DATA ANALYTICS

DR. BRENDA MULLALLY

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.

MISSING VALUES

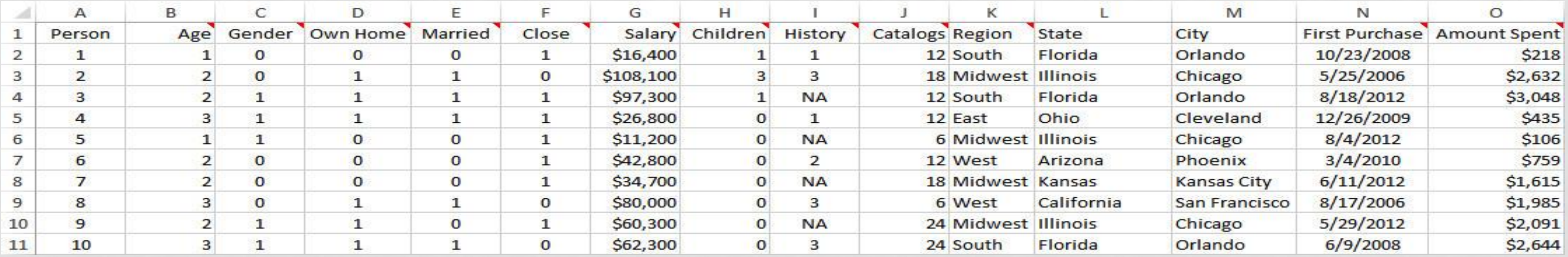
- Most real data sets have gaps in the data.
- There are two issues: how to detect these **missing values** and what to do about them.
- The more important issue is what to do about them:
  - One option is to simply ignore them. Then you will have to be aware of how the software deals with missing values.
  - Another option is to fill in missing values with the average of non missing values, but this isn’t usually a very good option.
  - A third option is to examine the nonmissing values in the *row* of a missing value; these values might provide clues on what the missing value should be.

EXCEL TABLES FOR FILTERING, SORTING, AND SUMMARIZING

- Tables are a tool introduced in Excel 2007.
- You now have the ability to designate a rectangular data set as a table and then employ a number of powerful tools for analyzing tables.
- These tools include:
  - Filtering
  - Sorting
  - Summarizing

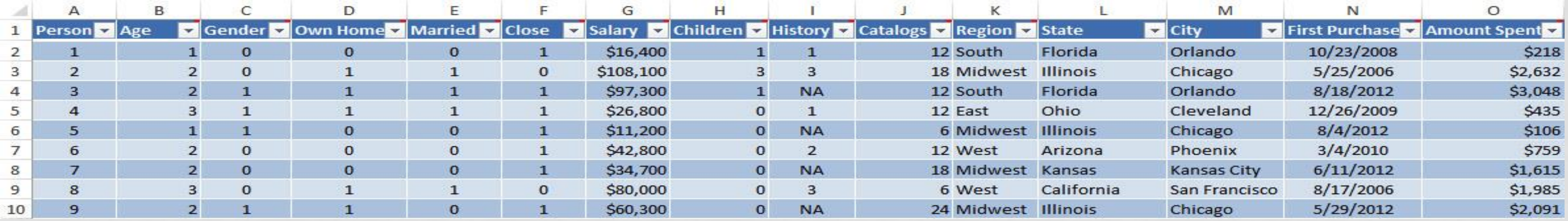
EXAMPLE 2.7:CATALOG MARKETING.XLSX

- **Objective**: To illustrate Excel tables for analyzing the HyTex data.
- **Solution**: Data set contains data on 1000 customers of HyTex, a fictional direct marketing company.
- Designate the data set as a table by selecting any cell in the data set and clicking the Table button on the Insert ribbon.
- Use the dropdown arrows next to the variable names to filter in many different ways.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2010	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2012	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2006	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2012	$2,091
11	10	3	1	1	1	0	$62,300	0	3	24	South	Florida	Orlando	6/9/2008	$2,644

