DATA ANALYTICS

DR. BRENDA MULLALLY

AGE OF TECHNOLOGY

- Technology has made it possible to collect and store huge amounts of data.
  - Retailers, credit agencies, investment companies, government agencies,
- It is difficult for businesses to make sense of all of the data collected.
- Many more people now have the power to analyse data and make decisions on the basis of quantitative analysis.
- Quantitative analysis is now conducted by people other than those that traditionally had done the number crunching.
- Most employees now have access to software to analyse data, particularly spreadsheet and database software.
- Quantitative analysis is now an integral part of these people’s job.

DATA DISCOVERY

- Basic data summaries and visualisations:
  - Summary statistics
  - Frequency tables
  - Histograms
  - Boxplots
  - Scatterplots
  - Correlation tables
  - Cross-tabulations

DATA DISCOVERY

- Typical employees today not just the managers and technical specialists have a wealth of easy-to-use tools at their disposal, and it is frequently up to them to summarize data in a way that is both meaningful and useful to their constituents: people within their company, their company’s suppliers, and their company’s customers. It takes some training and practice to do this effectively.

DATA DISCOVERY

- Data analysis in the real world is never done in a vacuum. It is done to solve a problem. Typically, there are four steps that are followed, whether the context is business, medical science, or any other field.

1. Recognise a problem that needs solving
2. Gather data to help understand and then solve the problem.
3. Analyse the data
4. Act on the analysis by changing policies, undertaking initiatives, publishing records etc.

DATA DISCOVERY

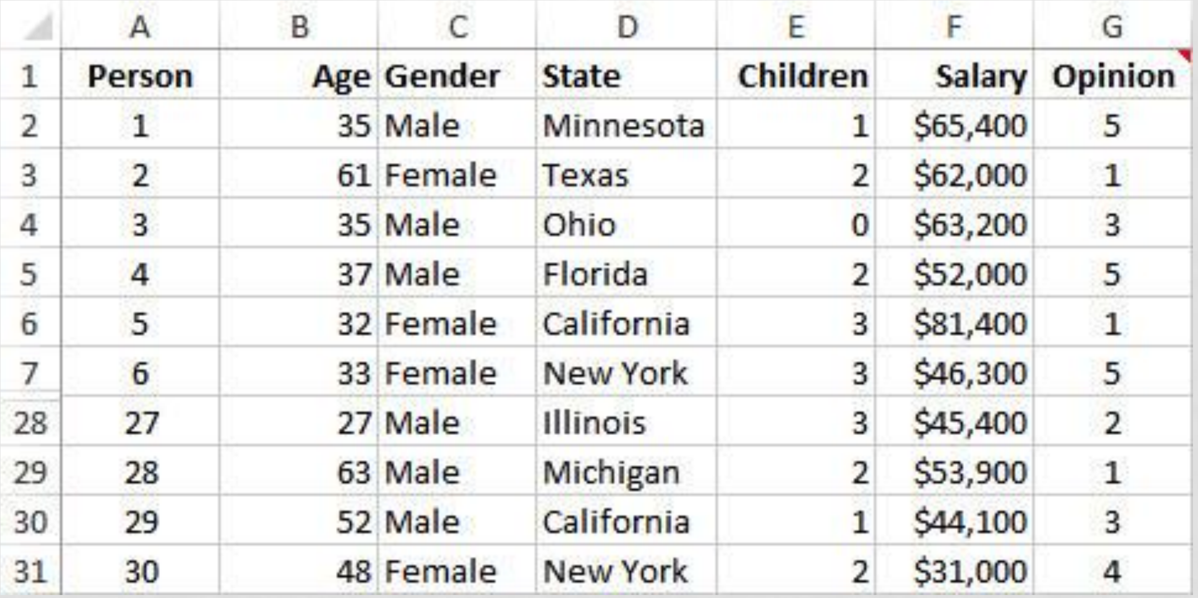
- Populations and Samples
  - Population includes all of the entities of interest: people, households, machines, or whatever.
  - Sample is a subset of a population, often randomly chosen and preferably representative of the population as a whole.
- It is very important that the sample is representative of the population. This means that any observed characteristics of the sample can be generalised to the population as a whole.

