CSX4202/ITX4202: Data Mining

Lecture 3

Asst. Prof. Dr. Rachsuda Setthawong
Computer Science Department
Assumption University

Outlines

- Data Exploration: Tasks and Techniques
  - Summary Statistics
  - Data Visualization
  - Online Analytical Processing (OLAP)
    - Pivot Table

What is data exploration?

- ***"A preliminary exploration of the data to better understand its characteristics"***

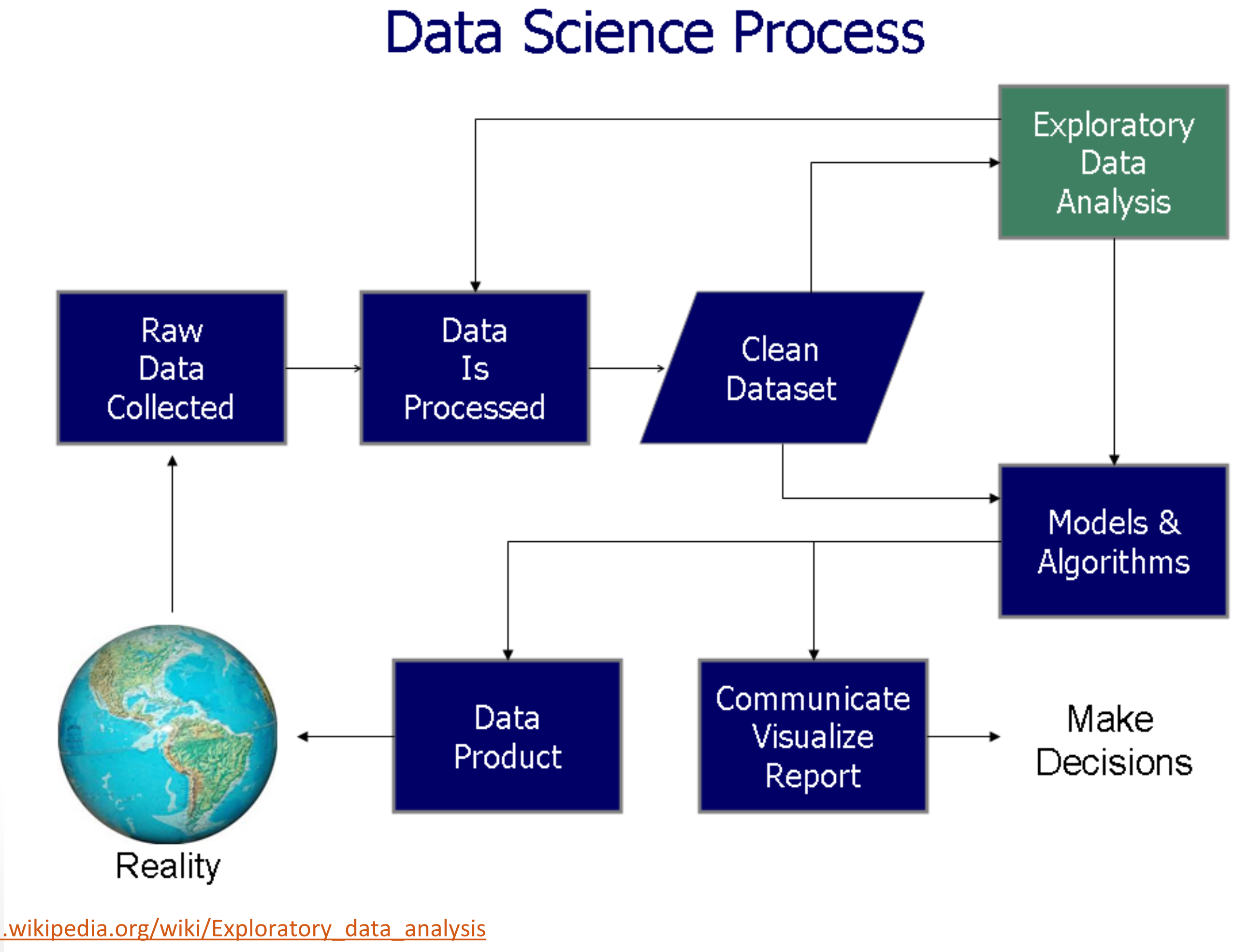
- Key motivations
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

- Characteristics of data
  - Size or amount of data
  - Completeness of the data
  - Correctness of the data
  - Possible relationships amongst data elements

Techniques Used In Data Exploration

- Summary statistics
- Visualization
- Online Analytical Processing (OLAP)

Exploratory Data Analysis (EDA) in Data Science Process



https://en.wikipedia.org/wiki/Exploratory_data_analysis

Asst. Prof. Dr. Rachsuda Setthawong

Five-number Summary in Exploratory Data Analysis (EDA)

- In statistics, John Tukey promoted the use of **five number summary** of numerical data, the two extremes (*maximum and minimum*), the median, and the quartiles in EDA.

- The **five-number summary** is a set of descriptive statistics that provide information about a dataset:
  - the sample minimum *(smallest observation)*
  - the lower quartile *(first quartile or 25th percentile)*
  - the median *(the middle value or 50th percentile)*
  - the upper quartile *(the third quartile or 75th percentile)*
  - the sample maximum *(largest observation)*

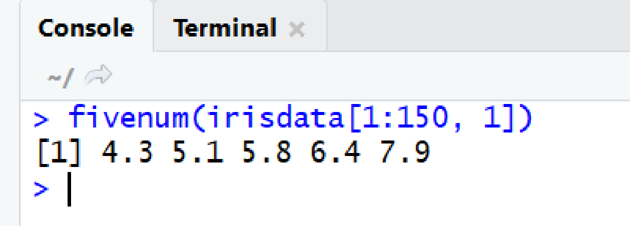
http://web.mnstate.edu/peil/MDEV102/U4/S36/S363.html#:~:text=The%20first%20quartile%2C%20denoted%20by,half%20of%20the%20data%20set.

Five-number Summary in R

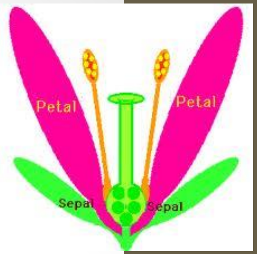
- #Read data from file into data frame
- irisdata<-read.table("C:/Users/noox/Downloads/iris.csv", header=TRUE, sep=",")

- #use help
- help (fivenum)

- #compute 5 Tukey's 5 descriptive statistics
- fivenum(irisdata[1:150, 1]) # or fivenum(irisdata[, 1])



```
> fivenum(irisdata[1:150, 1])
[1] 4.3 5.1 5.8 6.4 7.9
> |
```



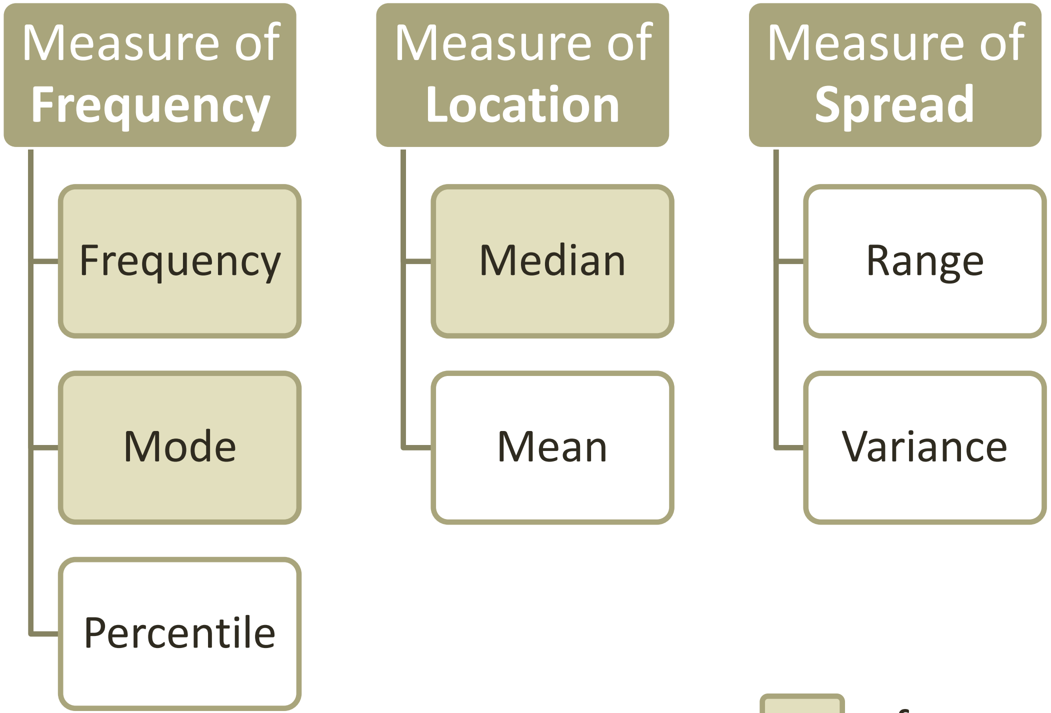
Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Source: UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/iris
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Summary Statistics

- Summary statistics are numbers that summarize properties of the data



Measure of **Frequency**

- Frequency
- Mode
- Percentile

Measure of **Location**

- Median
- Mean

Measure of **Spread**

- Range
- Variance

▨ : for categorical attributes

□ : for continuous attributes

Asst. Prof. Dr. Rachsuda Setthawong

Measure of **Frequency**

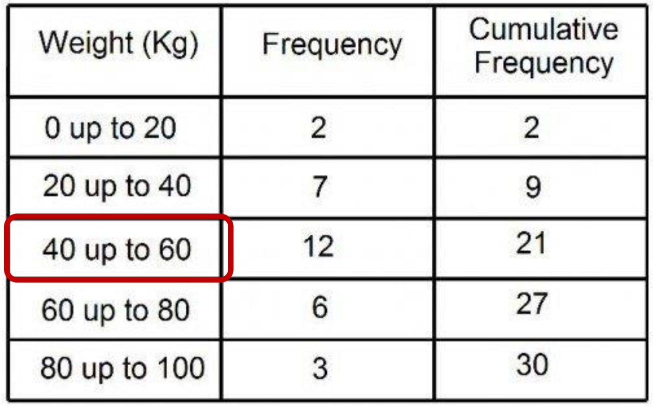
(Categorical attribute)

Frequency

- The **frequency** of an attribute value is *the percentage of time the value occurs in the data set*
- *E.g.,*

Mode

- The **mode** of an attribute is *the most frequent attribute value*



Weight (Kg)	Frequency	Cumulative Frequency
0 up to 20	2	2
20 up to 40	7	9
40 up to 60	12	21
60 up to 80	6	27
80 up to 100	3	30

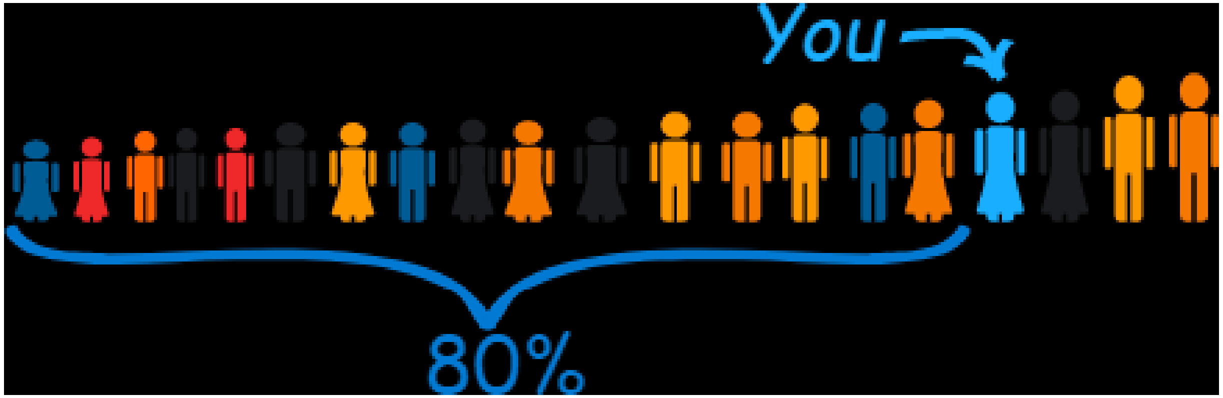
Mode

Percentiles

(Continuous attributes)

- Given an ordinal or continuous attribute *x* and a number *p* between 0 and 100, the p^{th} percentile is a value X_p of x such that *p*% of the observed values of x are less than X_p.

- E.g., if the 80th percentile $X_{80\%}$ is 165, 80% of all values of *X* are below *165*, and the other 20% are above *165.*



Percentiles

(Continuous attributes)

Number (*value*)	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

The rank (R) of the 50th percentile:

R = P/100 x (N + 1)

R = 50/100 x (8 + 1) = 0.5 x 9 = 4.5

If R is an integer, the P^{th} percentile is **the number** with rank R.

Otherwise, interpolate the P^{th} percentile as follows:

1. Define IR as the integer portion of R.

   IR = 4

2. Define FR as the fractional portion of R.

   FR = 0.5

3. Find the numbers (*values*) with Rank I_R and with Rank I_{R+1}.

   $I_R = 8$

   $I_{R+1} = 9$

4. Interpolate it:

   $$\mathbf{R_{Interpolate} = I_R + [\ FR * (I_{R+1} - I_R)\ ]}$$

   $$R_{Interpolate} = 8 + [(0.5)(9 - 8)] = 8.5$$

Ref: http://onlinestatbook.com/2/introduction/percentiles.html

Measures of **Location**:

Mean and Median

- The **mean** is the most common measure of the location of a set of points.
  - Disadv: the mean is very sensitive to outliers.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- The **median** (50th Percentile)

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- **Trimmed mean**
  - E.g., to trim the mean by 40%, we remove the lowest 20% and the highest 20% of values.

Asst. Prof. Dr. Rachsuda Setthawong

Measures of **Spread**:

Range and Variance

- **Range** is the difference between the max and min
- The **variance** or **standard deviation** ($\sqrt{variance}$) measures the spread of a set of points *(how far away the measurements are from the center).*

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Disadv: sensitive to outliers, so that other measures are often used.

Average Absolute Difference: $\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$

Median Absolute Difference: $\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Asst. Prof. Dr. Rachsuda Setthawong

Measures of Data Skewness

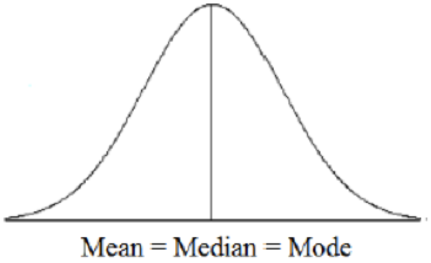
- Measures an imbalance and asymmetry from the mean of a data distribution

$$a_3 = \sum \frac{(X_i - \bar{X})^3}{ns^3}$$

s: the sample standard deviation

Note: different tools' package may use different formula

- **Normal distribution:**

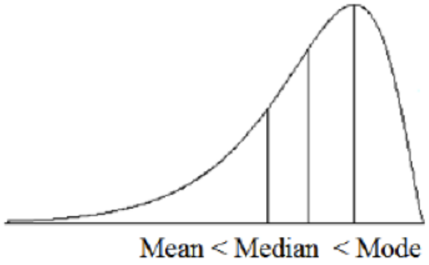


•If the skewness is between -0.5 and 0.5, the data are **fairly symmetrical**

•If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are **moderately skewed**

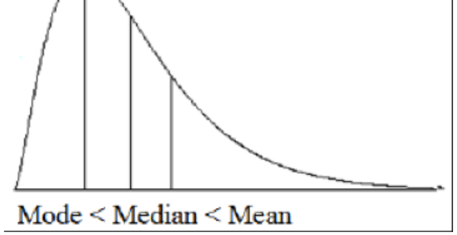
•If the skewness is less than -1 or greater than 1, the data are **highly skewed**

- **Skewness:**



Negative Skew

(longer left-hand tail)



Positive Skew

(longer right-hand tail)

Ref1: https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#skewness
Ref2: https://www.statisticshowto.com/skewness#:~:text=The%20formula%20given%20in%20most,could%20calculate%20skew%20by%20hand.
Article 'Transforming Skewed Data': https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16

What does the skewness of 'Sepal Length' tell us?

sepal length	
Mean	5.843333
Standard Error	0.067611
Median	5.8
Mode	5
Standard Deviation	0.828066
Sample Variance	0.685694
Kurtosis	-0.55206
Skewness	0.314911
Range	3.6
Minimum	4.3
Maximum	7.9
Sum	876.5
Count	150
Confidence Level(95.0%)	0.133601

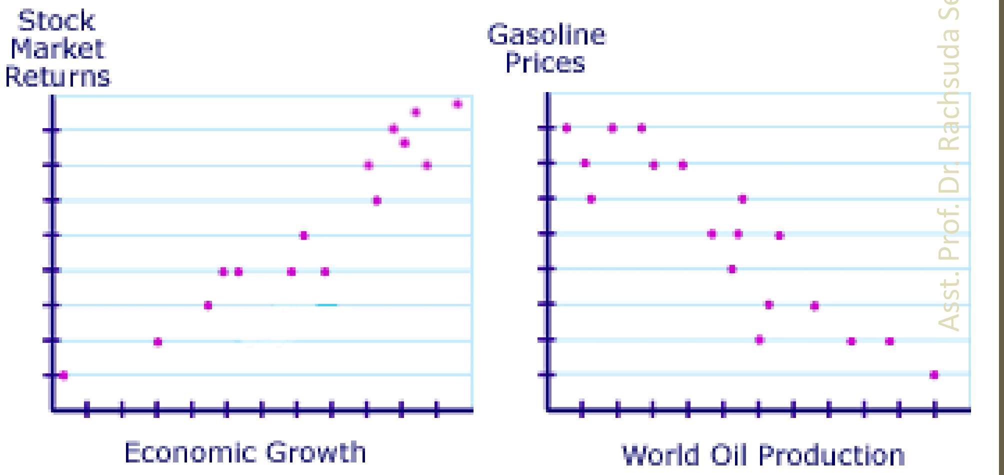
Measure of **Variables' Relationship**: Covariance and Correlation

- Indicate linear relationship of 2 variables (or objects).
- Correlation also shows the degree to which the variables tend to move together.

Covariance:

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Correlation: $r_{(x,y)} = \dfrac{COV(x,y)}{s_x s_y}$



x = the independent variable
y = the dependent variable
n = number of data points in the sample
$\overline{x}$ = the mean of the independent variable x
$\overline{y}$ = the mean of the dependent variable y

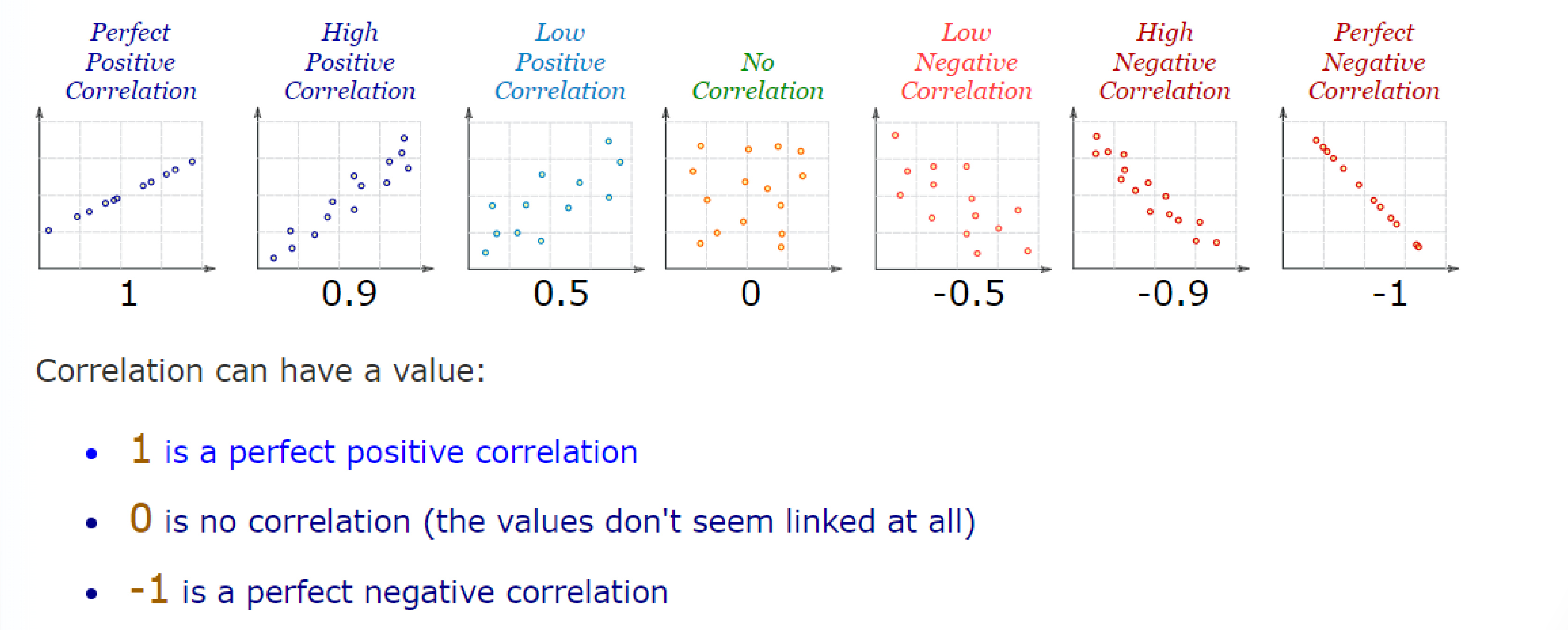
s_x = sample standard deviation of the random variable x
s_y = sample standard deviation of the random variable y

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Note: for Population standard deviation, the divisor is n (*not n-1*).

Ref: https://www0.gsb.columbia.edu/premba/analytical/s7/s7_5.cfm
https://keydifferences.com/difference-between-covariance-and-correlation.html

Correlation Interpretation



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

https://www.mathsisfun.com/data/correlation.html

Tools for Summary Statistics

- R
- Python
- Minitab
- SPSS
- MS Excel
- Weka
- Rapidminer
- Etc.