CATALOG MARKETING.XLSX



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
2	1	1	0	0	0	1	$16,400	1	1	12	South	Florida	Orlando	10/23/2008	$218
3	2	2	0	1	1	0	$108,100	3	3	18	Midwest	Illinois	Chicago	5/25/2006	$2,632
4	3	2	1	1	1	1	$97,300	1	NA	12	South	Florida	Orlando	8/18/2012	$3,048
5	4	3	1	1	1	1	$26,800	0	1	12	East	Ohio	Cleveland	12/26/2009	$435
6	5	1	1	0	0	1	$11,200	0	NA	6	Midwest	Illinois	Chicago	8/4/2012	$106
7	6	2	0	0	0	1	$42,800	0	2	12	West	Arizona	Phoenix	3/4/2010	$759
8	7	2	0	0	0	1	$34,700	0	NA	18	Midwest	Kansas	Kansas City	6/11/2012	$1,615
9	8	3	0	1	1	0	$80,000	0	3	6	West	California	San Francisco	8/17/2006	$1,985
10	9	2	1	1	0	1	$60,300	0	NA	24	Midwest	Illinois	Chicago	5/29/2012	$2,091

FILTERING

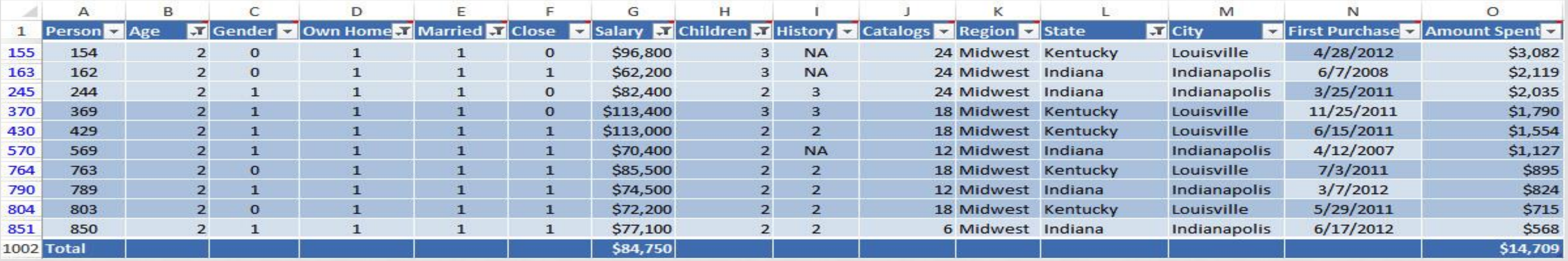
- Finding records that match particular criteria is called *filtering*.
- One way to filter is to create an Excel table, which automatically provides dropdown arrows next to the field names that allow you to filter.
- There are also three ways to filter on any rectangular data set with variable names:
  1. Use the Filter button from the Sort & Filter dropdown list on the Home ribbon.
  2. Use the Filter button from the Sort & Filter group on the Data ribbon.
  3. Right-click any cell in the data set and select Filter. You get several options, the most popular of which is Filter by Selected Cell's Value.

CATALOG MARKETING.XLSX

- **Objective**: To investigate the types of filters that can be applied to the HyTex data.
- **Solution**: There is almost no limit to the filters you can apply, but here are a few possibilities:
  - Filter on one or more values in a field.
  - Filter on more than one field.
  - Filter on a continuous numerical field.
  - *Top 10* and *Above/Below Average* filters.
  - Filter on a text field.
  - Filter on a date field.
  - Filter on color or icon.
  - Use a custom filter.

EXAMPLE 2.7
CATALOG MARKETING.XLSX

Results from a Typical Filter



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person	Age	Gender	Own Home	Married	Close	Salary	Children	History	Catalogs	Region	State	City	First Purchase	Amount Spent
155	154	2	0	1	1	0	$96,800	3	NA	24	Midwest	Kentucky	Louisville	4/28/2012	$3,082
163	162	2	0	1	1	1	$62,200	3	NA	24	Midwest	Indiana	Indianapolis	6/7/2008	$2,119
245	244	2	1	1	1	0	$82,400	2	3	24	Midwest	Indiana	Indianapolis	3/25/2011	$2,035
370	369	2	1	1	1	0	$113,400	3	3	18	Midwest	Kentucky	Louisville	11/25/2011	$1,790
430	429	2	1	1	1	1	$113,000	2	2	18	Midwest	Kentucky	Louisville	6/15/2011	$1,554
570	569	2	1	1	1	1	$70,400	2	NA	12	Midwest	Indiana	Indianapolis	4/12/2007	$1,127
764	763	2	0	1	1	1	$85,500	2	2	18	Midwest	Kentucky	Louisville	7/3/2011	$895
790	789	2	1	1	1	1	$74,500	2	2	12	Midwest	Indiana	Indianapolis	3/7/2012	$824
804	803	2	0	1	1	1	$72,200	2	2	18	Midwest	Kentucky	Louisville	5/29/2011	$715
851	850	2	1	1	1	1	$77,100	2	2	6	Midwest	Indiana	Indianapolis	6/17/2012	$568
1002	Total						$84,750								$14,709

