MT5125 MSc in Business Administration

2020/2021

Strat. Consultancy Project I & Data Analytics

DCU BUSINESS SCHOOL

Dr. Damien Dupré
damien.dupre@dcu.ie
Assistant Professor - Business Research Methods

dotLAB
The Irish Institute of Digital Business

Data Cleaning and Transformations

DCU BUSINESS SCHOOL

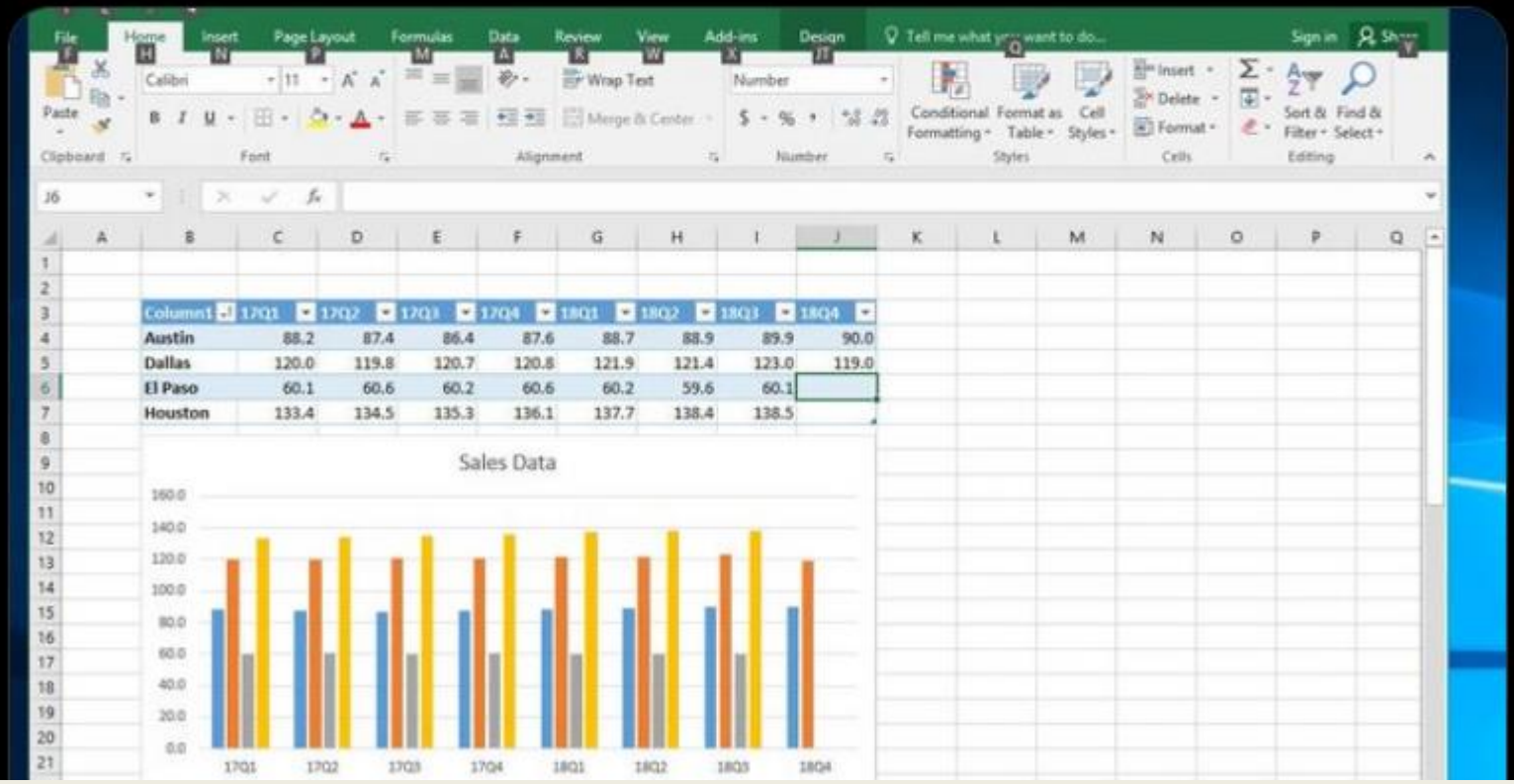
dotLAB
The Irish Institute of Digital Business

DATA CLEANING

Question from Kareem

Kareem Carr
@kareem_carr

dear data scientists, what is preventing your data analysis from looking like this?

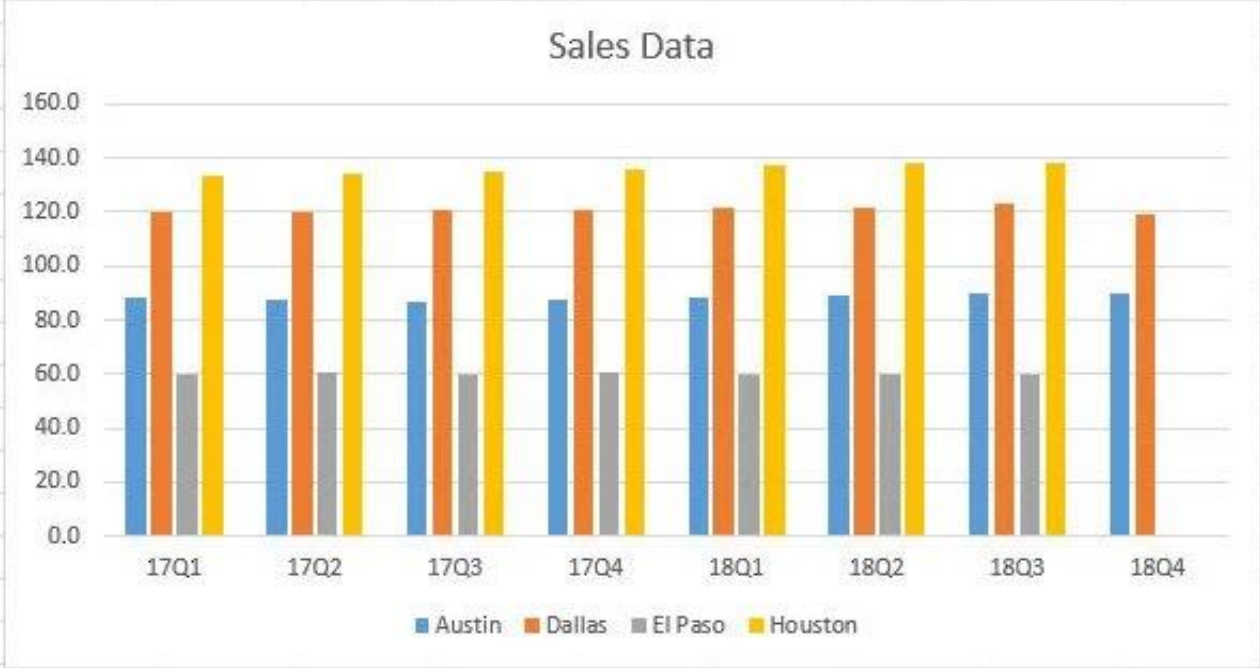


File Home Insert Page Layout Formulas Data Review View Add-ins Design Tell me what you want to do...