Tableau:
https://www.tableau.com/trial/data-visualization?utm_campaign_id=2017049&utm_campaign=Prospecting-CORE-ALL-ALL-ALL-ALL&utm_medium=Paid+Search&utm_source=Google+Search&utm_language=EN&utm_country=SEA&kw=%2Bvisualization%20%2Btableau&adgroup=CTX-Brand-Data+Visualization-EN-B&adused=324827239180&matchtype=b&placement=&gclid=EAIaIQobChMIocCWpKuS6gIVyhErCh0G6QaIEAAYASAAEgJ5CPD_BwE&gclsrc=aw.ds

Outlines

- Data Exploration: Tasks and Techniques
  - Summary Statistics
  - Data Visualization
  - Online Analytical Processing (OLAP)
    - Pivot Table

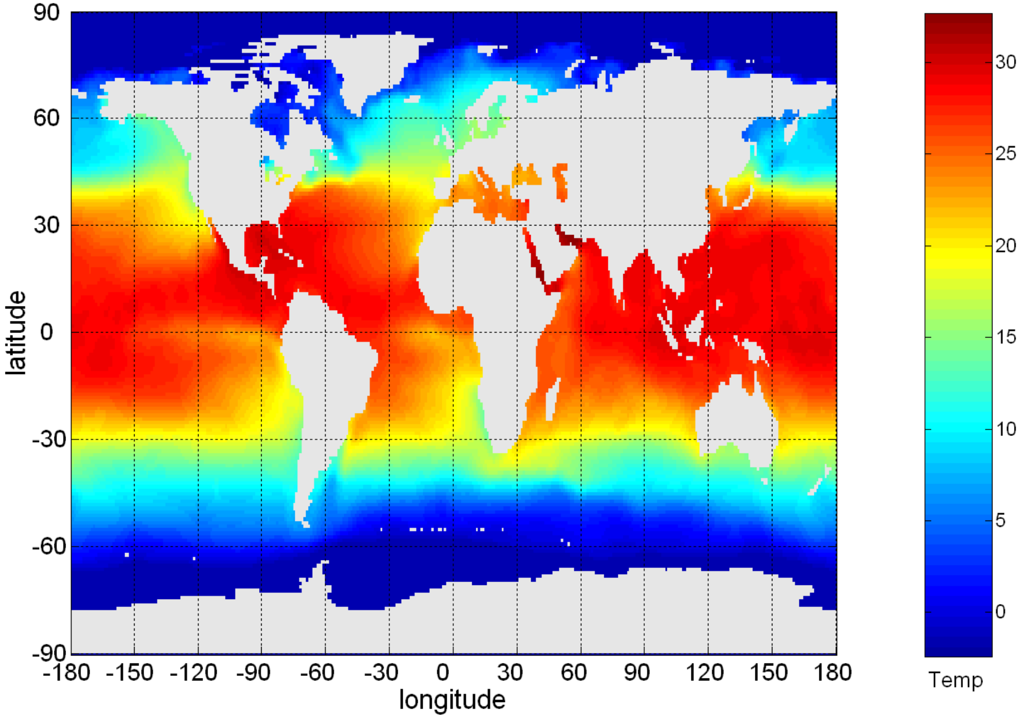
Visualization

- **The conversion of data into a visual or tabular format** so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- One of the most powerful and appealing techniques for data exploration:
  - Humans have a well-developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Example: Sea Surface Temperature

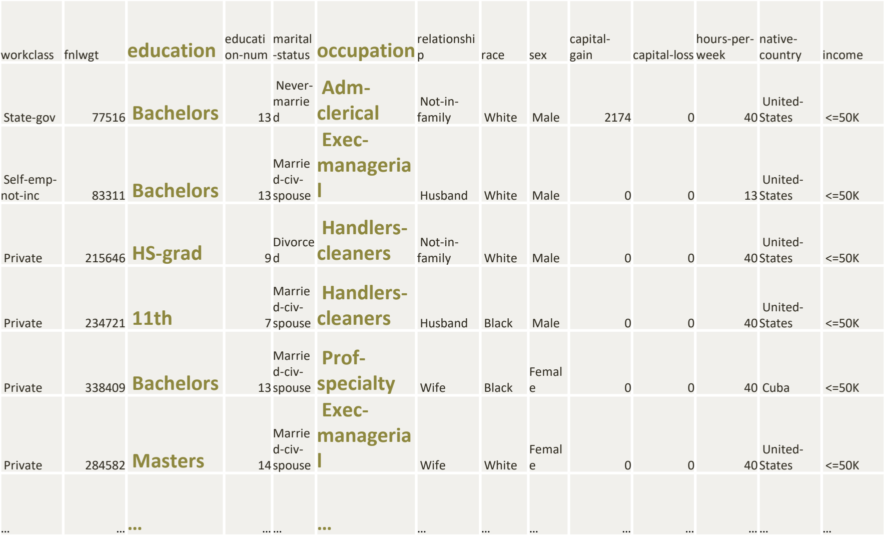
- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



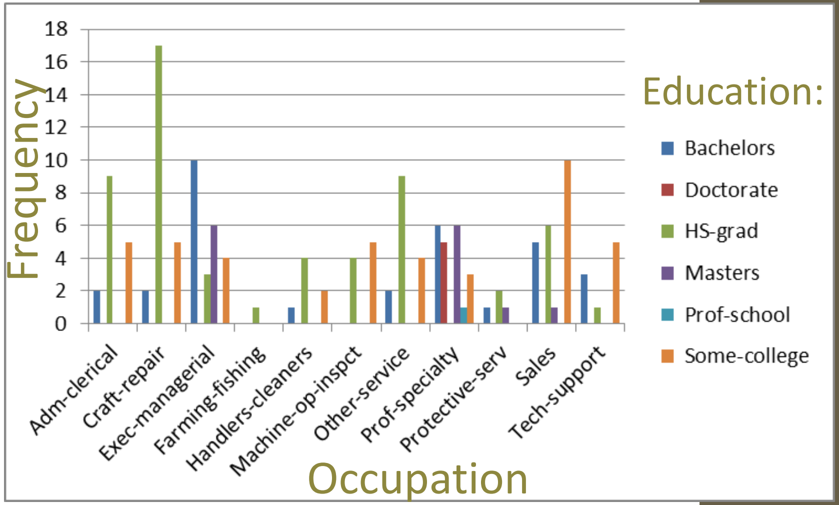
Asst. Prof. Dr. Rachsuda Setthawong

Representation

- Data (objects, their attributes, and their relationships) are translated into graphical elements (points, lines, shapes, and colors).

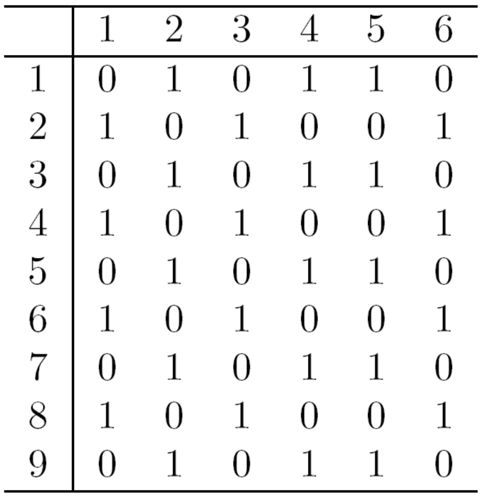


workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
…	…	…	…	…	…	…	…	…	…	…	…	…	…

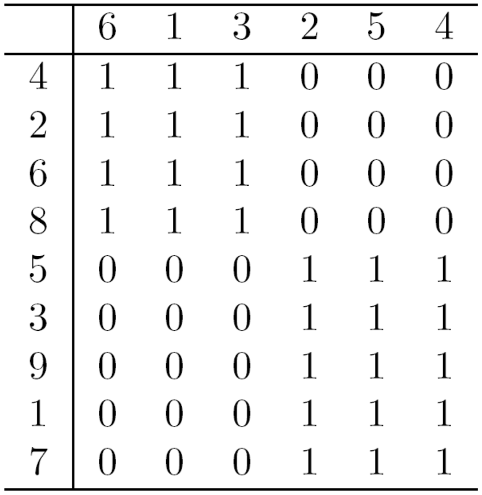


Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0



	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

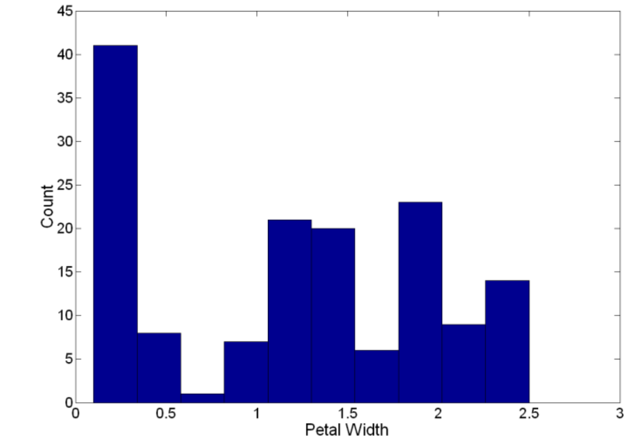
Selection

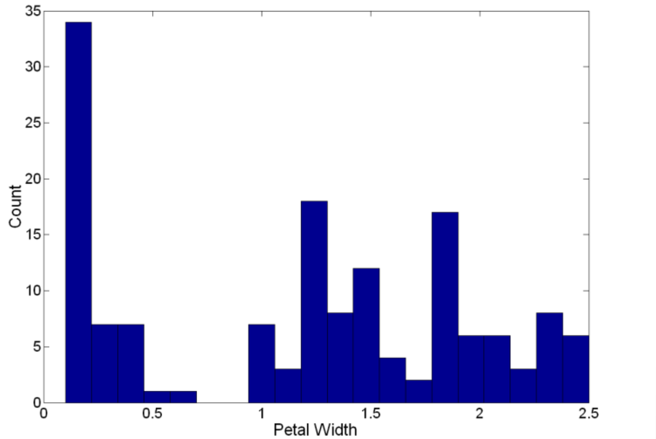
- Choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three

- Choosing a subset of objects
  - A region of the screen can only show limited no. of points
  - Can sample, but want to preserve points in sparse areas

Visualization Techniques: Histograms

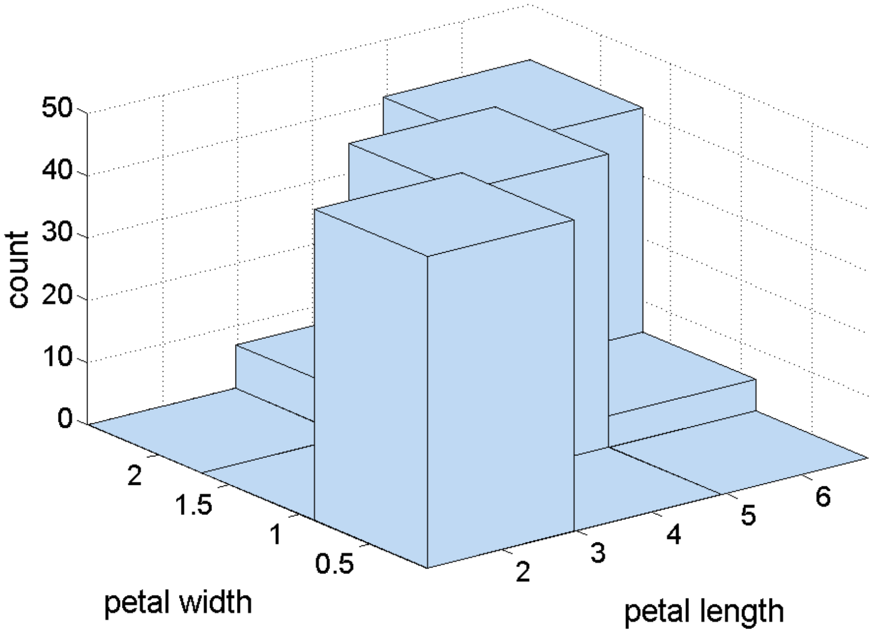
- **Histogram**
  - Usually **shows the distribution of values of a single variable**
  - Divide the values into **bins** and show a bar plot of the number of objects in each bin.
  - The **height** of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





Two-Dimensional Histograms

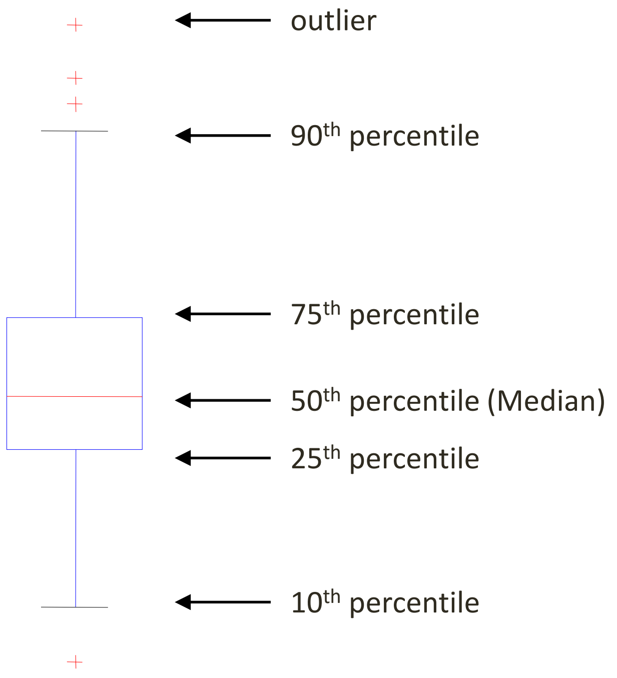
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



Asst. Prof. Dr. Rachsuda Setthawong

Visualization Techniques: Box Plots

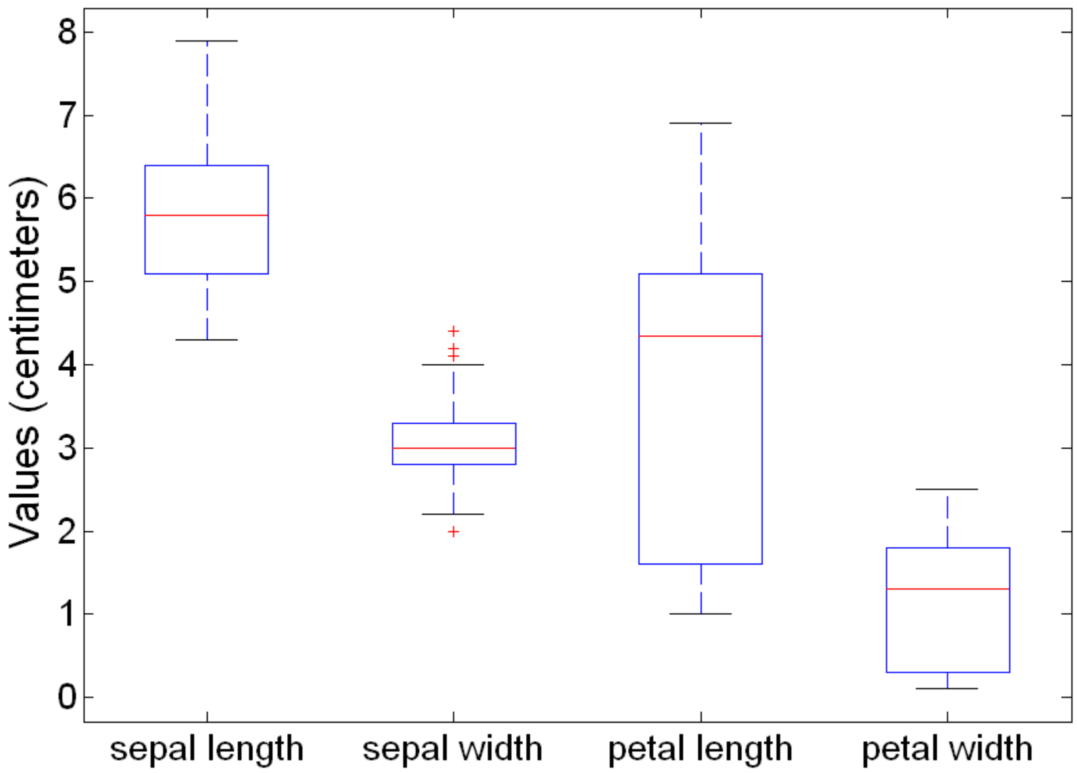
- Another way of displaying the distribution of data



How to create Box Plot: https://www.youtube.com/watch?v=ucWmfmXb1kk

Example of Box Plots

- Box plots can be used to compare attributes



https://medium.com/dayem-siddiqui/understanding-and-interpreting-box-plots-d07aab9d1b6c

Visualization Techniques: Scatter Plots

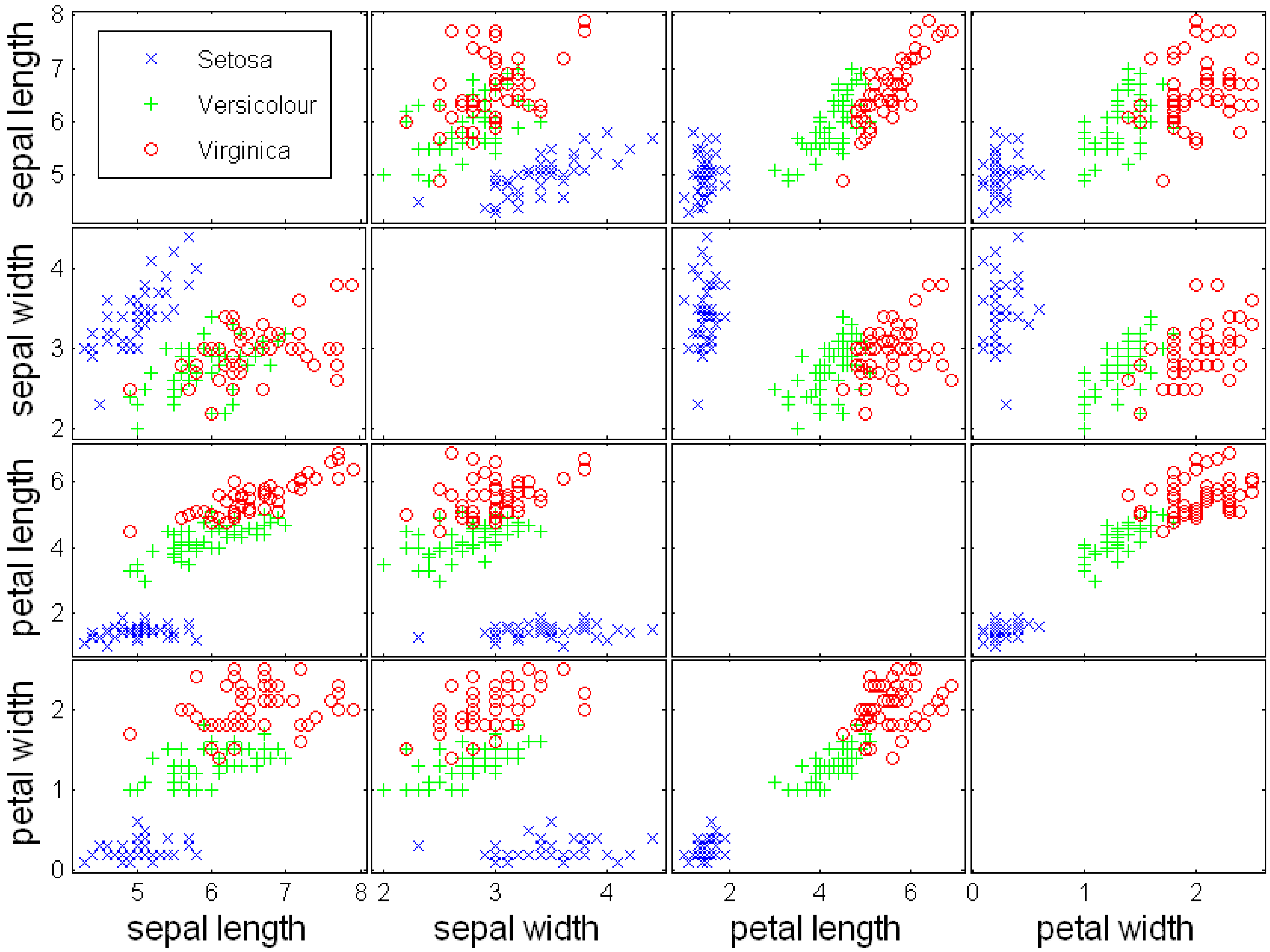
- Attributes values determine the position

- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots

- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects

- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
  - See example on the next slide

Scatter Plot Array of Iris Attributes

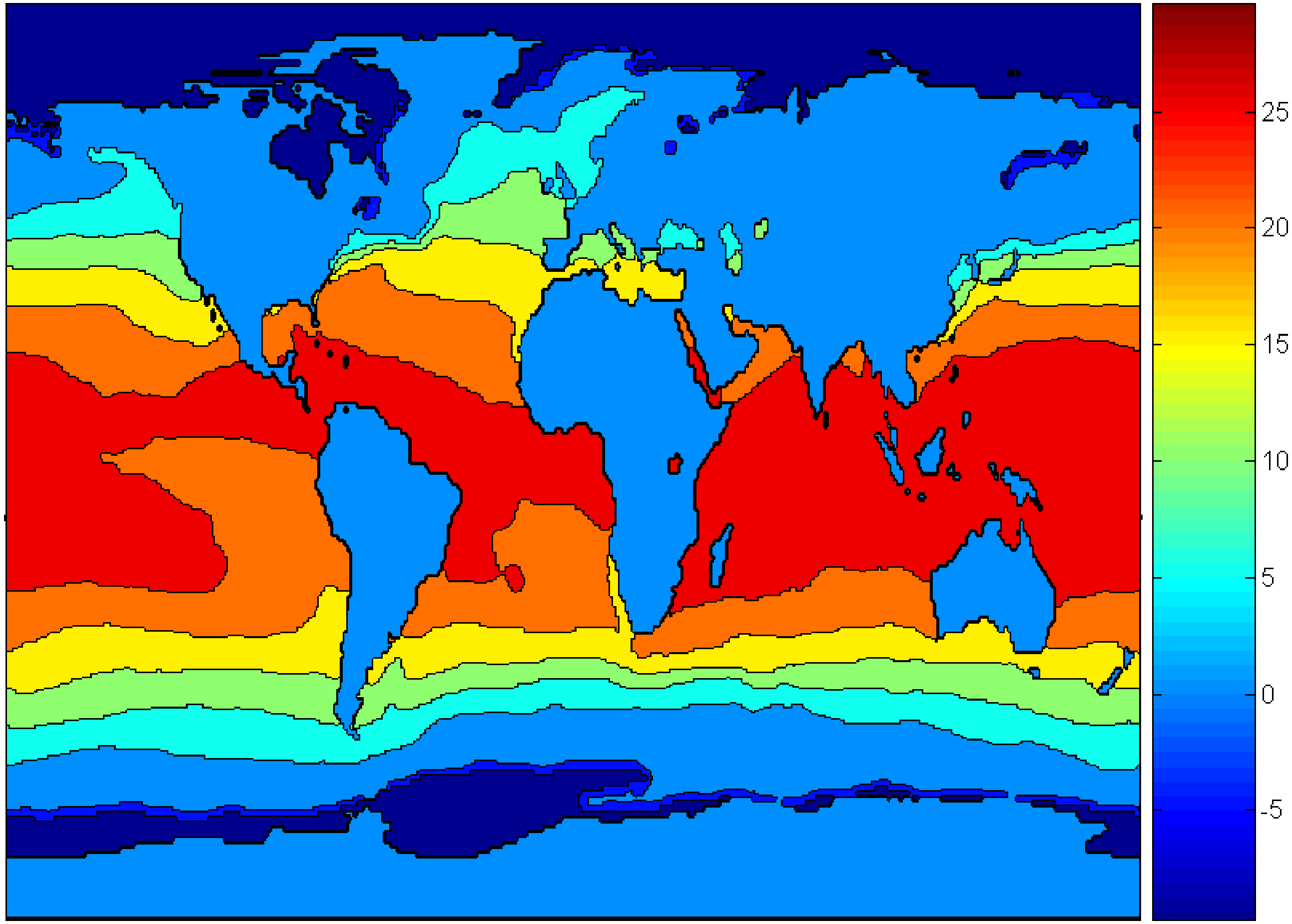


Visualization Techniques: Contour Plots

- Useful when a continuous attribute is measured on a spatial grid.
- Partition the plane into regions of similar values.
- The contour lines that form the boundaries of these regions connect points with equal values.