RELATIONSHIPS AMONG VARIABLES

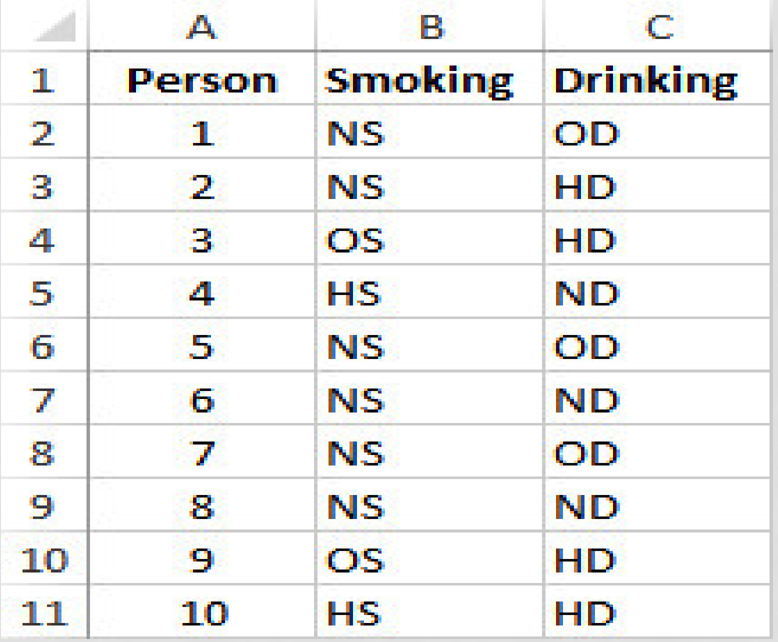
- The primary interest in data analysis is usually in *relationships* between variables.
  - The most useful numerical summary measure is correlation.
  - The most useful graph is a scatterplot.
  - To break down a numerical variable by a categorical variable, it is useful to create side-by-side box plots.
  - Excel's® pivot table breaks down one variable by others so that all sorts of relationships can be uncovered very quickly.

RELATIONSHIPS AMONG CATEGORICAL VARIABLES

- The most meaningful way to examine relationships between two categorical variables is with counts and corresponding charts of the counts.
  - You can find counts of the categories of either variable separately, as well as counts of the *joint* categories of the two variables.
  - Corresponding percentages of totals and charts help tell the story.
- It is customary to display all such counts in a table called a **crosstabs** (for crosstabulations). This is also sometimes called a **contingency table**.

SMOKING DRINKING.XLSX

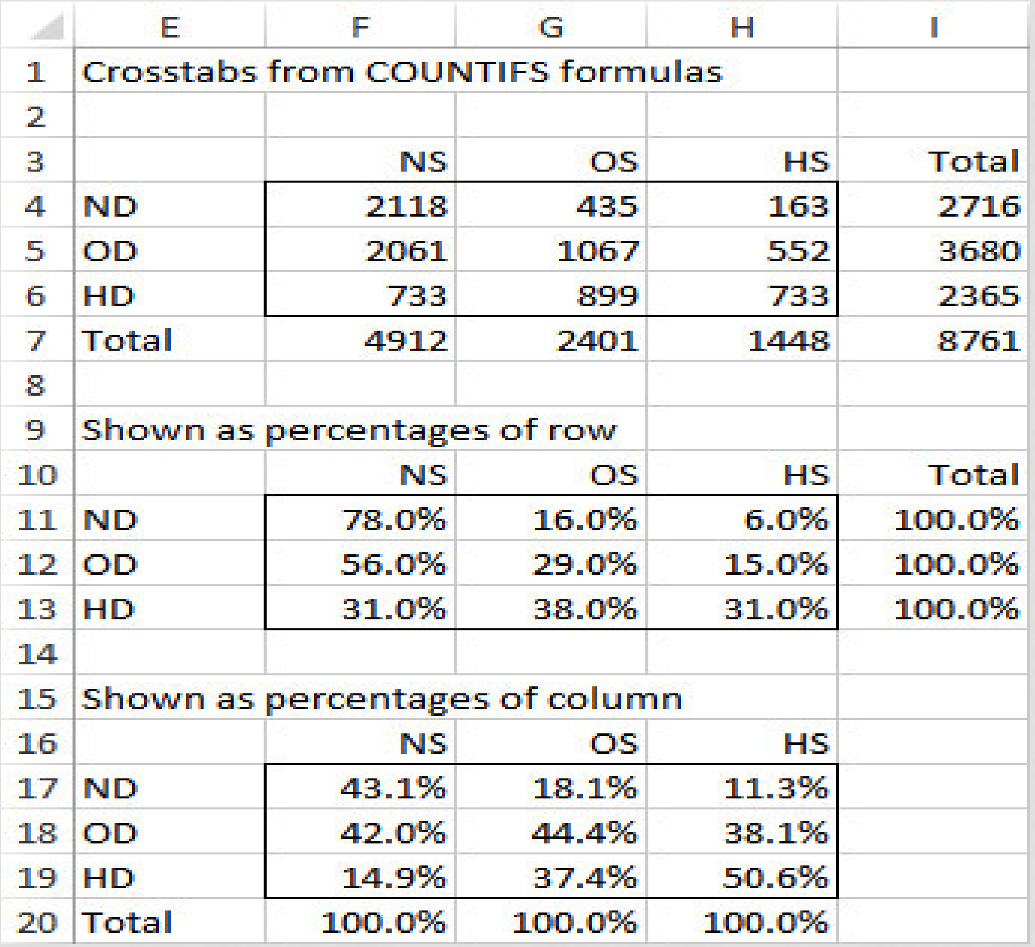
- **Objective:** To use a crosstabs to explore the relationship between smoking and drinking.
- **Solution:** Data set lists the smoking and drinking habits of 8761 adults.
- Categories have been coded “N,” “O,” “H,” “S,” and “D” for “Non,” “Occasional,” “Heavy,” “Smoker,” and “Drinker.”



	A	B	C
1	Person	Smoking	Drinking
2	1	NS	OD
3	2	NS	HD
4	3	OS	HD
5	4	HS	ND
6	5	NS	OD
7	6	NS	ND
8	7	NS	OD
9	8	NS	ND
10	9	OS	HD
11	10	HS	HD

EXAMPLE 3.1:
SMOKING DRINKING.XLSX (SLIDE 2 OF 2)

- To create the crosstabs, enter the category headings in Excel and use the *COUNTIFS* function to fill the table with counts of joint categories.
- Next, sum across rows and down columns to get totals.
- Then express the counts as percentages of row and percentages of column.



	E	F	G	H	I
1	Crosstabs from COUNTIFS formulas				
2					
3		NS	OS	HS	Total
4	ND	2118	435	163	2716
5	OD	2061	1067	552	3680
6	HD	733	899	733	2365
7	Total	4912	2401	1448	8761
8					
9	Shown as percentages of row				
10		NS	OS	HS	Total
11	ND	78.0%	16.0%	6.0%	100.0%
12	OD	56.0%	29.0%	15.0%	100.0%
13	HD	31.0%	38.0%	31.0%	100.0%
14					
15	Shown as percentages of column				
16		NS	OS	HS	
17	ND	43.1%	18.1%	11.3%	
18	OD	42.0%	44.4%	38.1%	
19	HD	14.9%	37.4%	50.6%	
20	Total	100.0%	100.0%	100.0%	

RELATIONSHIPS AMONG CATEGORICAL VARIABLES AND A NUMERICAL VARIABLE

- The **comparison problem** is an important problems in data analysis. It occurs whenever you want to compare a numerical measure across two or more subpopulations.
  - Examples:
    - The subpopulations are males and females, and the numerical measure is salary.
    - The subpopulations are different regions of the country, and the numerical measure is the cost of living.
    - The subpopulations are different days of the week, and the numerical measure is the number of customers going to a particular fast-food chain.