DATA DISCOVERY

- Data Sets, Variables, and Observations
  - Data set: a rectangular array of data where columns contain Variables, such as height, gender, and income.
  - Each row contains an observation.
  - Each observation contains the attributes of a particular member of a population: a person, a company, a city, a machine…
  - A variable (column) is often called a field or an attribute.
  - An observation (row) is often called a case or a record.

EXAMPLE 2.1: QUESTIONNAIRE DATA.XLSX

- **Objective:** To illustrate variables and observations in a typical data set.
- **Solution:** Data set includes observations on 30 people who responded to a questionnaire on the president’s environmental policies.
- Variables include: age, gender, state, children, salary, opinion.
- Include a row that lists variable names.
- Include a column that shows an index of the observation.



	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
7	6	33	Female	New York	3	$46,300	5
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

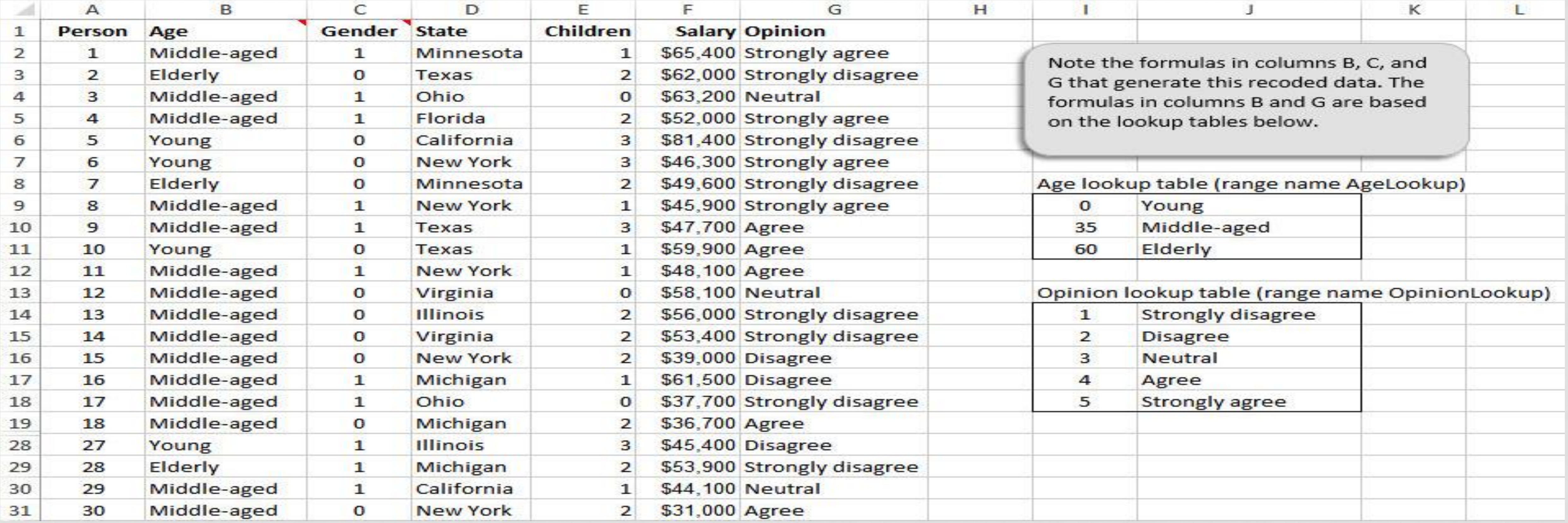
DATA DISCOVERY

- Data Types
  - Numerical and Categorical data
  - Do you want to do arithmetic on the data?
  - Can you average days of the week or gender?
  - What about a variable that has 1, 2, 3, 4, or 5 as its value?
  - Ordinal: a natural ordering to categories.
  - Nominal: no natural order to categories.
  - All categorical variables can be encoded with numbers but not all are, it is personal choice.
  - Dummy variable

DATA DISCOVERY

- Data Types
  - Sometimes a number variable is coded using a category.
  - binning (discretising)

ENVIRONMENTAL DATA USING A DIFFERENT CODING (SLIDE 3 OF 5)



