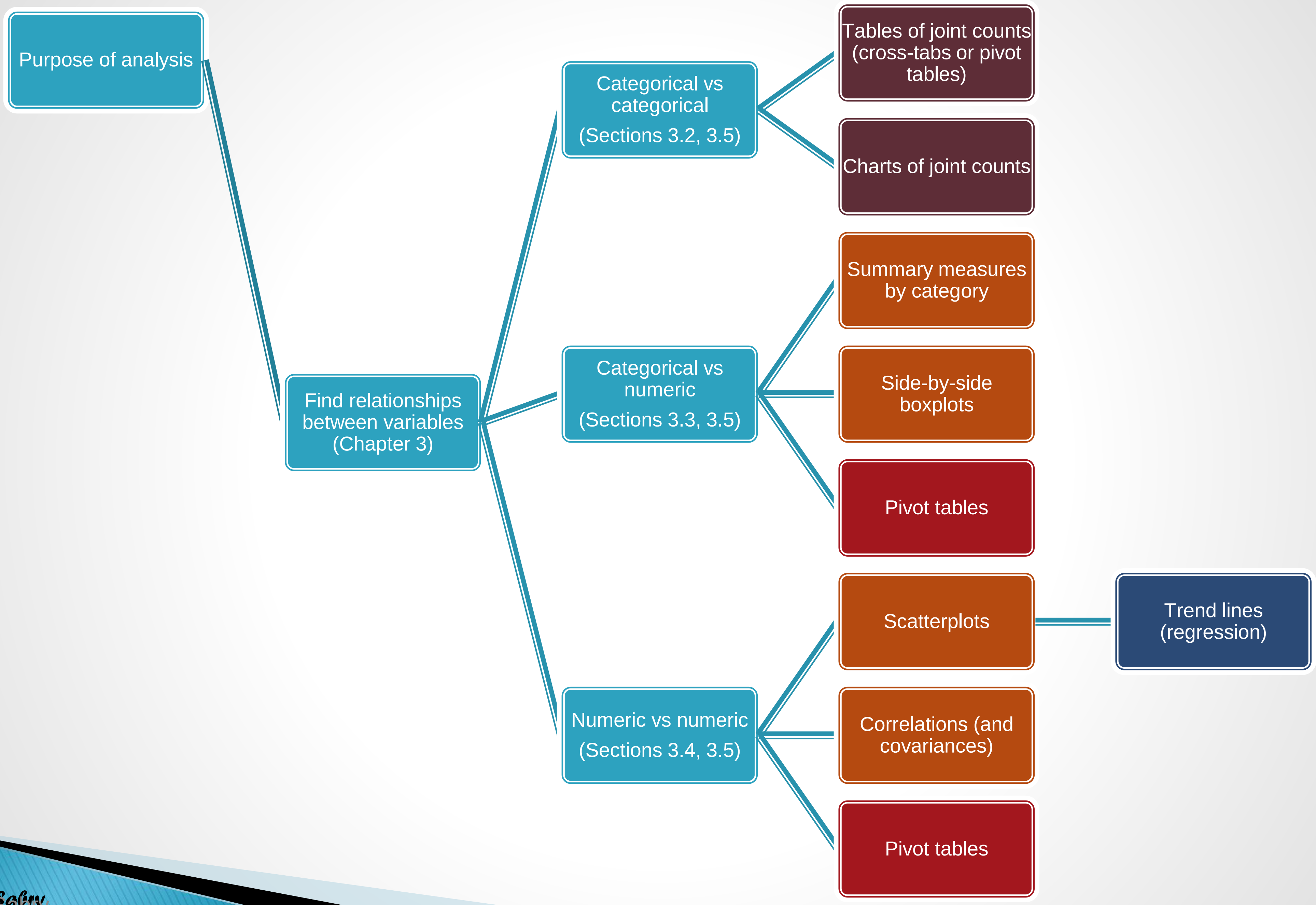
DS342 - Data Analytics

Lecture 4
Finding Relationships among Variables



- Purpose of analysis
  - Find relationships between variables (Chapter 3)
    - Categorical vs categorical (Sections 3.2, 3.5)
      - Tables of joint counts (cross-tabs or pivot tables)
      - Charts of joint counts
    - Categorical vs numeric (Sections 3.3, 3.5)
      - Summary measures by category
      - Side-by-side boxplots
      - Pivot tables
    - Numeric vs numeric (Sections 3.4, 3.5)
      - Scatterplots
        - Trend lines (regression)
      - Correlations (and covariances)
      - Pivot tables

Dr. Marwa Sabry

Introduction

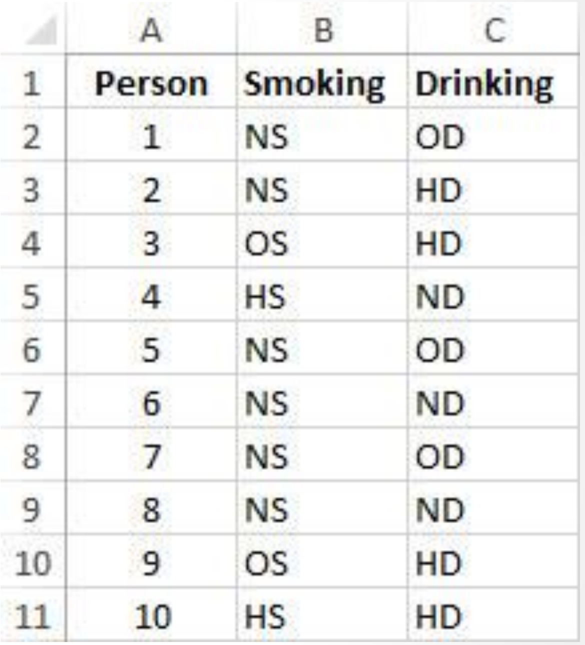
- The primary interest in data analysis is usually in *relationships* between variables.
  - The most useful numerical summary measure is correlation.
  - The most useful graph is a scatterplot.
  - To break down a numerical variable by a categorical variable, it is useful to create side-by-side box plots.
  - Excel's® pivot table breaks down one variable by others so that all sorts of relationships can be uncovered very quickly.
- The diagram in the file **Data Analysis Taxonomy.xlsx** gives you the big picture of which analyses are appropriate for which data types and which tools are best for performing the various analyses.

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.**

Example 3.1: Smoking Drinking.xlsx (slide 1 of 2)