Contour Plot Example:

Sea Surface Temperature (Dec, 1998)

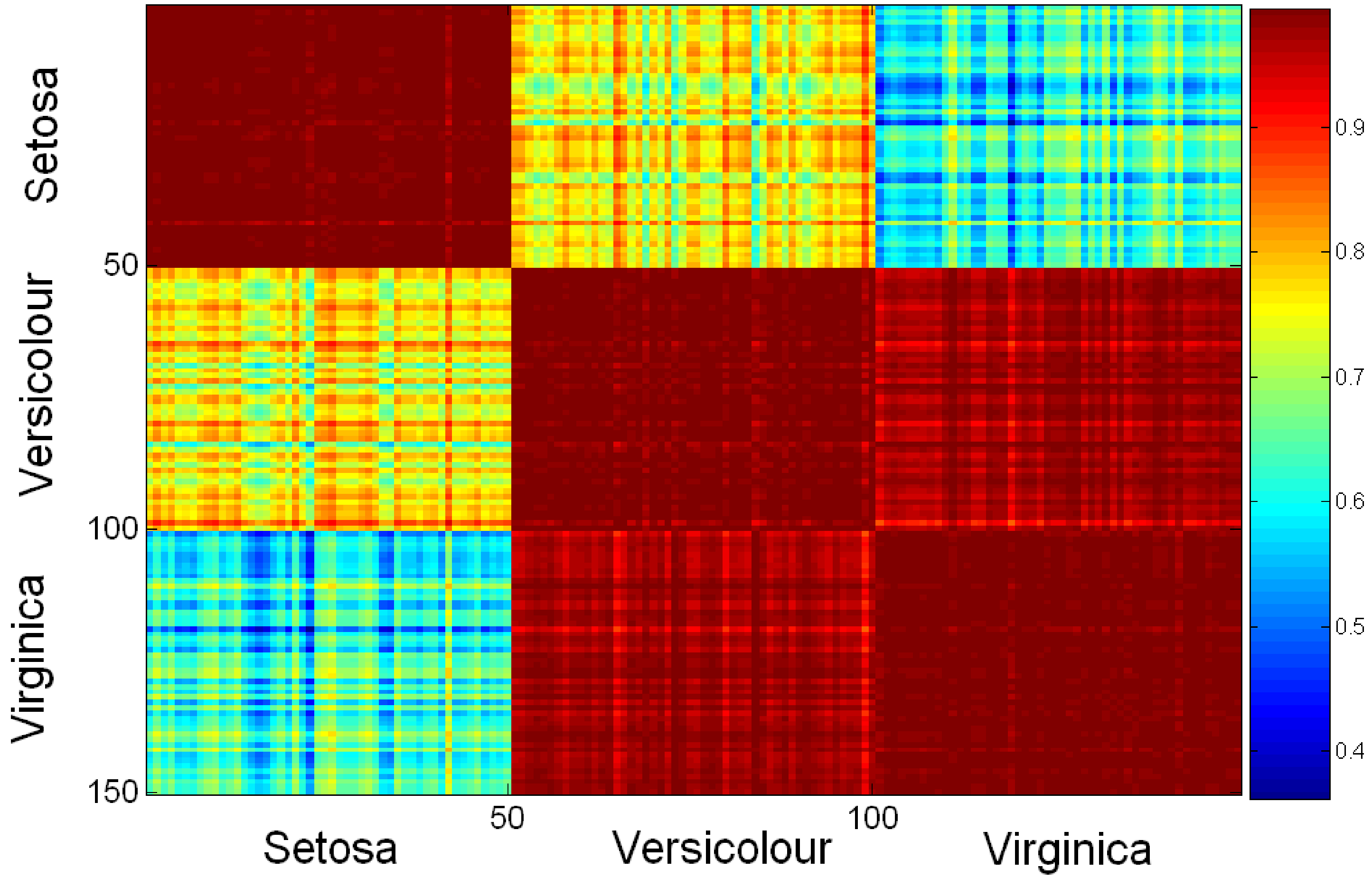


Celsius

Visualization Techniques: Matrix Plots

- Can plot the data matrix
- Useful when objects are sorted according to class
- Typically, the attributes are normalized to prevent one attribute from dominating the plot

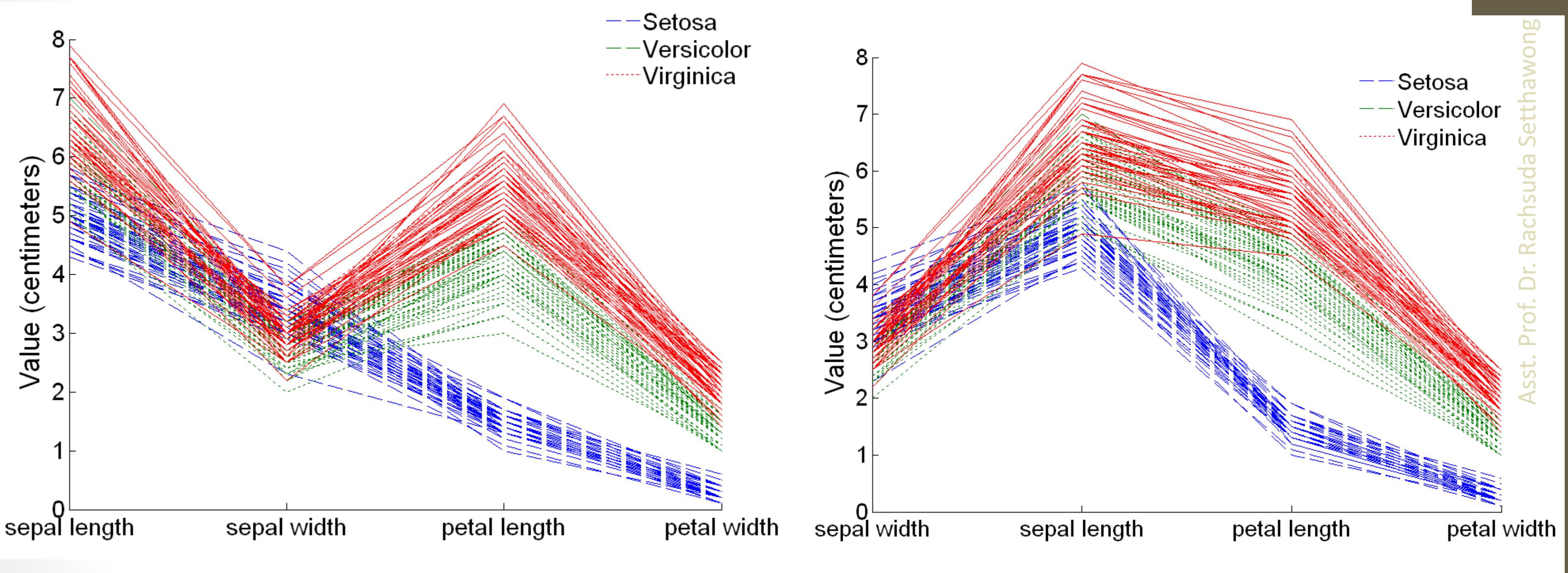
Visualization of the Iris Correlation Matrix



Visualization Techniques: Parallel Coordinates

- Used to plot the attribute values of high-dimensional data
- Use a set of parallel axes instead of using perpendicular axes.
- Thus, **each object is represented as a line**
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis, and
  - The points are connected by a line

Parallel Coordinates Plots for Iris Data



Can rearrange order of attributes

Other Visualization Techniques

- Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon

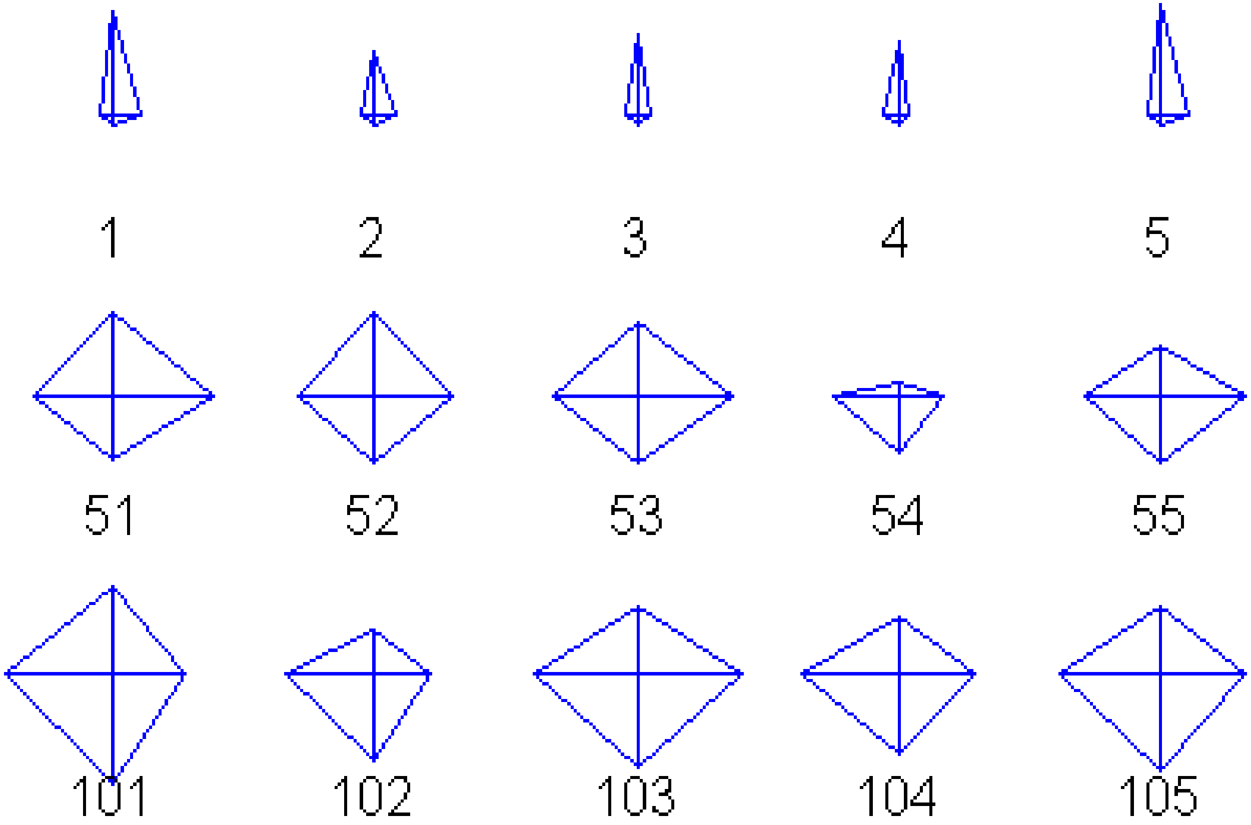
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces

Star Plots for Iris Data

Setosa

Versicolour

Virginica

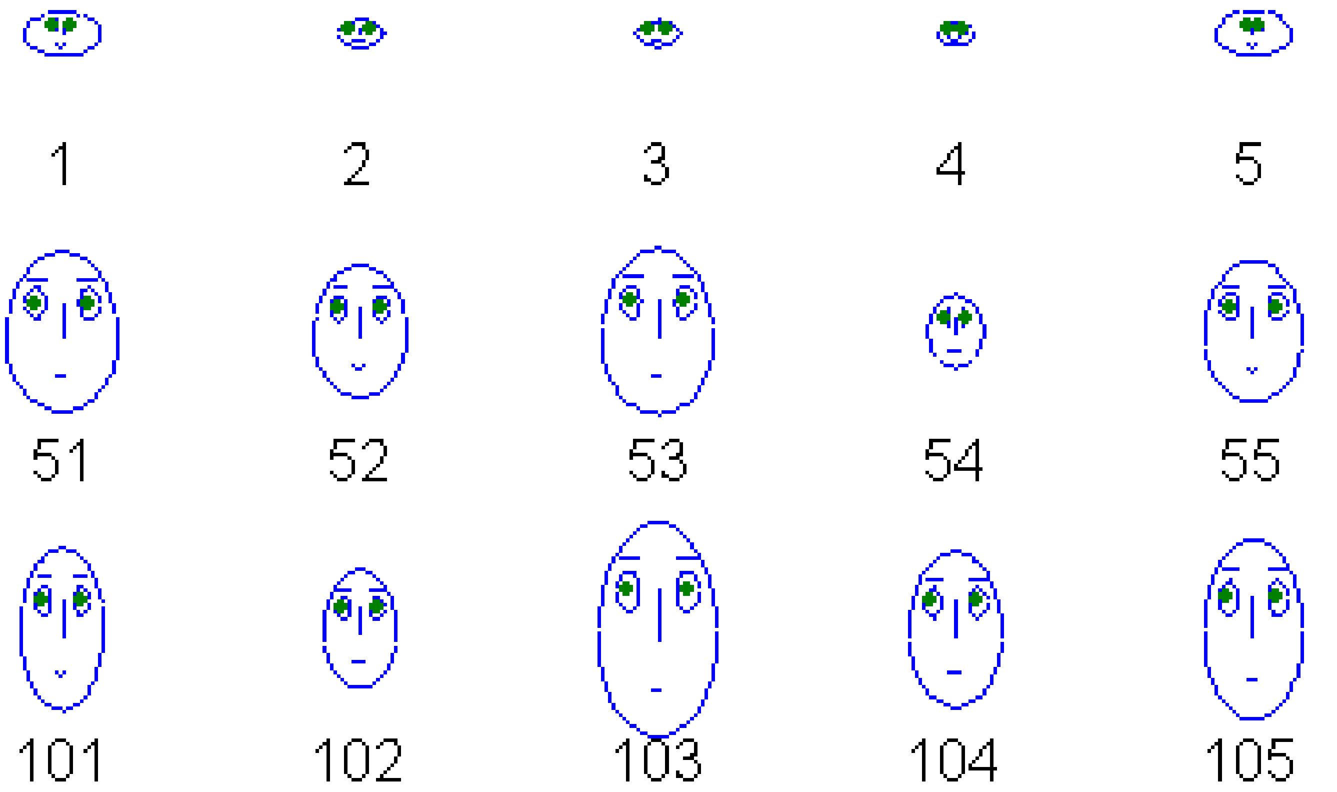


Chernoff Faces for Iris Data

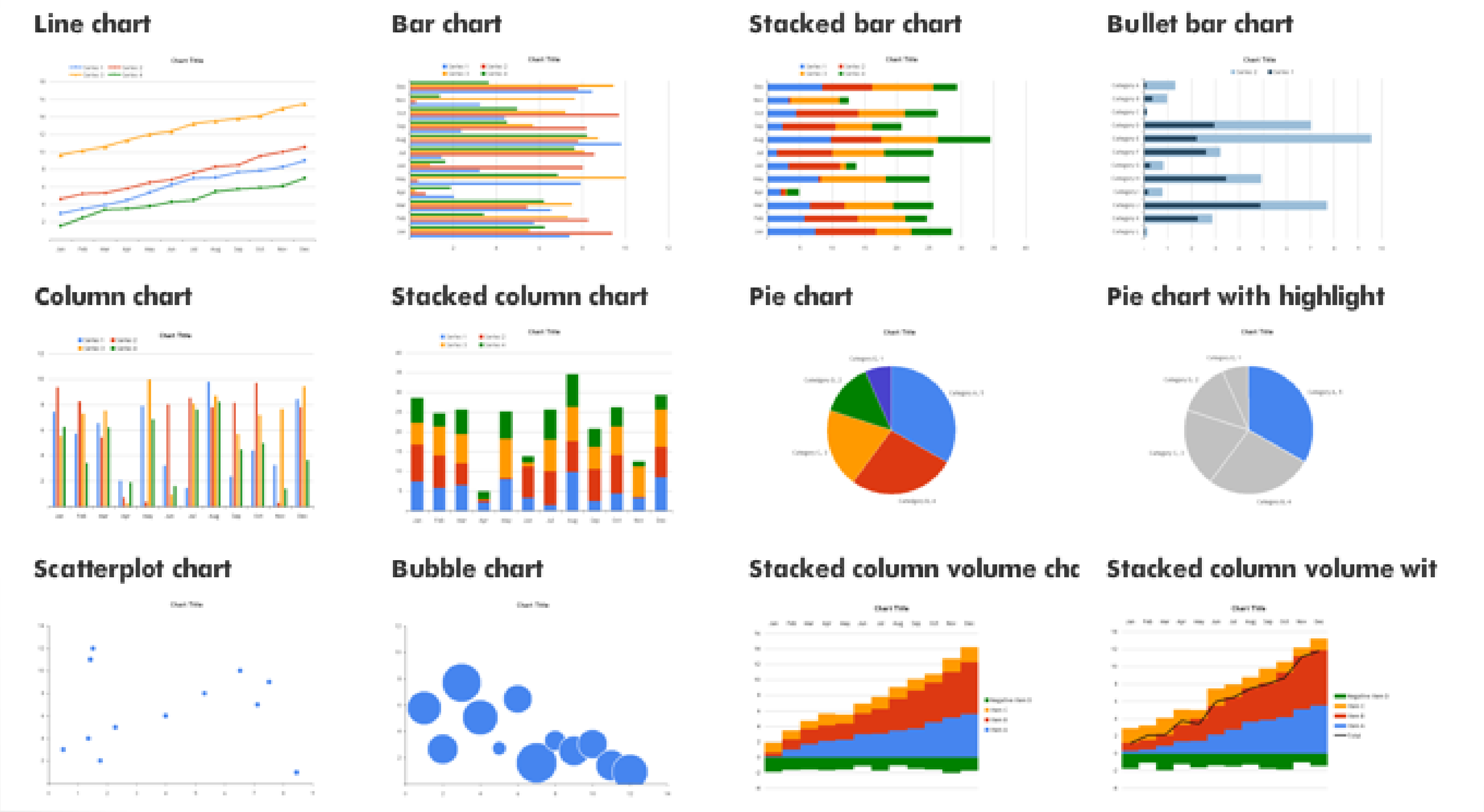
Setosa

Versicolour

Virginica



Sample Charts in MS Excel



Microsoft Tutorial – Creating a Graph from Start to Finish:
https://support.office.com/en-us/article/create-a-chart-from-start-to-finish-0baf399e-dd61-4e18-8a73-b3fd5d5680c2
Juicebox Chart Chooser: http://labs.juiceanalytics.com/chartchooser/

Outlines

- Data Exploration: Tasks and Techniques
  - Summary Statistics
  - Data Visualization
  - Online Analytical Processing (OLAP)
    - Pivot Table

On-Line Analytical Processing (OLAP)

- Relational databases put data into tables, while OLAP uses a multidimensional array representation.
  - Such representations of data previously existed in statistics and other fields

- There are a number of data analysis and data exploration operations that are easier with such a data representation.

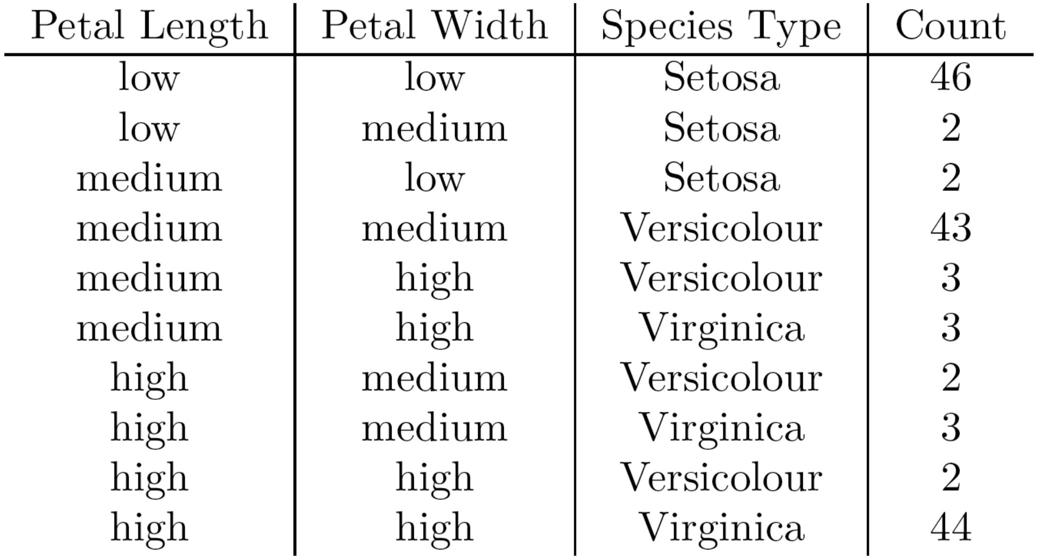
Creating a Multidimensional Array

- Two steps (Tabular data to a multidimensional array):
  - First, identify which attributes are to be the dimensions and which attribute is to be the target attribute whose values appear as entries in the multidimensional array.
    - The attributes used as **dimensions** must have **discrete values**
    - The **target value** is typically a **count** or **continuous value**, e.g., the cost of an item
    - Can have no target variable at all except the count of objects that have the same set of attribute values
  - Second, find the value of each entry in the multidimensional array by summing the values (of the target attribute) or count of all objects that have the attribute values corresponding to that entry.

Asst. Prof. Dr. Rachsuda Setthawong

Example: Iris data

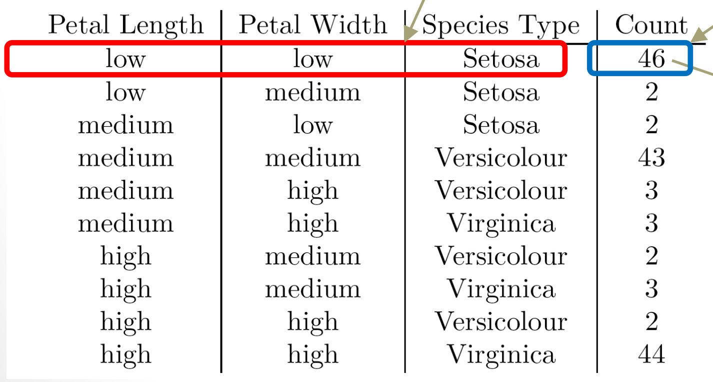
1. Define dimensions: A) petal width and B) petal length, and C) Species Type
    - **Discretize** A and B to have categorical values: *low*, *medium*, and *high*
2. Define target attribute: **count** attribute and count values of each case



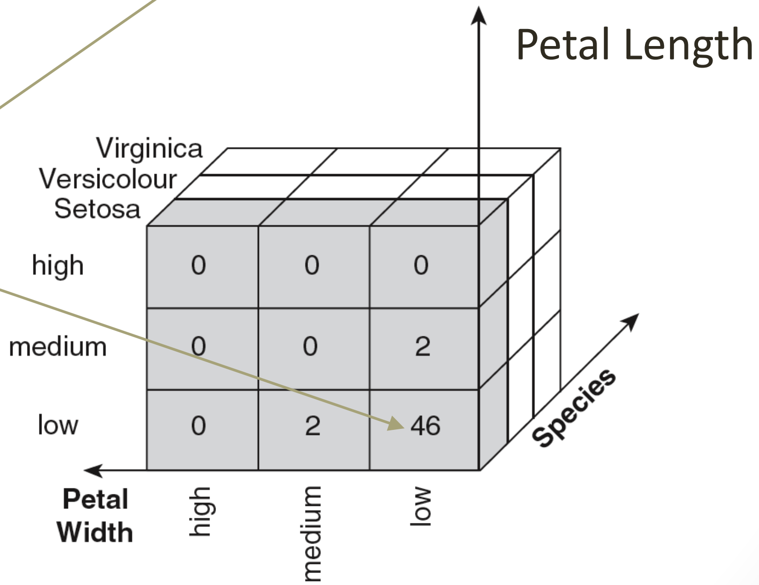
Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count value.
- All non-specified tuples are 0.