RELATIONSHIPS AMONG NUMERICAL VARIABLES

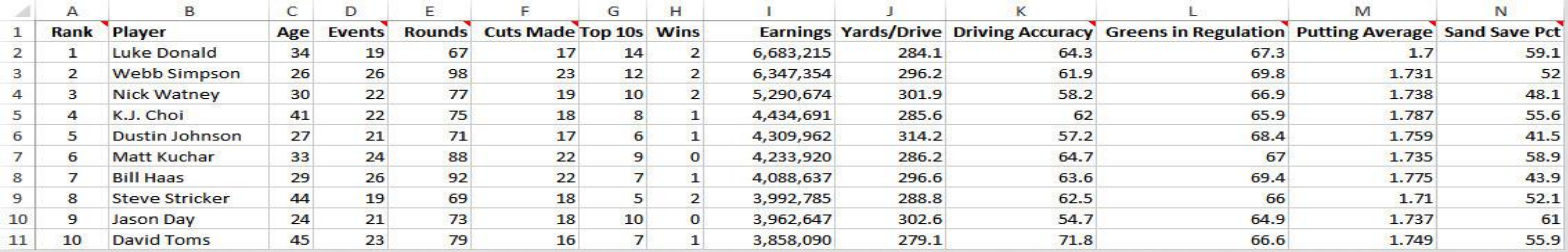
- To study relationships among numerical variables, a new type of chart, called a scatterplot, and two new summary measures, correlation and covariance, are used.
- These measures can be applied to any variables that are displayed numerically.
- However, they are appropriate only for truly numerical variables, not for categorical variables that have been coded numerically.

SCATTERPLOTS

- A **scatterplot** is a scatter of points, where each point denotes the values of an observation for two selected variables.
  - It is a graphical method for detecting relationships between two numerical variables.
  - The two variables are often labeled generically as *X* and *Y*, so a scatterplot is sometimes called an **X-Y chart**.
  - The purpose of a scatterplot is to make a relationship (or the lack of it) apparent.

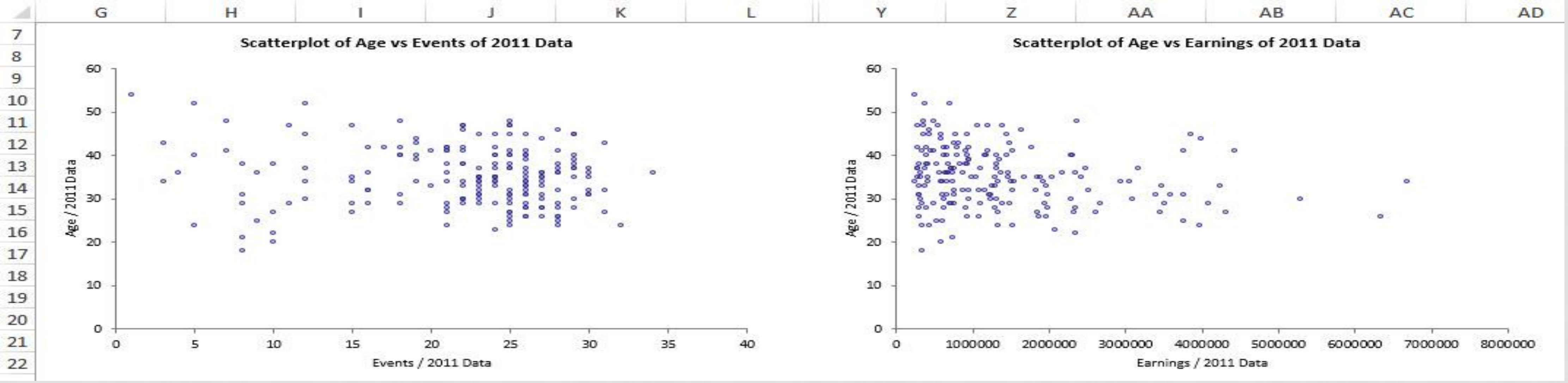
GOLFSTATS.XLSX

- **Objective**: To use scatterplots to search for relationships in the golf data.
- **Solution**: Data set includes an observation (stats) for each of the top 200 earners on the PGA Tour.
- Using Excel you can create a scatterplot for two variables such as Age and Events, or Age and Earnings.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rank	Player	Age	Events	Rounds	Cuts Made	Top 10s	Wins	Earnings	Yards/Drive	Driving Accuracy	Greens in Regulation	Putting Average	Sand Save Pct
2	1	Luke Donald	34	19	67	17	14	2	6,683,215	284.1	64.3	67.3	1.7	59.1
3	2	Webb Simpson	26	26	98	23	12	2	6,347,354	296.2	61.9	69.8	1.731	52
4	3	Nick Watney	30	22	77	19	10	2	5,290,674	301.9	58.2	66.9	1.738	48.1
5	4	K.J. Choi	41	22	75	18	8	1	4,434,691	285.6	62	65.9	1.787	55.6
6	5	Dustin Johnson	27	21	71	17	6	1	4,309,962	314.2	57.2	68.4	1.759	41.5
7	6	Matt Kuchar	33	24	88	22	9	0	4,233,920	286.2	64.7	67	1.735	58.9
8	7	Bill Haas	29	26	92	22	7	1	4,088,637	296.6	63.6	69.4	1.775	43.9
9	8	Steve Stricker	44	19	69	18	5	2	3,992,785	288.8	62.5	66	1.71	52.1
10	9	Jason Day	24	21	73	18	10	0	3,962,647	302.6	54.7	64.9	1.737	61
11	10	David Toms	45	23	79	16	7	1	3,858,090	279.1	71.8	66.6	1.749	55.9

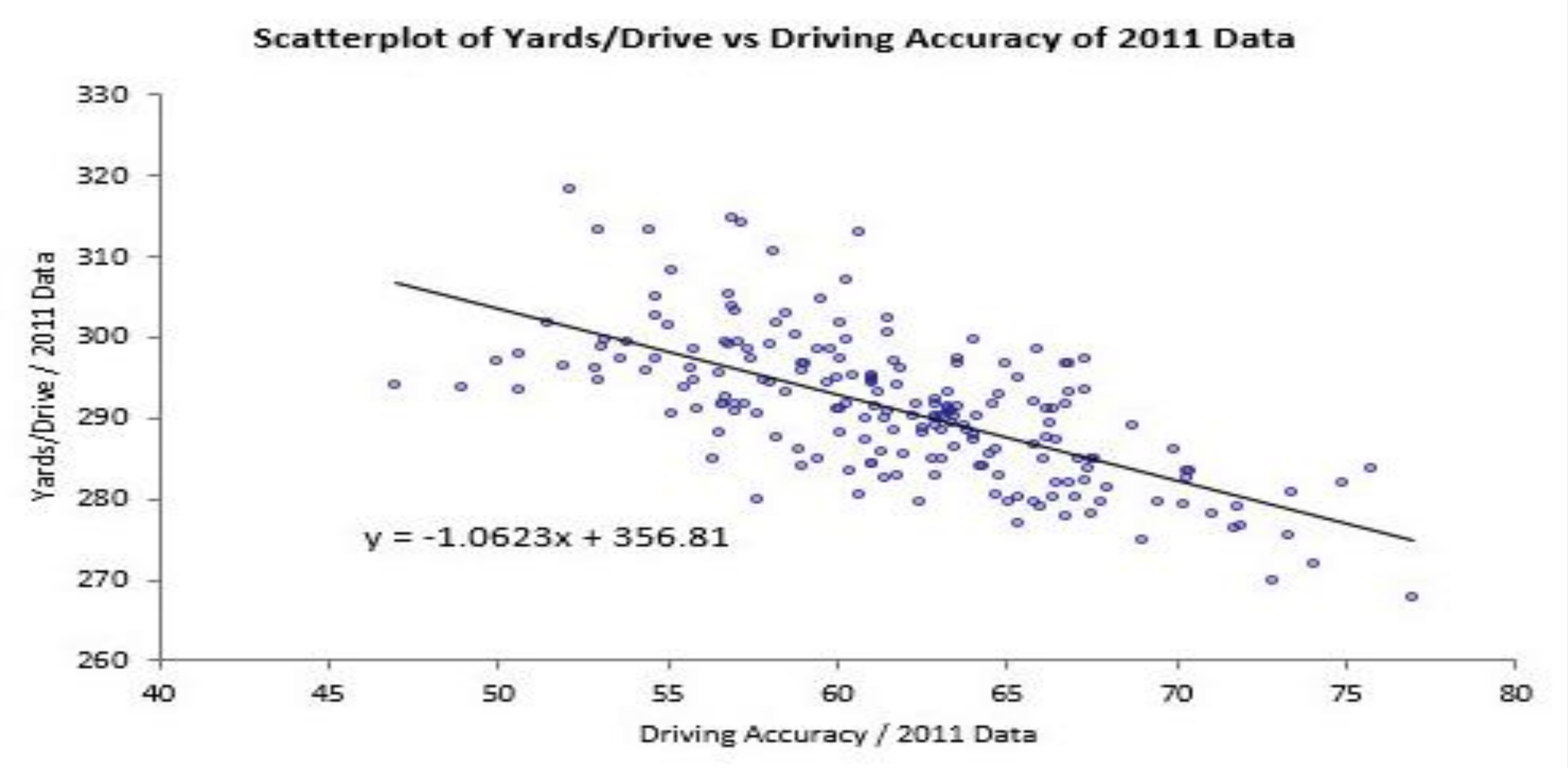
GOLFSTATS.XLSX



TREND LINES IN SCATTERPLOTS

- Once you have a scatterplot, Excel enables you to superimpose one of several trend lines on the scatterplot.
  - A **trend line** is a line or curve that “fits” the scatter as well as possible.
  - This could be a straight line, or it could be one of several types of curves.
- On the Layout tab for the scatterplot click on Trendline and choose the appropriate one. (in Excel 2013 on the design tab choose Add Chart Element).

SCATTERPLOT WITH TREND LINE AND EQUATION SUPERIMPOSED



CORRELATION AND COVARIANCE (SLIDE 1 OF 4)

- Correlation and covariance measure the strength and direction of a *linear* relationship between two numerical variables.
  - The relationship is “strong” if the points in a scatterplot cluster tightly around some straight line.
    - If this straight line rises from left to right, the relationship is *positive* and the measures will be positive numbers.
    - If it falls from left to right, the relationship is *negative* and the measures will be negative numbers.
  - The two numerical variables must be “paired” variables.
    - They must have the same number of observations, and the values for any observation should be naturally paired.

CORRELATION AND COVARIANCE
(SLIDE 2 OF 4)

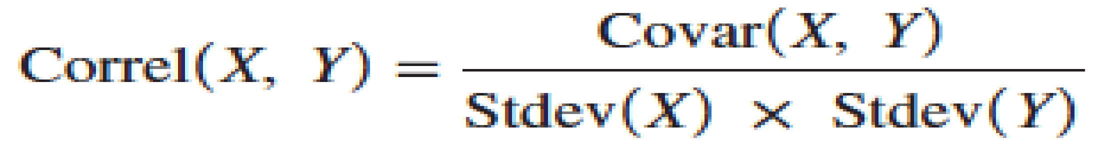
- **Covariance** is essentially an average of products of deviations from means.

$$\text{Covar}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

- Excel has a built-in *COVAR* function
- Covariance has a serious limitation as a descriptive measure because it is very sensitive to the *units* in which *X* and *Y* are measured.

CORRELATION AND COVARIANCE
(SLIDE 3 OF 4)

- **Correlation** is a unitless quantity that is unaffected by the measurement scale.



$$\text{Correl}(X,\ Y) = \frac{\text{Covar}(X,\ Y)}{\text{Stdev}(X)\ \times\ \text{Stdev}(Y)}$$

- The correlation is *always* between -1 and +1.
  - The closer it is to either of these two extremes, the closer the points in a scatterplot are to a straight line.
- Excel has a built-in *CORREL* function and the built in Add-In data analysis can calculate correlation on multiple variables.

CORRELATION AND COVARIANCE
(SLIDE 4 OF 4)

- Three important points about scatterplots, correlations, and covariances:
  - A correlation is a single-number summary of a scatterplot. It never conveys as much information as the full scatterplot.
  - You are usually on the lookout for large correlations, those near -1 or +1.
  - Do not even try to interpret covariances numerically except possibly to check whether they are positive or negative. For interpretive purposes, concentrate on correlations.