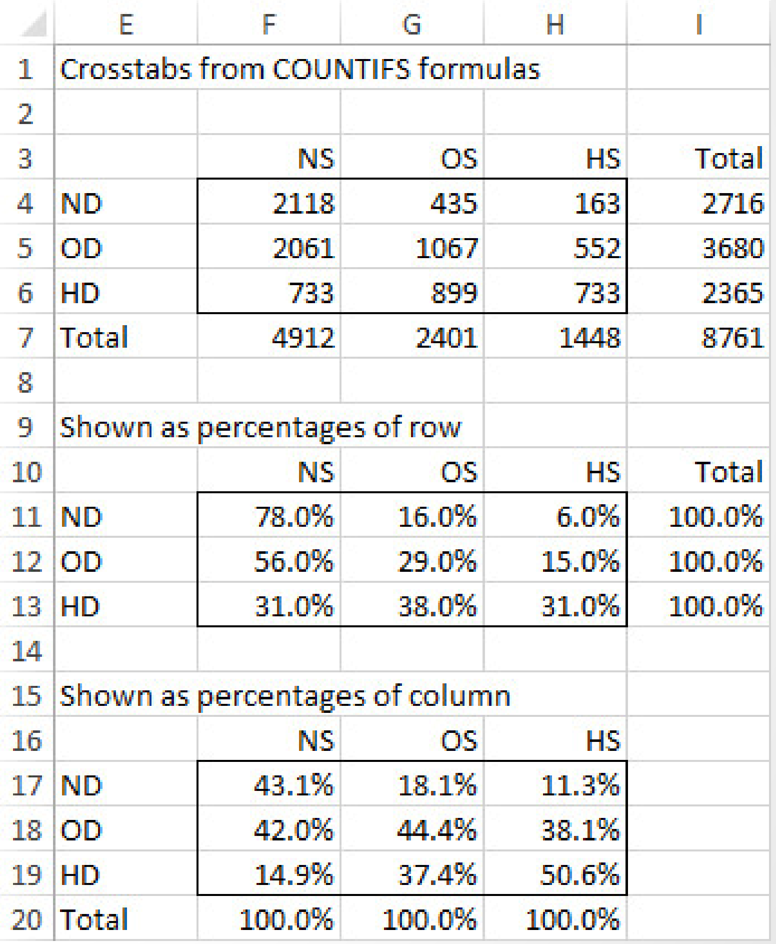
- **Objective:** To use a crosstabs to explore the relationship between smoking and drinking.
- **Solution:** Data set lists the smoking and drinking habits of 8761 adults.
- Categories have been coded “N,” “O,” “H,” “S,” and “D” for “Non,” “Occasional,” “Heavy,” “Smoker,” and “Drinker.”



	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.



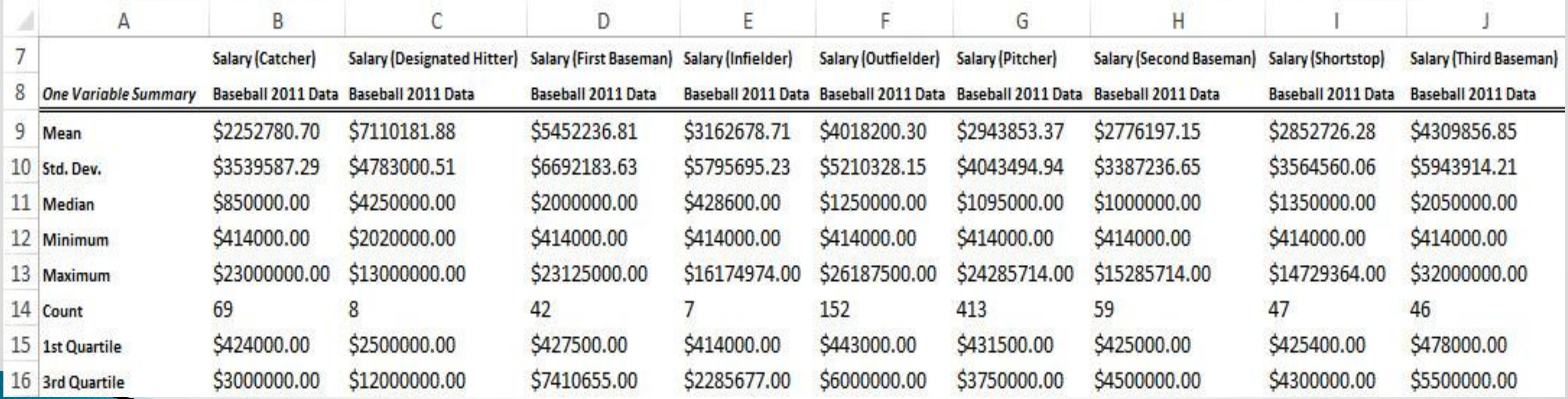
	E	F	G	H	I
1	Crosstabs from COUNTIFS formulas				
2					
3		NS	OS	HS	Total
4	ND	2118	435	163	2716
5	OD	2061	1067	552	3680
6	HD	733	899	733	2365
7	Total	4912	2401	1448	8761
8					
9	Shown as percentages of row				
10		NS	OS	HS	Total
11	ND	78.0%	16.0%	6.0%	100.0%
12	OD	56.0%	29.0%	15.0%	100.0%
13	HD	31.0%	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 one of the most 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.

Example 3.2:
Baseball Salaries 2011 Extra.xlsx (slide 1 of 2)

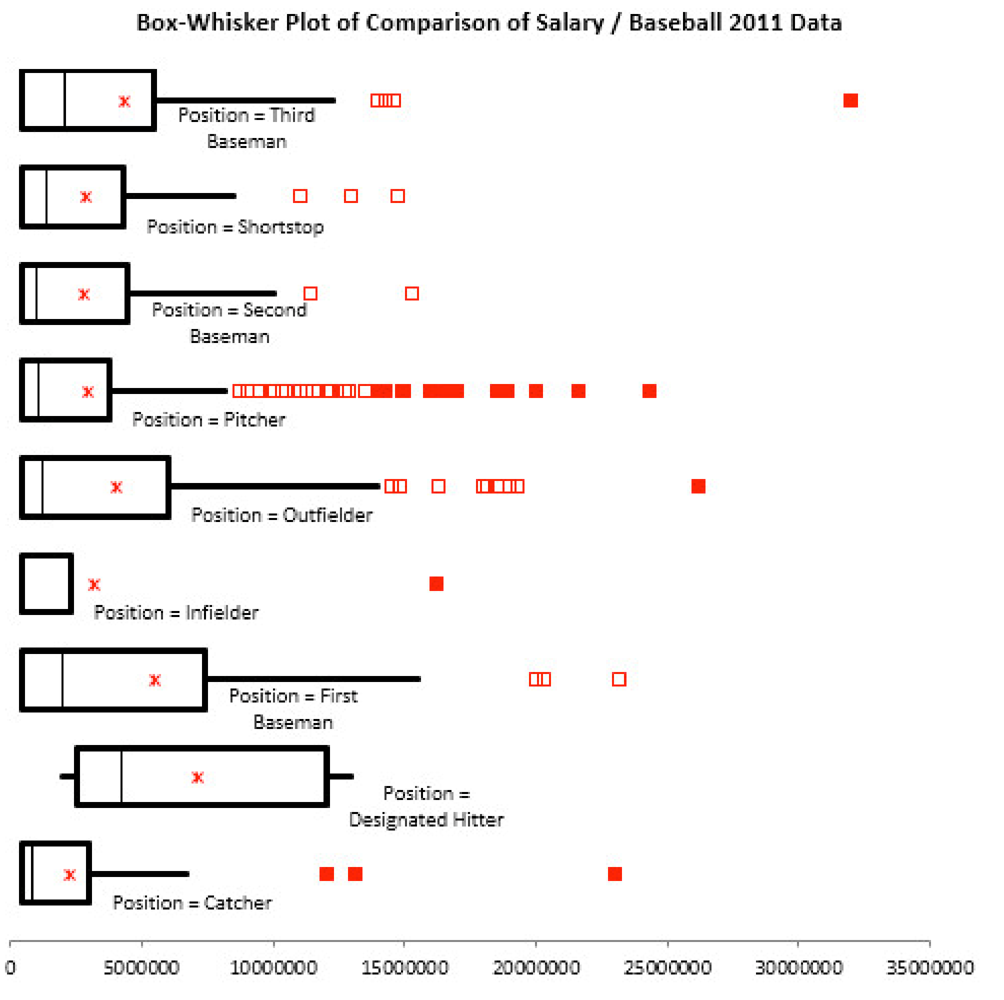
- **Objective**: To learn methods for breaking down baseball salaries by various categorical variables.
- **Solution**: Data set contains the same 2011 baseball data examined previously, as well as several extra categorical variables.