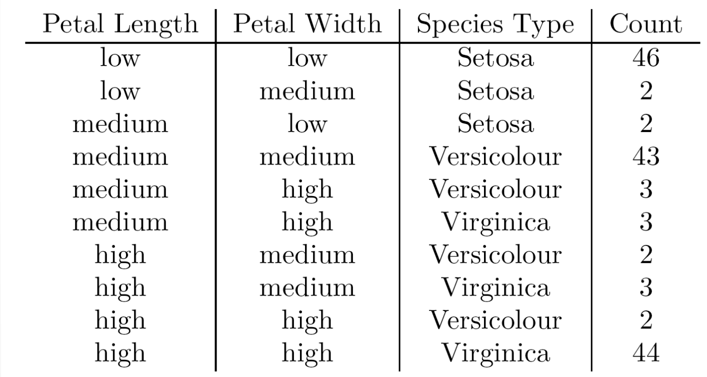


Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

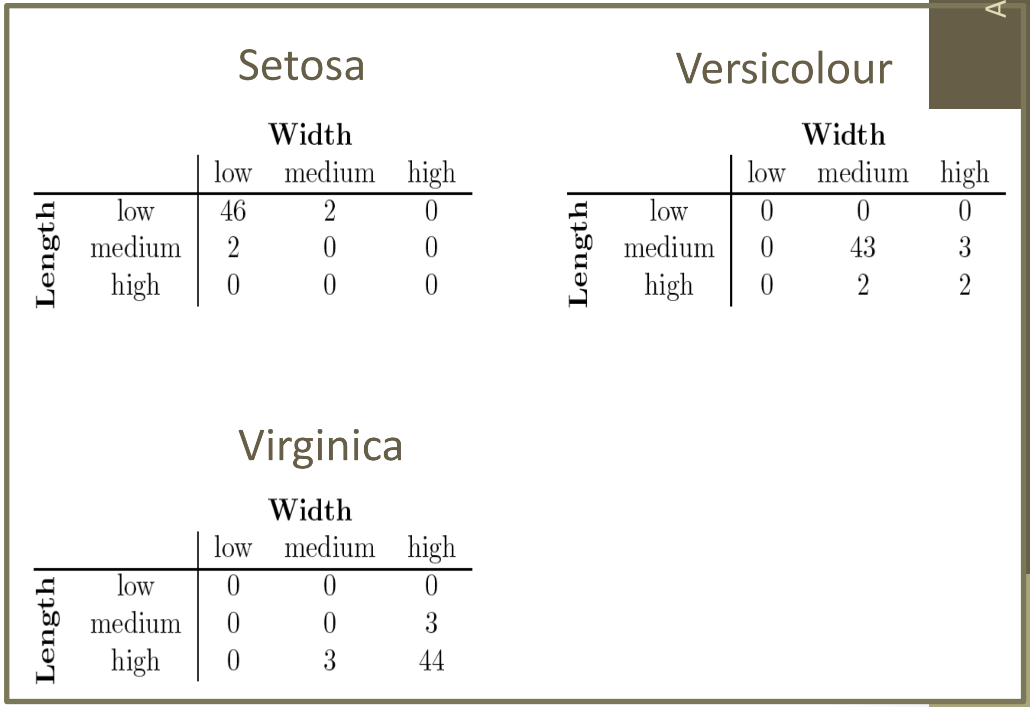


Example: Iris data (continued)

- Slices of the multidimensional array are shown by the following cross-tabulations



Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44



Setosa

		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

Versicolour

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

Virginica

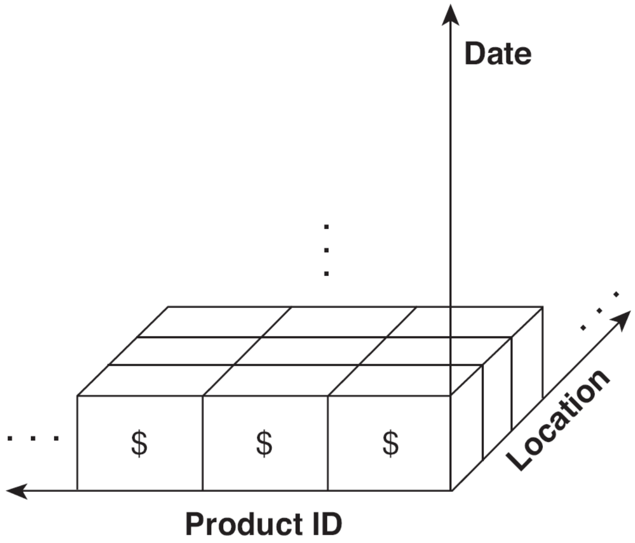
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a data cube
- A data cube is **a multidimensional representation of data**, together with all possible aggregates.
- By all possible aggregates, we mean the aggregates that result by **selecting a proper subset of the dimensions** and **summing over all remaining dimensions.**
- For example, if we choose the species type dimension of the Iris data and sum over all other dimensions, the result will be a one-dimensional entry with three entries, each of which gives the number of flowers of each type.

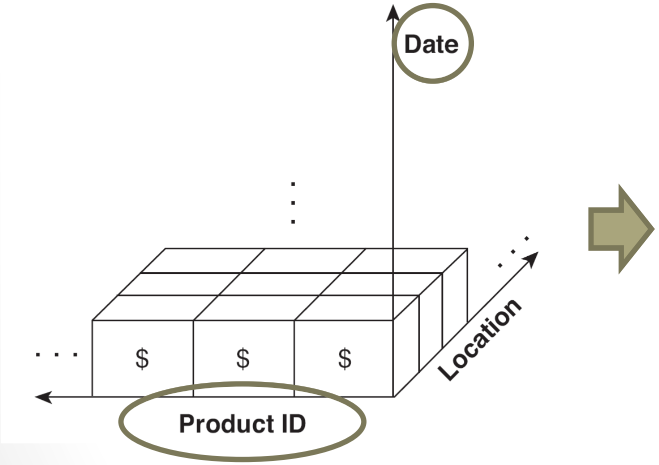
Data Cube Example

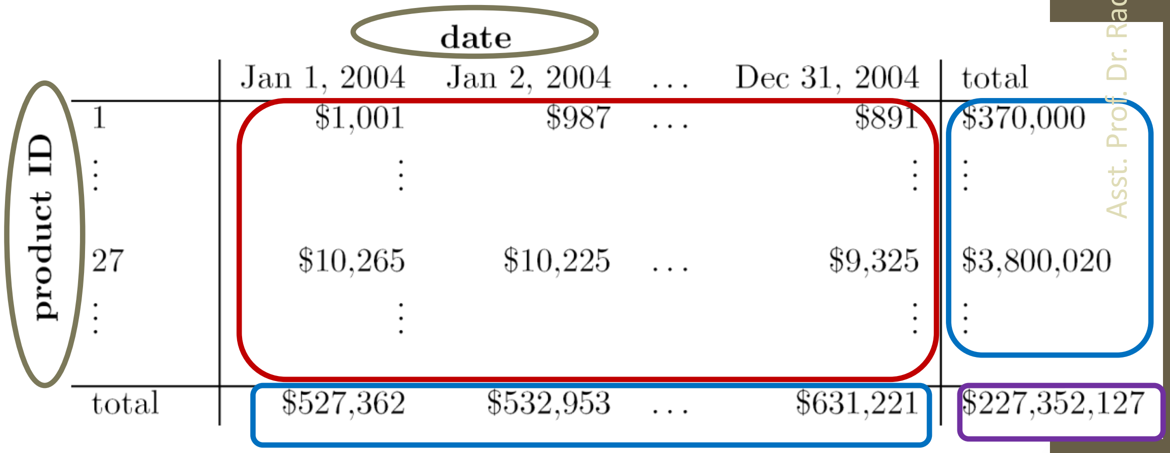
- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represented as a three-dimensional array
- There are 3 two-dimensional aggregates (3 choose 2), 3 one-dimensional aggregates, and 1 zero-dimensional aggregate (the overall total)



Data Cube Example (continued)

- The following figure table shows one of the two dimensional aggregates, along with two of the one-dimensional aggregates, and the overall total



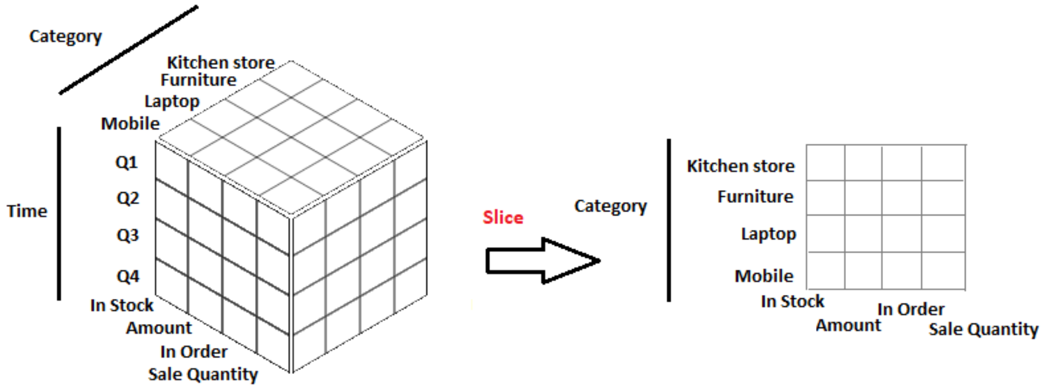


		date			
product ID	Jan 1, 2004	Jan 2, 2004	. . .	Dec 31, 2004	total
1	$1,001	$987	. . .	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	. . .	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	. . .	$631,221	$227,352,127

OLAP Operations:
Slicing and Dicing

- Slicing is **selecting a group of cells** from the entire multidimensional array **by specifying a specific value for one or more dimensions.**

E.g., Slice 1 quarter

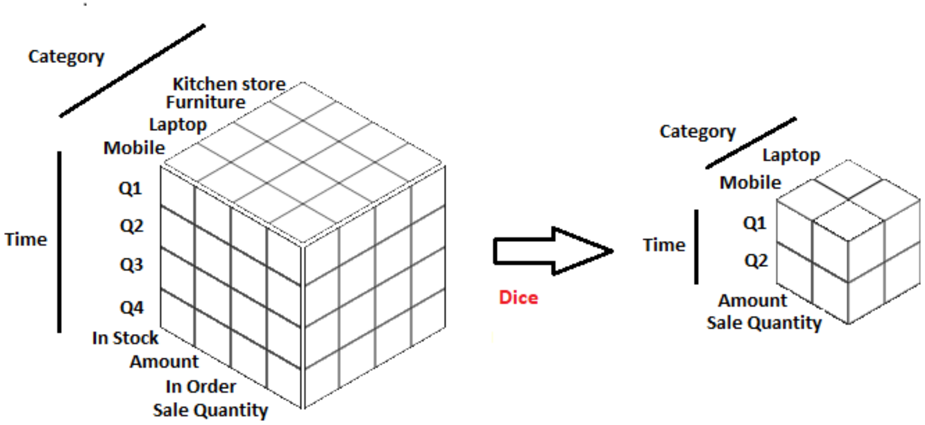


https://cracklogic.com/olap-tutorial/#Slice

OLAP Operations:

Slicing and Dicing

- Dicing involves **selecting a subset of cells** by specifying a **range of attribute values**.
  - This is equivalent to defining a subarray from the complete array.



- In practice, both operations can also be accompanied by aggregation over some dimensions.

https://cracklogic.com/olap-tutorial/#Slice

Asst. Prof. Dr. Rachsuda Setthawong

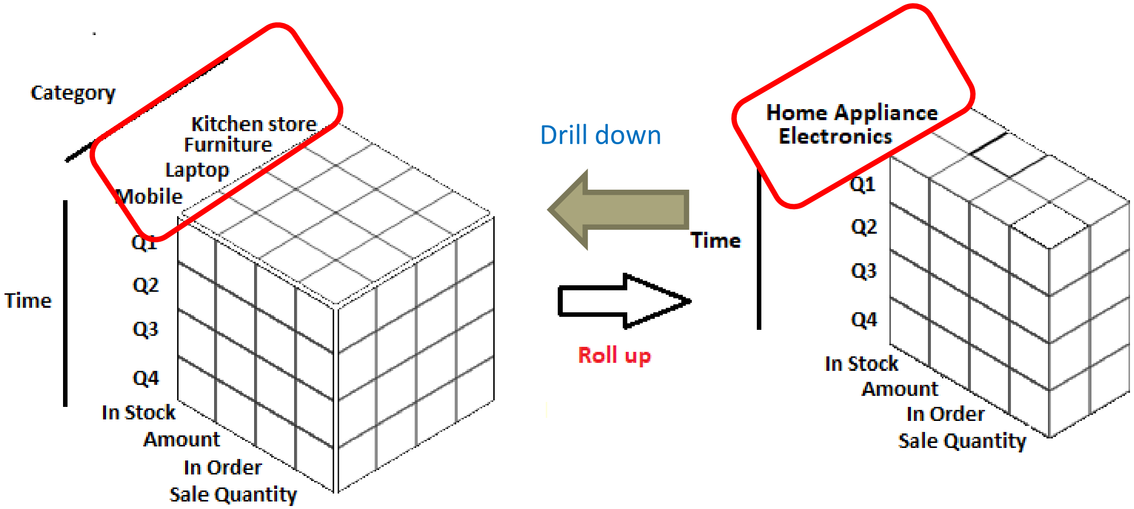
OLAP Operations: Roll-up and Drill-down

- **Attribute values** often have a **hierarchical structure**.
    - Each *date* is associated with a year, month, and week.
    - A *location* is associated with a continent, country, state (province, etc.), and city.
    - *Products* can be divided into various categories, such as clothing, electronics, and furniture.

- Note that these categories often nest and form a tree
    - A year contains months which contains day
    - A country contains a state which contains a city

OLAP Operations: Roll-up and Drill-down

- This **hierarchical structure** gives rise to the roll-up and drill-down operations.



https://cracklogic.com/olap-tutorial/#Slice

Asst. Prof. Dr. Rachsuda Setthawong

Outlines

- Data Exploration: Tasks and Techniques
  - Summary Statistics
  - Data Visualization
  - Online Analytical Processing (OLAP)
    - Pivot Table

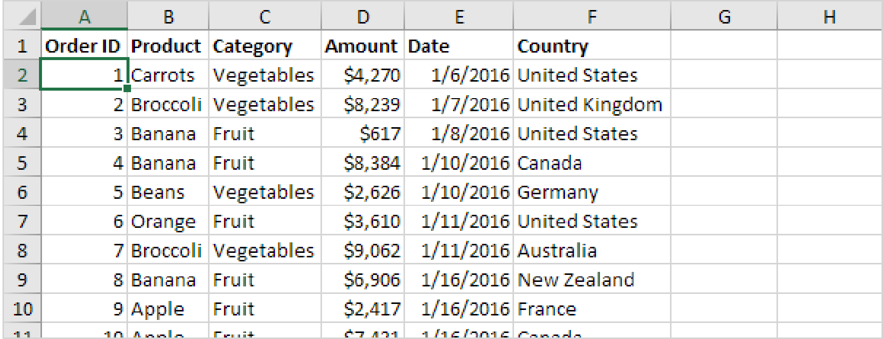
Pivot Table in MS Excel

https://exceljet.net/excel-pivot-tables

What is Pivot Table?

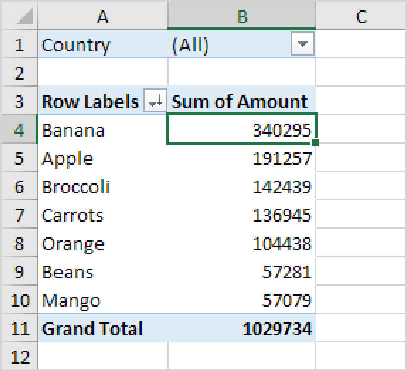
- A **table** of statistics that summarizes the data of a more extensive **table** (e.g., a database, spreadsheet).
- This summary include sums, averages, or other statistics, which the **pivot table** groups together in a meaningful way.

Raw Data

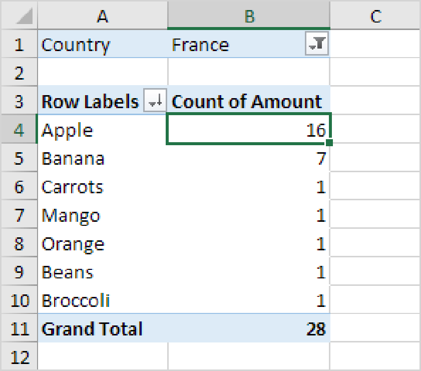


	A	B	C	D	E	F	G	H
1	Order ID	Product	Category	Amount	Date	Country		
2	1	Carrots	Vegetables	$4,270	1/6/2016	United States		
3	2	Broccoli	Vegetables	$8,239	1/7/2016	United Kingdom		
4	3	Banana	Fruit	$617	1/8/2016	United States		
5	4	Banana	Fruit	$8,384	1/10/2016	Canada		
6	5	Beans	Vegetables	$2,626	1/10/2016	Germany		
7	6	Orange	Fruit	$3,610	1/11/2016	United States		
8	7	Broccoli	Vegetables	$9,062	1/11/2016	Australia		
9	8	Banana	Fruit	$6,906	1/16/2016	New Zealand		
10	9	Apple	Fruit	$2,417	1/16/2016	France		
11	10	Apple	Fruit	$7,431	1/16/2016	Canada		

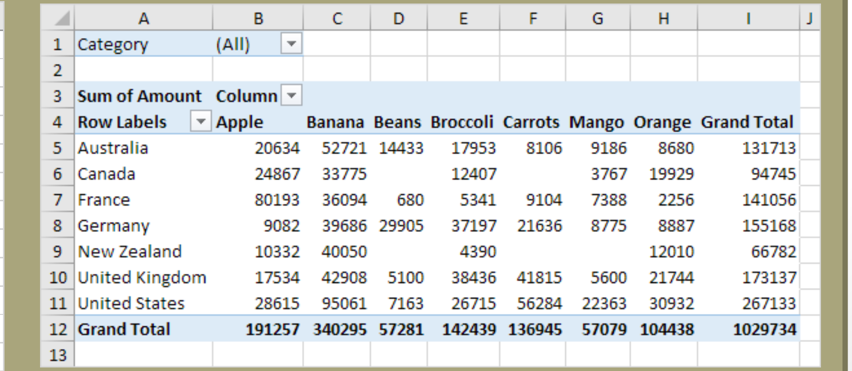
Pivot Tables



	A	B	C
1	Country	(All)	
2			
3	Row Labels	Sum of Amount	
4	Banana	340295	
5	Apple	191257	
6	Broccoli	142439	
7	Carrots	136945	
8	Orange	104438	
9	Beans	57281	
10	Mango	57079	
11	Grand Total	1029734	
12			



	A	B	C
1	Country	France	
2			
3	Row Labels	Count of Amount	
4	Apple	16	
5	Banana	7	
6	Carrots	1	
7	Mango	1	
8	Orange	1	
9	Beans	1	
10	Broccoli	1	
11	Grand Total	28	
12			

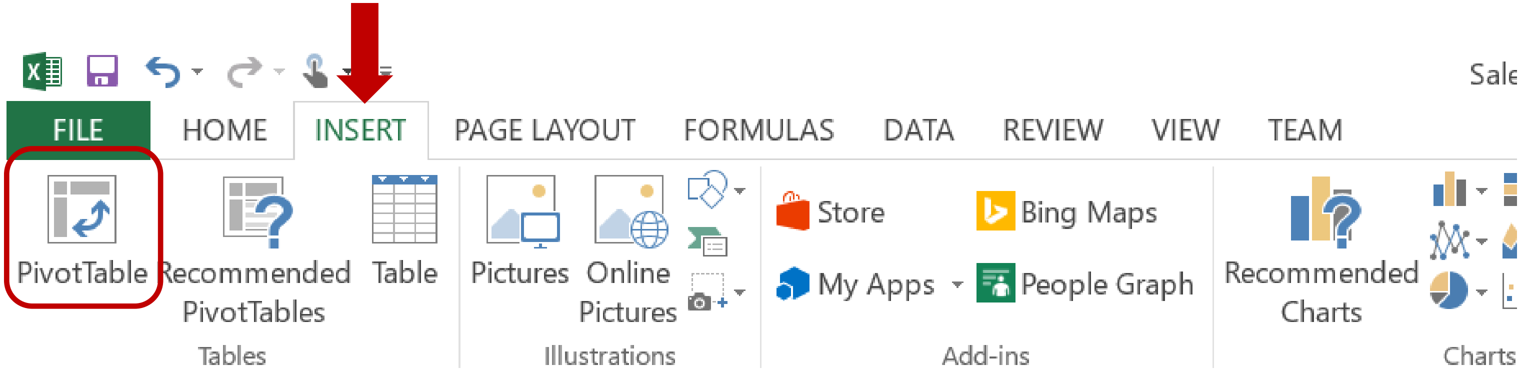


	A	B	C	D	E	F	G	H	I	J
1	Category	(All)								
2										
3	Sum of Amount	Column								
4	Row Labels	Apple	Banana	Beans	Broccoli	Carrots	Mango	Orange	Grand Total	
5	Australia	20634	52721	14433	17953	8106	9186	8680	131713	
6	Canada	24867	33775		12407		3767	19929	94745	
7	France	80193	36094	680	5341	9104	7388	2256	141056	
8	Germany	9082	39686	29905	37197	21636	8775	8887	155168	
9	New Zealand	10332	40050		4390			12010	66782	
10	United Kingdom	17534	42908	5100	38436	41815	5600	21744	173137	
11	United States	28615	95061	7163	26715	56284	22363	30932	267133	
12	Grand Total	191257	340295	57281	142439	136945	57079	104438	1029734	
13										

https://www.excel-easy.com/data-analysis/pivot-tables.html

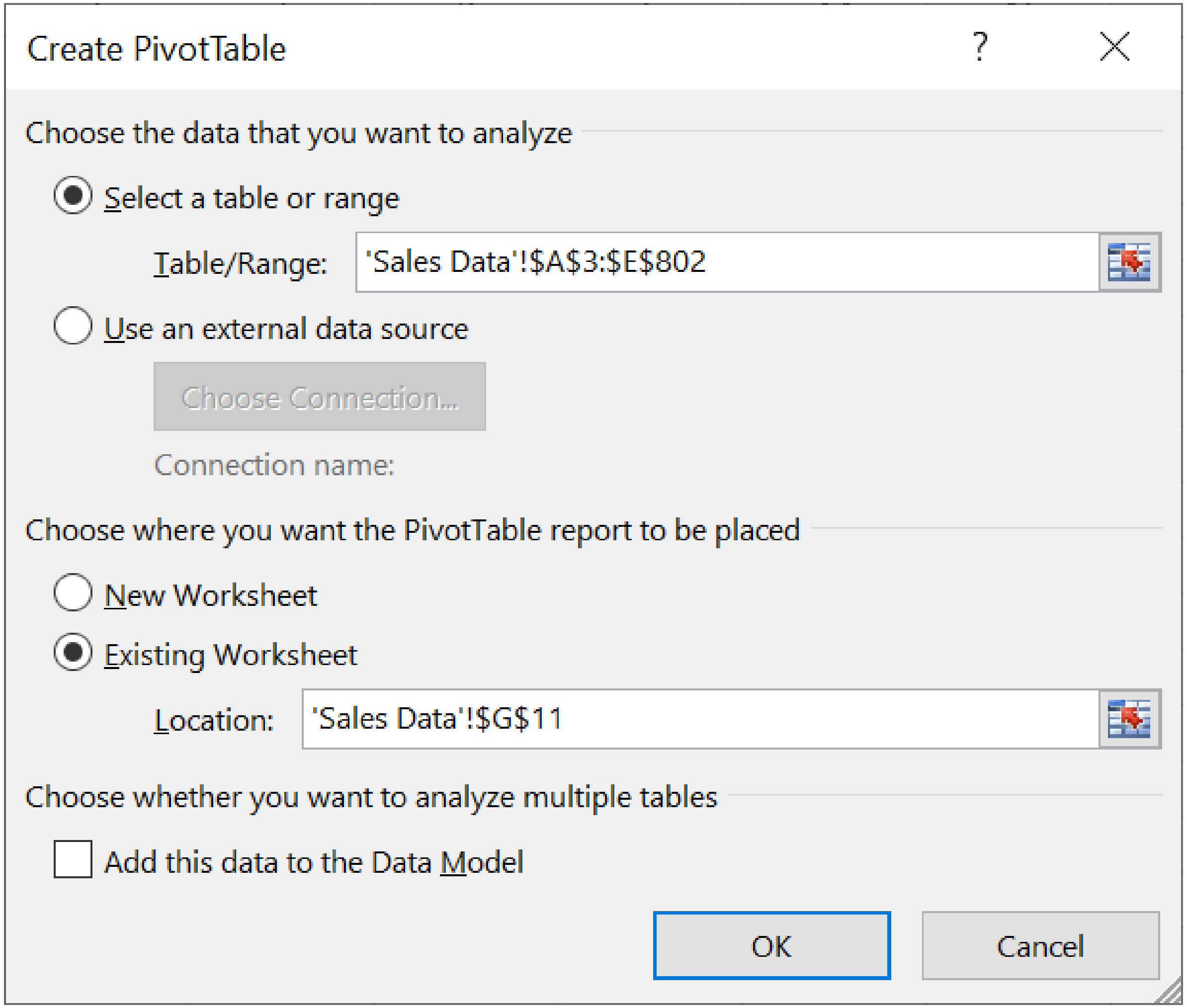
Creating Pivot Table: 1/6

1. Open the file **SalesData.xlsx**
2. Add PivotTable



Creating Pivot Table: 2/6

Selecting a Table or Range of Data



Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: 'Sales Data'!$A$3:$E$802

○ Use an external data source

Connection name:

Choose where you want the PivotTable report to be placed

○ New Worksheet

● Existing Worksheet

Location: 'Sales Data'!$G$11

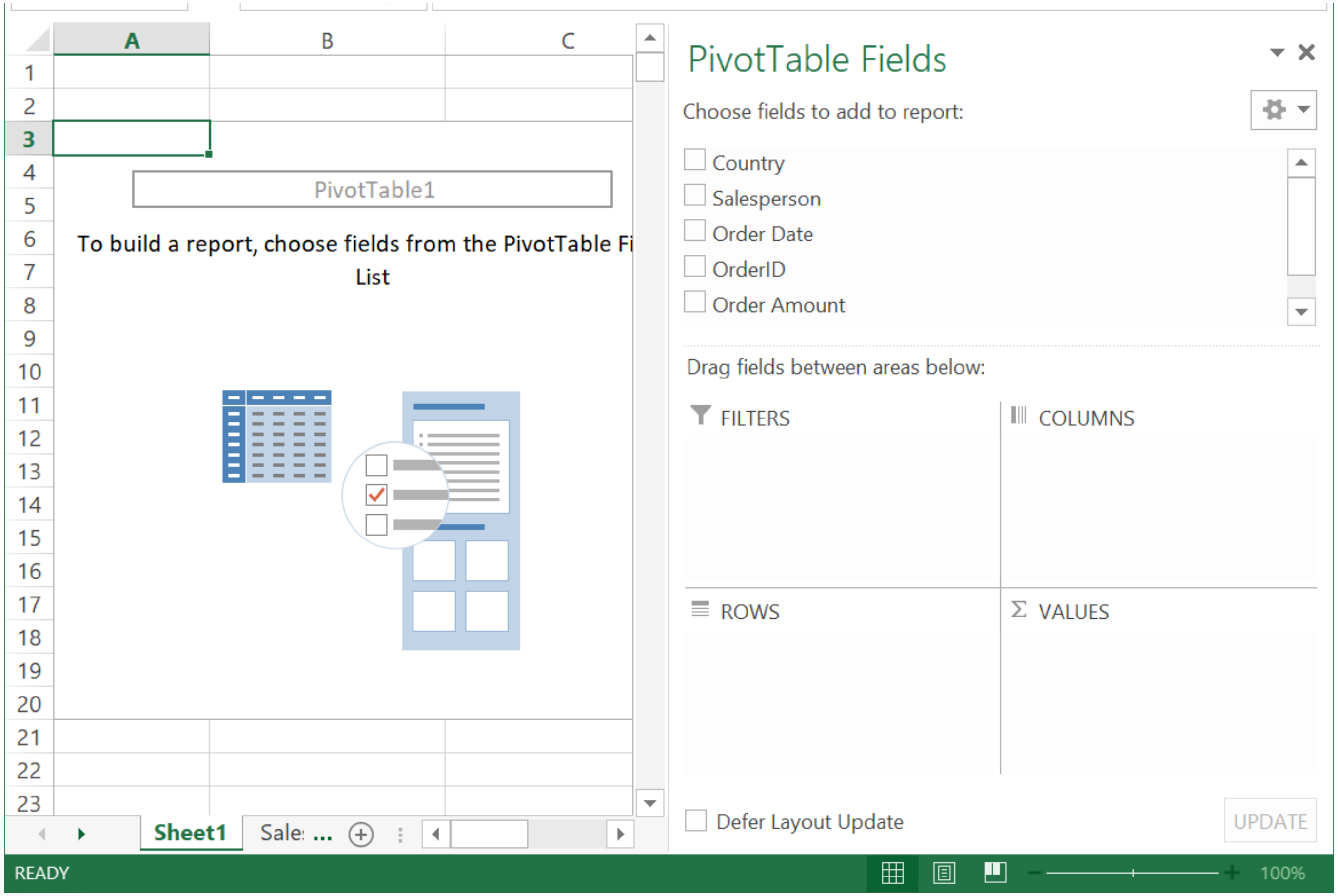
Choose whether you want to analyze multiple tables

□ Add this data to the Data Model

Asst. Prof. Dr. Rachsuda Setthawong

Creating Pivot Table: 3/6

The Worksheet with Pivot Table Added



PivotTable1

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

PivotTable Fields

Choose fields to add to report:

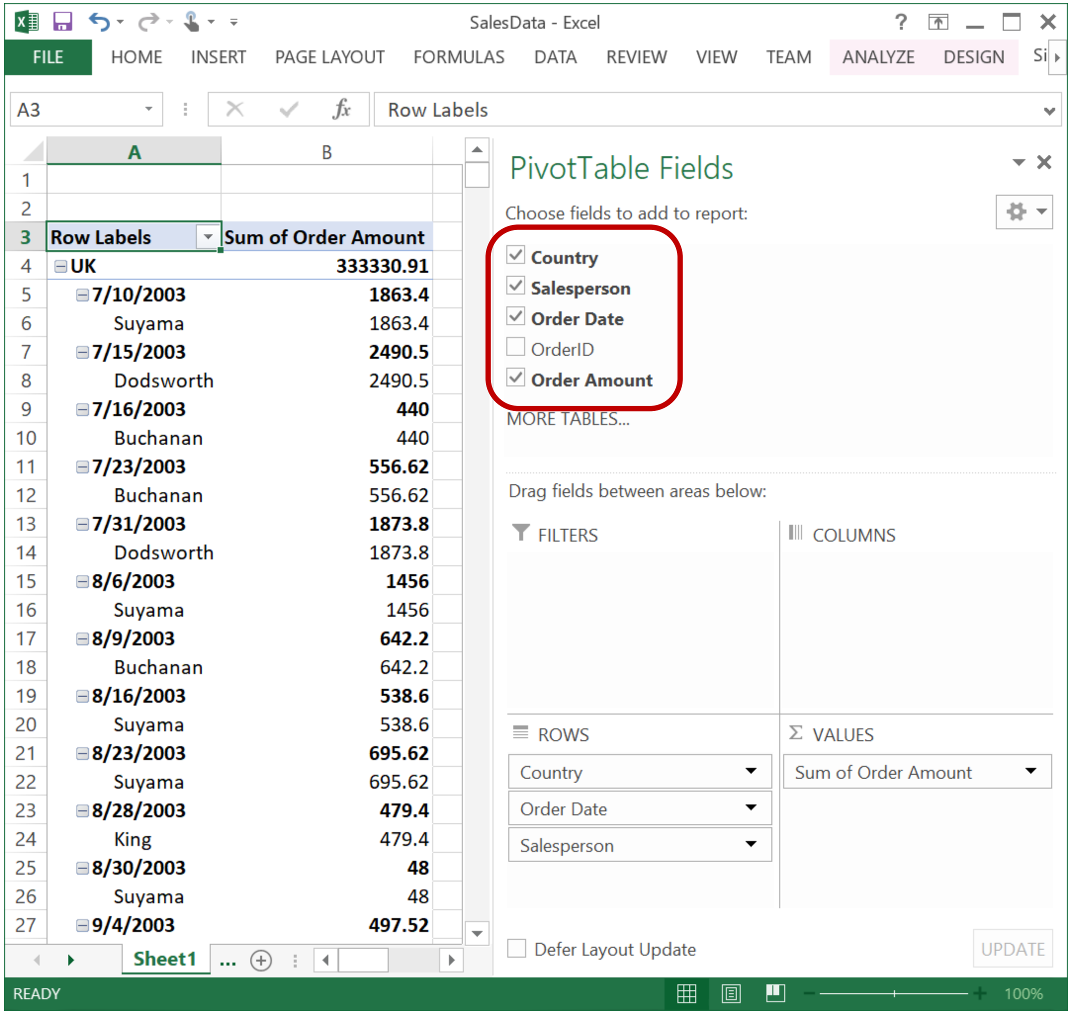
□ Country
□ Salesperson
□ Order Date
□ OrderID
□ Order Amount

Drag fields between areas below:

□ Defer Layout Update

Creating Pivot Table: 4/6

Selecting Fields to be Displayed in Pivot Table



Row Labels	Sum of Order Amount
⊟UK	333330.91
⊟7/10/2003	1863.4
Suyama	1863.4
⊟7/15/2003	2490.5
Dodsworth	2490.5
⊟7/16/2003	440
Buchanan	440
⊟7/23/2003	556.62
Buchanan	556.62
⊟7/31/2003	1873.8
Dodsworth	1873.8
⊟8/6/2003	1456
Suyama	1456
⊟8/9/2003	642.2
Buchanan	642.2
⊟8/16/2003	538.6
Suyama	538.6
⊟8/23/2003	695.62
Suyama	695.62
⊟8/28/2003	479.4
King	479.4
⊟8/30/2003	48
Suyama	48
⊟9/4/2003	497.52

PivotTable Fields

Choose fields to add to report:

- ☑ Country
- ☑ Salesperson
- ☑ Order Date
- ☐ OrderID
- ☑ Order Amount

MORE TABLES...

Drag fields between areas below:

FILTERS

COLUMNS

ROWS
- Country
- Order Date
- Salesperson

VALUES
- Sum of Order Amount

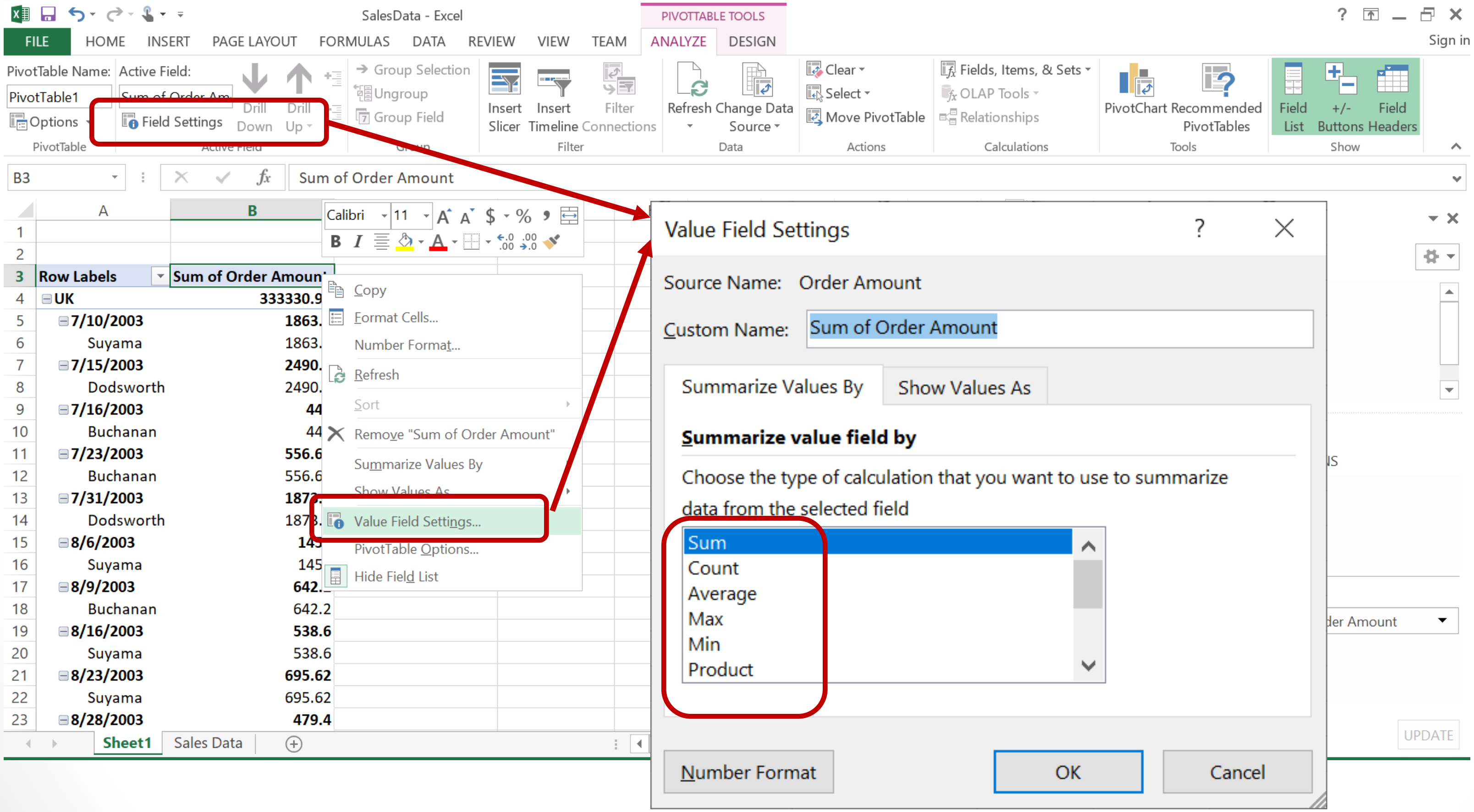
☐ Defer Layout Update

UPDATE

Asst. Prof. Dr. Rachsuda Setthawong

Creating Pivot Table: 5/6

Changing Summary Info.



	A	B
1		
2		
3	Row Labels	Sum of Order Amoun'
4	UK	333330.9
5	7/10/2003	1863.
6	Suyama	1863.
7	7/15/2003	2490.
8	Dodsworth	2490.
9	7/16/2003	44
10	Buchanan	44
11	7/23/2003	556.6
12	Buchanan	556.6
13	7/31/2003	1873
14	Dodsworth	1873.
15	8/6/2003	145
16	Suyama	145
17	8/9/2003	642.
18	Buchanan	642.2
19	8/16/2003	538.6
20	Suyama	538.6
21	8/23/2003	695.62
22	Suyama	695.62
23	8/28/2003	479.4

Value Field Settings

Source Name: Order Amount

Custom Name: Sum of Order Amount

Summarize Values By | Show Values As

Summarize value field by

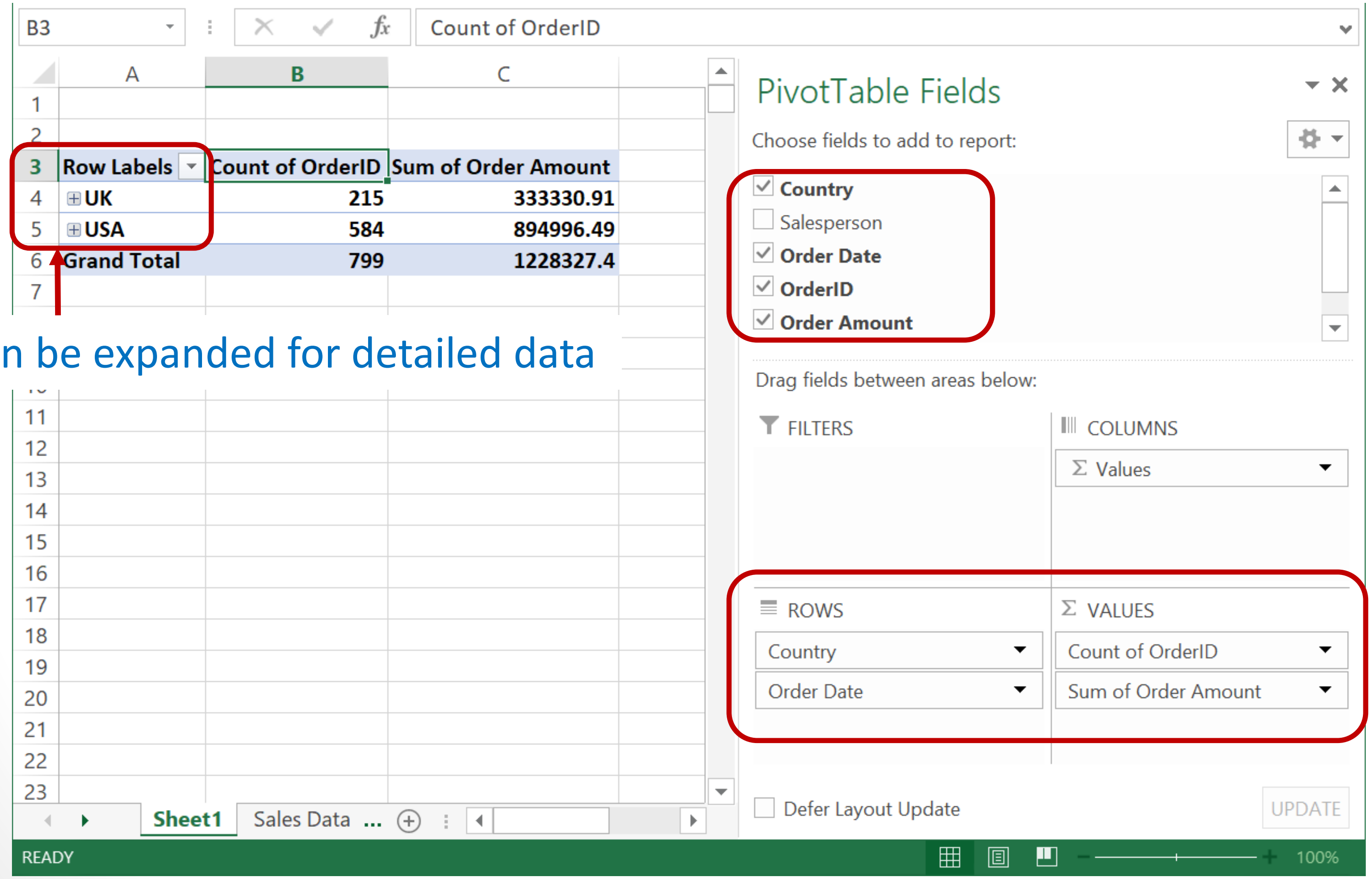
Choose the type of calculation that you want to use to summarize data from the selected field

- Sum
- Count
- Average
- Max
- Min
- Product

Number Format | OK | Cancel

Creating Pivot Table: 6/6

Rearranging Fields in Rows and Changing Summary Info.