Paste Clipboard Calibri 11 Font Wrap Text Merge & Center Alignment Number Conditional Formatting Format as Table Cell Styles Styles

J6

Column1	17Q1	17Q2	17Q3	17Q4	18Q1	18Q2	18Q3	18Q4
Austin	88.2	87.4	86.4	87.6	88.7	88.9	89.9	90.0
Dallas	120.0	119.8	120.7	120.8	121.9	121.4	123.0	119.0
El Paso	60.1	60.6	60.2	60.6	60.2	59.6	60.1	
Houston	133.4	134.5	135.3	136.1	137.7	138.4	138.5	



Answer from Michael

Michael @CoffeeCodeCrash · Oct 23

Replying to @kareem_carr

Ditch the chart, save as CSV and, at the very least, column headings should be in row 1, columns should start at column A. Any other spreadsheet is chaotic evil.

Clean a data table

1. Ditch the chart and all non values
   - Charts can mess up with other software
2. Save as .csv file
   - Better format and keeps only the current sheet
3. Column headings in row 1
   - No more than 1 heading row and remove blanks
4. Columns start at column A
   - Remove blanks before data

Any other spreadsheet is chaotic evil!

EXERCISE: CLEAN UNICEF.XLSX

Clean unicef.xlsx

- On the MT5125 Loop page, download and open the document “unicef.xlsx” located in:
  - Data Analytics Supplementary Information> Lecture 2
- Clean this data file

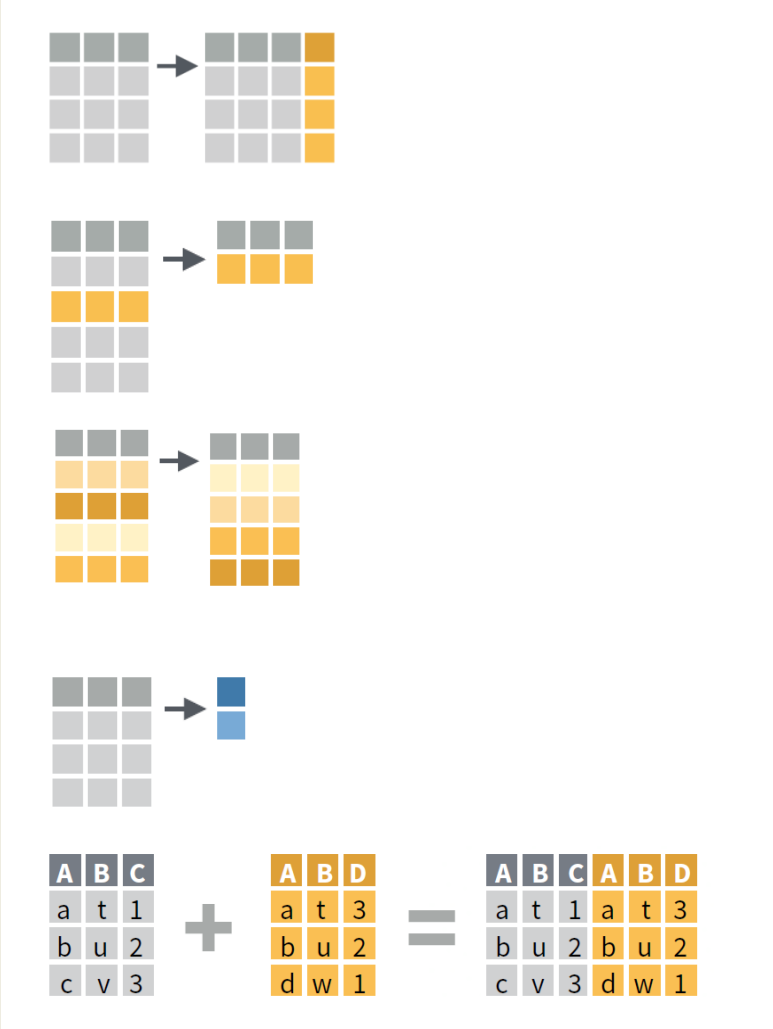
TRANSFORM DATA

Master the Key Movements

- Most important work is to tidy your data
  - Takes time to saves time and solves problems
  - 5 movements are necessary to master (almost)

Master the Key Movements

- Extension
- Reduction
- Direction
- Aggregation
- Combination



EXTENSION

Extension with MS Excel

- Extension = New Column
- In MS Excel:
  1. First row is row name (name convention)
  2. Second row is the function (starts with =)
  3. Following rows are applied (squared corner)

Extension with MS Excel

1

	A	B	C	D	E	F
	les for	30/04/2013				
	Product	Quantity	Price	Net sales	Sales tax	Gross sales
4	Latte	120	2.2	=B4*C4		
5	Mocha	90	2.5			
6	Capuchino	80	2			
7	Macchiato	150	2.8			
8	Americano	140	2.1			
9	Flat White	75	2.3			
10	Espresso	60	2.9			
11	Double Espresso	40	3.2			
12	Cappuccino	32	2.5			
13						
14						

Sales

2

	A	B	C	D	E	F
	ales for	30/04/2013				
	Product	Quantity	Price	Net sales	Sales tax	Gross sales
4	Latte	120	2.2	264.00		
5	Mocha	90	2.5			
6	Capuchino	80	2			
7	Macchiato	150	2.8			
8	Americano	140	2.1			
9	Flat White	75	2.3			
10	Espresso	60	2.9			
11	Double Espresso	40	3.2			
12	Cappuccino	32	2.5			
13						
14						

Sales

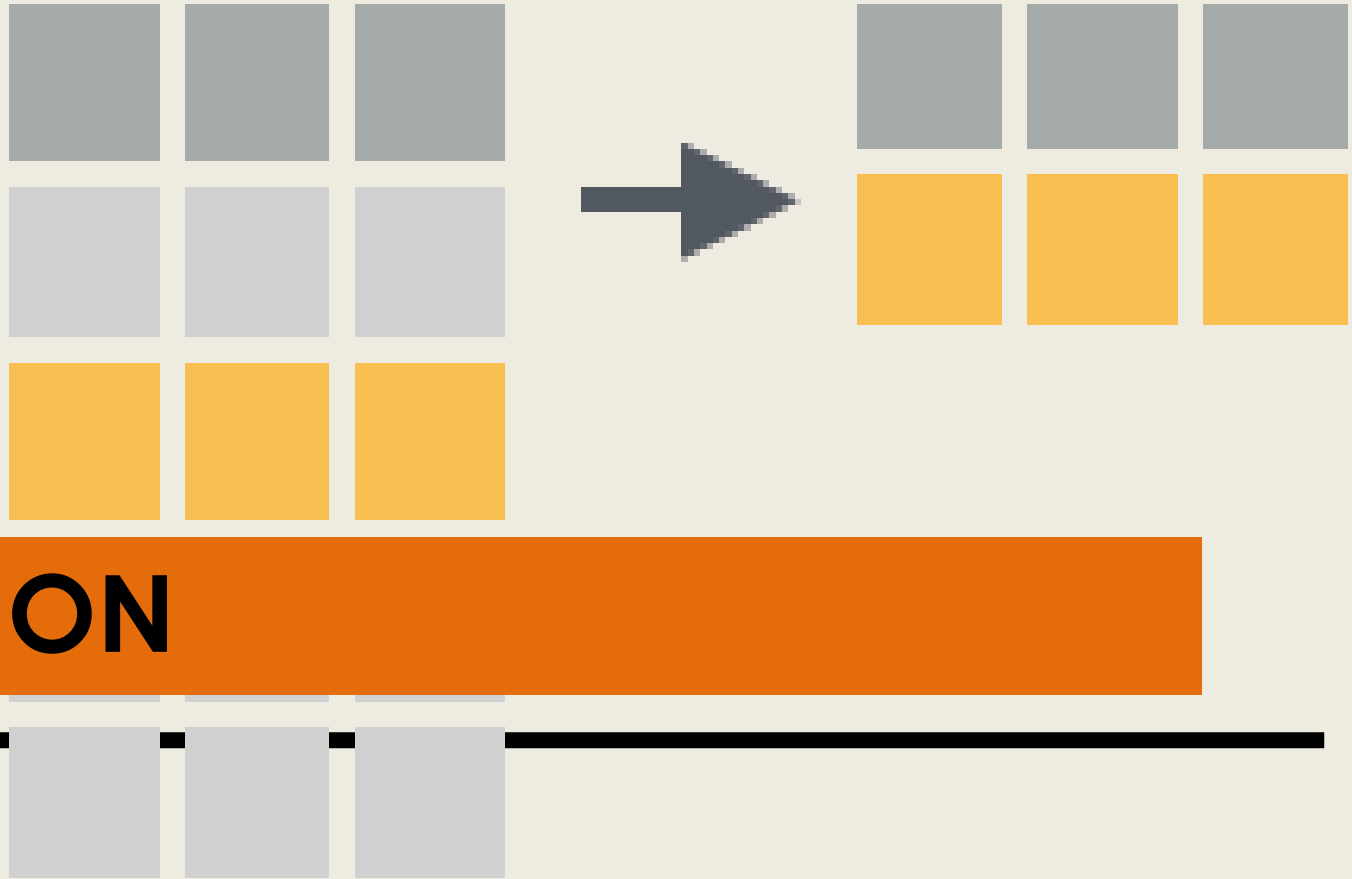
3

	A	B	C	D	E	F
	les for	30/04/2013				
	Product	Quantity	Price	Net sales	Sales tax	Gross sales
4	Latte	120	2.2	264.00		
5	Mocha	90	2.5	225.00		
6	Capuchino	80	2	160.00		
7	Macchiato	150	2.8	420.00		
8	Americano	140	2.1	294.00		
9	Flat White	75	2.3	172.50		
10	Espresso	60	2.9	174.00		
11	Double Espresso	40	3.2	128.00		
12	Cappuccino	32	2.5	80.00		
13						
14						

Sales

Excel Functions

- For numeric values
  - Numeric operator (+ - / *)
  - $ (fixed parameter)
  - COUNT(), MIN(), MAX(), SUM(), AVERAGE (), STDEV()
- For character strings
  - LEFT()
  - CONCATENATE()
- Extra function
  - IF(condition, value if true, value if false)



REDUCTION

Reduction with MS Excel

- Reduction = Filter Column
- In MS Excel:
  1. Select header row
  2. In Data tab, use Filter
  3. Click the drop-down arrow for the column you want to filter
  4. Choose values to filter

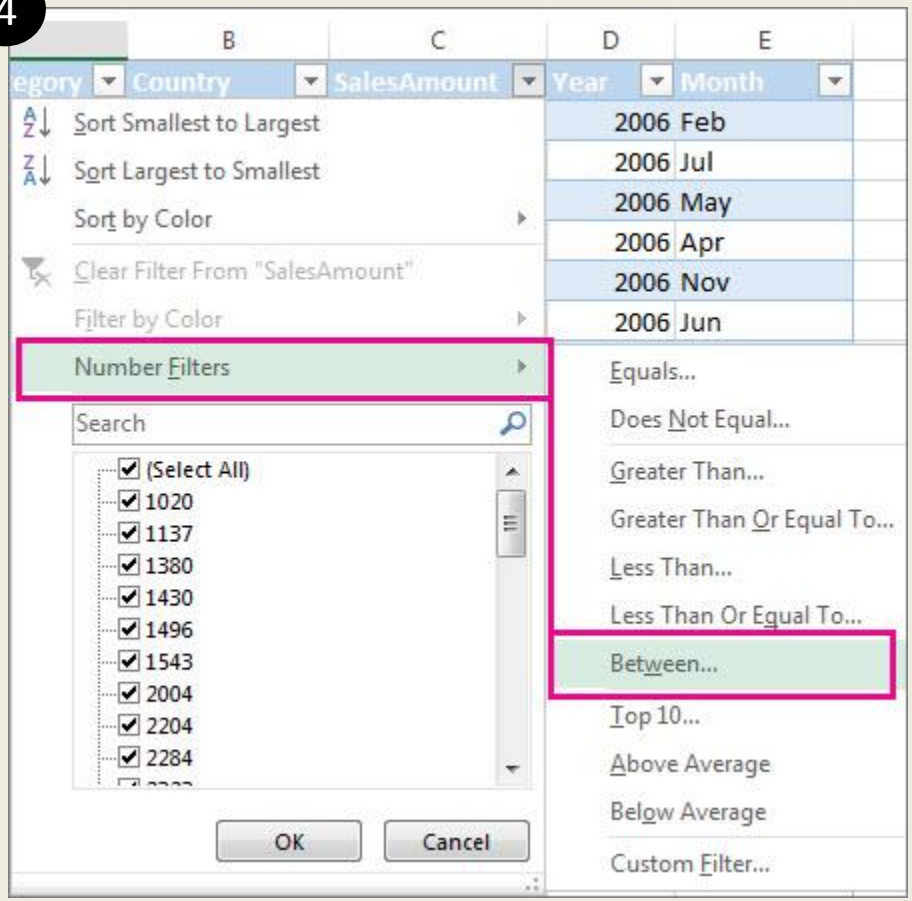
Reduction with MS Excel

2

3

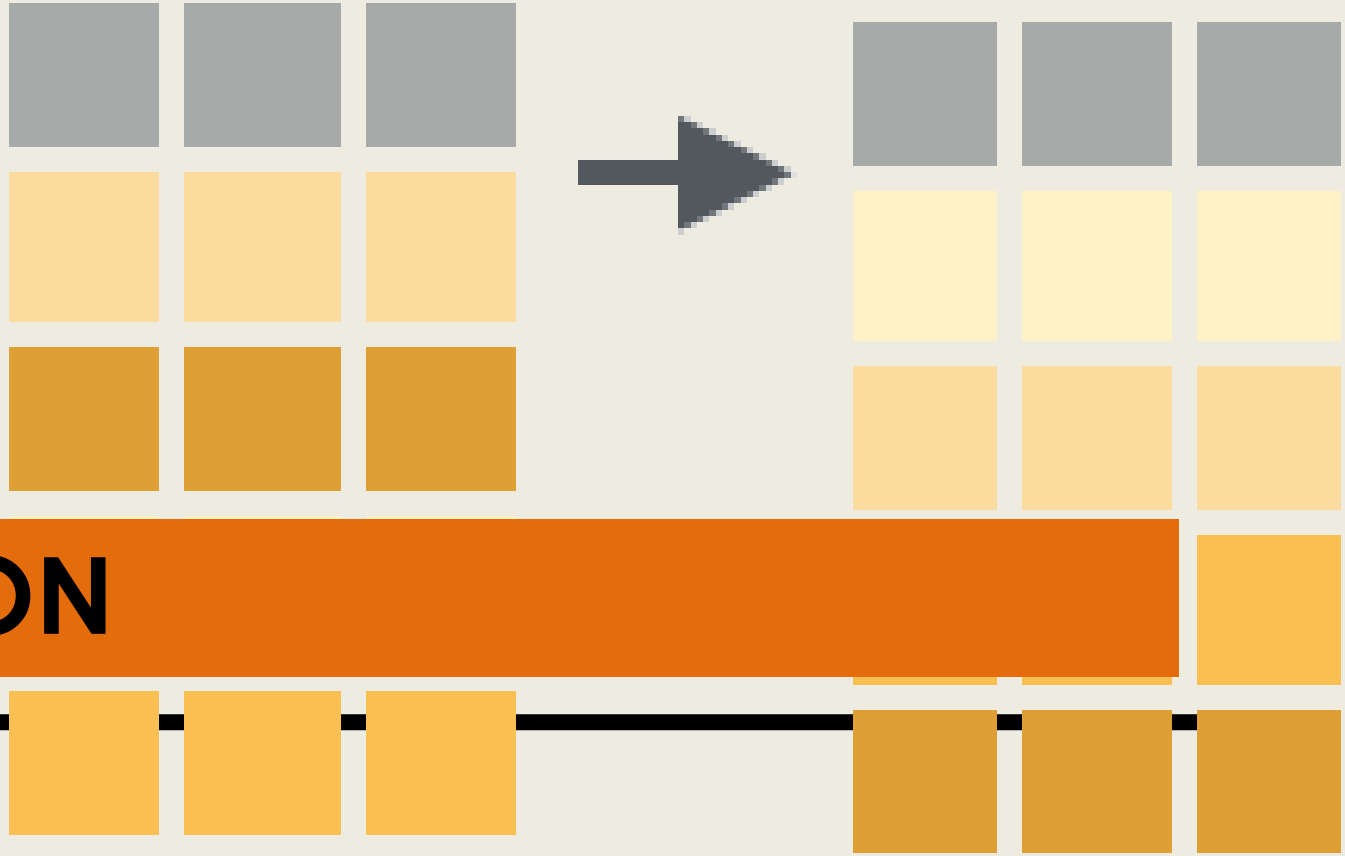
	A	B	C	D	E	F	G
1	Product	Qtr 1	Qtr 2	Qtr 3	Qtr 4	Grand Total	Average Sales
2	Chocolade	$ 744.60	$ 162.56	$ 68.85	$ 306.00	$ 1,282.01	$ [illegible]
3	Gumbär Gummibärchen	$ 5,079.60	$ 1,249.20	$ 2,061.17	$ 2,835.68	$ 11,225.65	$ 2,806.41
4	Maxilaku	$ 1,605.60	$ 620.00	$ 835.00	$ -	$ 3,060.60	$ 765.15
5	NuNuCa Nuß-Nougat-Crème	$ 193.20	$ 865.20		$ 493.50	$ 1,551.90	$ 517.30
6	Pavlova	$ 1,685.36	$ 2,646.08	$ 1,849.70	$ 999.01	$ 7,180.15	$ 1,795.04
7	Schoggi Chocolade	$ 1,755.00	$ 5,268.00	$ 2,195.00	$ 1,756.00	$ 10,974.00	$ 2,743.50
8	Scottish Longbreads	$ 1,267.50	$ 1,062.50	$ 492.50	$ 1,935.00	$ 4,757.50	$ 1,189.38
9	Sir Rodney's Marmalade		$ 4,252.50	$ 1,360.80	$ 1,701.00	$ 7,314.30	$ 2,438.10
10	Sir Rodney's Scones	$ 1,418.00	$ 756.00	$ 1,733.00	$ 1,434.00	$ 5,341.00	$ 1,335.25
11	Tarte au sucre	$ 4,728.00	$ 4,547.92	$ 5,472.30	$ 6,014.60	$ 20,762.82	$ 5,190.71
12	Teatime Chocolate Biscuits	$ 943.89	$ 349.60	$ 841.80	$ 204.70	$ 2,339.99	$ 585.00
13	Valkoinen suklaa	$ 845.00		$ 385.94	$ 942.50	$ 2,173.44	$ 724.48
14	Zaanse koeken	$ 817.00	$ 285.95	$ 668.80	$ 1,159.00	$ 2,930.75	$ 732.69
15	Total	$21,082.75	$22,065.51	$17,964.86	$19,780.99	$ 80,894.11	$ 1,626.42

4



Reduction with MS Excel

- Rows already filtered have a row index are coloured in blue
- Copy-Paste filtered table in a new document if you want to work only on these values
- More about Excel filters:
  - https://edu.gcfglobal.org/en/excel2010/filtering-data/1/
  - https://support.office.com/en-ie/article/filter-data-in-a-range-or-table-01832226-31b5-4568-8806-38c37dcc180e

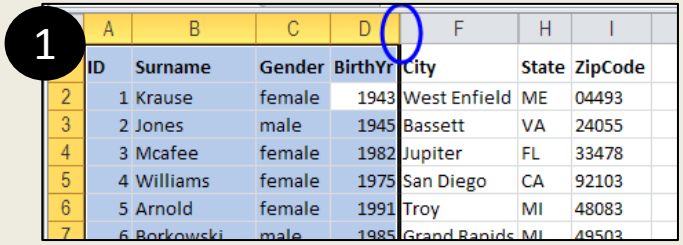


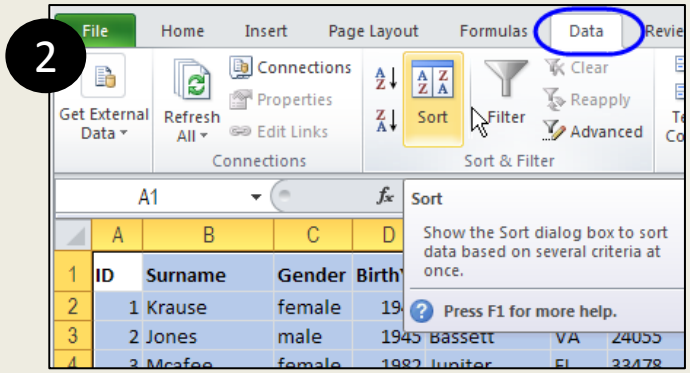
DIRECTION

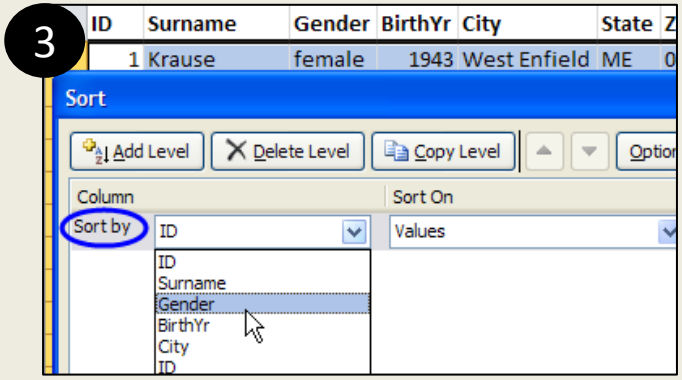
Direction with MS Excel

- Direction = Arrange Row Order
- In MS Excel:
  1. Select table
  2. In Data tab, use Sort
  3. Choose column to sort and how to sort

Direction with MS Excel

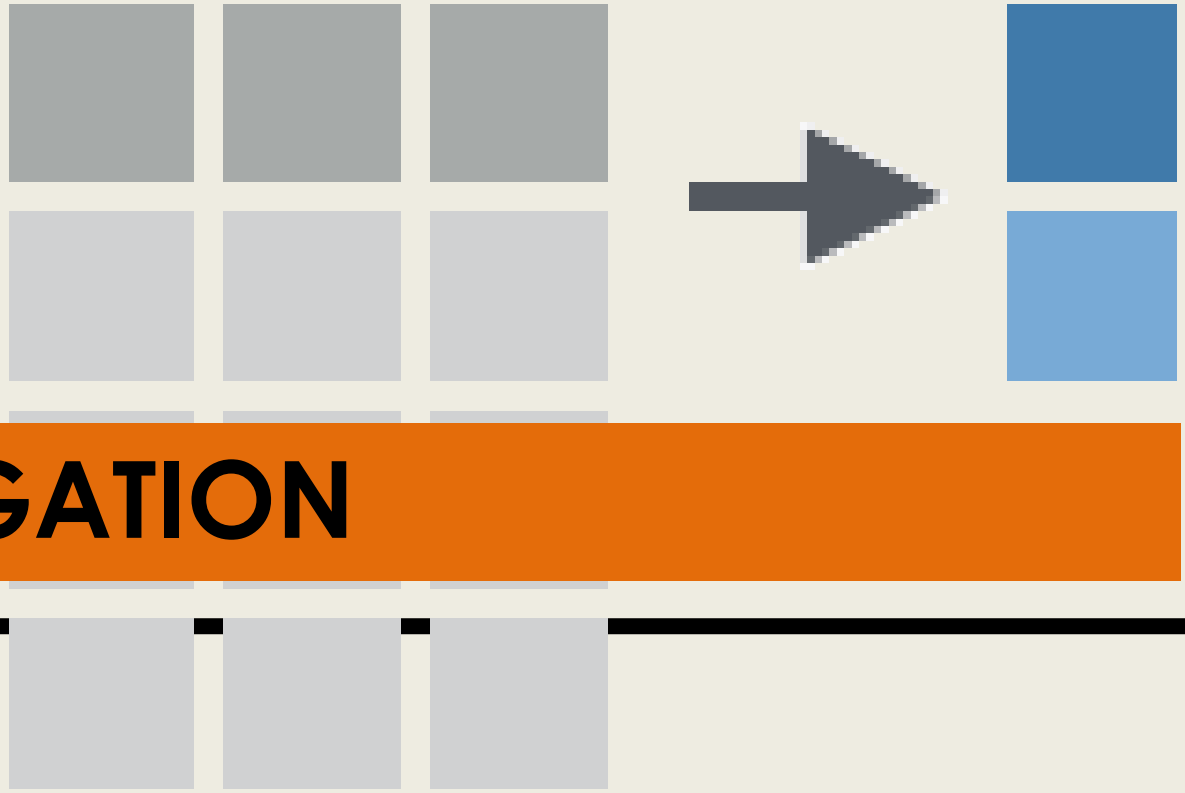






Direction with MS Excel

- Be careful of taking into account all the table
  - All rows
  - All columns
- Double check if all columns changed
- More about Excel sorting:
  - https://www.contextures.com/xlSort01.html
  - https://support.office.com/en-ie/article/sort-data-in-a-range-or-table-62d0b95d-2a90-4610-a6ae-2e545c4a4654



AGGREGATION

Aggregation with MS Excel

- Aggregation = Summary of Column
- In MS Excel:
  1. Simple = use function at the end of a table
  2. Complex = use pivot table

Simple Aggregation

Function	Calculation
=COUNT(A1:A10)	Total number of values
=MIN(A1:A10)	Minimum value
=MAX(A1:A10)	Maximum value
=SUM(A1:A10)	Sum of all values
=AVERAGE (A1:A10)	Sum of all values divided by total number
=STDEV(A1:A10)	Average distance of values to the average

Complex Aggregation

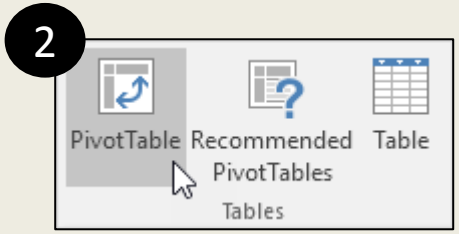
- Pivot Table
  1. Select data
  2. In Insert, use Pivot Table
  3. Drag columns to sort by row/column
  4. Choose value column to be aggregated
  5. Choose type of aggregation

Complex Aggregation

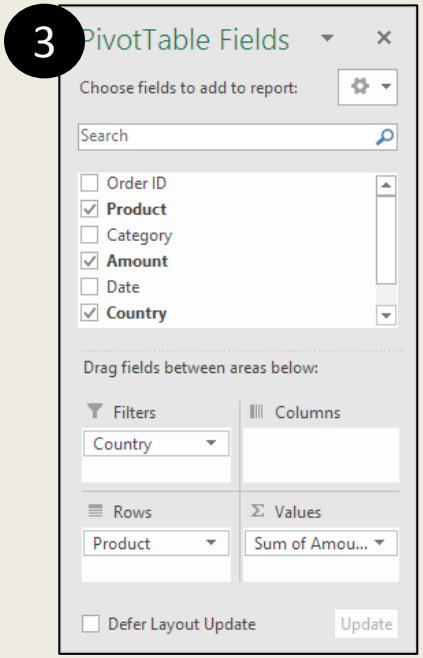
1

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

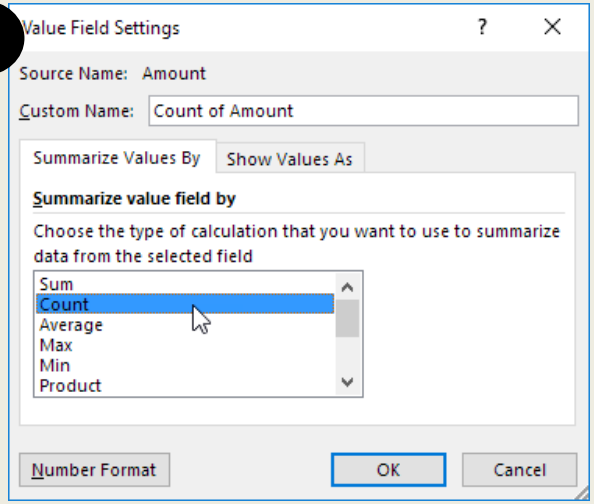
2



3



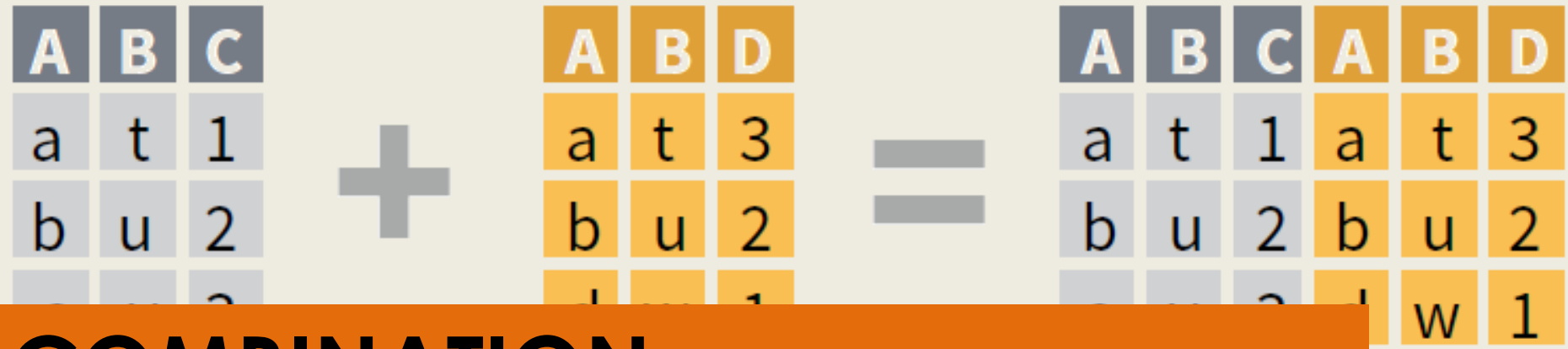
4



Country	France
Row Labels	**Count of Amount**
Apple	16
Banana	7
Carrots	1
Mango	1
Orange	1
Beans	1
Broccoli	1
Grand Total	**28**

Complex Aggregation

- If you want to use the Pivot Table for further analysis
  - Copy-Paste it in another document
  - Paste as value (removes dynamic link)
- More about Excel pivot table:
  - https://www.excel-easy.com/data-analysis/pivot-tables.html
  - https://support.office.com/en-us/article/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576



COMBINATION

Combination with MS Excel

- Combination = Join two tables
- In MS Excel:
  1. One Column = vlookup function
  2. Multiple Columns = Power Query (Windows only)

One Column Combination

- =VLOOKUP(value, table, col_index, [range_lookup])
  - value - The value to look for in the first column of a table
  - table - The table from which to retrieve a value
  - col_index - The column in the table from which to retrieve a value
  - range_lookup - [optional] TRUE = approximate match (default). FALSE = exact match

One Cell Example:

H4 =VLOOKUP(H3,B4:E13,2,FALSE)

1 2 3 4

ID	First	Last	Email
610	Janet	Farley	j.farley@ace.com
798	Steven	Batista	s.batista@ace.com
841	Evelyn	Monet	e.monet@ace.com
886	Marilyn	Bradley	m.bradley@ace.com
622	Jonathan	Adder	j.adder@ace.com
601	Adrian	Birt	a.birt@ace.com
869	Julie	Irons	j.irons@ace.com
867	Erica	Tan	e.tan@ace.com
785	Harold	Clayton	h.clayton@ace.com
648	Sharyn	Castor	s.castor@ace.com

ID	622	
First	Jonathan	2
Last	Adder	3
Email	j.adder@ace.com	4

=VLOOKUP(H4,B5:E9,2,FALSE)

One Column Example:

D5 =VLOOKUP(C5,$G$5:$H$10,2,TRUE)

VLOOKUP - approximate match on sales

Name	Sales	Comm %	Commission
Applebee	$ 171,900	6%	$ 10,314
Bueller	$ 93,500	4%	$ 3,740
Chung	$ 151,200	6%	$ 9,072
Crawford	$ 119,850	5%	$ 5,993
Joyce	$ 89,450	4%	$ 3,578
King	$ 124,500	5%	$ 6,225
MacDonald	$ 131,100	6%	$ 7,866
Tanaka	$ 201,500	9%	$ 18,135

Sales	Comm %
$ 50,000	3%
$ 75,000	4%
$ 100,000	5%
$ 125,000	6%
$ 175,000	8%
$ 200,000	9%

VLOOKUP drops back to next largest value

=VLOOKUP(C5,$G$5:$H$10,2,TRUE)

Multiple Columns Combination

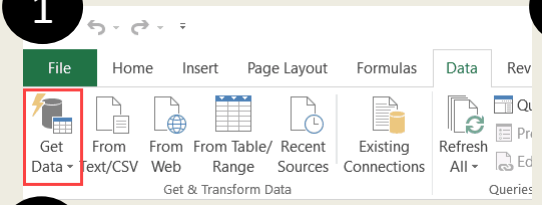
- MS Excel Power Query (Windows Only)
  - Add-on in Excel 2010-2013
  - Built-in in Excel 2016-2019
- Requirements
  - Tables to combine have to be saved in a document
  - Combine in a new document

Multiple Columns Combination

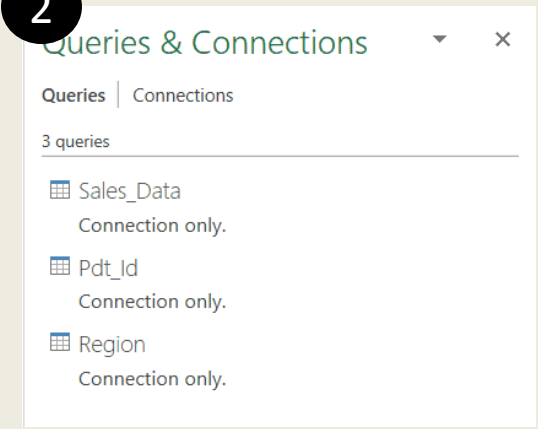
- MS Excel Power Query, in Data tab:
  1. File 1:
     - Get Data/New Query> From File > From [Workbook/CSV]
     - Select your first file and the corresponding sheet > Load
  2. File 2:
     - Get Data/New Query> From File > From [Workbook/CSV]
     - Select your second file and the corresponding sheet > Load
  3. Get Data/New Query > Combine Queries > Merge
  4. Identify the tables to merge/append and select the column to use
  5. Click Columns to Expend
  6. Untick "Use original column name as prefix"
  7. OK > Close and Load

Multiple Columns Combination

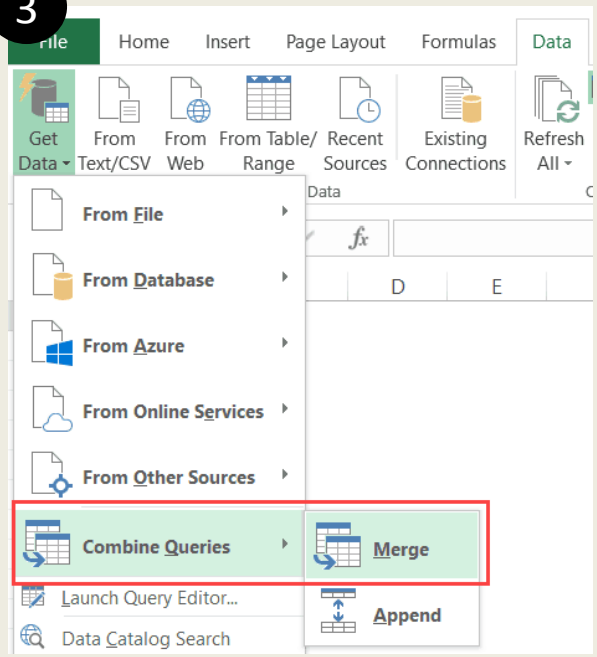
1



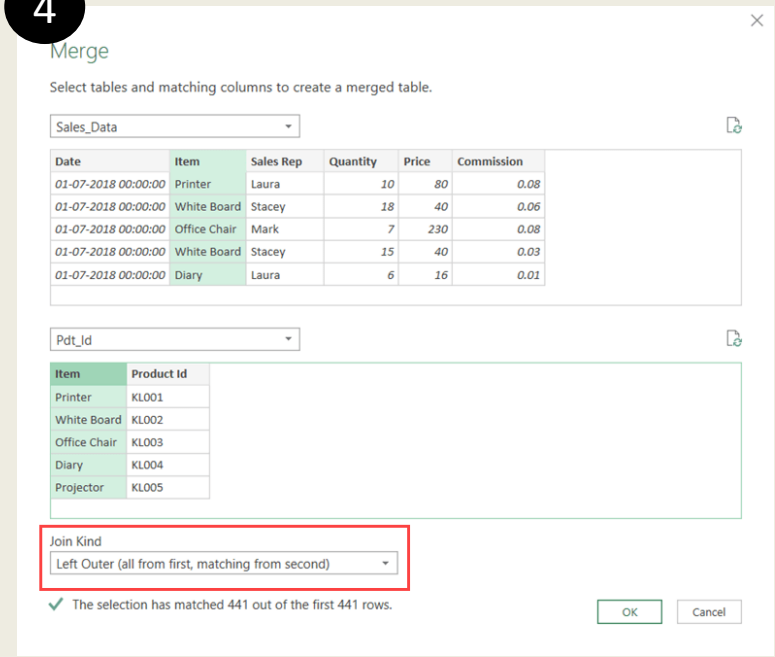
2



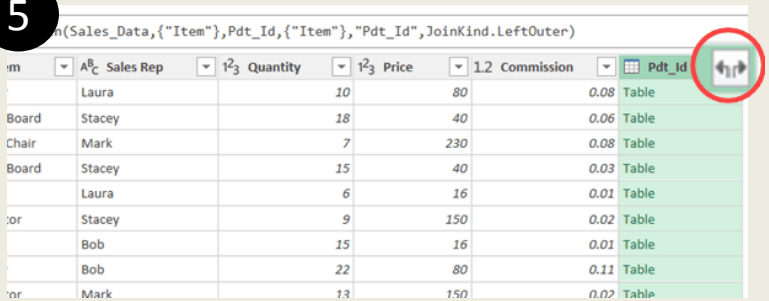
3



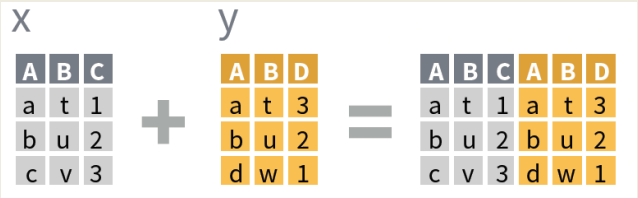
4



5



Types of Combination



- Left Join

A	B	C	D
a	t	1	3
b	u	2	2
c	v	3	NA

- Right Join

A	B	C	D
a	t	1	3
b	u	2	2
d	w	NA	1

- Inner Join

A	B	C	D
a	t	1	3
b	u	2	2

- Full Join

A	B	C	D
a	t	1	3
b	u	2	2
c	v	3	NA
d	w	NA	1

EXERCISE: TRANSFORMATIONS

Exercise: Transformations

On the MT5125 Loop page, download and open the document "employee_d&d_excerpt.xls" located in:

- Data Analytics Supplementary Information> Lecture 2

1. **Extension**: Create a new variable/column which is the average response to all the questions from the survey for each employee (q1 to q9)

Exercise: Transformations

2. **Reduction**: Filter employee's 2019 salary to keep only employees with a salary higher than 30k
3. **Aggregation**: Calculate the average salary by gender and by location
4. **Combination**: Using the VLOOKUP function, add to the table a column corresponding to the 2017 salary located in the 2nd sheet

EXTRA ANALYTIC TIP

Tidy Data

Levels

Database

1

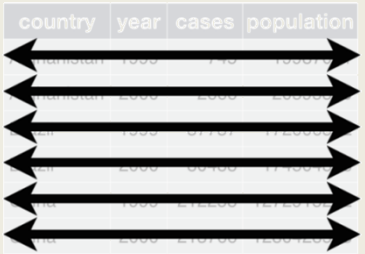
2 Tables/Files

3 Records/Observations

4 Fields/Variables

5 Characters (bytes)

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

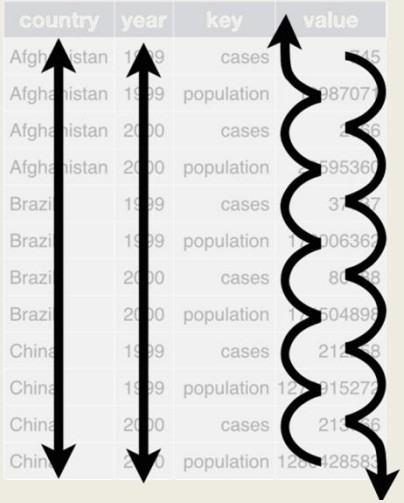


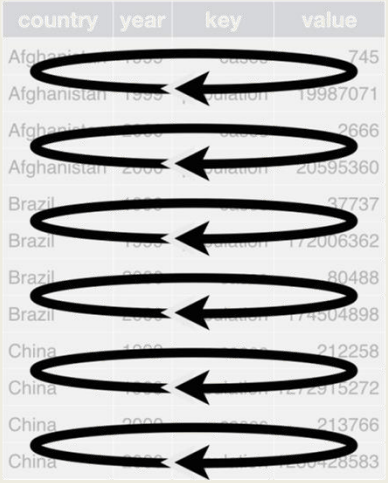
1. Each variable has its own column
2. Each observation is placed in its own row
3. Each value is placed in its own cell

Long or Wide?

- Long format

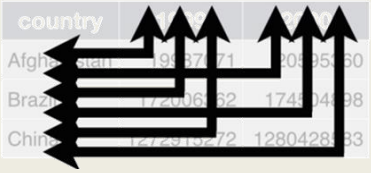
country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583





- Wide format

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583



Long or Wide?

- Which type is this table?

	A	B	C	D	E
1	Region	Qtr1	Qtr2	Qtr3	Qtr4
2	Mid West	2924300	3422700	2318100	2234000
3	North East	1455100	1422700	498200	1786900
4	South	4684000	6220500	5202600	5118700
5	West	2625200	3161400	2810000	2972900

Long or Wide?

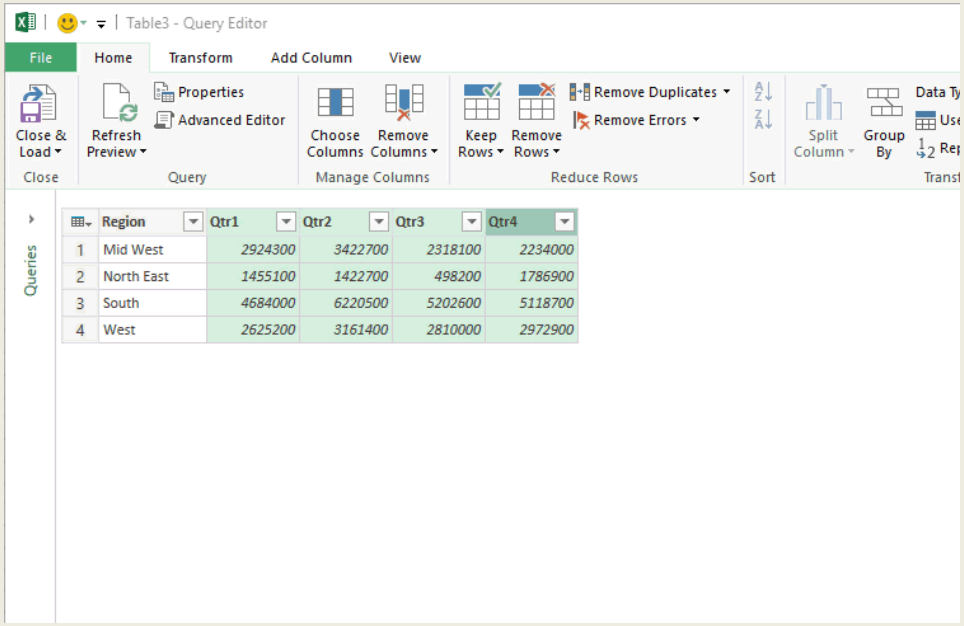
	A	B	C	D	E	F	G	H	I
1	Region	Qtr1	Qtr2	Qtr3	Qtr4		Region	Quarter	Sales
2	Mid West	2924300	3422700	2318100	2234000		Mid West	Qtr1	2924300
3	North East	1455100	1422700	498200	1786900		Mid West	Qtr2	3422700
4	South	4684000	6220500	5202600	5118700		Mid West	Qtr3	2318100
5	West	2625200	3161400	2810000	2972900		Mid West	Qtr4	2234000
6							North East	Qtr1	1455100
7							North East	Qtr2	1422700
8							North East	Qtr3	498200
9							North East	Qtr4	1786900
10							South	Qtr1	4684000
11							South	Qtr2	6220500
12							South	Qtr3	5202600
13							South	Qtr4	5118700
14							West	Qtr1	2625200
15							West	Qtr2	3161400
16							West	Qtr3	2810000
17							West	Qtr4	2972900

Reshape Table

- In Data tab
  - Get Data/New Query> From File > From [Workbook/CSV]
  - Select your file > Edit
  - Select columns to be reshaped
  - Transform
    - Pivot Columns: from long table to wide table
    - Unpivot Columns: from wide to long table

Reshape Table

- Example: from wide table to long table



	Region	Qtr1	Qtr2	Qtr3	Qtr4
1	Mid West	2924300	3422700	2318100	2234000
2	North East	1455100	1422700	498200	1786900
3	South	4684000	6220500	5202600	5118700
4	West	2625200	3161400	2810000	2972900

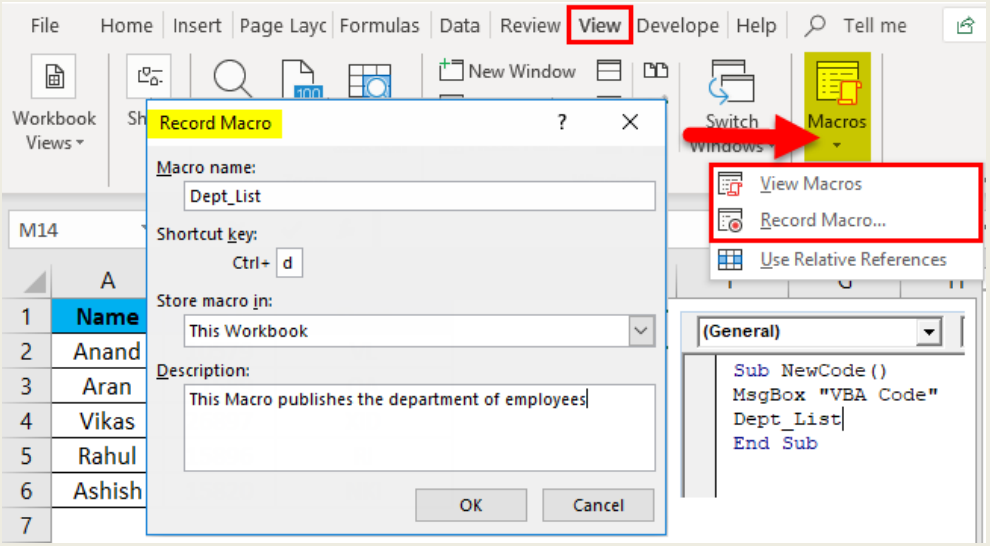
- See for more details:
  - https://trumpexcel.com/source-data-for-pivot-table/

MS Excel's Macro

- The Macro button allows to record a sequence of actions and to reproduce these actions
- VBA Code automatically recorded
- Useful to process similar data files

MS Excel's Macro

- In View tab
  1. Use Macros > Record Macro
  2. Do your actions
  3. Save the macro with a keyboard shortcut
  4. Use the macro again to reproduce your actions




DATABASE
USERS
ID
NAME
SURNAME
NICKNAME
EMAIL
PASSWORD
ORDERS
ID
ORDER_TIME
USER_ID
PRODUCT_ID
PRODUCTS
ID
NAME
DESCRIPTION
PRICE
PICTURE
ENABLED
CODE
IN-STOCK
CATEGORY_ID
CATEGORY
ID
NAME
DESCRIPTION
WEB SERVER
FIREWALL
QUESTIONS?