	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	Middle-aged	1	Minnesota	1	$65,400	Strongly agree
3	2	Elderly	0	Texas	2	$62,000	Strongly disagree
4	3	Middle-aged	1	Ohio	0	$63,200	Neutral
5	4	Middle-aged	1	Florida	2	$52,000	Strongly agree
6	5	Young	0	California	3	$81,400	Strongly disagree
7	6	Young	0	New York	3	$46,300	Strongly agree
8	7	Elderly	0	Minnesota	2	$49,600	Strongly disagree
9	8	Middle-aged	1	New York	1	$45,900	Strongly agree
10	9	Middle-aged	1	Texas	3	$47,700	Agree
11	10	Young	0	Texas	1	$59,900	Agree
12	11	Middle-aged	1	New York	1	$48,100	Agree
13	12	Middle-aged	0	Virginia	0	$58,100	Neutral
14	13	Middle-aged	0	Illinois	2	$56,000	Strongly disagree
15	14	Middle-aged	0	Virginia	2	$53,400	Strongly disagree
16	15	Middle-aged	0	New York	2	$39,000	Disagree
17	16	Middle-aged	1	Michigan	1	$61,500	Disagree
18	17	Middle-aged	1	Ohio	0	$37,700	Strongly disagree
19	18	Middle-aged	0	Michigan	2	$36,700	Agree
28	27	Young	1	Illinois	3	$45,400	Disagree
29	28	Elderly	1	Michigan	2	$53,900	Strongly disagree
30	29	Middle-aged	1	California	1	$44,100	Neutral
31	30	Middle-aged	0	New York	2	$31,000	Agree

Note the formulas in columns B, C, and G that generate this recoded data. The formulas in columns B and G are based on the lookup tables below.

Age lookup table (range name AgeLookup)

0	Young
35	Middle-aged
60	Elderly

Opinion lookup table (range name OpinionLookup)

1	Strongly disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly agree

TYPES OF DATA

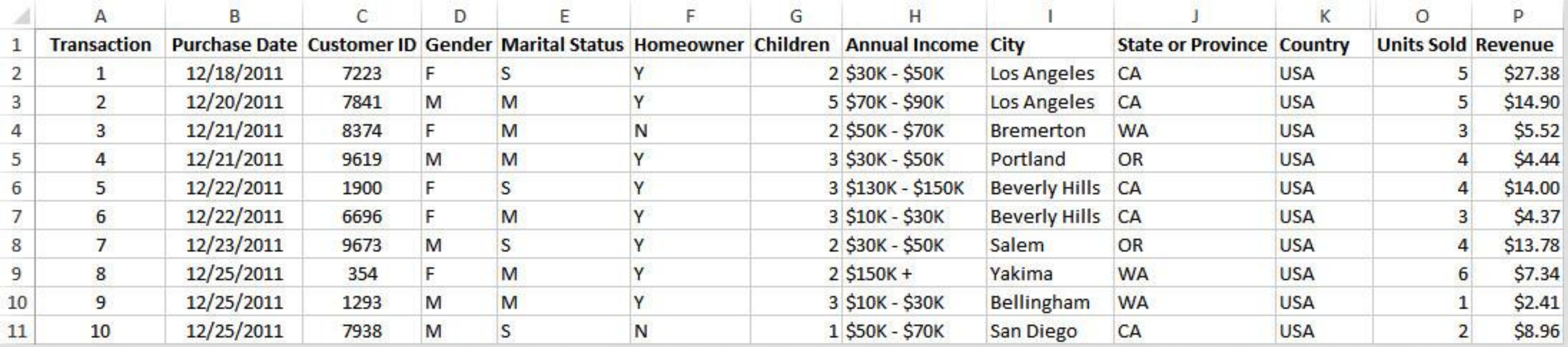
- A numerical variable is **discrete** if it results from a count, such as the number of children.
- A **continuous** variable is the result of an essentially continuous measurement, such as weight or height.
- Data Set:
- **Cross-sectional** data are data on a cross section of a population at a distinct point in time.
- **Time series** data are data collected over time.

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.