B3 | fx | Count of OrderID

Row Labels	Count of OrderID	Sum of Order Amount
⊞UK	215	333330.91
⊞USA	584	894996.49
Grand Total	799	1228327.4

Can be expanded for detailed data

PivotTable Fields

Choose fields to add to report:

- ☑ Country
- □ Salesperson
- ☑ Order Date
- ☑ OrderID
- ☑ Order Amount

Drag fields between areas below:

FILTERS

COLUMNS: Values

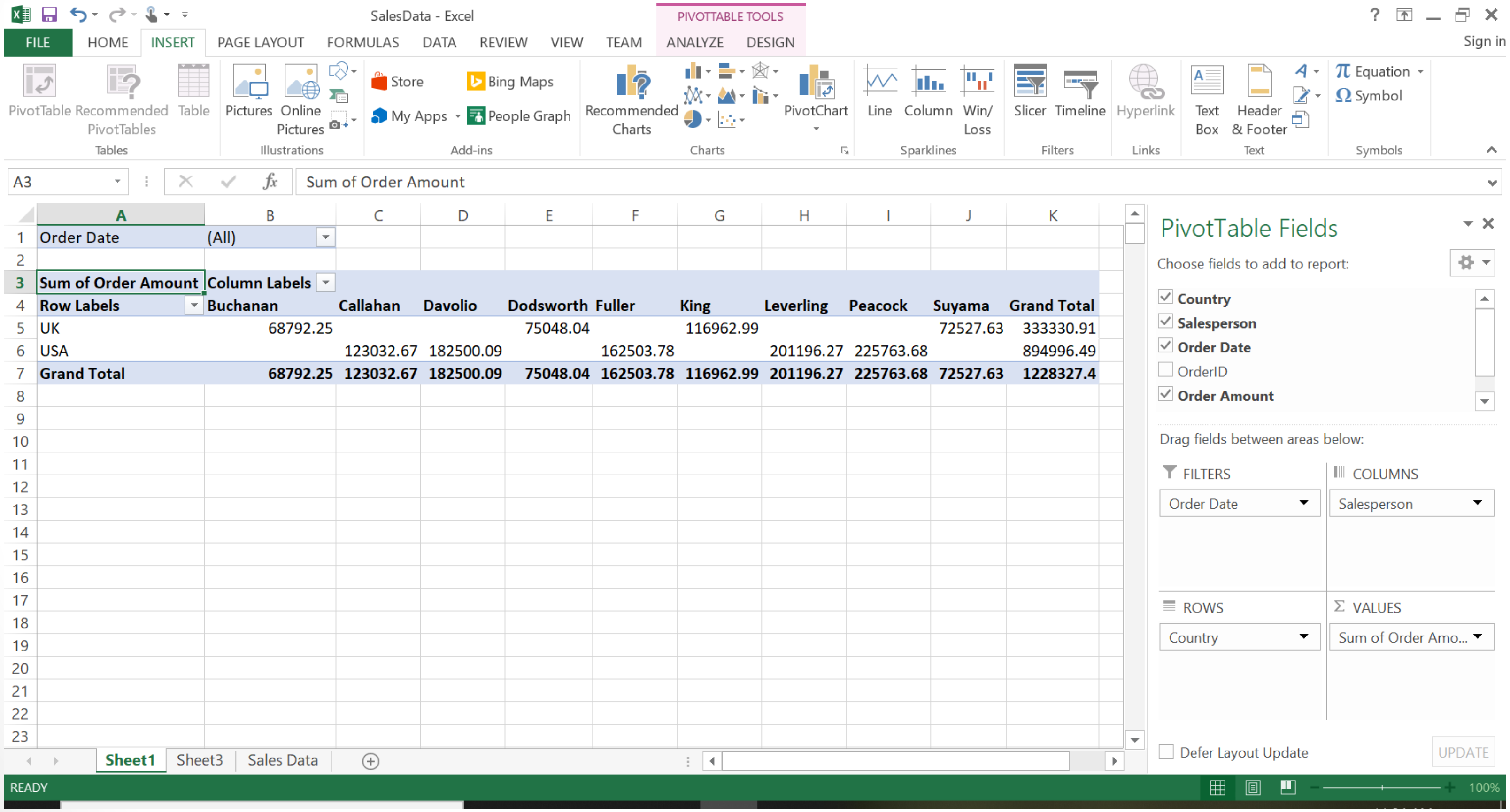
ROWS: Country, Order Date

VALUES: Count of OrderID, Sum of Order Amount

□ Defer Layout Update

Asst. Prof. Dr. Rachsuda Setthawong

Example of Summary 1:
Order Amount per Salesperson in Different Countries



A3 · fx · Sum of Order Amount

Order Date	(All)									
Sum of Order Amount	**Column Labels**									
Row Labels	**Buchanan**	**Callahan**	**Davolio**	**Dodsworth**	**Fuller**	**King**	**Leverling**	**Peacock**	**Suyama**	**Grand Total**
UK	68792.25			75048.04		116962.99			72527.63	333330.91
USA		123032.67	182500.09		162503.78		201196.27	225763.68		894996.49
Grand Total	**68792.25**	**123032.67**	**182500.09**	**75048.04**	**162503.78**	**116962.99**	**201196.27**	**225763.68**	**72527.63**	**1228327.4**

PivotTable Fields

Choose fields to add to report:

☑ **Country**
☑ **Salesperson**
☑ **Order Date**
□ OrderID
☑ **Order Amount**

Drag fields between areas below:

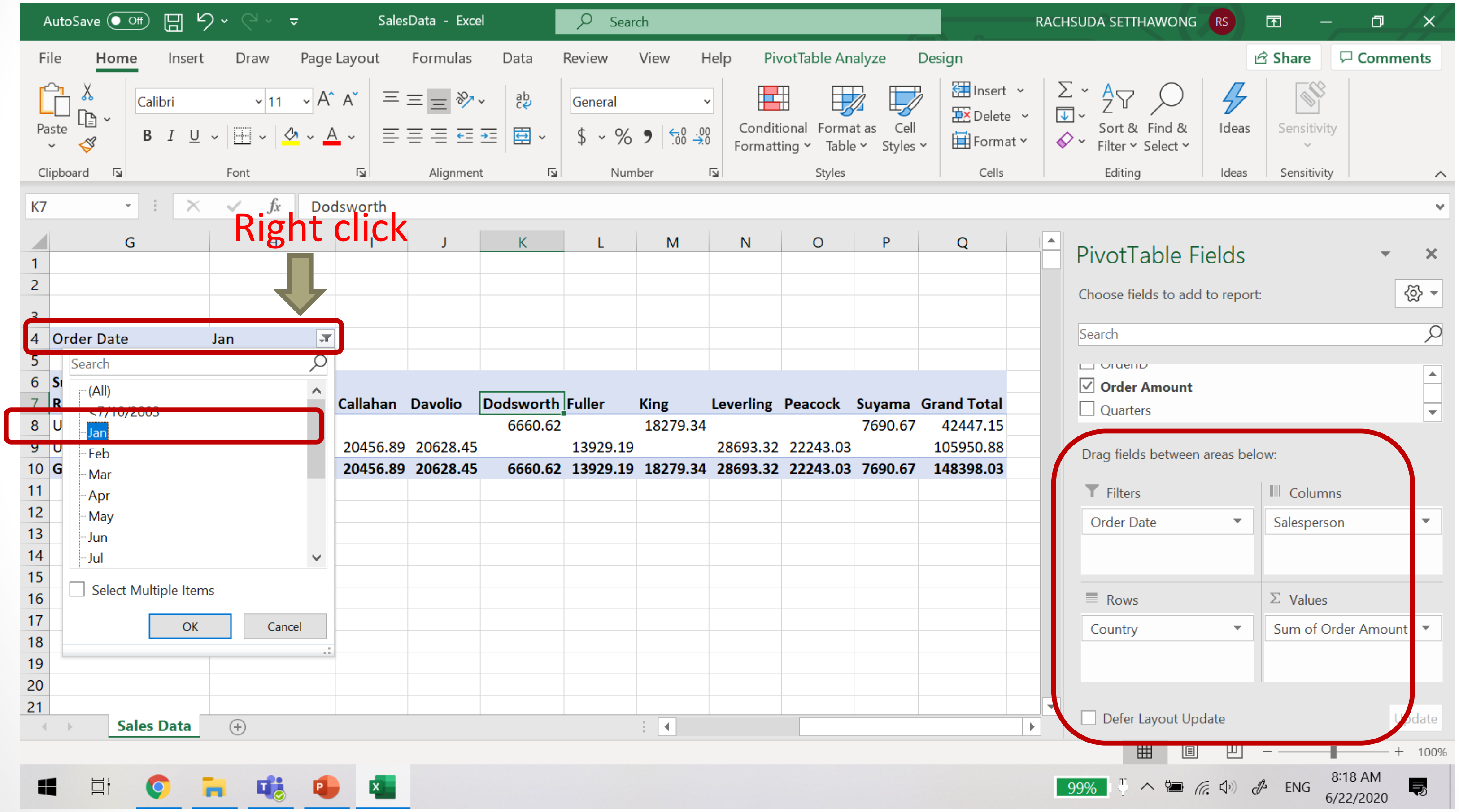
FILTERS	COLUMNS
Order Date	Salesperson

ROWS	VALUES
Country	Sum of Order Amo...

□ Defer Layout Update · UPDATE

Example of Summary 2 (using Filter):

Order Amount in 2003 per Salesperson in Different Countries



K7 | Dodsworth

	Callahan	Davolio	Dodsworth	Fuller	King	Leverling	Peacock	Suyama	Grand Total
U			6660.62		18279.34			7690.67	42447.15
U	20456.89	20628.45		13929.19		28693.32	22243.03		105950.88
G	20456.89	20628.45	6660.62	13929.19	18279.34	28693.32	22243.03	7690.67	148398.03

Order Date: Jan

- (All)
- <7/10/2003
- Jan
- Feb
- Mar
- Apr
- May
- Jun
- Jul

□ Select Multiple Items

OK | Cancel

PivotTable Fields

Choose fields to add to report:

- ☑ Order Amount
- □ Quarters

Drag fields between areas below:

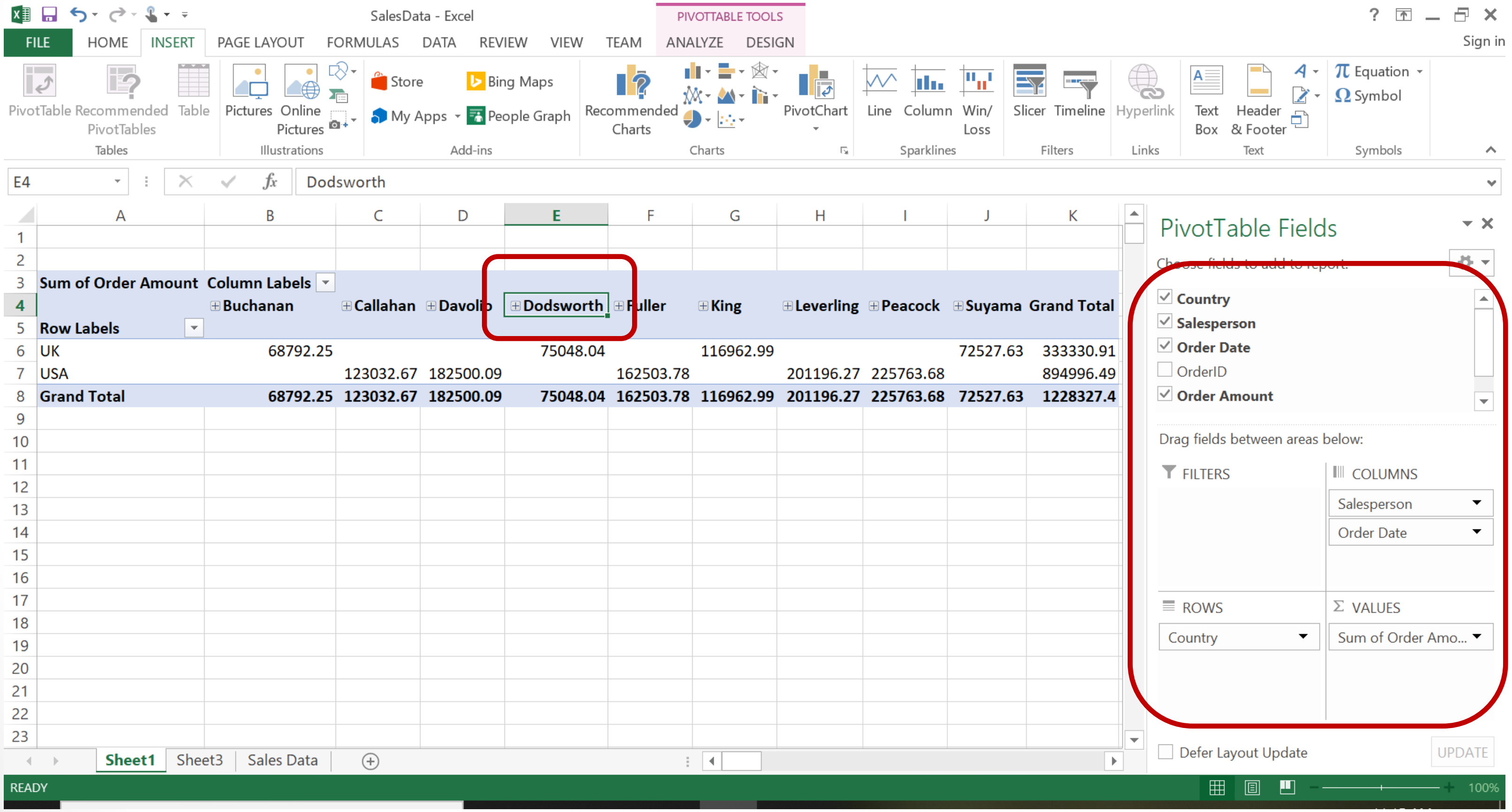
Filters	Columns
Order Date	Salesperson

Rows	Values
Country	Sum of Order Amount

□ Defer Layout Update

Example of Summary 3 (Aggregation using Group):

Grouping Order Amount by Year per Salesperson in Different Countries



E4 | Dodsworth

Sum of Order Amount	Column Labels									
	⊞Buchanan	⊞Callahan	⊞Davolio	⊞Dodsworth	⊞Fuller	⊞King	⊞Leverling	⊞Peacock	⊞Suyama	Grand Total
Row Labels										
UK	68792.25			75048.04		116962.99			72527.63	333330.91
USA		123032.67	182500.09		162503.78		201196.27	225763.68		894996.49
Grand Total	68792.25	123032.67	182500.09	75048.04	162503.78	116962.99	201196.27	225763.68	72527.63	1228327.4

PivotTable Fields

Choose fields to add to report:

- ☑ Country
- ☑ Salesperson
- ☑ Order Date
- ☐ OrderID
- ☑ Order Amount

Drag fields between areas below:

FILTERS

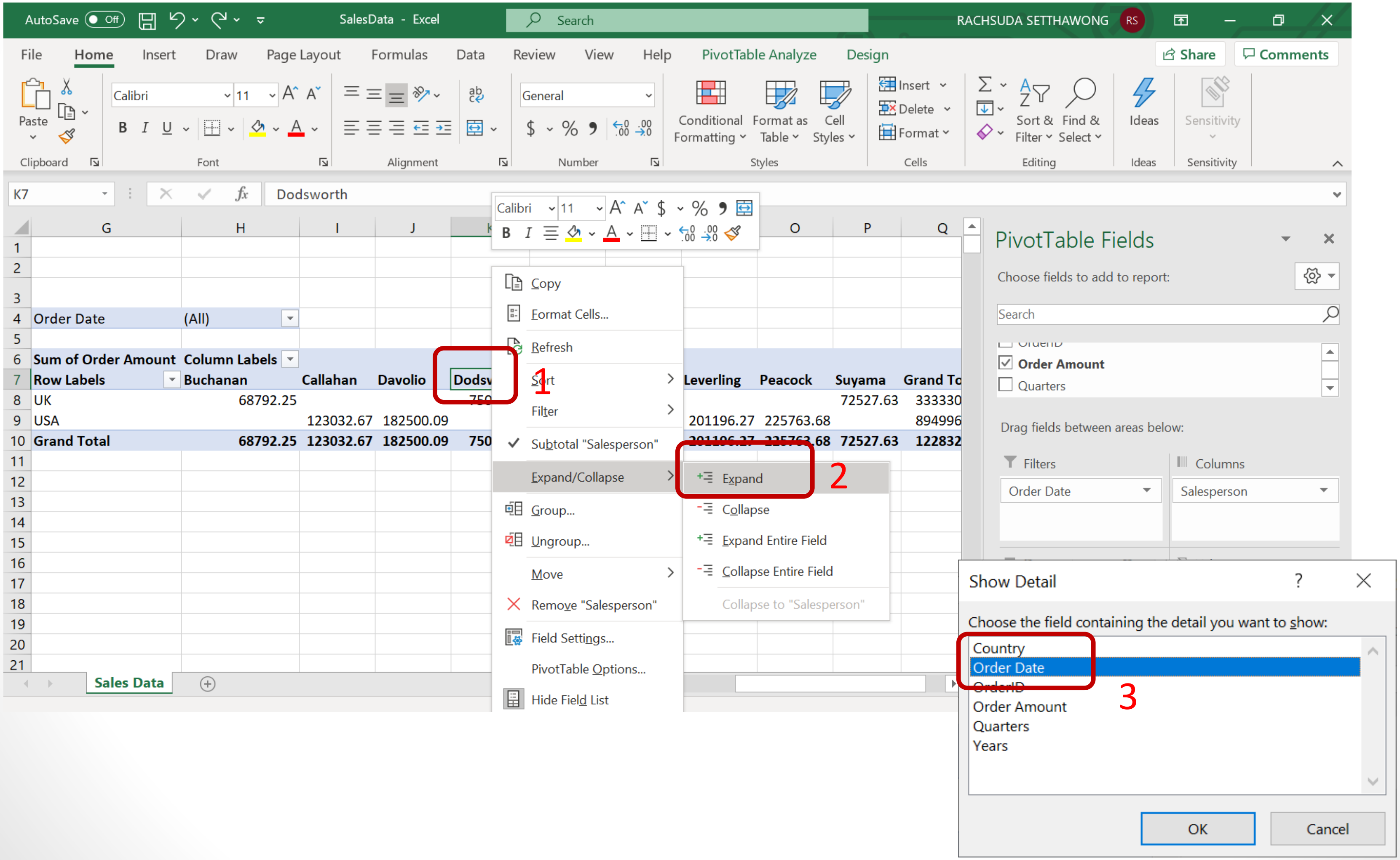
COLUMNS: Salesperson, Order Date

ROWS: Country

VALUES: Sum of Order Amo...

☐ Defer Layout Update UPDATE

Example of Summary 3 (Aggregation using Group): 1/3



	G	H	I	J	K	O	P	Q
4	Order Date	(All)						
6	Sum of Order Amount	Column Labels						
7	Row Labels	Buchanan	Callahan	Davolio	Dodsw	Leverling	Peacock	Suyama	Grand To
8	UK	68792.25			750			72527.63	333330
9	USA		123032.67	182500.09		201196.27	225763.68		894996
10	Grand Total	68792.25	123032.67	182500.09	750	201196.27	225763.68	72527.63	122832

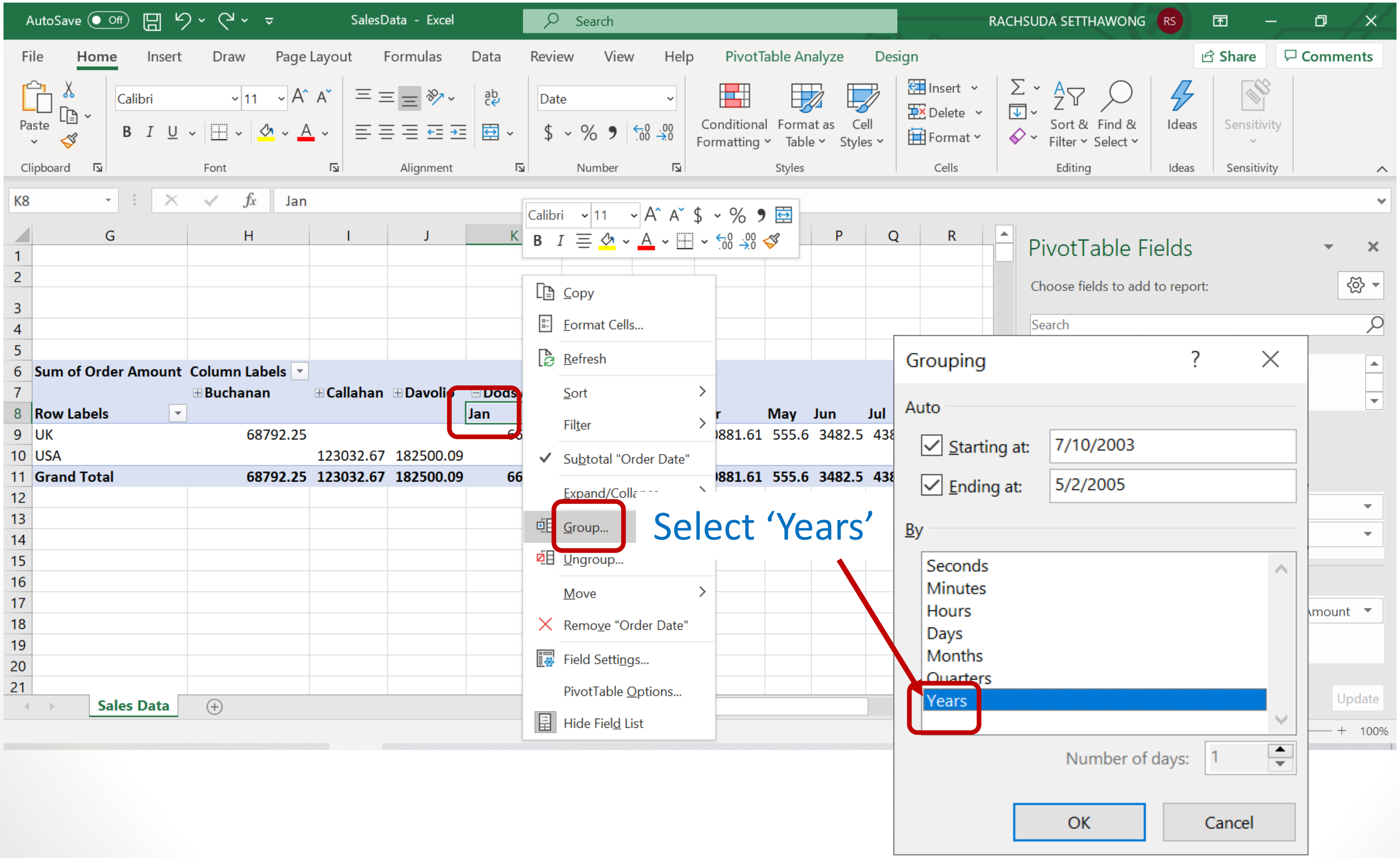
Show Detail

Choose the field containing the detail you want to show:

- Country
- Order Date
- OrderID
- Order Amount
- Quarters
- Years

OK Cancel

Example of Summary 3 (Aggregation using Group): 2/3



	G	H	I	J	K	…			
6	Sum of Order Amount	Column Labels							
7		⊞Buchanan	⊞Callahan	⊞Davolio	⊟Dods				
8	Row Labels				Jan	r	May	Jun	Jul
9	UK	68792.25			6	881.61	555.6	3482.5	438
10	USA		123032.67	182500.09					
11	Grand Total	68792.25	123032.67	182500.09	66	881.61	555.6	3482.5	438

Grouping

Auto

☑ Starting at: 7/10/2003

☑ Ending at: 5/2/2005

By

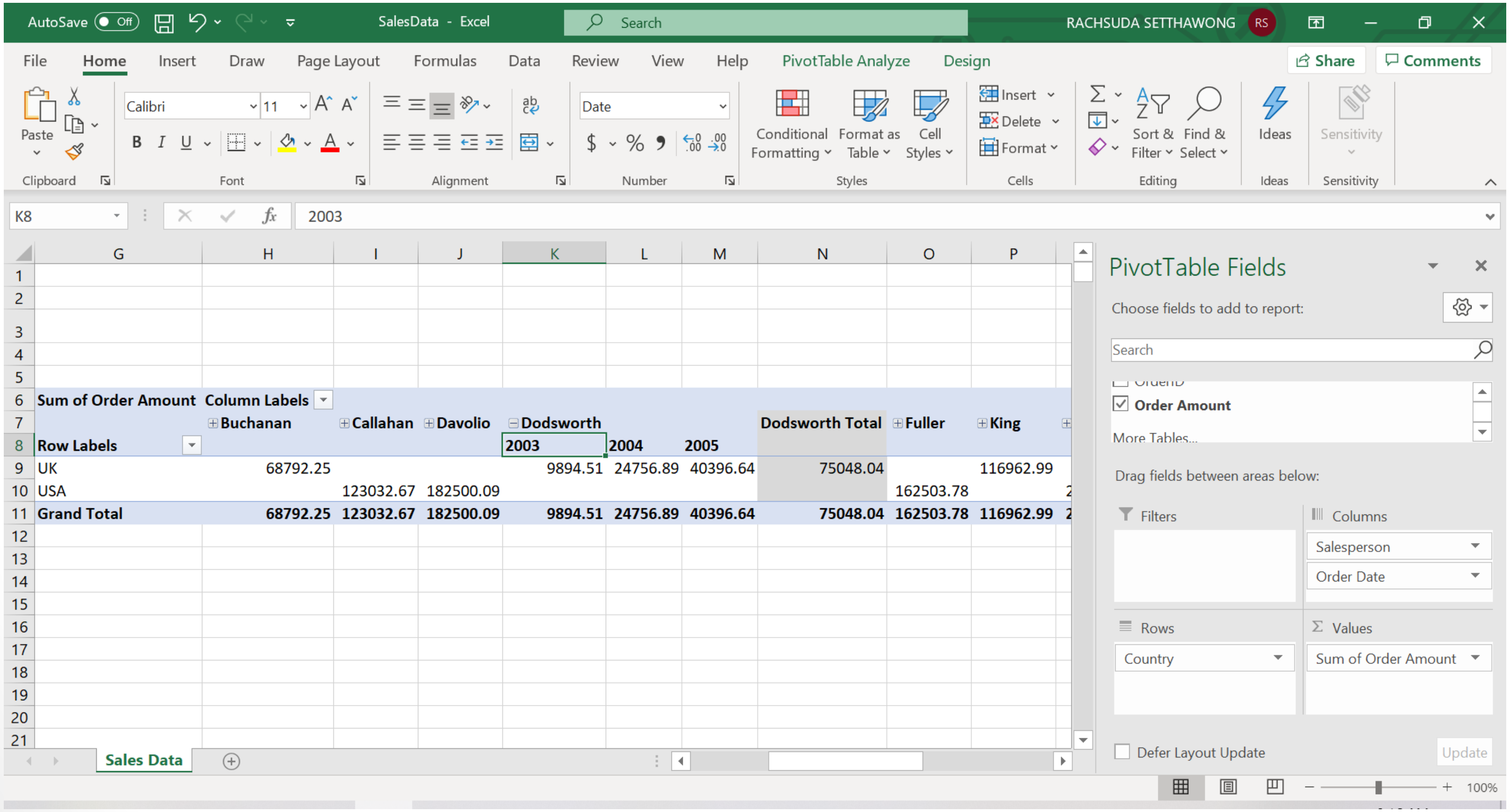
- Seconds
- Minutes
- Hours
- Days
- Months
- Quarters
- Years

Number of days: 1

OK Cancel

Select 'Years'

Example of Summary 3 (Aggregation using Group): 3/3



Sum of Order Amount	Column Labels									
	⊞Buchanan	⊞Callahan	⊞Davolio	⊟Dodsworth			Dodsworth Total	⊞Fuller	⊞King	⊞
Row Labels				2003	2004	2005				
UK	68792.25			9894.51	24756.89	40396.64	75048.04		116962.99	
USA		123032.67	182500.09					162503.78		2
Grand Total	68792.25	123032.67	182500.09	9894.51	24756.89	40396.64	75048.04	162503.78	116962.99	2

PivotTable Fields

Choose fields to add to report:

Search

OrderID

☑ Order Amount

More Tables...