	A	B	C	D	E	F	G	H	I	J
7		Salary (Catcher)	Salary (Designated Hitter)	Salary (First Baseman)	Salary (Infielder)	Salary (Outfielder)	Salary (Pitcher)	Salary (Second Baseman)	Salary (Shortstop)	Salary (Third Baseman)
8	*One Variable Summary*	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data	Baseball 2011 Data
9	Mean	$2252780.70	$7110181.88	$5452236.81	$3162678.71	$4018200.30	$2943853.37	$2776197.15	$2852726.28	$4309856.85
10	Std. Dev.	$3539587.29	$4783000.51	$6692183.63	$5795695.23	$5210328.15	$4043494.94	$3387236.65	$3564560.06	$5943914.21
11	Median	$850000.00	$4250000.00	$2000000.00	$428600.00	$1250000.00	$1095000.00	$1000000.00	$1350000.00	$2050000.00
12	Minimum	$414000.00	$2020000.00	$414000.00	$414000.00	$414000.00	$414000.00	$414000.00	$414000.00	$414000.00
13	Maximum	$23000000.00	$13000000.00	$23125000.00	$16174974.00	$26187500.00	$24285714.00	$15285714.00	$14729364.00	$32000000.00
14	Count	69	8	42	7	152	413	59	47	46
15	1st Quartile	$424000.00	$2500000.00	$427500.00	$414000.00	$443000.00	$431500.00	$425000.00	$425400.00	$478000.00
16	3rd Quartile	$3000000.00	$12000000.00	$7410655.00	$2285677.00	$6000000.00	$3750000.00	$4500000.00	$4300000.00	$5500000.00

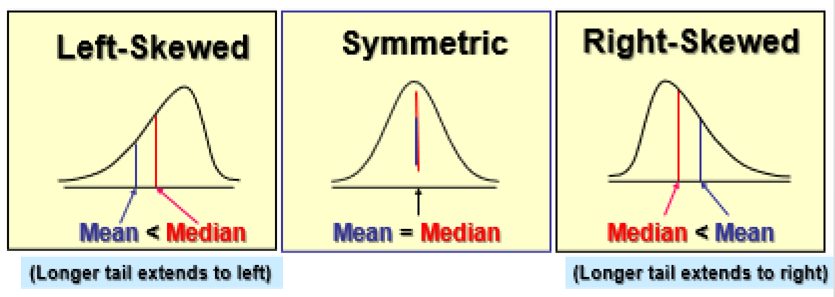
Example 3.2: Baseball Salaries 2011 Extra.xlsx (slide 2 of 2)

- Create side-by-side boxplots, by selecting Box-Whisker Plot from the Summary Graphs dropdown list and filling in the resulting dialog box.



Elements of Boxplots

1. The box itself, from bottom to top, extends from the first quartile to the third quartile, so it contains the middle 50% of the data. (Box plots from some software packages are shown horizontally rather than vertically. Then the *width* of the box indicates the middle 50% of the data.)
2. The horizontal line inside the box represents the median, and the x inside the box represents the mean.
3. The lines from either end of the box, also called whiskers, extend as far as the most distant data value within 1.5 IQRs (interquartile ranges) of the box.
4. Any data values beyond the whiskers are called outliers and are shown as individual points.



Relationships Among Numerical Variables

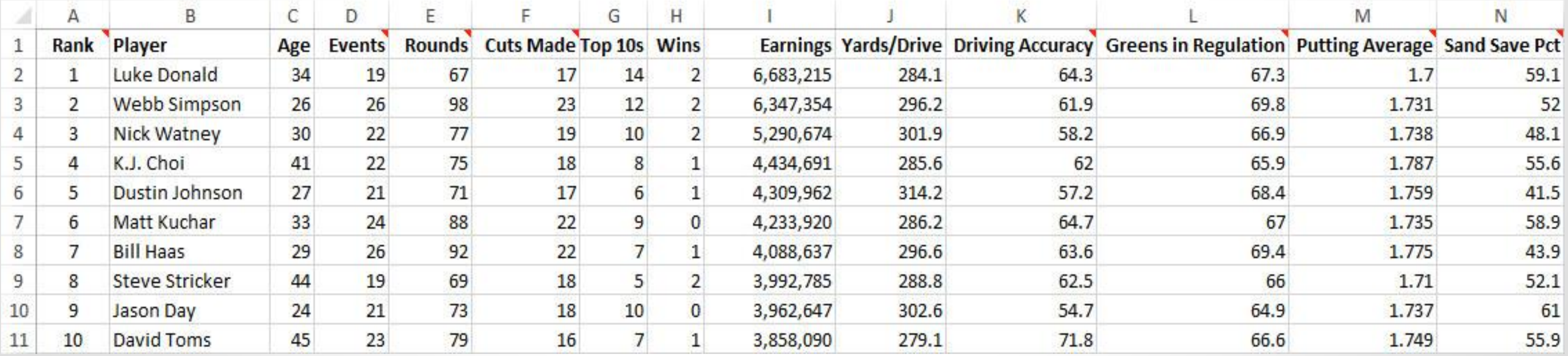
- To study relationships among numerical variables, a new type of chart, called a scatterplot, and two new summary measures, correlation and covariance, are used.
- These measures can be applied to any variables that are displayed numerically.
- However, they are appropriate only for truly numerical 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.

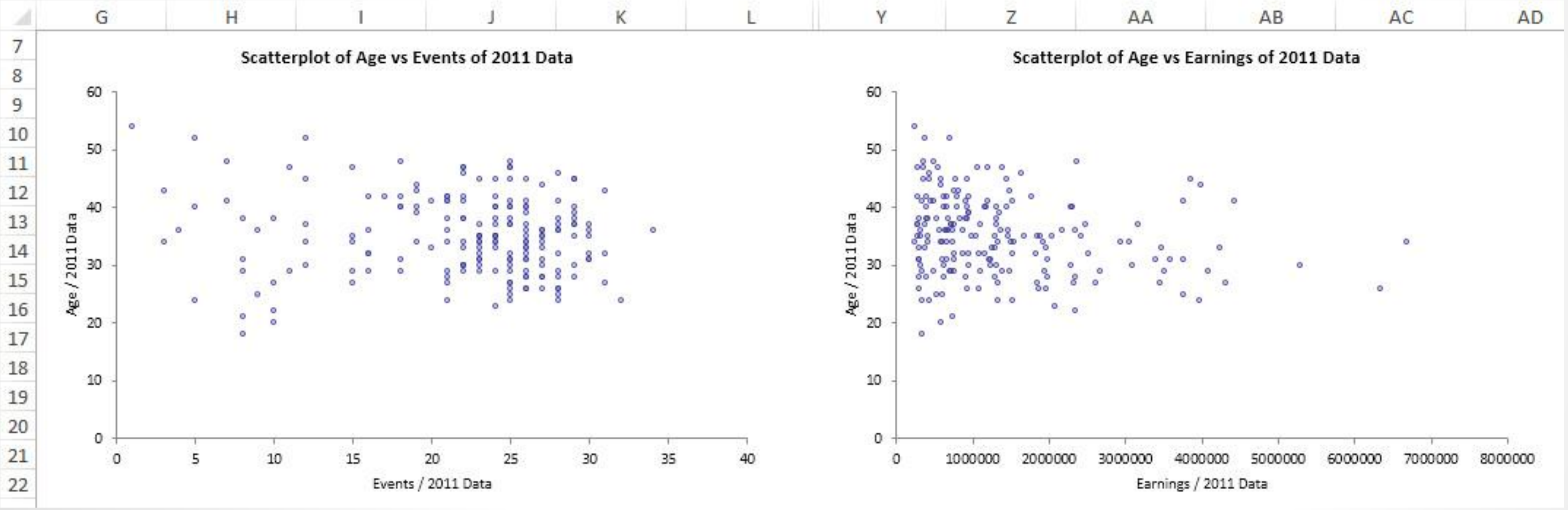
Example 3.3: GolfStats.xlsx (slide 1 of 2)

