA	USA	1	$2.41
11	10	12/25/2011	7938	M	S	N	1	$50K - $70K	San Diego	CA	USA	2	$8.96

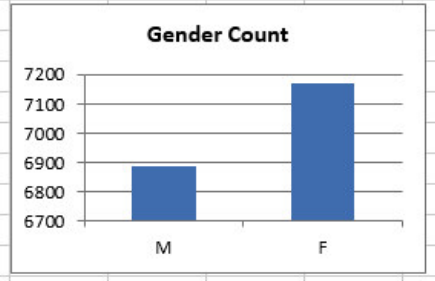
EXAMPLE 2.2: SUPERMARKET TRANSACTIONS.XLSX

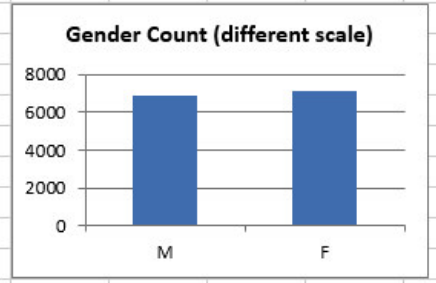
- TO GET THE COUNTS IN COLUMN S, USE EXCEL'S *COUNTIF* FUNCTION.

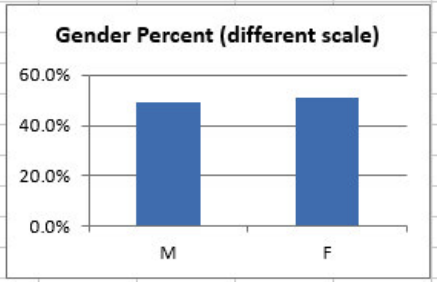
☐ TO GET THE PERCENTAGES IN COLUMN T, DIVIDE EACH COUNT BY THE TOTAL NUMBER OF OBSERVATIONS.

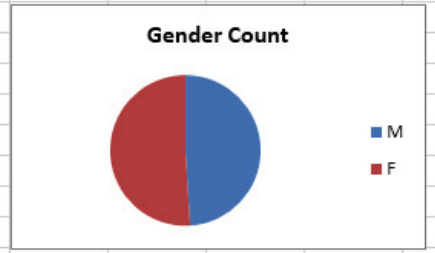
☐ WHEN CREATING CHARTS, BE CAREFUL TO USE APPROPRIATE SCALES.

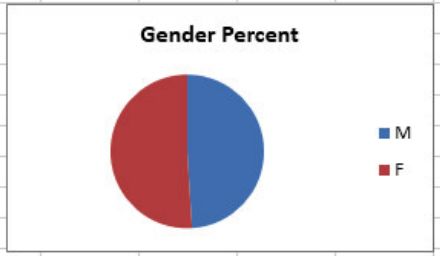
	R	S	T
1	**Categorical summaries**		
2	Gender	Count	Percent
3	M	6889	49.0%
4	F	7170	51.0%
5			100.0%
6			
7	Marital Status	Count	Percent
8	S	7193	51.2%
9	M	6866	48.8%
10			100.0%
11			
12	Homeowner	Count	Percent
13	Y	8444	60.1%
14	N	5615	39.9%
15			100.0%
16			
17	Annual Income	Count	Percent
18	$10K - $30K	3090	22.0%
19	$30K - $50K	4601	32.7%
20	$50K - $70K	2370	16.9%
21	$70K - $90K	1709	12.2%
22	$90K - $110K	613	4.4%
23	$110K - $130K	643	4.6%
24	$130K - $150K	760	5.4%
25	$150K +	273	1.9%
26			100.0%
27			
28			
29			
30			











EXAMPLE 2.2: SUPERMARKET TRANSACTIONS.XLSX

- ANOTHER EFFICIENT WAY TO FIND COUNTS FOR A CATEGORICAL VARIABLE IS TO USE DUMMY (0–1) VARIABLES.
  - RECODE EACH VARIABLE SO THAT ONE CATEGORY IS REPLACED BY 1 AND ALL OTHERS BY 0.
    - THIS CAN BE DONE USING A SIMPLE IF FORMULA.
  - FIND THE COUNT OF THAT CATEGORY BY SUMMING THE 0S AND 1S.
  - FIND THE PERCENTAGE OF THAT CATEGORY BY AVERAGING THE 0S AND 1S.

	A	B	C	D	E
1	Transaction	Purchase Date	Customer ID	Gender	Gender Dummy for M
2	1	12/18/2011	7223	F	0
3	2	12/20/2011	7841	M	1
4	3	12/21/2011	8374	F	0
5	4	12/21/2011	9619	M	1
6	5	12/22/2011	1900	F	0
7	6	12/22/2011	6696	F	0
8	7	12/23/2011	9673	M	1
9	8	12/25/2011	354	F	0
10	9	12/25/2011	1293	M	1
11	10	12/25/2011	7938	M	1
14055	14054	12/29/2013	2032	F	0
14056	14055	12/29/2013	9102	F	0
14057	14056	12/29/2013	4822	F	0
14058	14057	12/31/2013	250	M	1
14059	14058	12/31/2013	6153	F	0
14060	14059	12/31/2013	3656	M	1
14061				Count	6889
14062				Percent	49.0%

HOW TO DESCRIBE NUMERICAL VARIABLES?

- THERE ARE MANY WAYS TO SUMMARIZE NUMERICAL VARIABLES, BOTH WITH NUMERICAL SUMMARY MEASURES AND WITH CHARTS.
- TO LEARN HOW THE VALUES OF A VARIABLE ARE DISTRIBUTED, ASK:
  - WHAT ARE THE MOST "TYPICAL" VALUES?
  - HOW SPREAD OUT ARE THE VALUES?
  - WHAT ARE THE "EXTREME" VALUES ON EITHER END?
  - IS THE CHART OF THE VALUES SYMMETRIC ABOUT SOME MIDDLE VALUE, OR IS IT SKEWED IN SOME DIRECTION? DOES IT HAVE ANY OTHER PECULIAR FEATURES BESIDES POSSIBLE SKEWNESS?

EXAMPLE 2.3: BASEBALL SALARIES 2011.XLSX

- **OBJECTIVE**: TO LEARN HOW SALARIES ARE DISTRIBUTED ACROSS ALL 2011 MLB PLAYERS.
- **SOLUTION**: DATA SET CONTAINS DATA ON 843 MAJOR LEAGUE BASEBALL PLAYERS IN THE 2011 SEASON.
- VARIABLES ARE PLAYER'S NAME, TEAM, POSITION, AND SALARY.
- CREATE SUMMARY MEASURES OF BASEBALL SALARIES USING EXCEL FUNCTIONS.

	A	B	C	D
1	**Player**	**Team**	**Position**	**Salary**
2	A.J. Burnett	New York Yankees	Pitcher	$16,500,000
3	A.J. Ellis	Los Angeles Dodgers	Catcher	$421,000
4	A.J. Pierzynski	Chicago White Sox	Catcher	$2,000,000
5	Aaron Cook	Colorado Rockies	Pitcher	$9,875,000
6	Aaron Crow	Kansas City Royals	Pitcher	$1,400,000
7	Aaron Harang	San Diego Padres	Pitcher	$3,500,000
8	Aaron Heilman	Arizona Diamondbacks	Pitcher	$2,000,000
9	Aaron Hill	Toronto Blue Jays	Second Baseman	$5,000,000
10	Aaron Laffey	Seattle Mariners	Pitcher	$431,600
11	Aaron Miles	Los Angeles Dodgers	Second Baseman	$500,000
12	Aaron Rowand	San Francisco Giants	Outfielder	$13,600,000
13	Adam Dunn	Chicago White Sox	Designated Hitter	$12,000,000
14	Adam Everett	Cleveland Indians	Shortstop	$700,000

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			

WHAT ARE THE MOST TYPICAL VALUES?

MEASURES OF CENTRAL TENDENCY

- THE **MEAN** IS THE AVERAGE OF ALL VALUES.
  - IF THE DATA SET REPRESENTS A SAMPLE FROM SOME LARGER POPULATION, THIS MEASURE IS CALLED THE **SAMPLE MEAN** AND IS DENOTED BY $\bar{X}$.
  - IF THE DATA SET REPRESENTS THE ENTIRE POPULATION, THIS MEASURE IS CALLED THE **POPULATION MEAN** AND IS DENOTED BY M.

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

- IN EXCEL, THE MEAN CAN BE CALCULATED WITH THE *AVERAGE* FUNCTION.

WHAT ARE THE MOST TYPICAL VALUES?

MEASURES OF CENTRAL TENDENCY

- THE **MEDIAN** IS THE MIDDLE OBSERVATION WHEN THE DATA ARE SORTED FROM SMALLEST TO LARGEST.
  - IF THE NUMBER OF OBSERVATIONS IS ODD, THE MEDIAN IS LITERALLY THE MIDDLE OBSERVATION.
  - IF THE NUMBER OF OBSERVATIONS IS EVEN, THE MEDIAN IS USUALLY DEFINED AS THE AVERAGE OF THE TWO MIDDLE OBSERVATIONS.
- IN EXCEL, THE MEDIAN CAN BE CALCULATED WITH THE *MEDIAN* FUNCTION.

WHAT ARE THE MOST TYPICAL VALUES?

MEASURES OF CENTRAL TENDENCY

- THE **MODE** IS THE VALUE THAT APPEARS MOST OFTEN.
  - IN MOST CASES WHERE A VARIABLE IS ESSENTIALLY CONTINUOUS, THE MODE IS NOT VERY INTERESTING BECAUSE IT IS OFTEN THE RESULT OF A FEW LUCKY TIES.
  - HOWEVER, IT IS NOT ALWAYS A RESULT OF LUCK AND MAY REVEAL INTERESTING INFORMATION.
- IN EXCEL, THE MODE CAN BE CALCULATED WITH THE *MODE.SNGL* FUNCTION.

MINIMUM, MAXIMUM, PERCENTILES, AND QUARTILES

- FOR ANY PERCENTAGE *P*, THE *P*TH **PERCENTILE** IS THE VALUE SUCH THAT A PERCENTAGE *P* OF ALL VALUES ARE LESS THAN IT.
- THE **QUARTILES** DIVIDE THE DATA INTO FOUR GROUPS, EACH WITH (APPROXIMATELY) A QUARTER OF ALL OBSERVATIONS.
  - THE FIRST, SECOND AND THIRD QUARTILES ARE THE PERCENTILES CORRESPONDING TO $P = 25\%$, $P = 50\%$, AND $P = 75\%$.
  - BY DEFINITION, THE SECOND QUARTILE ($P = 50\%$) IS EQUAL TO THE MEDIAN.
- THE **MINIMUM** AND **MAXIMUM** VALUES CAN BE CALCULATED WITH EXCEL'S *MIN* AND *MAX* FUNCTIONS, AND THE PERCENTILES AND QUARTILES WITH EXCEL'S *PERCENTILE* AND *QUARTILE* FUNCTIONS.

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			

HOW SPREAD OUT ARE THE VALUES? 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 $(X_I - MEAN)^2$.

HOW SPREAD OUT ARE THE VALUES? MEASURES OF VARIABILITY

- THE **SAMPLE VARIANCE** IS DENOTED BY S^2, AND THE **POPULATION VARIANCE** BY σ^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.

HOW SPREAD OUT ARE THE VALUES? 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 Σ, 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