Drag fields between areas below:

Filters

Columns: Salesperson, Order Date

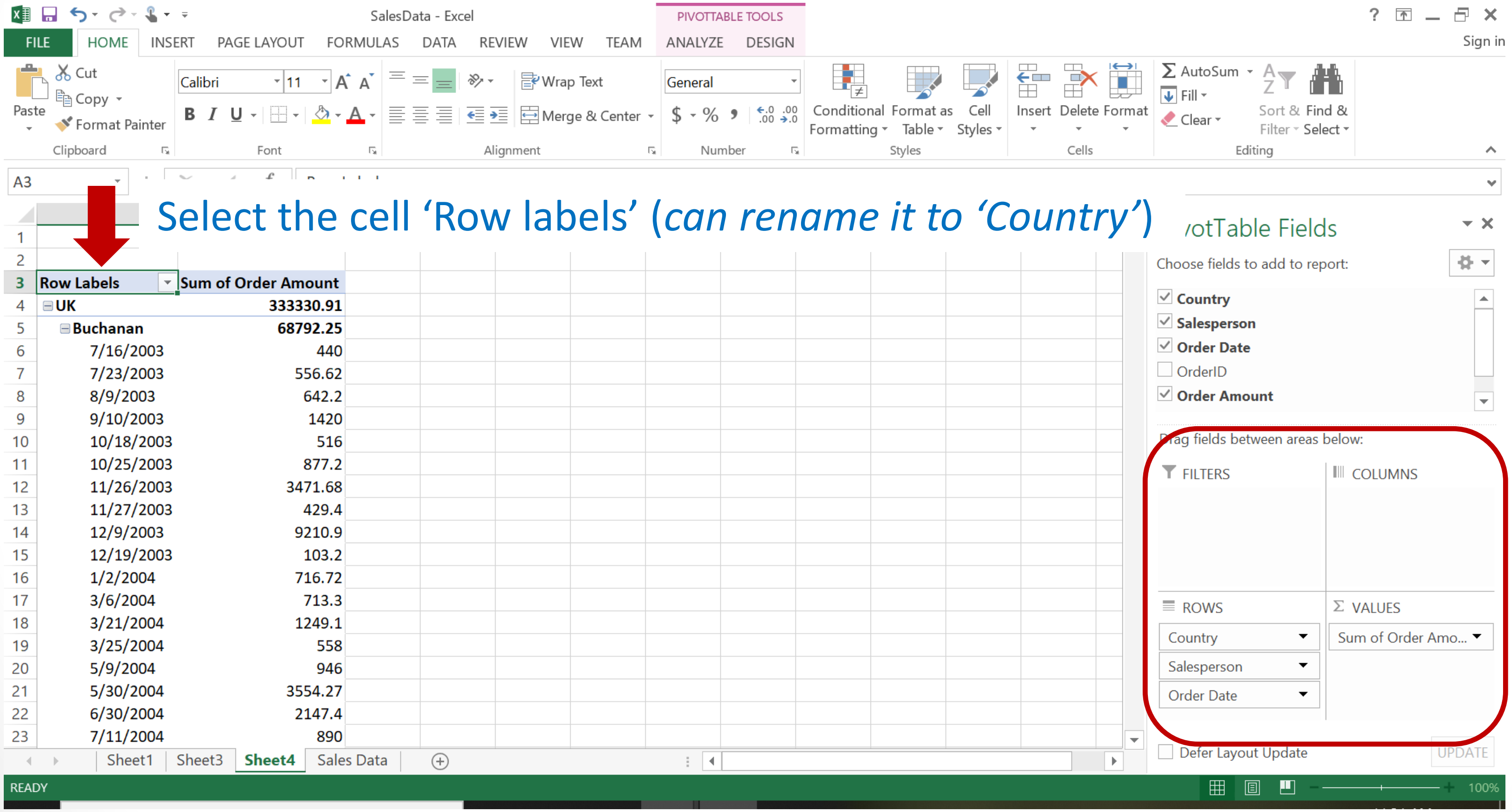
Rows: Country

Values: Sum of Order Amount

□ Defer Layout Update

Update

Example 4: Slicing – 1/3



Select the cell ‘Row labels’ (*can rename it to ‘Country’*)

3	Row Labels	Sum of Order Amount
4	⊟UK	333330.91
5	⊟Buchanan	68792.25
6	7/16/2003	440
7	7/23/2003	556.62
8	8/9/2003	642.2
9	9/10/2003	1420
10	10/18/2003	516
11	10/25/2003	877.2
12	11/26/2003	3471.68
13	11/27/2003	429.4
14	12/9/2003	9210.9
15	12/19/2003	103.2
16	1/2/2004	716.72
17	3/6/2004	713.3
18	3/21/2004	1249.1
19	3/25/2004	558
20	5/9/2004	946
21	5/30/2004	3554.27
22	6/30/2004	2147.4
23	7/11/2004	890

/otTable Fields

Choose fields to add to report:

- ☑ Country
- ☑ Salesperson
- ☑ Order Date
- □ OrderID
- ☑ Order Amount

rag fields between areas below:

FILTERS

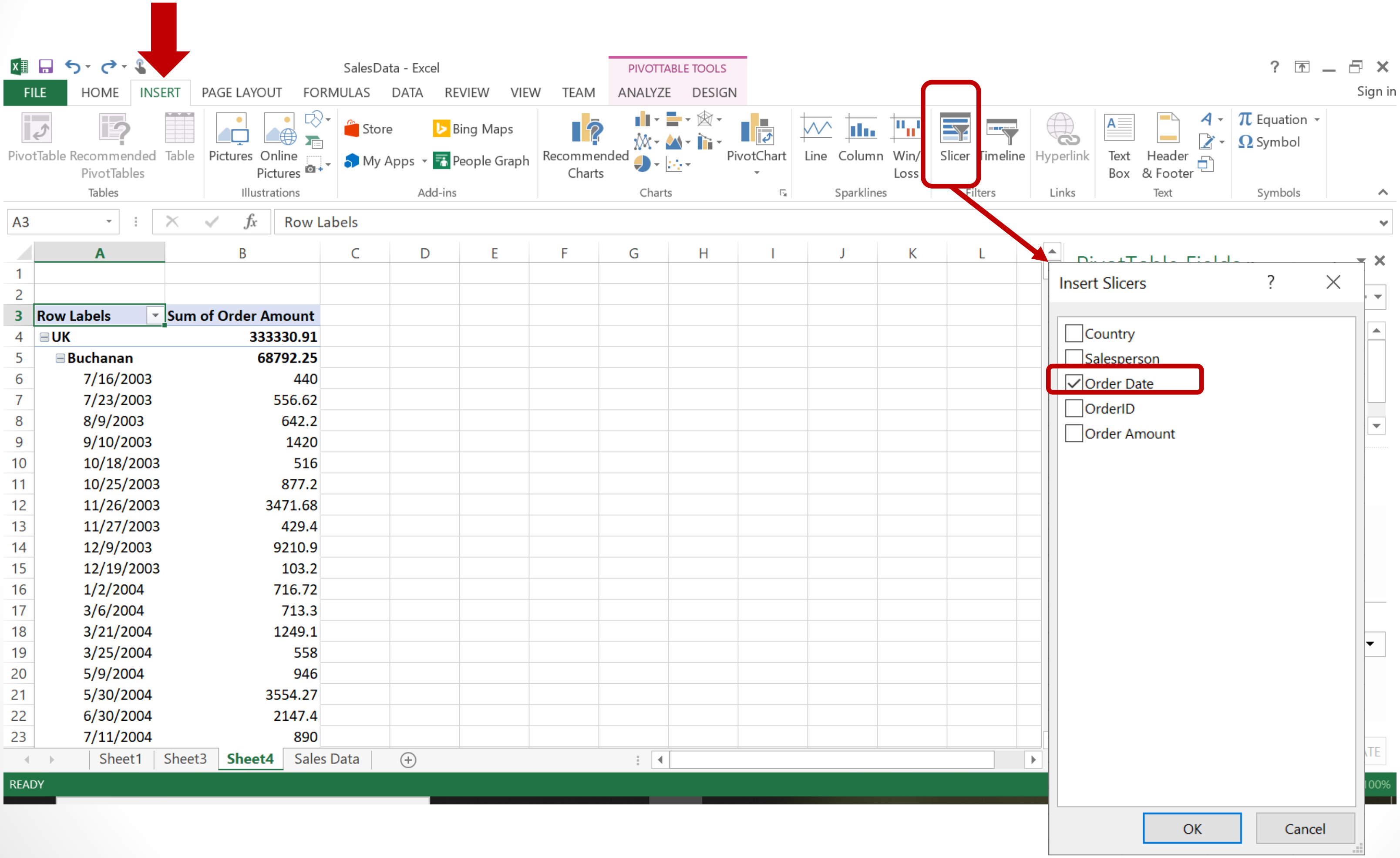
COLUMNS

ROWS
- Country
- Salesperson
- Order Date

VALUES
- Sum of Order Amo...

□ Defer Layout Update UPDATE

Example 4: Slicing – 2/3



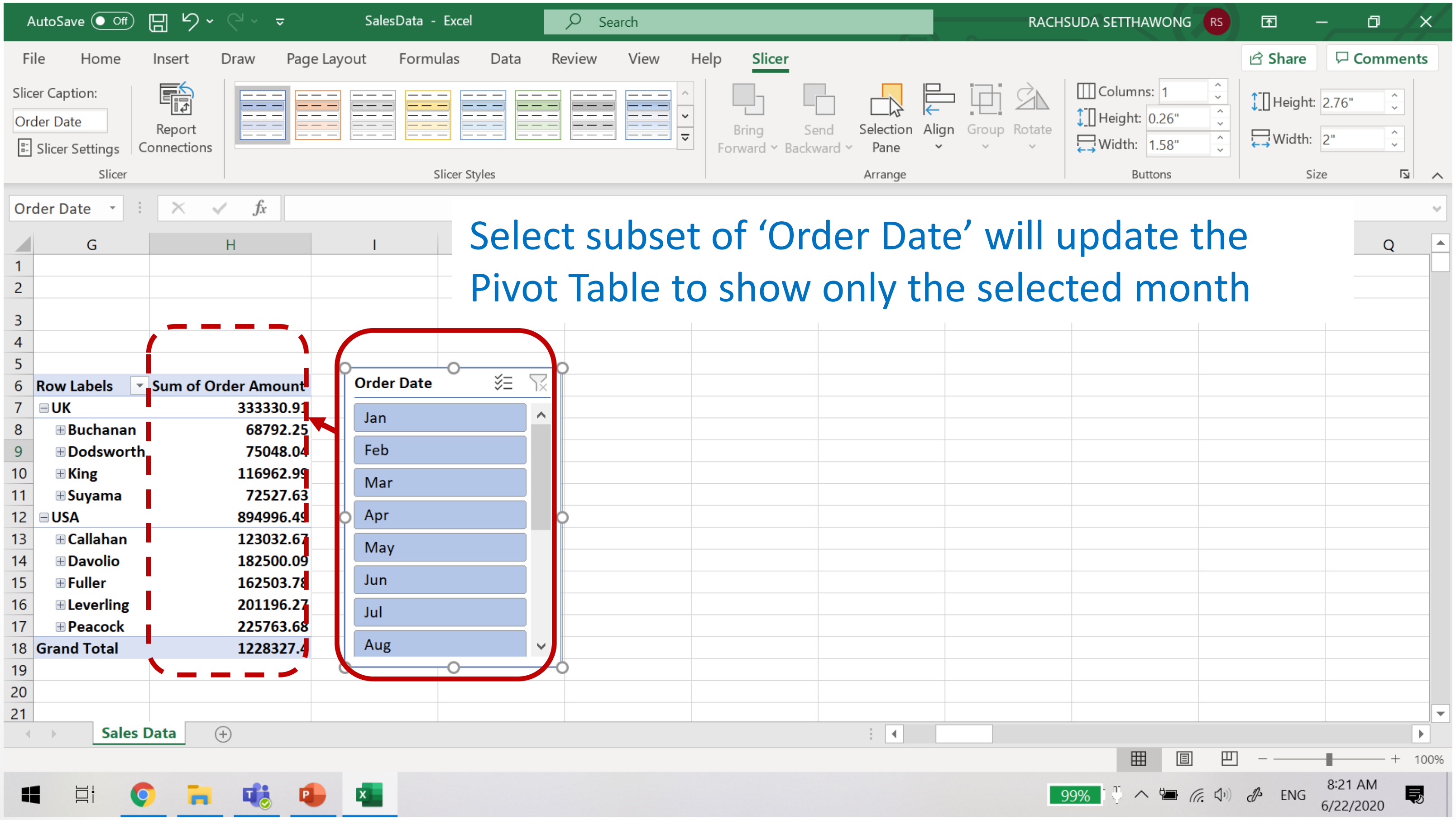
	A	B
1		
2		
3	Row Labels	Sum of Order Amount
4	⊟UK	333330.91
5	⊟Buchanan	68792.25
6	7/16/2003	440
7	7/23/2003	556.62
8	8/9/2003	642.2
9	9/10/2003	1420
10	10/18/2003	516
11	10/25/2003	877.2
12	11/26/2003	3471.68
13	11/27/2003	429.4
14	12/9/2003	9210.9
15	12/19/2003	103.2
16	1/2/2004	716.72
17	3/6/2004	713.3
18	3/21/2004	1249.1
19	3/25/2004	558
20	5/9/2004	946
21	5/30/2004	3554.27
22	6/30/2004	2147.4
23	7/11/2004	890

Insert Slicers

□ Country
□ Salesperson
☑ Order Date
□ OrderID
□ Order Amount

OK Cancel

Example 4: Slicing – 3/3



Slicer Caption: Order Date

Columns: 1

Height: 0.26"

Width: 1.58"

Height: 2.76"

Width: 2"

Select subset of ‘Order Date’ will update the Pivot Table to show only the selected month

Row Labels	Sum of Order Amount
⊟UK	333330.91
⊞Buchanan	68792.25
⊞Dodsworth	75048.04
⊞King	116962.99
⊞Suyama	72527.63
⊟USA	894996.49
⊞Callahan	123032.67
⊞Davolio	182500.09
⊞Fuller	162503.78
⊞Leverling	201196.27
⊞Peacock	225763.68
Grand Total	1228327.4

Order Date

Jan

Feb

Mar

Apr

May

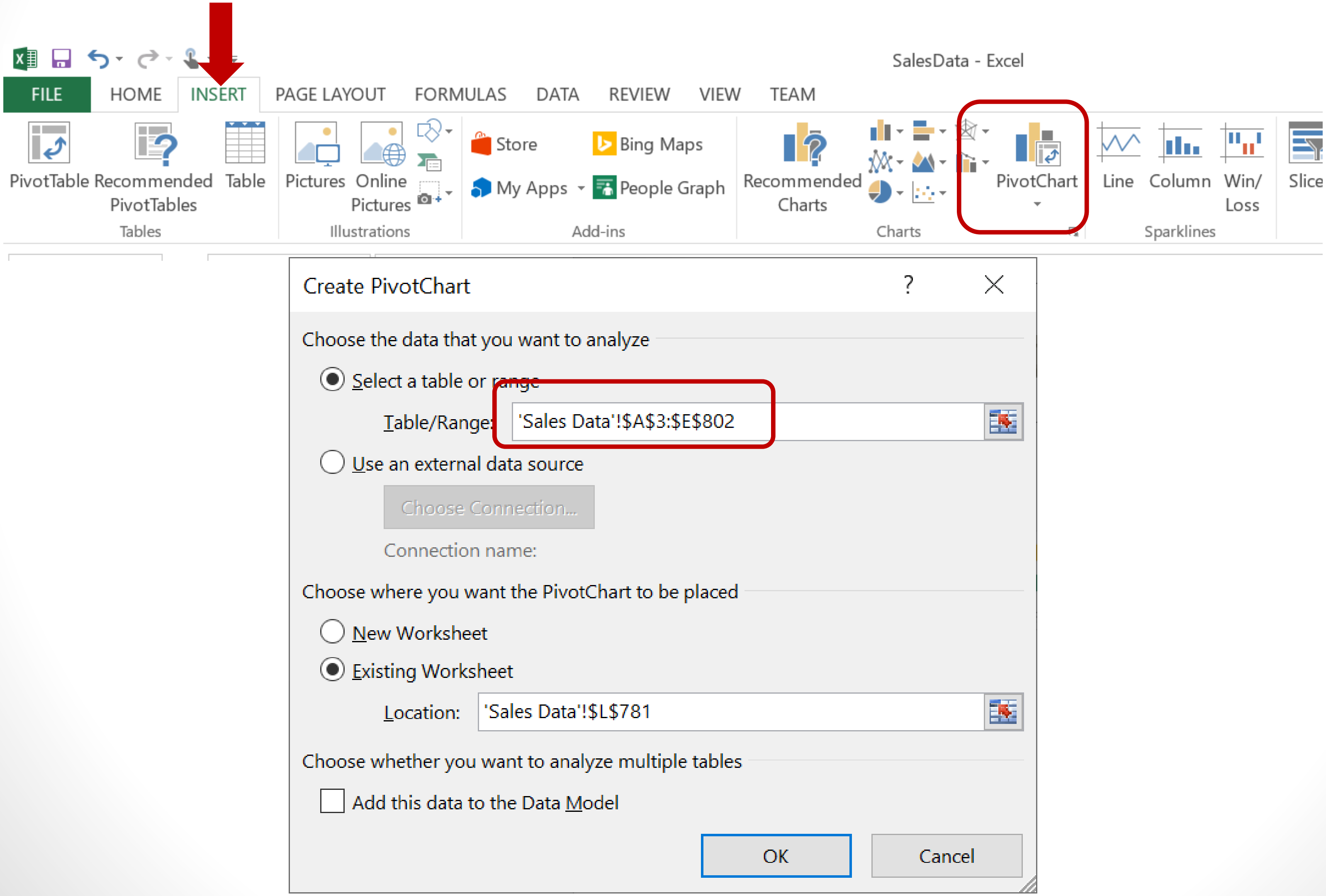
Jun

Jul

Aug

Asst. Prof. Dr. Rachsuda Setthawong

Creating Pivot Chart



Create PivotChart

Choose the data that you want to analyze

● Select a table or range

Table/Range: 'Sales Data'!$A$3:$E$802

○ Use an external data source

Connection name:

Choose where you want the PivotChart to be placed

○ New Worksheet

● Existing Worksheet

Location: 'Sales Data'!$L$781

Choose whether you want to analyze multiple tables

□ Add this data to the Data Model

Asst. Prof. Dr. Rachsuda Setthawong

Pivot Chart Added

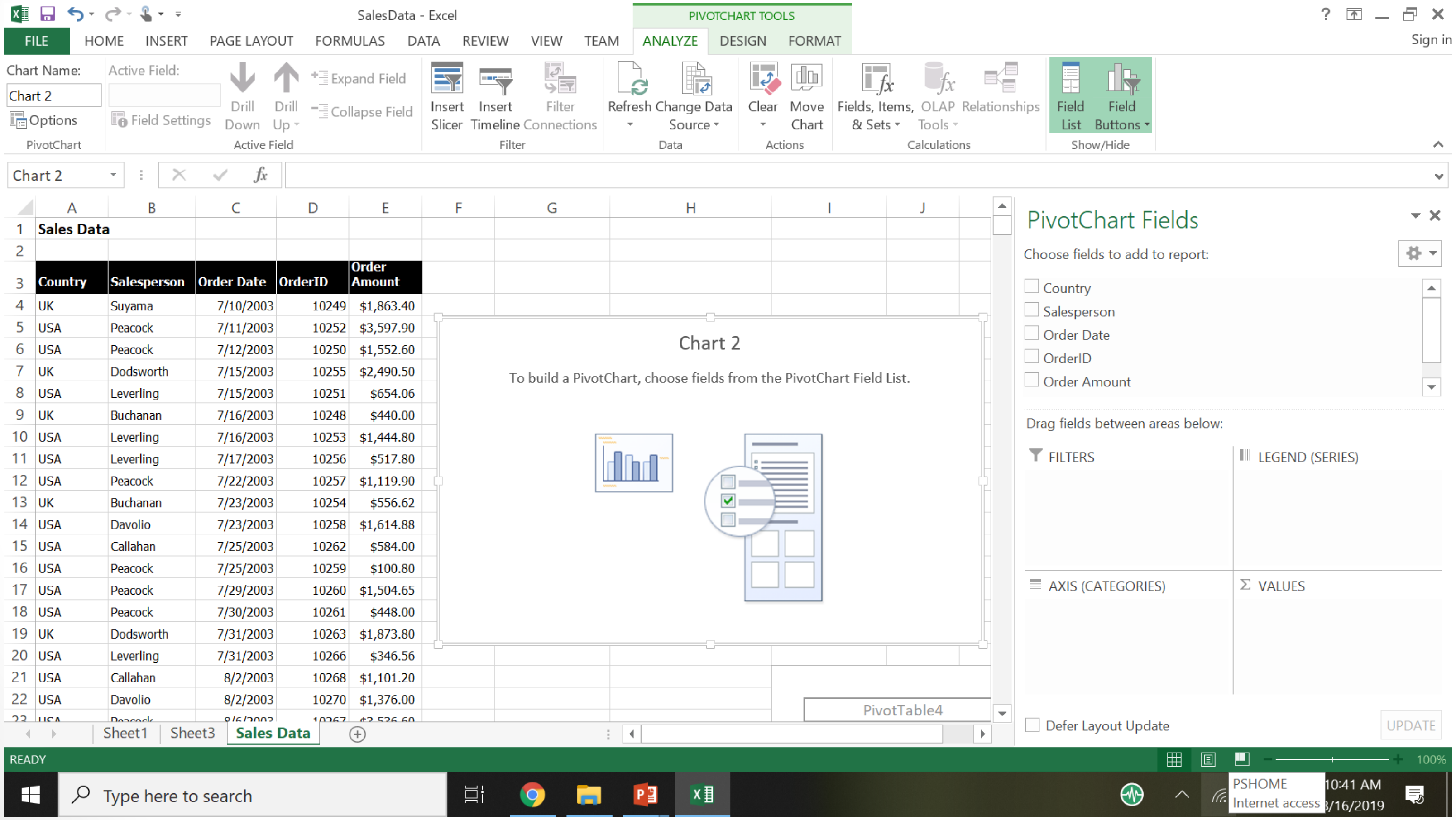


Chart Name: Chart 2

Chart 2

	A	B	C	D	E
1	**Sales Data**				
2					
3	**Country**	**Salesperson**	**Order Date**	**OrderID**	**Order Amount**
4	UK	Suyama	7/10/2003	10249	$1,863.40
5	USA	Peacock	7/11/2003	10252	$3,597.90
6	USA	Peacock	7/12/2003	10250	$1,552.60
7	UK	Dodsworth	7/15/2003	10255	$2,490.50
8	USA	Leverling	7/15/2003	10251	$654.06
9	UK	Buchanan	7/16/2003	10248	$440.00
10	USA	Leverling	7/16/2003	10253	$1,444.80
11	USA	Leverling	7/17/2003	10256	$517.80
12	USA	Peacock	7/22/2003	10257	$1,119.90
13	UK	Buchanan	7/23/2003	10254	$556.62
14	USA	Davolio	7/23/2003	10258	$1,614.88
15	USA	Callahan	7/25/2003	10262	$584.00
16	USA	Peacock	7/25/2003	10259	$100.80
17	USA	Peacock	7/29/2003	10260	$1,504.65
18	USA	Peacock	7/30/2003	10261	$448.00
19	UK	Dodsworth	7/31/2003	10263	$1,873.80
20	USA	Leverling	7/31/2003	10266	$346.56
21	USA	Callahan	8/2/2003	10268	$1,101.20
22	USA	Davolio	8/2/2003	10270	$1,376.00
23	USA	Peacock	8/6/2003	10267	$3,536.60

Chart 2

To build a PivotChart, choose fields from the PivotChart Field List.