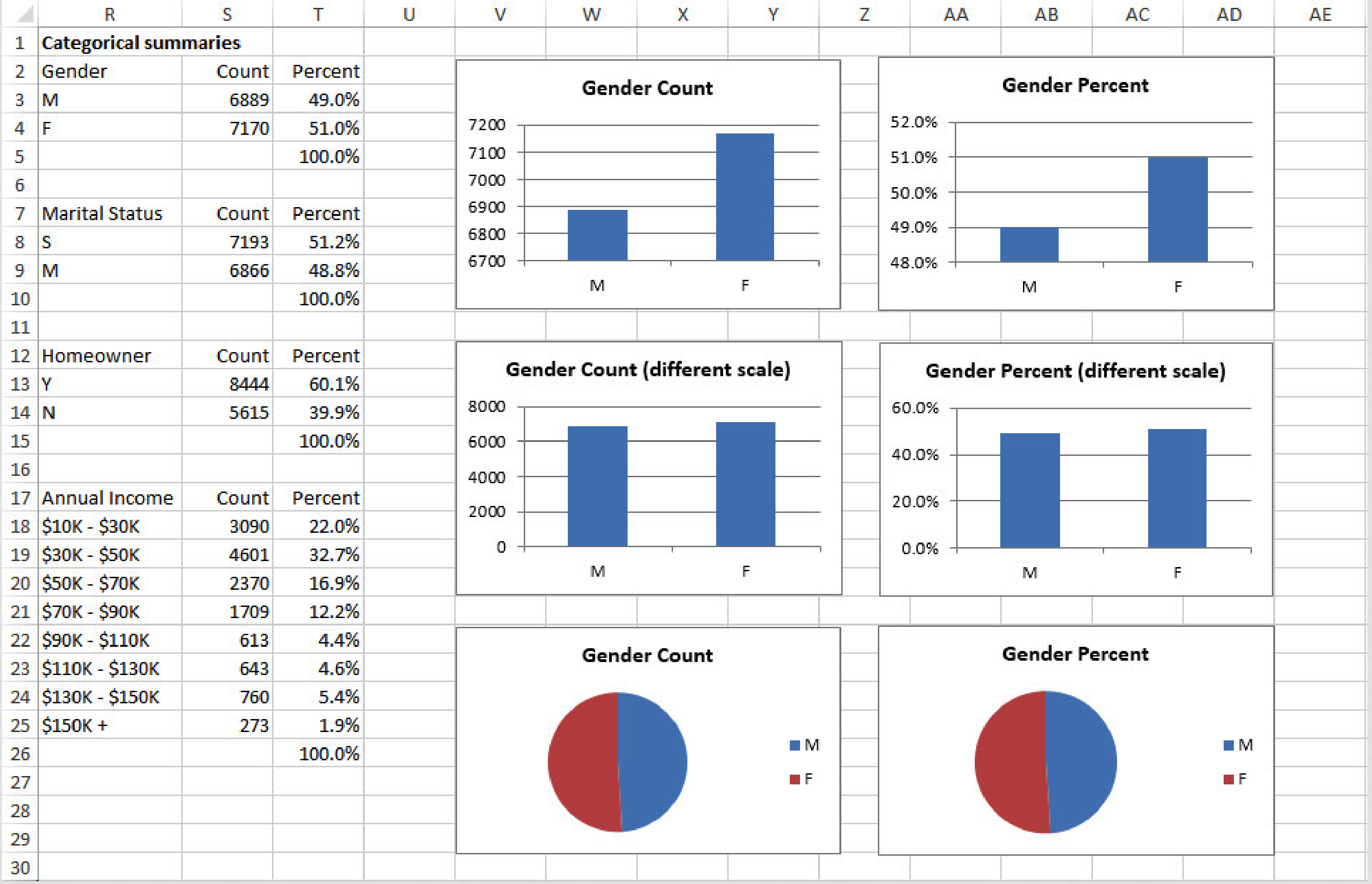
	A	B	C	D	E	F	G	H	I	J	K	O	P
1	Transaction	Purchase Date	Customer ID	Gender	Marital Status	Homeowner	Children	Annual Income	City	State or Province	Country	Units Sold	Revenue
2	1	12/18/2011	7223	F	S	Y	2	$30K - $50K	Los Angeles	CA	USA	5	$27.38
3	2	12/20/2011	7841	M	M	Y	5	$70K - $90K	Los Angeles	CA	USA	5	$14.90
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.

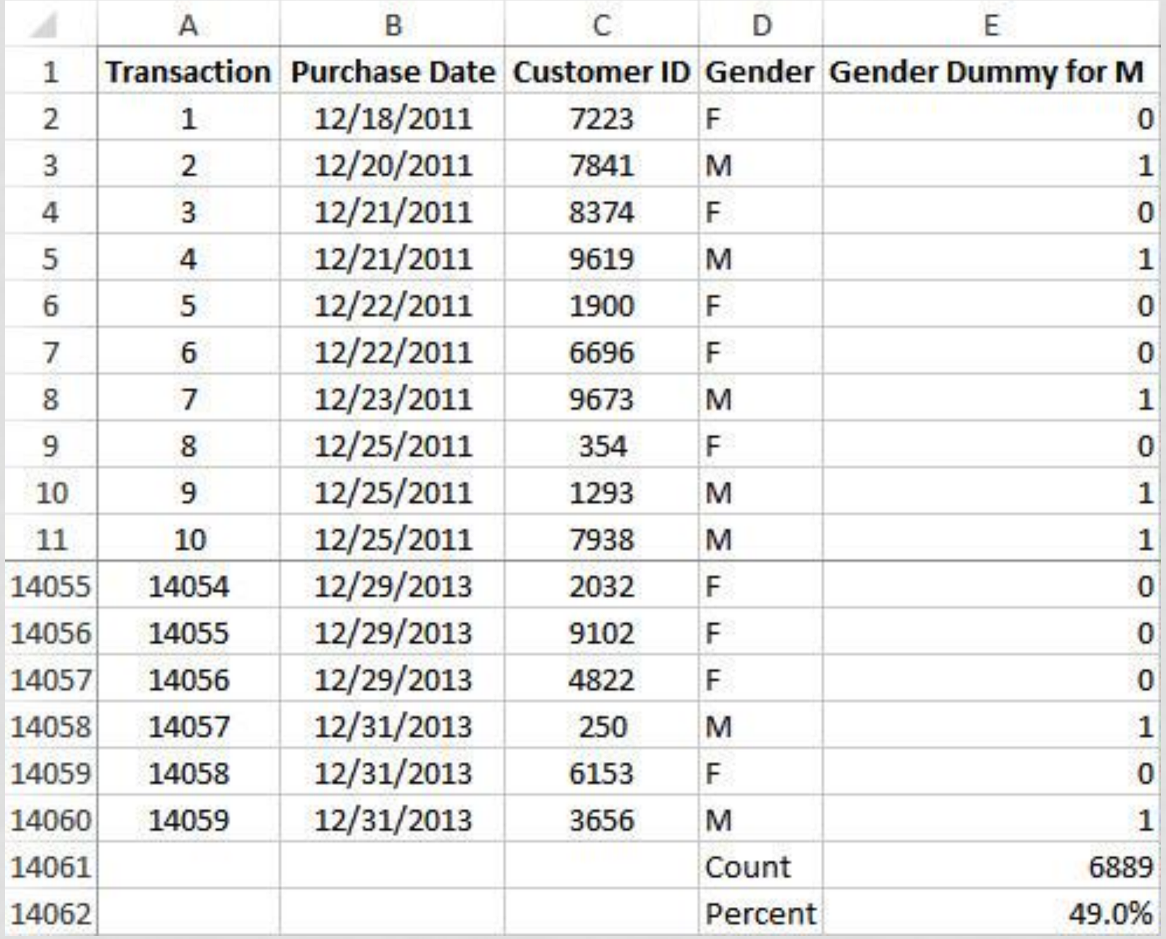
□ When creating charts, be careful to use appropriate scales.



	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.



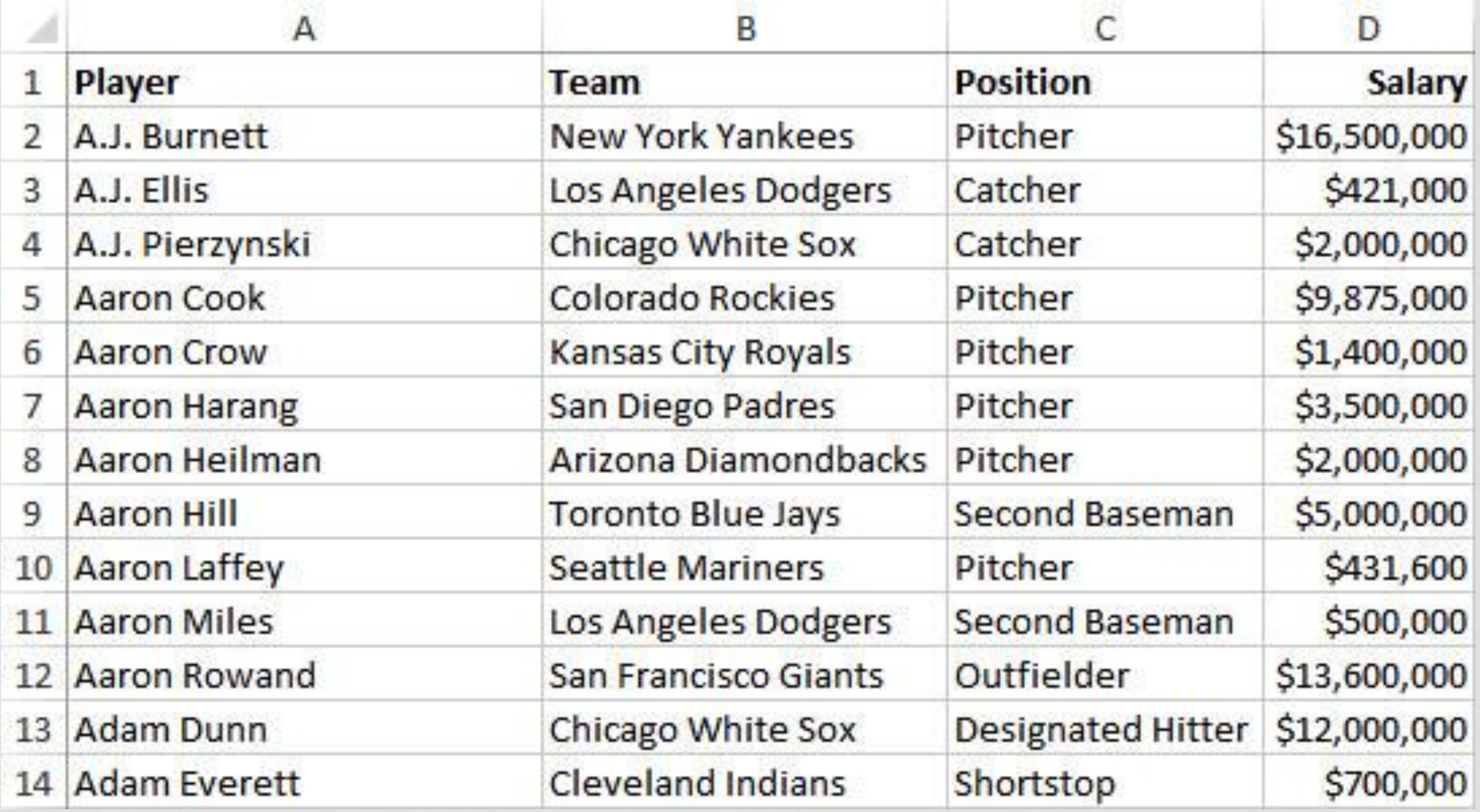
	A	B	C	D	E
1	Transaction	Purchase Date	Customer ID	Gender	Gender Dummy for M
2	1	12/18/2011	7223	F	0
3	2	12/20/2011	7841	M	1
4	3	12/21/2011	8374	F	0
5	4	12/21/2011	9619	M	1
6	5	12/22/2011	1900	F	0
7	6	12/22/2011	6696	F	0
8	7	12/23/2011	9673	M	1
9	8	12/25/2011	354	F	0
10	9	12/25/2011	1293	M	1
11	10	12/25/2011	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%

DESCRIPTIVE MEASURES FOR 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?

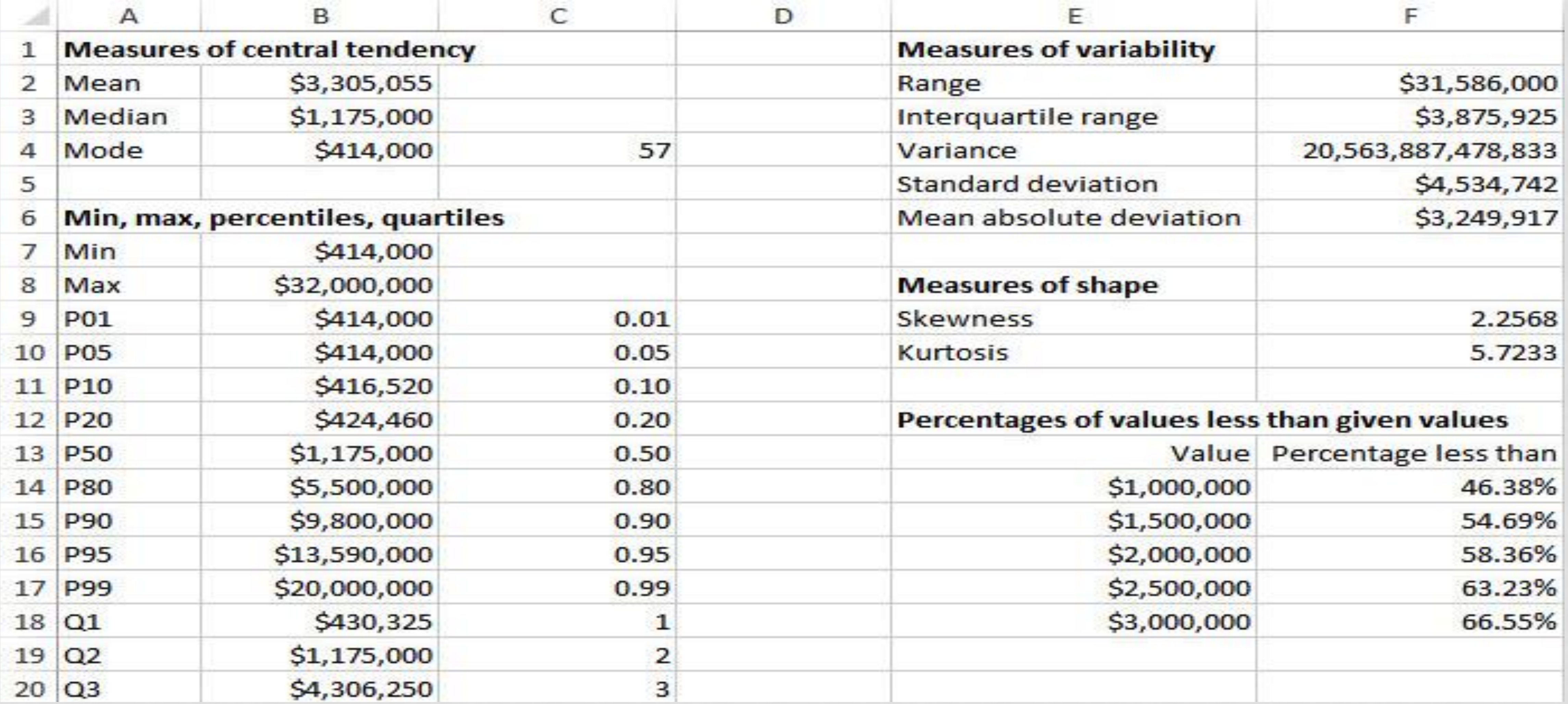
EXAMPLE 2.3:
BASEBALL SALARIES 2011.XLSX

- **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			

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 *X*.
  - If the data set represents the entire population, it is called the **population mean** and is denoted by *μ*.

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

- In Excel, the mean can be calculated with the *AVERAGE* function.

MEASURES OF CENTRAL TENDENCY

(SLIDE 2 OF 3)

- 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.

MEASURES OF CENTRAL TENDENCY
(SLIDE 3 OF 3)

- 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.