- **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.
- From Charts, select scatterplot. (Age Vs Events and Age Vs. Earnings of 2011)



	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

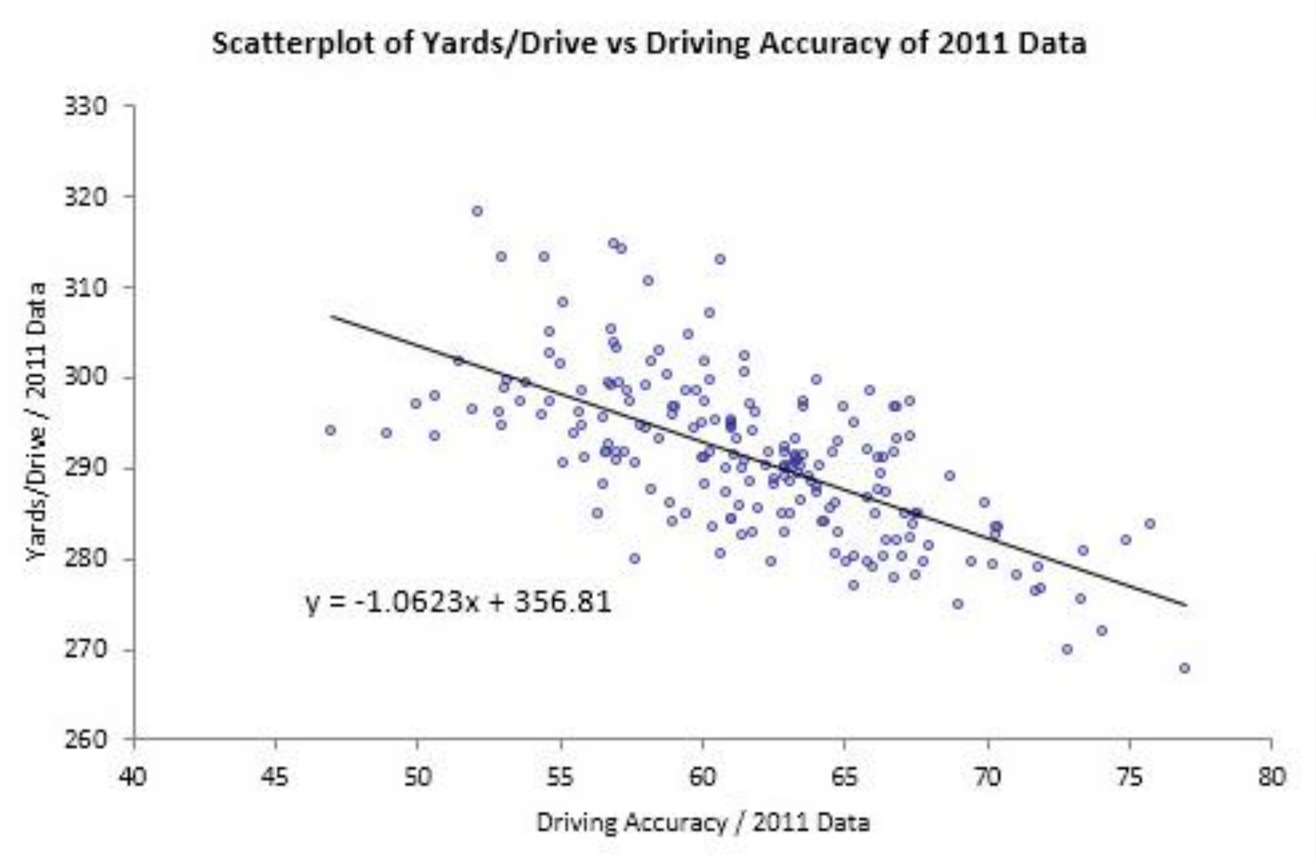
Example 3.3: GolfStats.xlsx (slide 2 of 2)



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.
- To do this, right-click on any point in the chart, select Add Trendline, and fill out the resulting dialog box.

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, that calculates covariances automatically.
- 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.

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

- 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, that calculates correlations automatically.

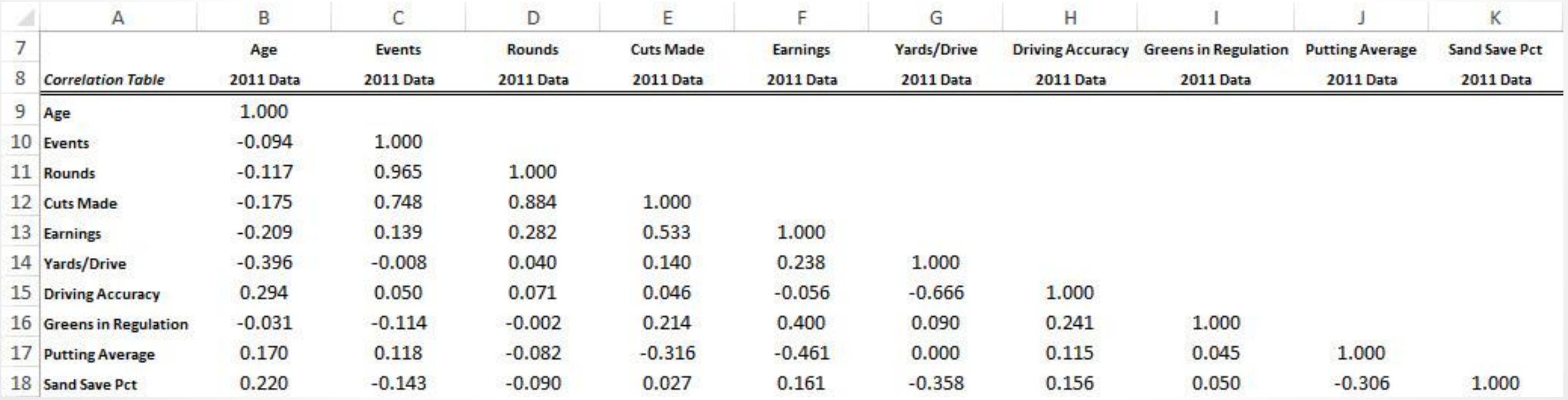
Correlation and Covariance

(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.

Example 3.3 (Continued) GolfStats.xlsx (slide 1 of 2)

- **Objective**: To use correlations to understand relationships in the golf data.
- **Solution**: Create a table of correlations by selecting Correlation and Covariance from the Summary Statistics dropdown list.
- Fill in the resulting dialog box and check Correlations.

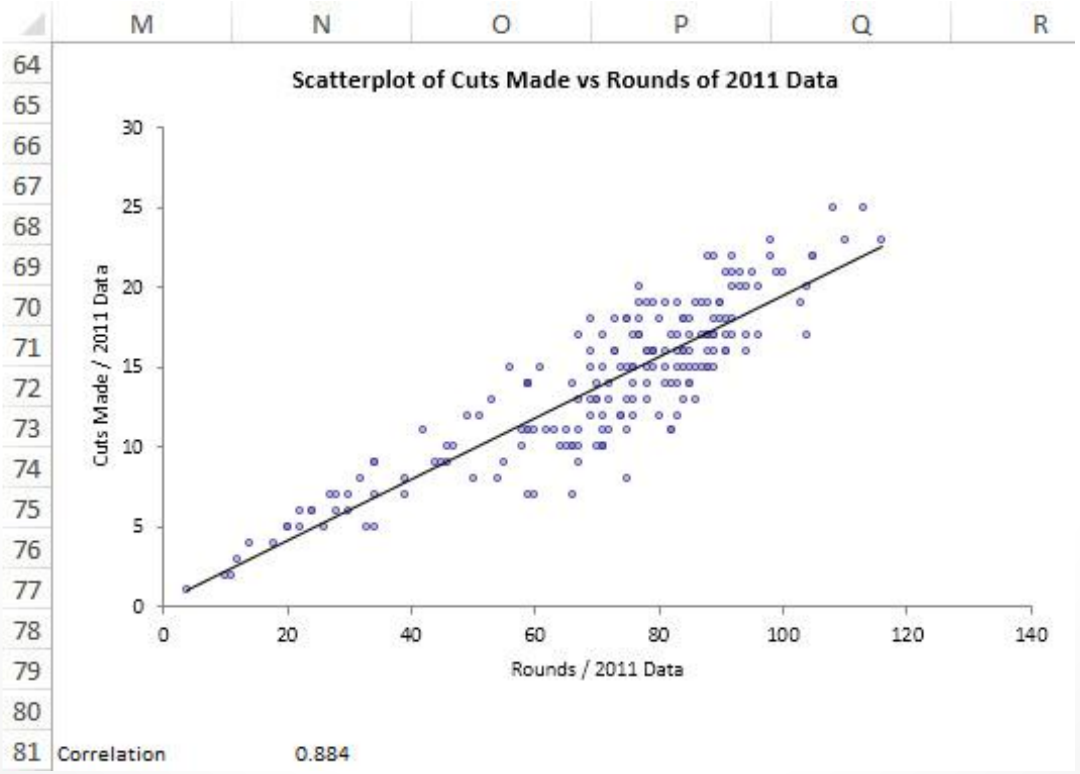


	A	B	C	D	E	F	G	H	I	J	K
7		Age	Events	Rounds	Cuts Made	Earnings	Yards/Drive	Driving Accuracy	Greens in Regulation	Putting Average	Sand Save Pct
8	*Correlation Table*	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data
9	Age	1.000									
10	Events	-0.094	1.000								
11	Rounds	-0.117	0.965	1.000							
12	Cuts Made	-0.175	0.748	0.884	1.000						
13	Earnings	-0.209	0.139	0.282	0.533	1.000					
14	Yards/Drive	-0.396	-0.008	0.040	0.140	0.238	1.000				
15	Driving Accuracy	0.294	0.050	0.071	0.046	-0.056	-0.666	1.000			
16	Greens in Regulation	-0.031	-0.114	-0.002	0.214	0.400	0.090	0.241	1.000		
17	Putting Average	0.170	0.118	-0.082	-0.316	-0.461	0.000	0.115	0.045	1.000	
18	Sand Save Pct	0.220	-0.143	-0.090	0.027	0.161	-0.358	0.156	0.050	-0.306	1.000

Example 3.3 (Continued)

GolfStats.xlsx (slide 2 of 2)

- You can learn more about a correlation by creating the corresponding scatterplot.



Correlation 0.884

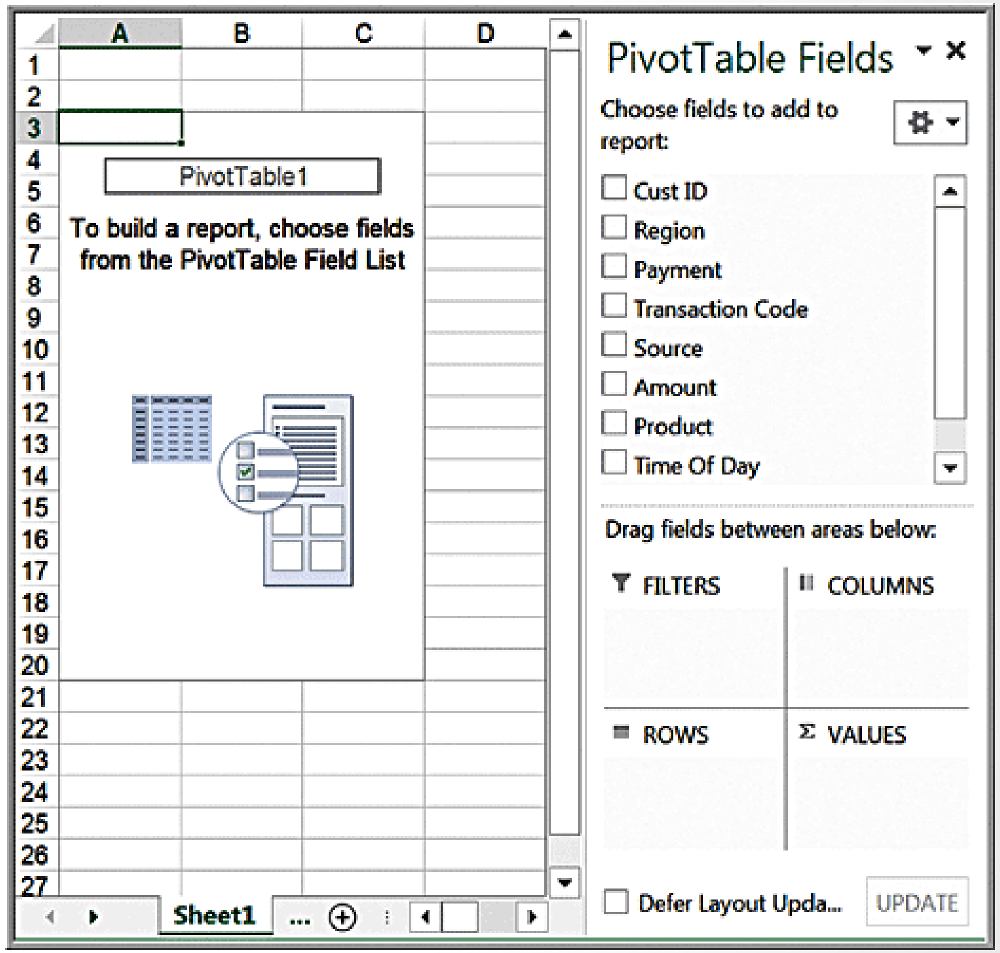
Pivot Tables

- The **pivot table** is an Excel tool that allows you to break data down by categories.
- Sometimes pivot tables are used to display tables of counts, often called crosstabs or contingency tables.
- However, crosstabs typically list only counts, whereas pivot tables can list counts, sums, averages, and other summary measures.
- PivotTables allows you to create custom summaries and charts of key information in the data.

Constructing PivotTables

Click inside your database
Insert >
Tables >
PivotTable

The wizard creates a blank PivotTable as shown.


PivotTable1

To build a report, choose fields from the PivotTable Field List

PivotTable Fields

Choose fields to add to report:

- □ Cust ID
- □ Region
- □ Payment
- □ Transaction Code
- □ Source
- □ Amount
- □ Product
- □ Time Of Day

Drag fields between areas below:

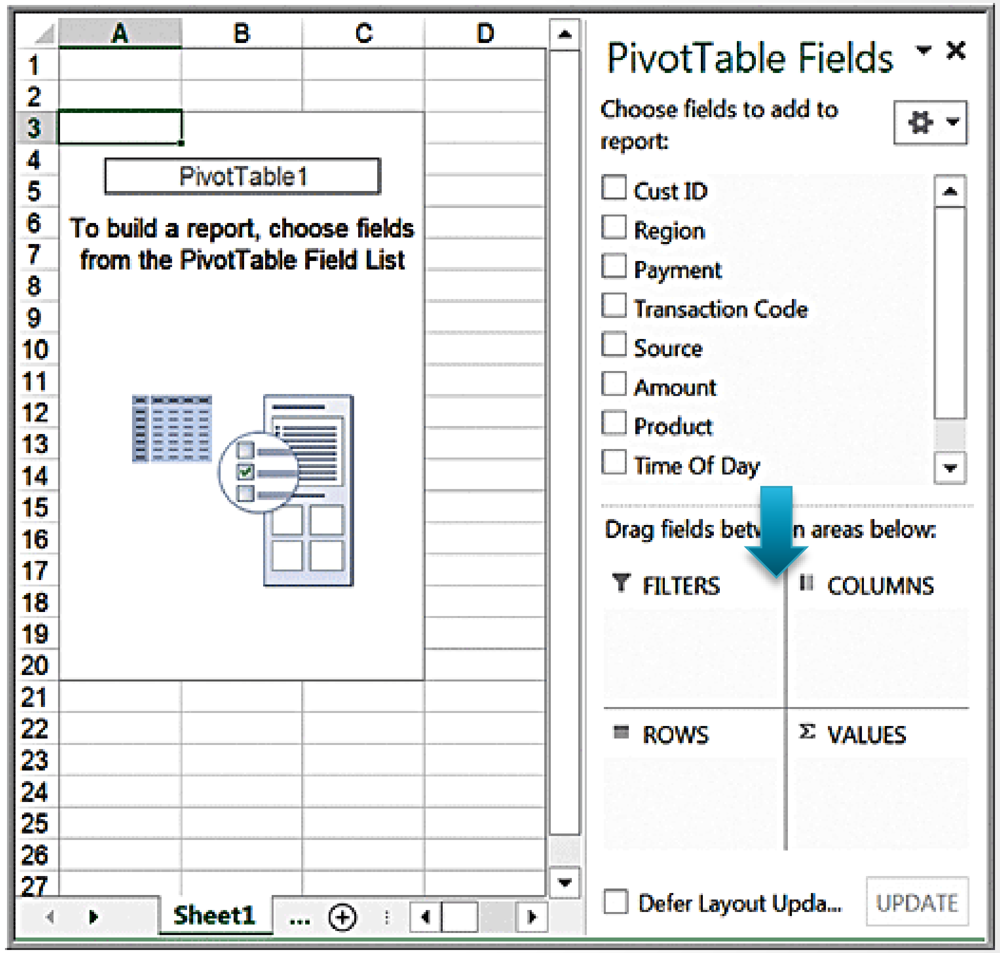
FILTERS | COLUMNS | ROWS | VALUES

□ Defer Layout Upda...

PivotTable Field List

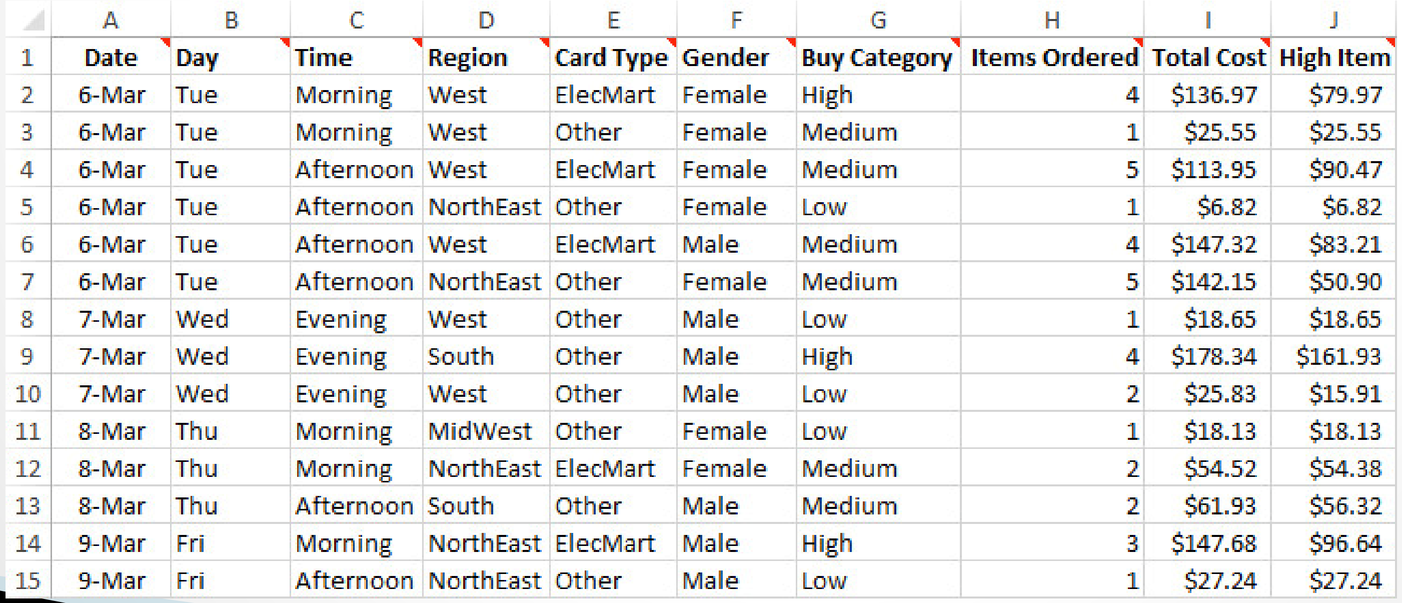
Select and drag the fields to one of the PivotTable areas:

- *Report Filter*
- *Column Labels*
- *Row Labels*
- *Σ Values*



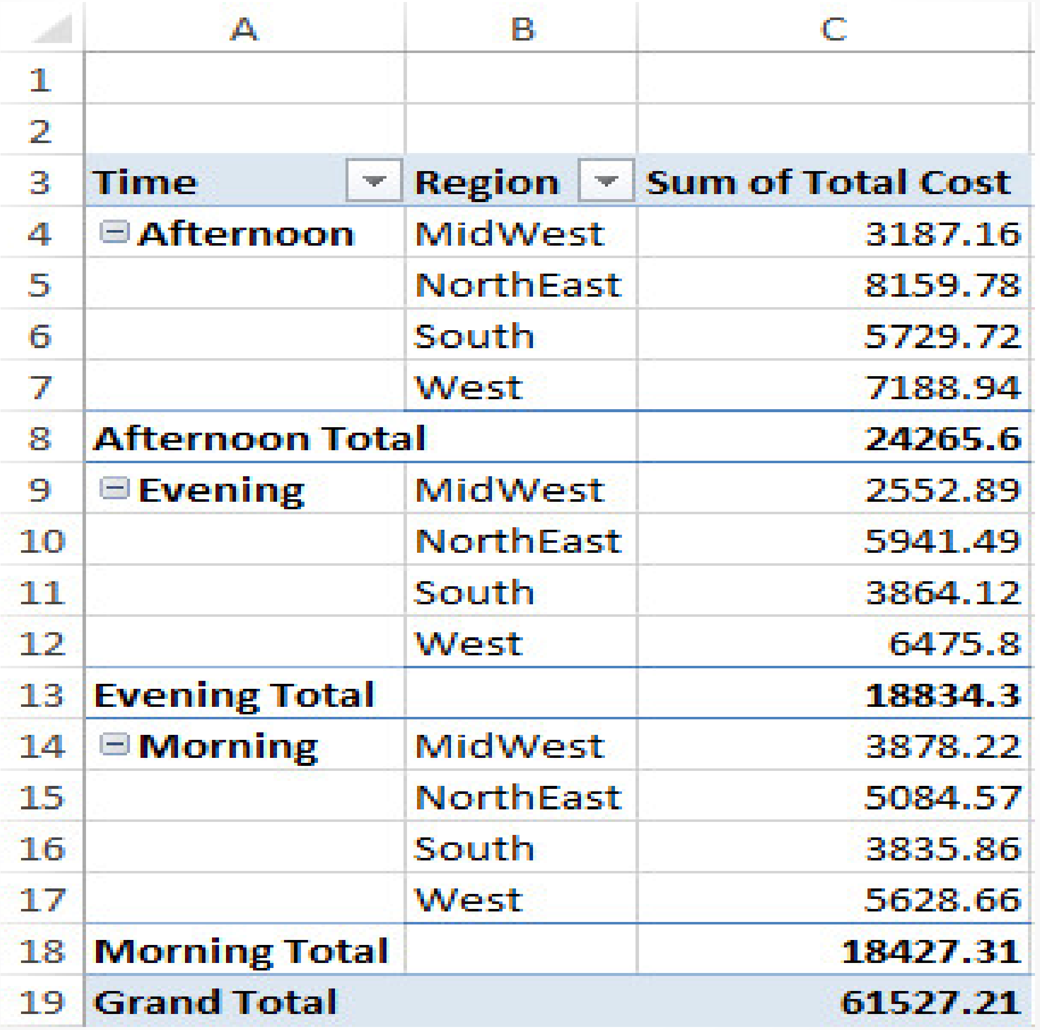
Example 3.4: Elecmart Sales.xlsx (slide 1 of 2)

- **Objective**: To use pivot tables to break down the customer order data by a number of categorical variables.
- **Solution**: Data set contains data on 400 customer orders during several months for Elecmart company.
- Create a pivot table by clicking the PivotTable button on the Insert ribbon.



	A	B	C	D	E	F	G	H	I	J
1	Date	Day	Time	Region	Card Type	Gender	Buy Category	Items Ordered	Total Cost	High Item
2	6-Mar	Tue	Morning	West	ElecMart	Female	High	4	$136.97	$79.97
3	6-Mar	Tue	Morning	West	Other	Female	Medium	1	$25.55	$25.55
4	6-Mar	Tue	Afternoon	West	ElecMart	Female	Medium	5	$113.95	$90.47
5	6-Mar	Tue	Afternoon	NorthEast	Other	Female	Low	1	$6.82	$6.82
6	6-Mar	Tue	Afternoon	West	ElecMart	Male	Medium	4	$147.32	$83.21
7	6-Mar	Tue	Afternoon	NorthEast	Other	Female	Medium	5	$142.15	$50.90
8	7-Mar	Wed	Evening	West	Other	Male	Low	1	$18.65	$18.65
9	7-Mar	Wed	Evening	South	Other	Male	High	4	$178.34	$161.93
10	7-Mar	Wed	Evening	West	Other	Male	Low	2	$25.83	$15.91
11	8-Mar	Thu	Morning	MidWest	Other	Female	Low	1	$18.13	$18.13
12	8-Mar	Thu	Morning	NorthEast	ElecMart	Female	Medium	2	$54.52	$54.38
13	8-Mar	Thu	Afternoon	South	Other	Male	Medium	2	$61.93	$56.32
14	9-Mar	Fri	Morning	NorthEast	ElecMart	Male	High	3	$147.68	$96.64
15	9-Mar	Fri	Afternoon	NorthEast	Other	Male	Low	1	$27.24	$27.24

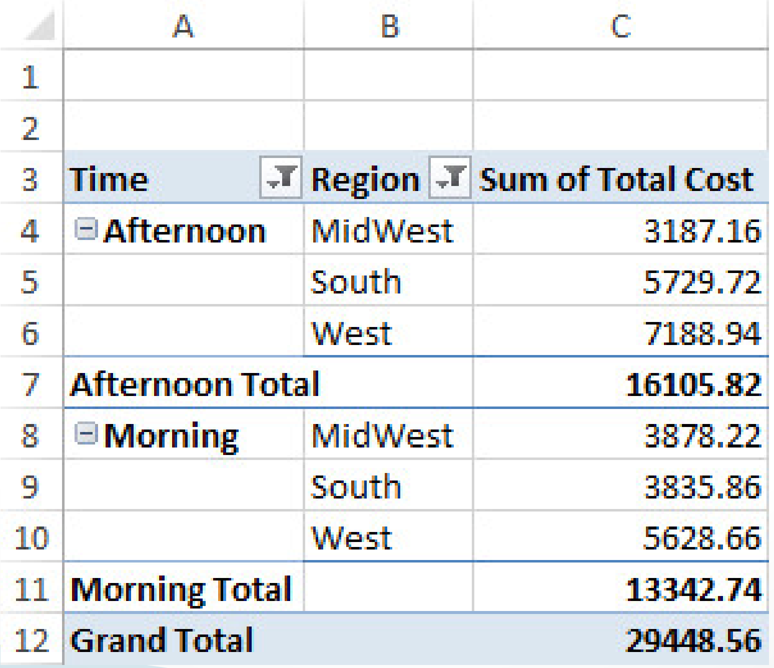
Example 3.4: Elecmart Sales.xlsx (slide 2 of 2)



	A	B	C
1			
2			
3	**Time** ▾	**Region** ▾	**Sum of Total Cost**
4	⊟ **Afternoon**	MidWest	3187.16
5		NorthEast	8159.78
6		South	5729.72
7		West	7188.94
8	**Afternoon Total**		**24265.6**
9	⊟ **Evening**	MidWest	2552.89
10		NorthEast	5941.49
11		South	3864.12
12		West	6475.8
13	**Evening Total**		**18834.3**
14	⊟ **Morning**	MidWest	3878.22
15		NorthEast	5084.57
16		South	3835.86
17		West	5628.66
18	**Morning Total**		**18427.31**
19	**Grand Total**		**61527.21**

Hiding Categories (Filtering)

- You can filter out any items in a pivot table that you don't want to see.
  - Click the Row Labels dropdown arrow of the active field and check the items you want to filter on.
  - A pivot table with hidden categories is shown below.



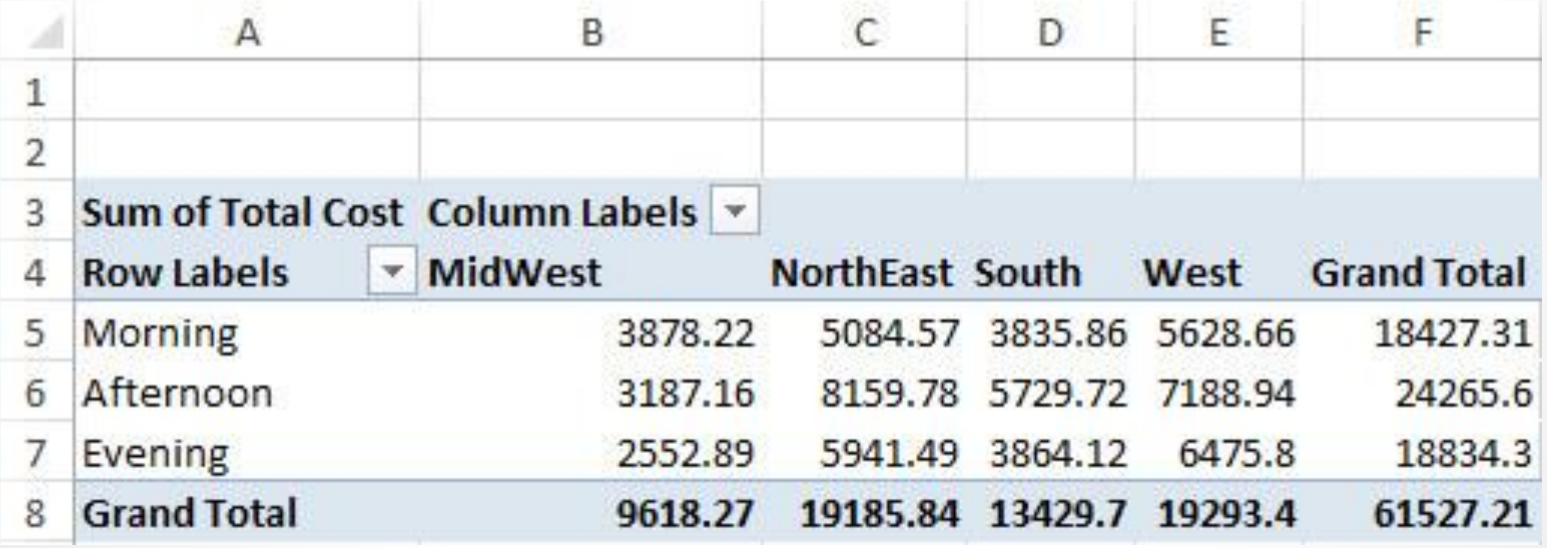
	A	B	C
1			
2			
3	**Time**	**Region**	**Sum of Total Cost**
4	⊟**Afternoon**	MidWest	3187.16
5		South	5729.72
6		West	7188.94
7	**Afternoon Total**		**16105.82**
8	⊟**Morning**	MidWest	3878.22
9		South	3835.86
10		West	5628.66
11	**Morning Total**		**13342.74**
12	**Grand Total**		**29448.56**

Sorting on Values or Categories

- It is easy to sort in a pivot table, either by the numbers in the Values area or by the labels in a Rows or Columns field.
  - To sort by the numbers in the Values area, right-click any number and select Sort.
  - To sort on the labels of a Rows or Columns field, right-click any of the categories and select Sort.
    - You can also click the dropdown arrow for the field and get the dialog box that allows both sorting and filtering.

Changing Locations of Fields (Pivoting)

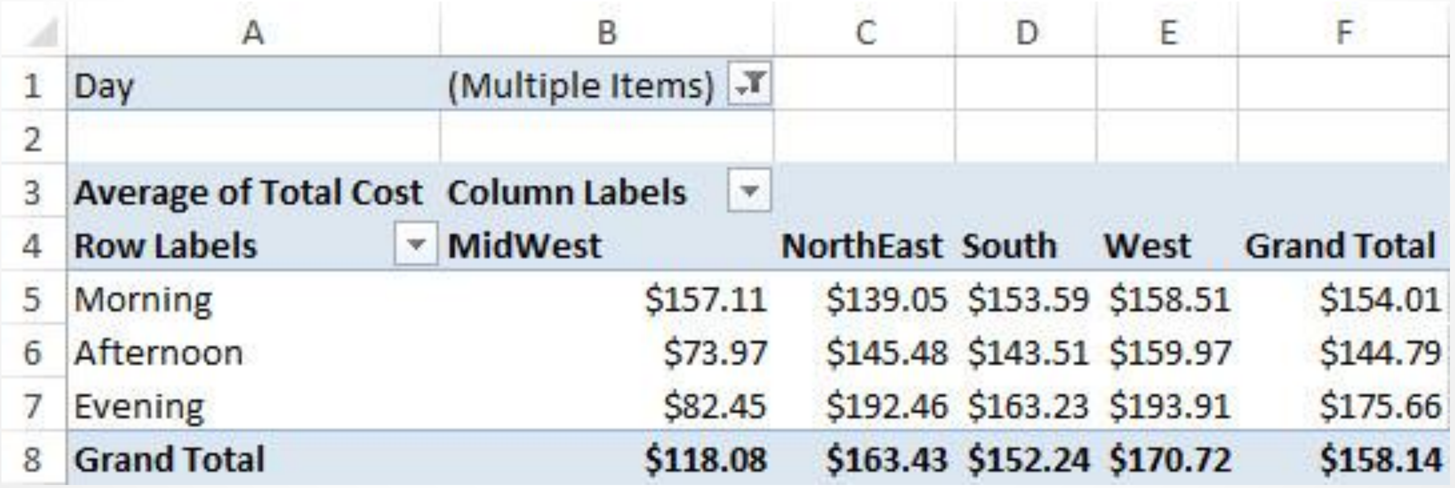
- You can choose where to place variables in a pivot table.
  - For example, to place the Region variable in the Columns area, drag the Region button from the Rows area of the PivotTable Fields pane to the Columns area.



	A	B	C	D	E	F
1						
2						
3	**Sum of Total Cost**	**Column Labels** ▾				
4	**Row Labels** ▾	**MidWest**	**NorthEast**	**South**	**West**	**Grand Total**
5	Morning	3878.22	5084.57	3835.86	5628.66	18427.31
6	Afternoon	3187.16	8159.78	5729.72	7188.94	24265.6
7	Evening	2552.89	5941.49	3864.12	6475.8	18834.3
8	**Grand Total**	**9618.27**	**19185.84**	**13429.7**	**19293.4**	**61527.21**

Changing Field Settings

- You can change various settings in the Field Settings dialog box.
  - To get to this dialog box:
    - Click the Field Setting button on the Analyze/Options ribbon.
    - OR right-click any of the pivot table cells and select the Field Settings item.
  - The pivot table with Value Field Settings changed to Average is shown below.



	A	B	C	D	E	F
1	Day	(Multiple Items)				
2						
3	**Average of Total Cost**	**Column Labels**				
4	**Row Labels**	**MidWest**	**NorthEast**	**South**	**West**	**Grand Total**
5	Morning	$157.11	$139.05	$153.59	$158.51	$154.01
6	Afternoon	$73.97	$145.48	$143.51	$159.97	$144.79
7	Evening	$82.45	$192.46	$163.23	$193.91	$175.66
8	**Grand Total**	**$118.08**	**$163.43**	**$152.24**	**$170.72**	**$158.14**