PivotTable4

PivotChart Fields

Choose fields to add to report:

☐ Country
☐ Salesperson
☐ Order Date
☐ OrderID
☐ Order Amount

Drag fields between areas below:

FILTERS

LEGEND (SERIES)

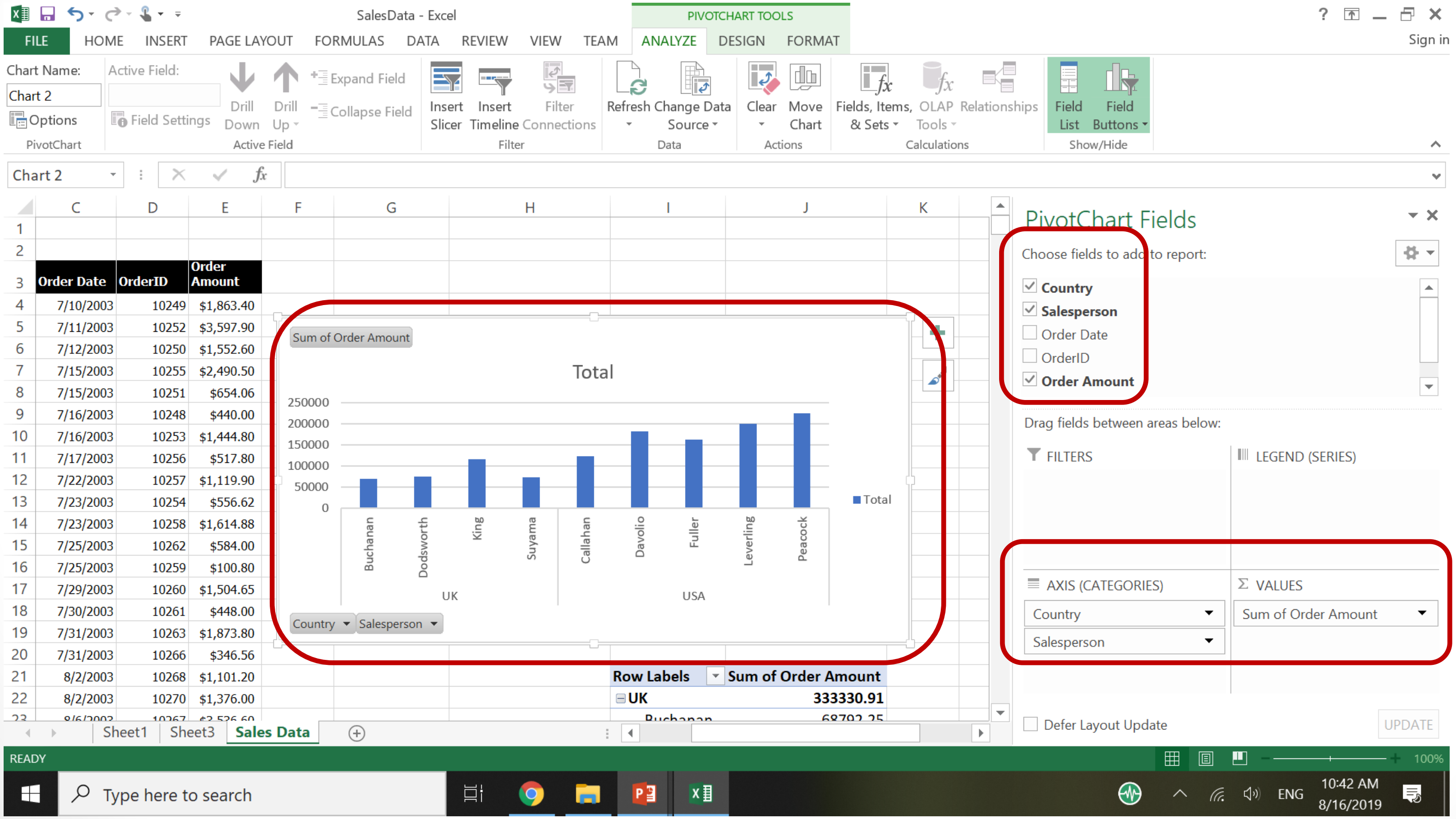
AXIS (CATEGORIES)

Σ VALUES

☐ Defer Layout Update

UPDATE

Adding Fields to the Chart



	C	D	E
3	Order Date	OrderID	Order Amount
4	7/10/2003	10249	$1,863.40
5	7/11/2003	10252	$3,597.90
6	7/12/2003	10250	$1,552.60
7	7/15/2003	10255	$2,490.50
8	7/15/2003	10251	$654.06
9	7/16/2003	10248	$440.00
10	7/16/2003	10253	$1,444.80
11	7/17/2003	10256	$517.80
12	7/22/2003	10257	$1,119.90
13	7/23/2003	10254	$556.62
14	7/23/2003	10258	$1,614.88
15	7/25/2003	10262	$584.00
16	7/25/2003	10259	$100.80
17	7/29/2003	10260	$1,504.65
18	7/30/2003	10261	$448.00
19	7/31/2003	10263	$1,873.80
20	7/31/2003	10266	$346.56
21	8/2/2003	10268	$1,101.20
22	8/2/2003	10270	$1,376.00

Row Labels	Sum of Order Amount
⊟UK	333330.91
Buchanan	68792.25

PivotChart Fields

Choose fields to add to report:

- ☑ Country
- ☑ Salesperson
- □ Order Date
- □ OrderID
- ☑ Order Amount

Drag fields between areas below:

FILTERS

LEGEND (SERIES)

AXIS (CATEGORIES): Country, Salesperson

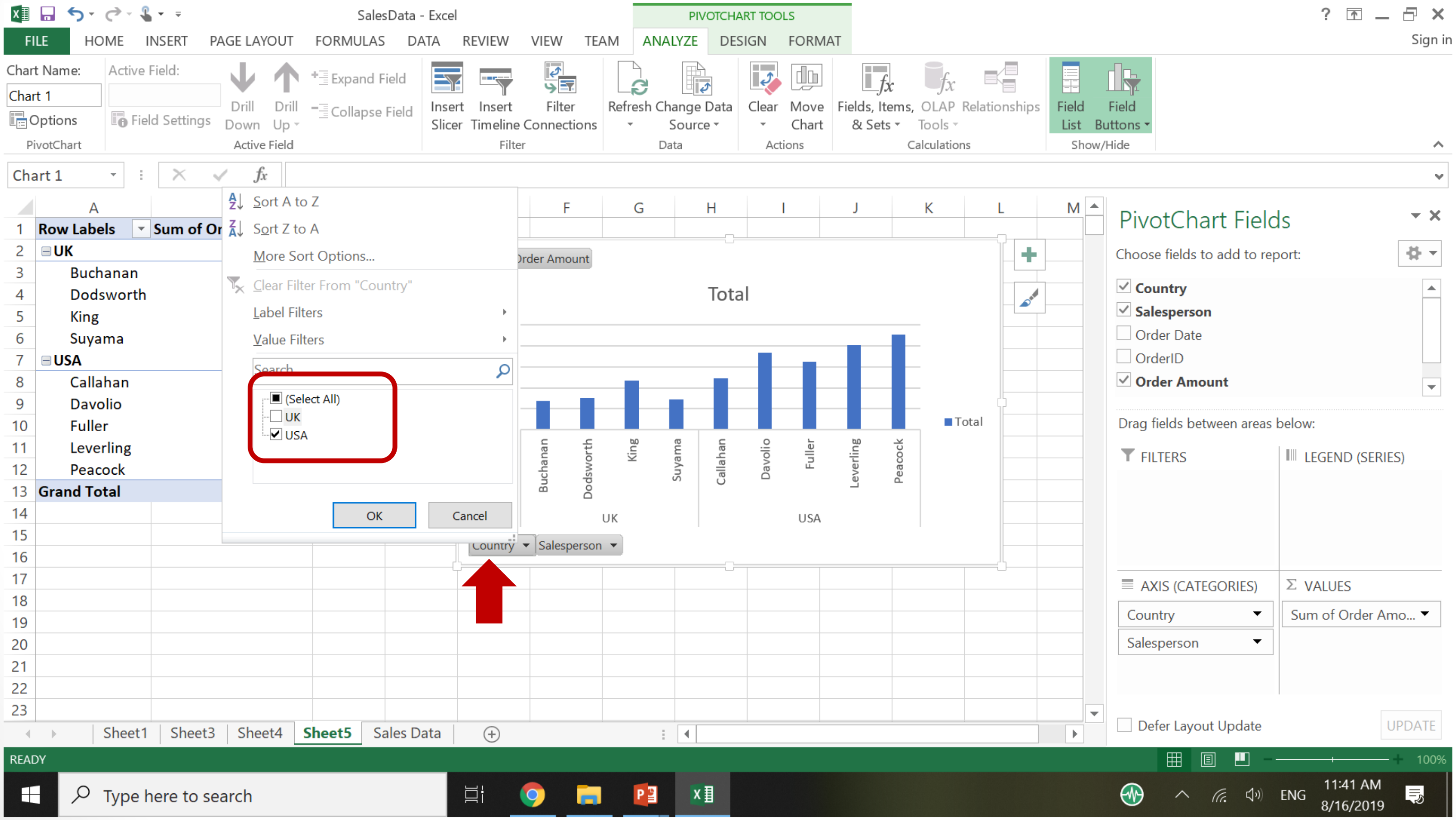
VALUES: Sum of Order Amount

□ Defer Layout Update

UPDATE

More Action on Chart:

Interactively Update the Chart using Filter



Row Labels	Sum of Or
UK	
Buchanan	
Dodsworth	
King	
Suyama	
USA	
Callahan	
Davolio	
Fuller	
Leverling	
Peacock	
Grand Total	

■ (Select All)
□ UK
☑ USA

☑ Country
☑ Salesperson
□ Order Date
□ OrderID
☑ Order Amount

AXIS (CATEGORIES): Country, Salesperson

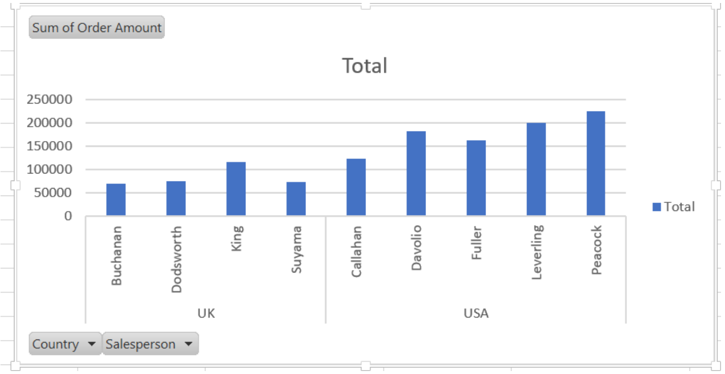
VALUES: Sum of Order Amo...

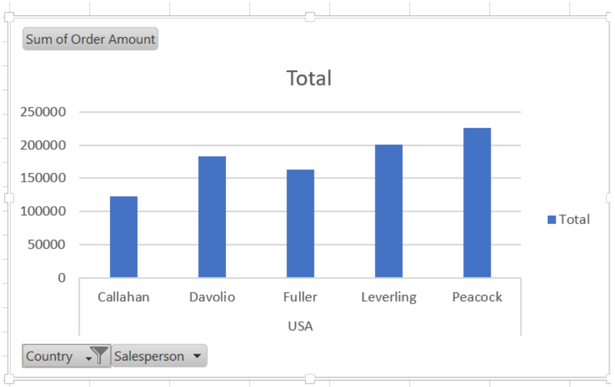
□ Defer Layout Update

Asst. Prof. Dr. Rachsuda Setthawong

More Action on Chart:

Interactively Update the Chart using Filter: Result





Asst. Prof. Dr. Rachsuda Setthawong

More Action on Pivot Table (Chart):

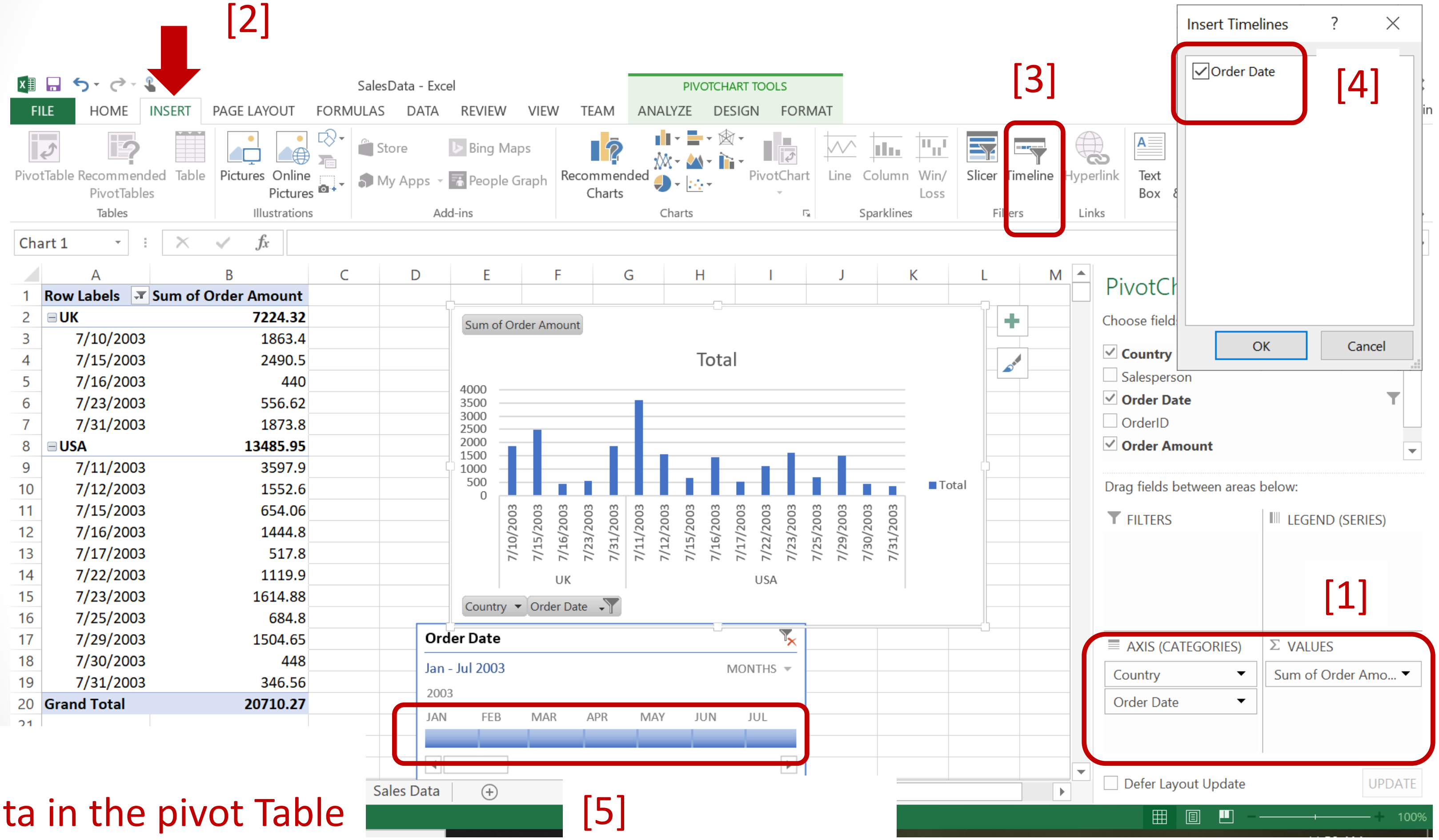
Using Timeline to Filter Data

(Must be ‘Date/Time’ Field)

[2]

[3]

[4]



Insert Timelines

☑ Order Date

OK Cancel

Row Labels	Sum of Order Amount
⊟UK	7224.32
7/10/2003	1863.4
7/15/2003	2490.5
7/16/2003	440
7/23/2003	556.62
7/31/2003	1873.8
⊟USA	13485.95
7/11/2003	3597.9
7/12/2003	1552.6
7/15/2003	654.06
7/16/2003	1444.8
7/17/2003	517.8
7/22/2003	1119.9
7/23/2003	1614.88
7/25/2003	684.8
7/29/2003	1504.65
7/30/2003	448
7/31/2003	346.56
Grand Total	20710.27

Order Date

Jan - Jul 2003 MONTHS

2003

JAN FEB MAR APR MAY JUN JUL

☑ Country
□ Salesperson
☑ Order Date
□ OrderID
☑ Order Amount

AXIS (CATEGORIES): Country, Order Date

VALUES: Sum of Order Amo...

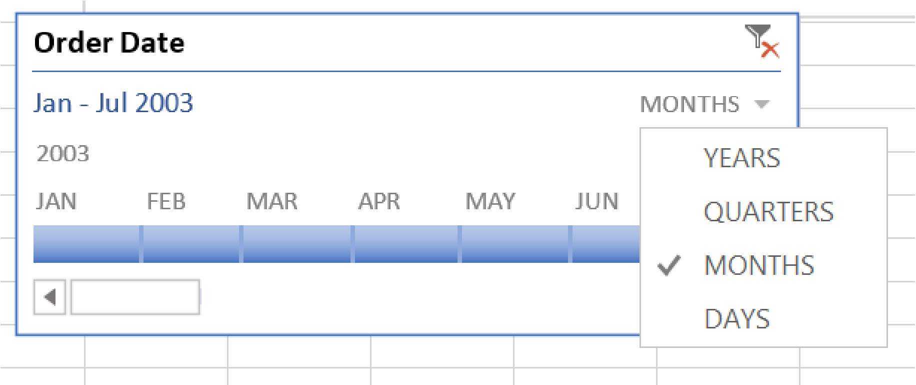
[1]

[6]
Data in the pivot Table (and graph if any) are updated, accordingly.

[5]
Select JAN – JULY

More Action on Pivot Table (Chart):

Changing Timeline's Granularity



Order Date

Jan - Jul 2003

MONTHS

Asst. Prof. Dr. Rachsuda Setthawong

Analysis ToolPak in MS Excel

Loading and the Activating Analysis ToolPak

1. Click the **File** tab, click **Options**, and then click the **Add-Ins** category.

2. If you're using Excel 2007, click the **Microsoft Office Button** , and then click **Excel Options**

3. In the **Manage** box, select **Excel Add-ins** and then click **Go**.

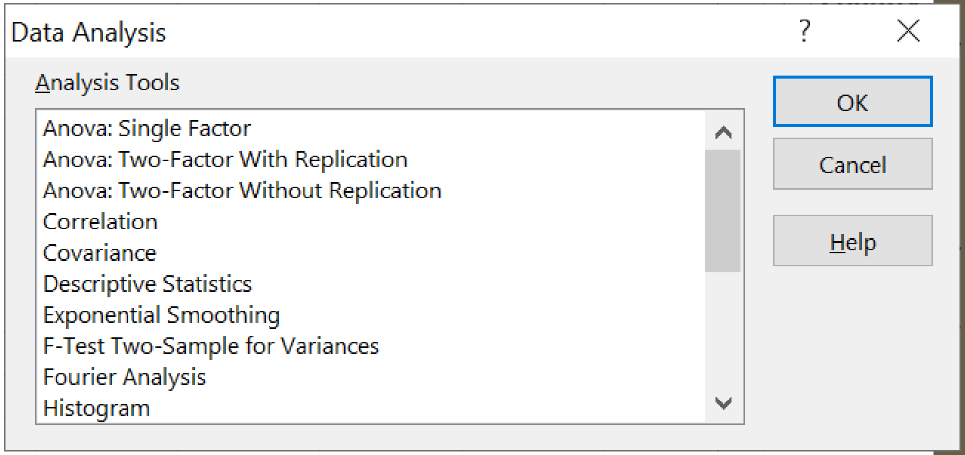
4. If you're using Excel for Mac, in the file menu go to **Tools** > **Excel Add-ins**.

5. In the **Add-Ins** box, check the **Analysis ToolPak** check box, and then click **OK**.
   - If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.
   - If you are prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

https://support.office.com/en-us/article/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6?fbclid=IwAR1mBggWEPwoUFv4pTRHnOLpQS7V5rk8aCHpNHE4RIMhMMvXs-V2K59pZ20&ui=en-US&rs=en-US&ad=US

Click the **Data** tab,
in Analysis section click **Data Analysis**

- Anova
- Correlation
- Covarience
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test
- z-Test



Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram