MITS6005

Big Data

Session 2

Acquire & Transform Data

Class objectives

- Demonstrate knowledge of data quality concepts
- Demonstrate knowledge of key terms and capabilities in Excel
- Demonstrate how to use Excel to acquire, transform, analyze, and visualize data
- Demonstrate knowledge of leading practices for presenting findings in Excel

Why excel?

- We have no choice… everybody uses it!
- Easy-to-use
- Many different functions
- Advanced capabilities through add-ins
- Easy to explore and manipulate the data
- Can present data and analysis together
- Integrated with common desktop applications

But…

- Significant potential for human error
- Limited scalability

Keep it simple and ask the Internet!

Key terms

Workbook	An Excel file is referred to as a workbook
Worksheet	A single “tab” or “sheet” within a Workbook
Cell	The boxes within the worksheet where information is stored. Cells are referenced by column letters (A, B, C, …) and row numbers (1, 2, 3, …) like a map: A1 BX800 EEE20
Range	A contiguous set of cells referenced by the top left and bottom right cells, separated by a colon (:): A1:D23 BA2:CT8

Cell references

- Excel formulas can reference different cells
- Changes to the referenced cells result in updates to value calculated by the formula

	A	B	C	D
1	**Outgoings:**	*01-Dec-11*		
2				
3	Income	2,500		
4			2,500	
5				
6	Rent	-1,200		
7	Utilities	-165		
8	Travel	-100		
9	Food	-420		
10				
11	Outgoings		=B6+B7+B8+B9	
12				
13	Remainder		615	

Cells referenced by a formula

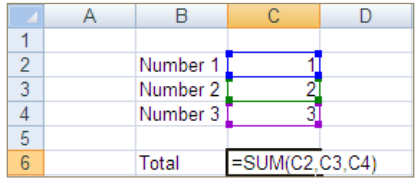
Formula

Cell references *(continued)*

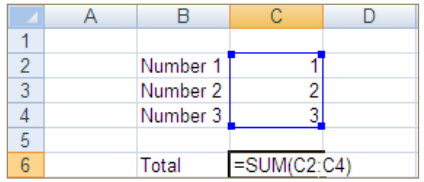
Inputs within functions can be either a single cell reference or a block of cells referred to as a range.

The two example below provide the same result:

=SUM(C1,C2,C3) Single cells selection – use the Ctrl Key



=SUM(C1:C3) Range selection – use Shift Key or Drag with mouse



Note: should a row be inserted between C1 and C3, then example 1 will still provide the same results, however example two will extend to add 4 cells.

Absolute cell references

Absolute referencing is a way of referring to cells within formulas so that once copied, the cell reference remains fixed to a particular cell

	A	B	C	D
1	Item	Price	VAT	Total
2	A	£ 4.99	£	=B2+C2
3	B	£ 5.99	£	
4	C	£ 7.99	£	
5				

Relative referencing

As a formula is copied along, the row/column numbers adjust accordingly

	A	B	C	D
1	Item	Price	VAT	Total
2	A	£ 4.99	=B2*$B$6	£ 5.49
3	B	£ 5.99	£ 0.60	
4	C	£ 7.99	£ 0.80	
5				
6	Vat Rate	10%		

Absolute referencing

As a formula is copied along, it continues to refer to the same cell as before

Absolute cell referencing

With absolute cell referencing, a dollar sign ($) appears in the cell reference:

=B4 Refers to column B and row 4,

but this will vary if the formula is copied across a range of cells

(relative reference)

=$B4 Will always refer to column B, but row reference can vary

=B$4 Will always refer to row 4, but column reference can vary

=$B$4 Will always refer to column B and row 4

(absolute reference)

Keyboard shortcut F4 cycles through the four absolute cell reference options

Formulas

- Excel calculations are specified with formulas in each cell
- To create a formula, type an equals sign (=) followed by the function and required arguments

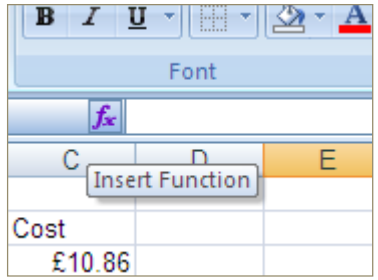
Order of operations:

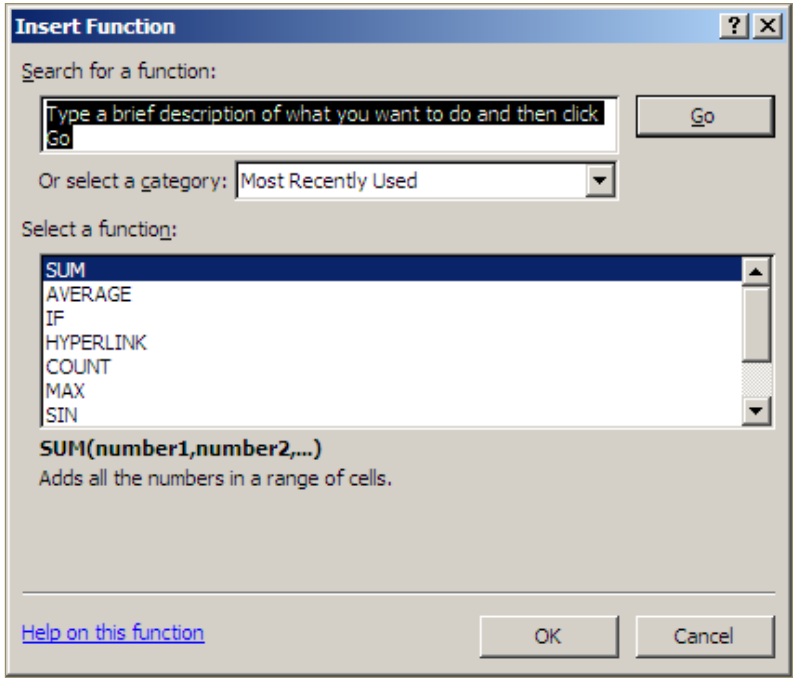
- Brackets ( )
- Exponent ^
- Division/Multiplication / *
- Addition/Subtraction + -

	A	B	C	D
1	**Outgoings:**	*01-Dec-11*		
2				
3	Income	2,500		
4			2,500	
5				
6	Rent	-1,200		
7	Utilities	-165		
8	Travel	-100		
9	Food	-420		
10				
11	Outgoings		=-1200-165-100-420	
12				
13	Remainder		615	

Functions

- The ***fx*** button to the left of the formula bar opens a list of all available functions





Function arguments

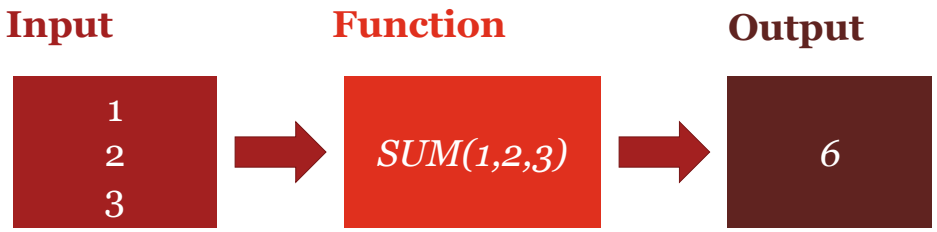
Functions take zero or more inputs or arguments

=FUNCTION*(Input1,Input2,…)*

Arguments can be:

Values	1, 2, 3 or 1.2, 3.6, 2.1 or “x”, “y”, “z” or TRUE, FALSE
Cell References	B91, A452, C3

	A	B	C	D	E
1					
2		Number 1	1		
3		Number 2	2		
4		Number 3	3		
5					
6		Total	6	=SUM(C2,C3,C4)	
7					



Navigation shortcuts (Windows)

Ctrl + → (← ↑ ↓)	navigate to last non-empty cell in that direction
Ctrl + Shift + → (← ↑ ↓)	select block of cells to last non-empty cell in that direction
Ctrl + Space	select the entire column of the cell you have selected
Shift + Space	select the entire row of the cell you have selected

Acquire data

Getting data into excel

- Excel files have the extension .xlsx or .xls (older versions)
- Excel can also import data from delimited text files
  - Some comma-separated values files can be opened directly by Excel
  - Tab-separated values can be copy-and-pasted directly into an Excel worksheet
  - Delimited text can also be copy-and-pasted into Excel and then separated with text-to-columns

Name	Date modified	Type	Size
General_Ledger_Account_Balances.csv	4/3/2017 4:50 PM	CSV File	3,649 KB
Rodent_Inspection.csv	4/3/2017 4:51 PM	CSV File	262,799 KB
New_York_City_Farmers_Markets.csv	4/3/2017 4:51 PM	CSV File	22 KB
FY_2017_PMMR_Data_Extract.csv	4/3/2017 4:51 PM	CSV File	210 KB
Inspections.csv	4/3/2017 4:51 PM	CSV File	1,892 KB
Bid_Tabulations.csv	4/3/2017 4:51 PM	CSV File	2,409 KB

Example – Delimited text

header

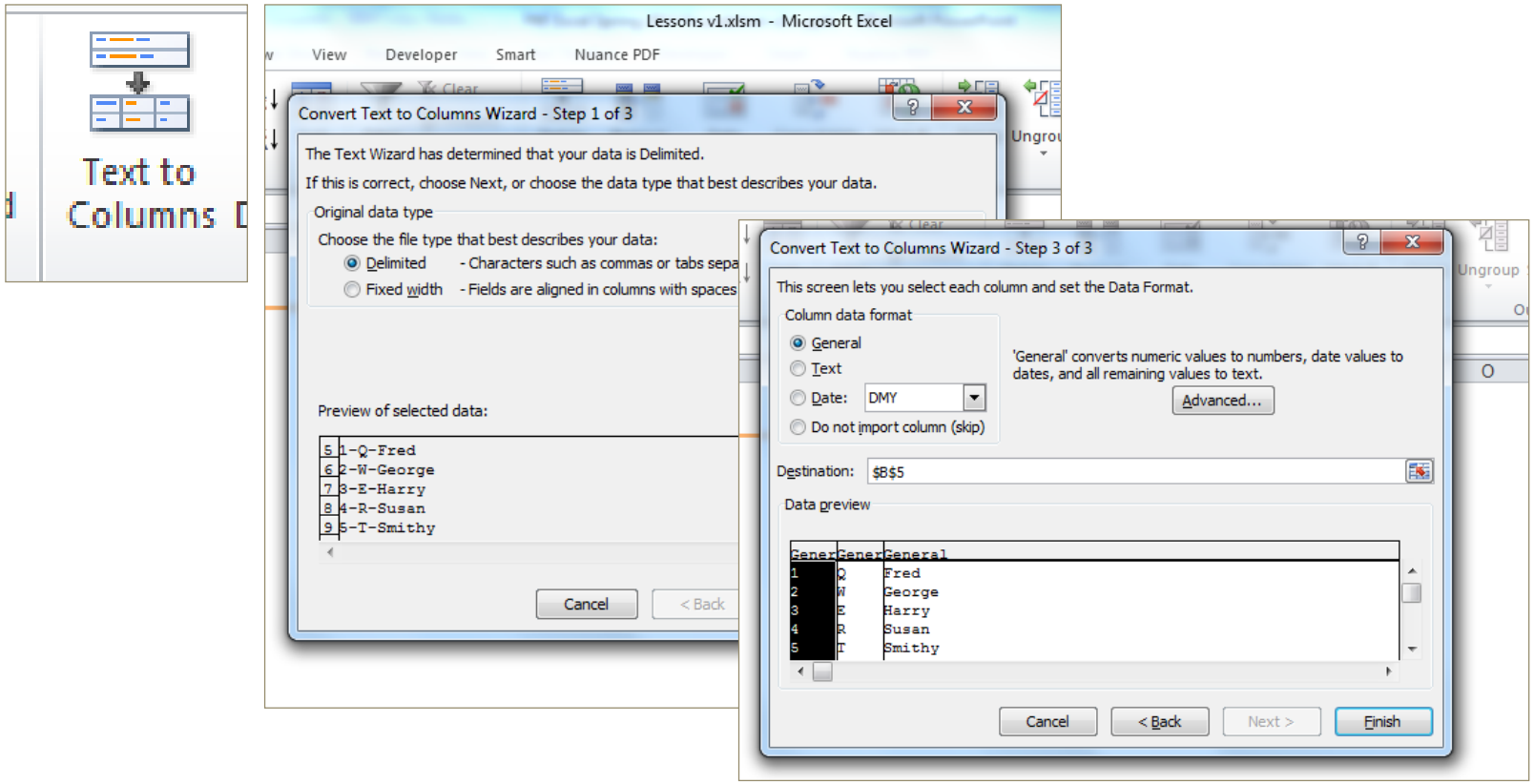
```
DBN,SCHOOL NAME,Num of SAT Test Takers,SAT Critical Reading Avg. Score,SAT Math Avg. Score,SAT Writing Avg. Score
01M292,HENRY STREET SCHOOL FOR INTERNATIONAL STUDIES,29,355,404,363
01M448,UNIVERSITY NEIGHBORHOOD HIGH SCHOOL,91,383,423,366
01M450,EAST SIDE COMMUNITY SCHOOL,70,377,402,370
01M458,FORSYTH SATELLITE ACADEMY,7,414,401,359
01M509,MARTA VALLE HIGH SCHOOL,44,390,433,384
01M515,LOWER EAST SIDE PREPARATORY HIGH SCHOOL,112,332,557,316
01M539,"NEW EXPLORATIONS INTO SCIENCE, TECHNOLOGY AND MATH HIGH SCHOOL",159,522,574,525
01M650,CASCADES HIGH SCHOOL,18,417,418,411
01M696,BARD HIGH SCHOOL EARLY COLLEGE,130,624,604,628
02M047,47 THE AMERICAN SIGN LANGUAGE AND ENGLISH SECONDARY SCHOOL,16,395,400,387
02M288,FOOD AND FINANCE HIGH SCHOOL,62,409,393,392
02M294,ESSEX STREET ACADEMY,53,394,384,378
02M296,HIGH SCHOOL OF HOSPITALITY MANAGEMENT,58,374,375,362
02M298,PACE HIGH SCHOOL,85,423,438,432
02M300,"URBAN ASSEMBLY SCHOOL OF DESIGN AND CONSTRUCTION, THE",48,404,449,416
02M303,"FACING HISTORY SCHOOL, THE",76,353,358,340
02M305,"URBAN ASSEMBLY ACADEMY OF GOVERNMENT AND LAW, THE",50,375,388,385
02M308,LOWER MANHATTAN ARTS ACADEMY,40,403,392,405
02M313,"JAMES BALDWIN SCHOOL, THE: A SCHOOL FOR EXPEDITIONARY LEARNING",69,408,390,390
02M316,"URBAN ASSEMBLY SCHOOL OF BUSINESS FOR YOUNG WOMEN, THE",42,373,370,384
02M374,GRAMERCY ARTS HIGH SCHOOL,60,391,391,394
02M376,NYC ISCHOOL,92,473,483,479
```

delimiter

text qualifier

Text to columns

- You might have text in one column that should be split across multiple columns, such as "1-Q-Fred" in the example below
- You can use the "Text to Columns" wizard to accomplish this



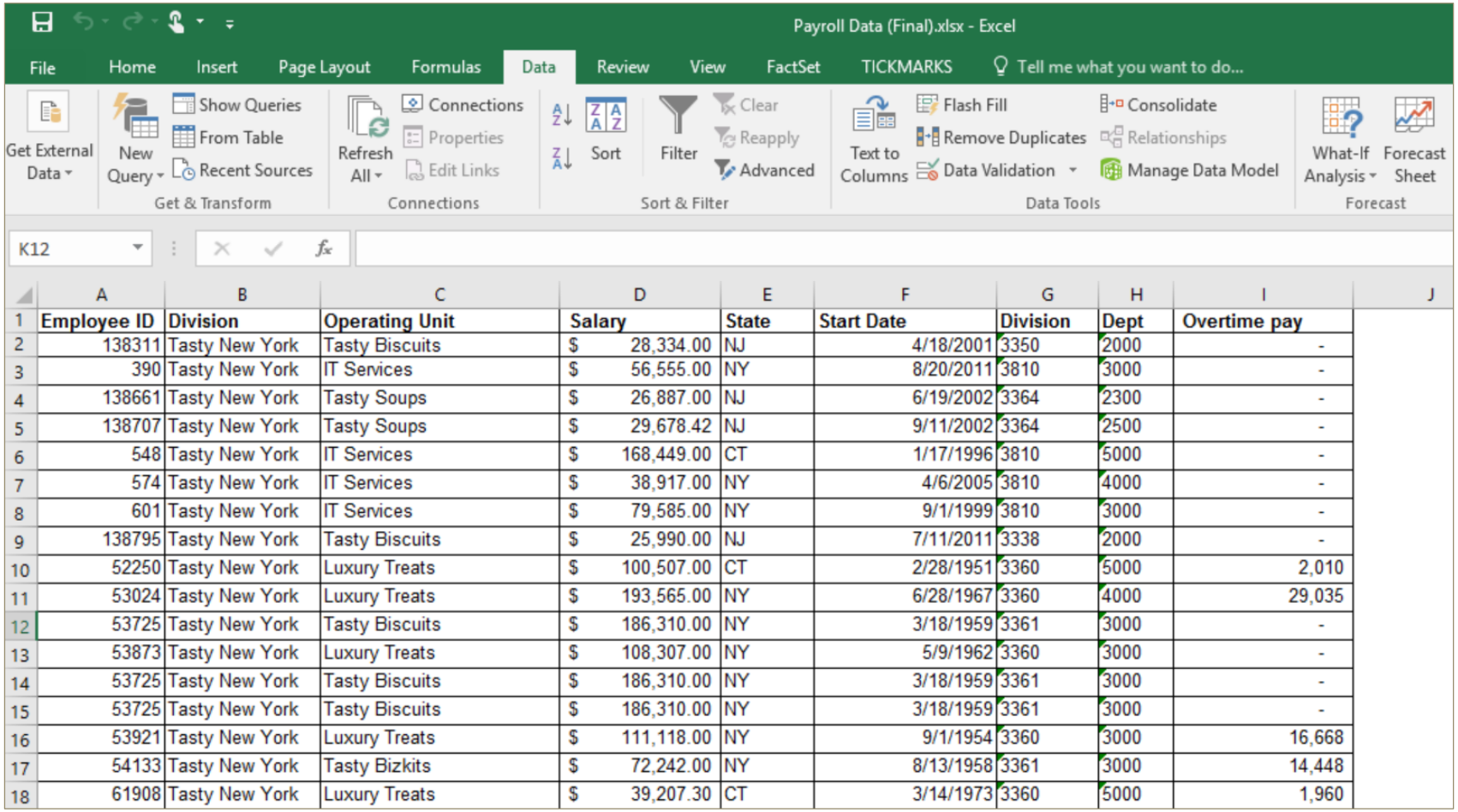
References to worksheets and workbooks

- You can also use data from other worksheets or other workbooks in your formulas
- Another worksheet in the same workbook is represented by “Sheet!”
- A worksheet in a different workbook is represented by “[Workbook.xlsx]Sheet!”

Company	=Data!A59

Company	Visa Inc.
Income	=[External_Data.xlsx]Sheet1!$D$19

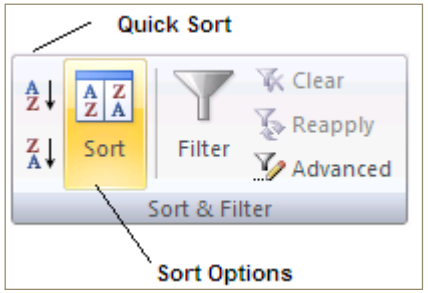
Sample spreadsheet

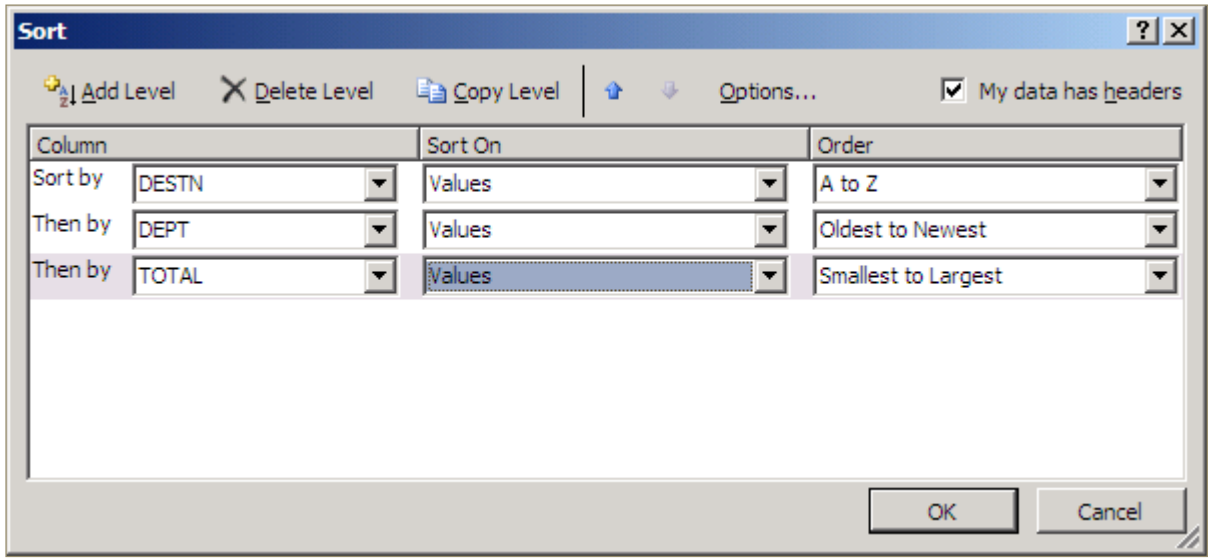


	A	B	C	D	E	F	G	H	I	J
1	Employee ID	Division	Operating Unit	Salary	State	Start Date	Division	Dept	Overtime pay	
2	138311	Tasty New York	Tasty Biscuits	$ 28,334.00	NJ	4/18/2001	3350	2000	-	
3	390	Tasty New York	IT Services	$ 56,555.00	NY	8/20/2011	3810	3000	-	
4	138661	Tasty New York	Tasty Soups	$ 26,887.00	NJ	6/19/2002	3364	2300	-	
5	138707	Tasty New York	Tasty Soups	$ 29,678.42	NJ	9/11/2002	3364	2500	-	
6	548	Tasty New York	IT Services	$ 168,449.00	CT	1/17/1996	3810	5000	-	
7	574	Tasty New York	IT Services	$ 38,917.00	NY	4/6/2005	3810	4000	-	
8	601	Tasty New York	IT Services	$ 79,585.00	NY	9/1/1999	3810	3000	-	
9	138795	Tasty New York	Tasty Biscuits	$ 25,990.00	NJ	7/11/2011	3338	2000	-	
10	52250	Tasty New York	Luxury Treats	$ 100,507.00	CT	2/28/1951	3360	5000	2,010	
11	53024	Tasty New York	Luxury Treats	$ 193,565.00	NY	6/28/1967	3360	4000	29,035	
12	53725	Tasty New York	Tasty Biscuits	$ 186,310.00	NY	3/18/1959	3361	3000	-	
13	53873	Tasty New York	Luxury Treats	$ 108,307.00	NY	5/9/1962	3360	3000	-	
14	53725	Tasty New York	Tasty Biscuits	$ 186,310.00	NY	3/18/1959	3361	3000	-	
15	53725	Tasty New York	Tasty Biscuits	$ 186,310.00	NY	3/18/1959	3361	3000	-	
16	53921	Tasty New York	Luxury Treats	$ 111,118.00	NY	9/1/1954	3360	3000	16,668	
17	54133	Tasty New York	Tasty Bizkits	$ 72,242.00	NY	8/13/1958	3361	3000	14,448	
18	61908	Tasty New York	Luxury Treats	$ 39,207.30	CT	3/14/1973	3360	5000	1,960	

Sorting

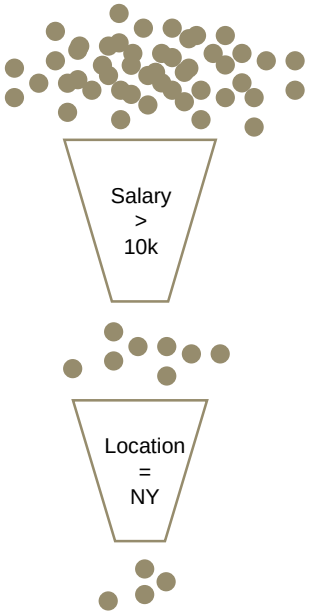
- Highlight a range and use the Sort option to order records by one or more columns
- Make sure to select an entire table to avoid sorting only part of it

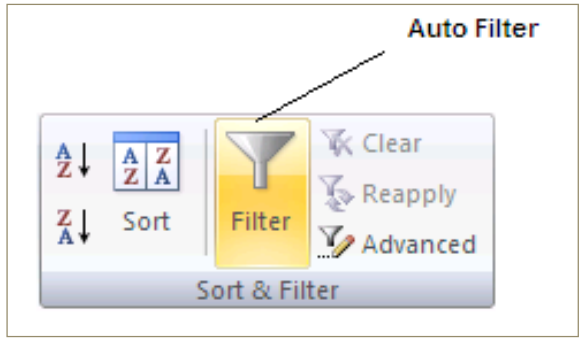




Filters

- Filtering allows you to hide rows in a range if they don't match select criteria
- Highlight the range and click the "Filter" button

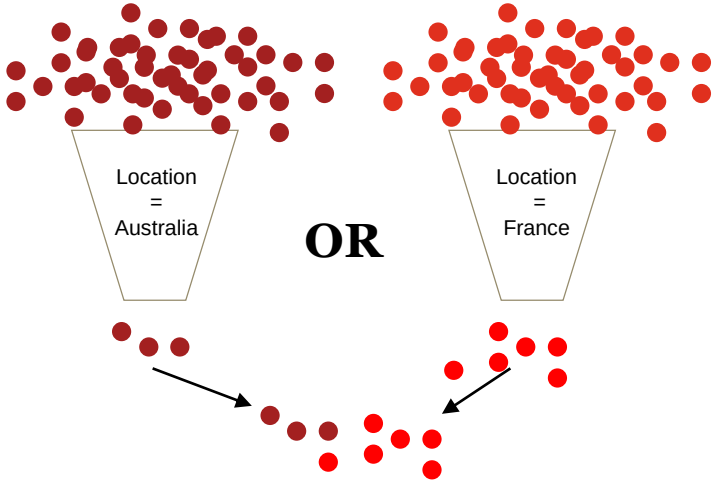


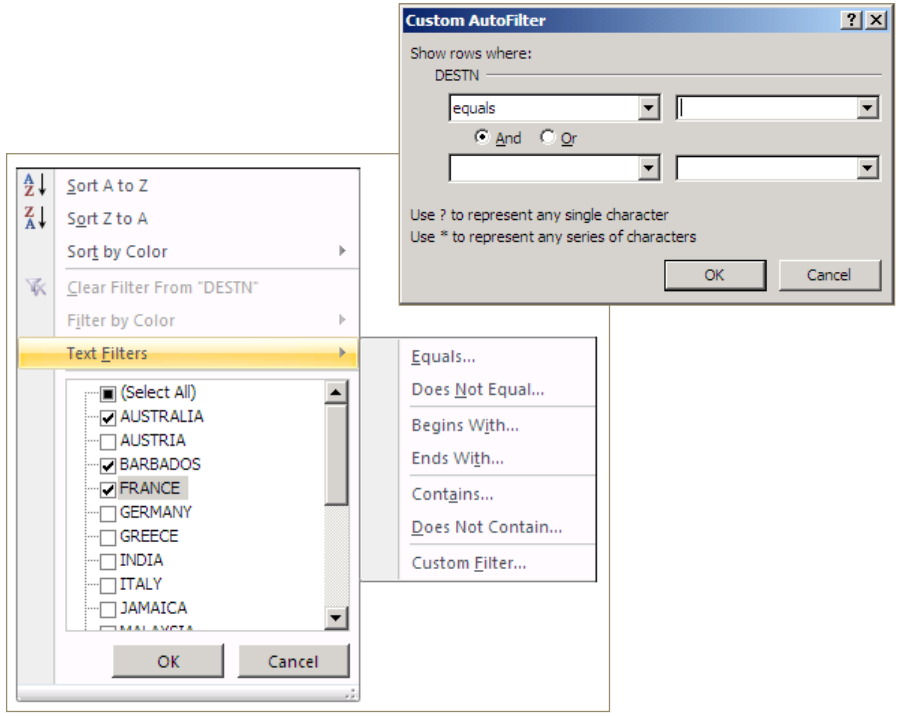


Custom filters

Custom filters allow you to use multiple criteria such as AND or OR

- Click on drop down arrow
- Click “Text Filters”
- Choose Custom





Summary statistics

COUNT()
Counts numbers

COUNTBLANK()
Counts empty cells

COUNTA()
Counts non-empty cells

MIN()
Returns minimum

MAX()
Returns maximum

AVERAGE()
Returns average/mean

ROWS/COLUMNS()
Return number of rows/columns

	A	B	C	D	E	F	G
1							
2		X	Y		A	B	
3		1	2	3	4	5	
4							
5							
6		Number of Values		5	=COUNT(B3:F3)		
7							
8							
9							
10		Number of Labels		4	=COUNTA(B2:F2)		
11							
12							
13							
14		Minimum Value		1	=MIN(B3:F3)		
15							
16							
17							
18		Maximum Value		5	=MAX(B3:F3)		
19							
20							
21							
22		Average Value		3	=AVERAGE(B3:F3)		
23							
24							

Status bar

- Highlighting a range of cells returns summary statistics on the status bar in the lower right corner
- This is a quick way to get information about part of your data

Operating Unit	Basic Pay	State	Start Date	Div
Tasty Biscuits	$ 28,334.00	NJ	4/18/2001	335
IT Services	$ 56,555.00	NY	8/20/2011	381
Tasty Soups	$ 26,887.00	NJ	6/19/2002	336
Tasty Soups	$ 29,678.42	NJ	9/11/2002	336
IT Services	$ 168,449.00	CT	1/17/1996	381
IT Services	$ 38,917.00		4/6/2005	381
IT Services	$ 79,585.00	NY	9/1/1999	381
Tasty Biscuits	$ 25,990.00	NJ	7/11/2011	333

100010	$ 11,190.07	$ 5,595.03	18,586
100118	$ 4,609.63	$ 2,304.82	10,386
100014	$ 19,871.14	$ 9,935.57	-

AVERAGE: $61,980.68 COUNT: 5 SUM: $309,903.42

Can we rely on the data?

Why does 'Booking Country' have two country codes while 'Originator Country' is blank ?

Booking Country	Originator Countr
USUK	
USUK	
CHUK	
USUK	
USUK	
SGCH	
SG	
SG	
SGUS	
USCH	
CHCH	
USUS	
USUK	
US	
UKCH	
SGCH	
CHCH	
SG	
CHCH	
USCH	
USUS	

K	L	M
Originator Name	Beneficiary Country	Beneficiary Nam
70. Wilbert Floyd	CH	28. Dixie Lane
2. Alfredo Lyons	SG	33. Elbert Castro
60. Elvira Ramirez	US	100. Jeanette Hicks
65. Arthur Owens	CH	56. Stephen Mitchell
14. Alan Gardner	SG	46. Lowell Stephens
62. Jessie Cruz	UK	70. Velma Drake
66. Ivan Norman	SG	95. Celia Mullins
93. Michelle Dennis	SG	65. Eileen Munoz
16. Gwendolyn Fitzgerald		79. Lawrence Rodger
7. Marguerite Scott	XX	49. Marsha Holloway
17. Trevor Bailey	US	33. Elbert Castro
5. Leslie Hunt	US	66. Lyle May
44. Clayton Oliver	SG	62. David Ramsey
78. Tami Sullivan	US	14. Jason Hoffman
86. Emma Dean	SG	54. Nathan Turner

Does 'Beneficiary Country' have a valid country code ?

ary Cust	Transaction Date
7	04/12/2005
7	23/01/2010
21	26/11/2009
13	14/04/2005
1	02/07/2008
10	30/11/2004
10	09/09/2004
19	03/10/2005
26	02/07/2009
26	08/01/2008
5	15/04/2013
2	02/09/2009
26	12/10/1900
19	08/10/2008
15	22/01/2004
4	30/10/2004

Does this look like a valid transaction date?

Does this look like a valid customer name?

A	B	C	
Customer ID	First name	Last Name	Ge
1	Echo	Demming	F
2	Telma	Saulnier	F
3	Dorthey	Trotter	M
4	Mrs	J	F
5	Zachary	Goodall	F
6	Maribeth	Olson	F

Data quality dimensions

Exercise – Data quality dimensions

Identify which of the six DQ dimensions applies to the issues described below:

1. ‘Gender’ field has the special characters like ~!@#$%^*();
2. ‘First name’ is blank or Null
3. ‘Last name’ field has only designators such as LLP, LLC, Mr., Mrs., etc
4. ‘Address’ field has only numbers
5. ‘Account Type’ field does not have pre-defined list of values
6. ‘Account Number’ field have duplicate values
7. ‘Forex rate’ field does not have up to date exchange rate

Exercise – Excel #1

Load the data from ‘payroll_data.txt’ and ‘reference_data.txt’ into two tabs in an Excel spreadsheet and save it as ‘Payroll Data.xlsx’:

1. How did you import the data?
2. How many rows are there?
3. How many columns?
4. What is the average salary?
5. How many missing values are in each column?
6. What potential data quality issues do you find? Which dimensions are they related to?

Transform data

Data formats

Data entered an Excel worksheet can be represented in different formats, for example:

- Number
- Currency
- Percentage
- Text
- Date

Dates are stored as the number of days from a fixed historical date

	A	B	C	D	E
1	Salary (Currency)	Salary (Number)	State (Text)	Start Date (Short Date)	Start Date (Number)
2	$ 28,334.00	28334.00	NJ	4/18/2001	36999
3	$ 56,555.00	56555.00	NY	8/20/2011	40775
4	$ 26,887.00	26887.00	NJ	6/19/2002	37426
5	$ 29,678.42	29678.42	NJ	9/11/2002	37510
6	$ 168,449.00	168449.00	CT	1/17/1996	35081
7	$ 38,917.00	38917.00	NY	4/6/2005	38448
8	$ 79,585.00	79585.00	NY	9/1/1999	36404

Text functions

=CONCATENATE (text1, text2, ...) or &

Appends two or more strings

=MID(text, start_num, num_chars)

Extracts a specific number of characters starting from a given position

=LEFT(text, num_chars)

Extracts a specific number of characters from the left of a cell

=RIGHT(text, num_chars)

Extracts a specific number of characters from the right of a cell

=LEN(text)

Return the number of characters (including spaces) in a cell

=SUBSTITUTE(text, old_text, new_text)

Replaces a string with another string

=FIND(find_text, within_text, start_position)

Return the position of a match and error if no match

Number functions

=ROUND(number, num_digits)

Rounds a figure to a specified number of digits

=LARGE(array, k)

Returns the kth largest number

=SMALL(array, k)

Returns the kth smallest number

=PRODUCT(number1, number2, ...)

Multiplies several values together

=EXP(number)

Returns Euler's number (e) raised to a number

=RAND()

Returns a random number between 0 and 1

=RANDBETWEEN(bottom, top)

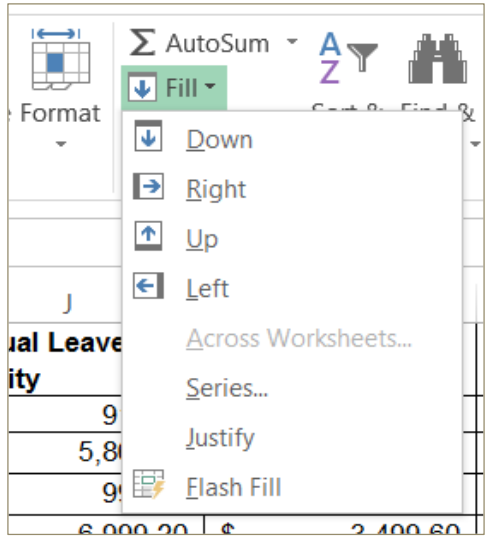
Returns a random integer between the specified values

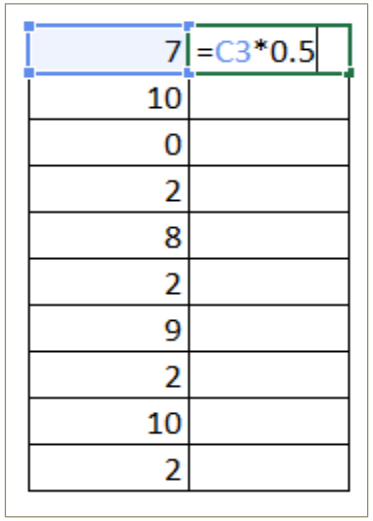
Date/time functions

=TODAY()	Displays current date
=NOW()	Displays current date and time
=DATE(year, month, day)	Generates a date given day, month, and year
=DAY(serial_number)	Determines day of date, e.g. 31
=MONTH(serial_number)	Determines month of date, e.g. 12
=YEAR(serial_number)	Determines year of date, e.g. 2001
=WEEKNUM(serial_number)	Returns the week number in the year
=WEEKDAY(serial_number)	Returns the position of the day in a workweek
=DAYS(end_date, start_date)	
Returns number of days between two dates	
=NETWORKDAYS(start_date, end_date)	
Returns the number of working days between dates	

Fill

- You can apply a formula to an entire column by using the Fill options
- You can also Fill Down by double-clicking a cell with a formula in a table





7	3.5
10	5
0	0
2	1
8	4
2	1
9	4.5
2	1
10	5
2	1

Error trapping

=ISNA(value)

Determine if cell or result of formula is showing #N/A!

=ISERROR(value)

Determine if cell or result of formula is an error

=IFERROR(value, value_if_error)

Give an alternative result if the formula produces an error

Total Fee	Days Worked	Rate per Day	*(using isError)*
£20,000.00	240	£ 83.33	=IFERROR(G11/H11,"No Data")
£31,000.00	320	£ 96.88	£ 96.88
£29,500.00	110	£ 268.18	£ 268.18
£ -	0	#DIV/0!	No Data

IF statements

- **IF()** analyses the contents of one or more cells to determine whether a condition is TRUE or FALSE
- Given the value of the condition, different values can be returned

Syntax

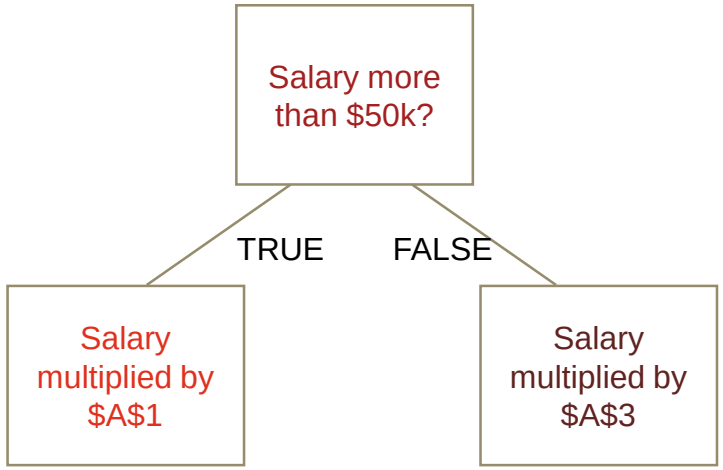
=IF (logical_test , value_if_true , value_if_false)

Examples

=IF (B1 > 50 , B1 * $A$1, B1 * $A$3)

=IF (B1 > B2 , B1 , B2)

=IF (B1 = “Male” , B1 , B2)



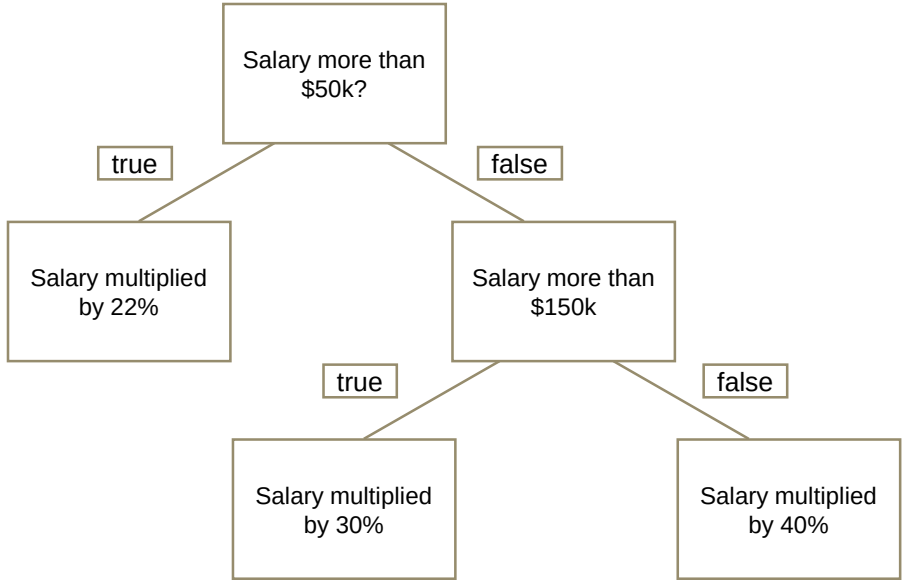
Conditional operators

Any function can be used with conditional operators in IF statements to test conditions

>	Greater than
<	Less than
>=	Greater than and equal to
<=	Less than and equal to
<>	Not equal to
=	Equal to

Nested conditional statements

=IF (logical_test , value_if_true , IF (logical_test , value_if_true , value_if_false))



=IF (B1 > 50 , B1 * $A$1, IF (B1 > 150 , B1 * $A$2, B1 * $A$3))

Logical operators

=AND(logical1, logical2, ...)	Gives **True** if all the conditions are met
=OR(logical1, logical2, ...)	Gives **True** if at least one condition is met
=NOT(logical)	Gives **True** if the condition is not met

=IF(AND(B2="Global",C2="Executive"),"Yes","No")

=IF(OR(C2="Consultant",C2="Director"),"Yes","No")

=IF(NOT(A2=$G$27),"Yes","No")

	A	B	C	D	E	F
1	Name	Market Unit	Grade	Global Executive?	Consultant or Director?	Not I Parker
2	C Anderson Barker	CIPS - FE	Consultant	No	Yes	Yes
3	I Parker	Global	Consultant	No	Yes	Yes
4	J Bloggs	CIPS - FE	Director	No	Yes	Yes
5	P Smith	Global	Executive	Yes	No	Yes
6	K Oscar	Global	Executive	Yes	No	Yes
7	P Jones	CIPS - FE	Executive	No	No	Yes
8	A Catford	CIPS - FE	Consultant	No	Yes	Yes
9	B Collins	CIPS - FE	Director	No	Yes	Yes
10	J Simmons	CIPS - FE	Director	No	Yes	Yes
11	K Jenkins	Global	Consultant	No	Yes	Yes

VLOOKUP

Often we need to add data from one table to another

This requires one column in each table to act as the link between them

VLOOKUP searches vertically down the left-hand column of a table to find a match, then returns the corresponding value from a specified column of the table

In most cases, the fourth argument should be 'FALSE' to require exact matches

Always use absolute cell references to specify the range containing the lookup data

Syntax

=VLOOKUP(lookup_value, table_array, col_index_num, range_lookup)

Example

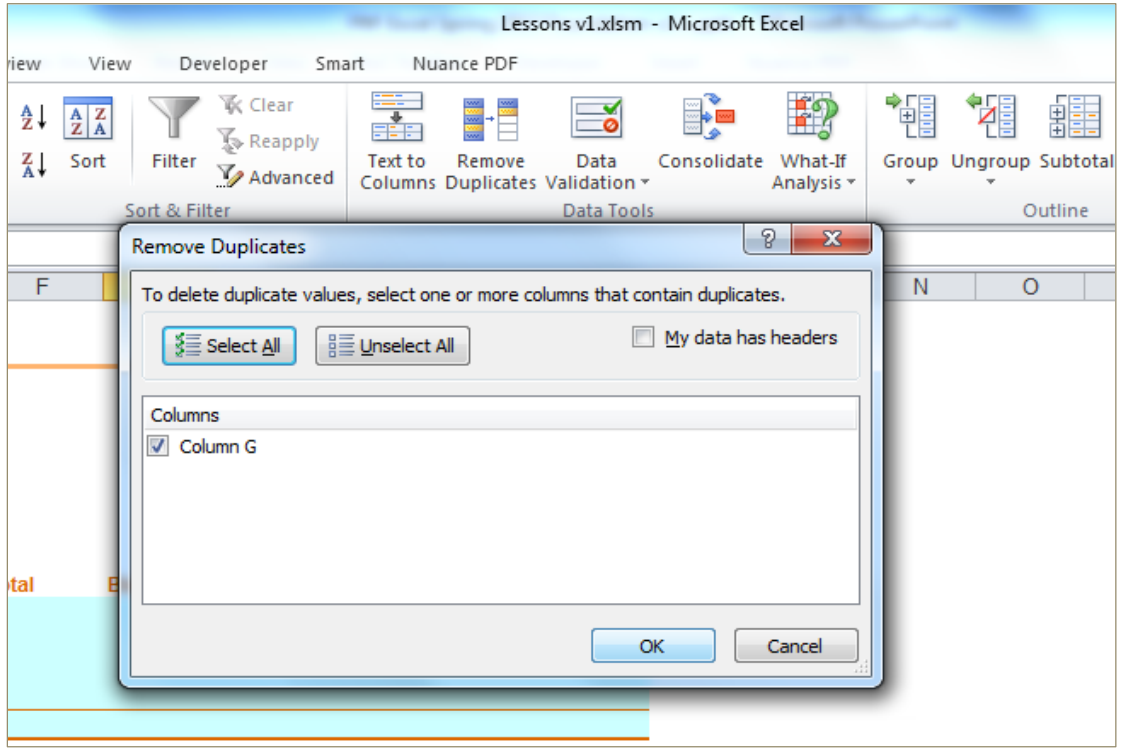
=VLOOKUP("Grade 5", $A$2:$B$6, 2, FALSE)

...returns a value of 46000, because this is the value in column 2 of the table alongside "Grade 5"

	A	B	C
1		Salary	
2	Grade 1	£30,000	
3	Grade 2	£32,000	
4	Grade 3	£35,000	
5	Grade 4	£40,000	
6	Grade 5	£46,000	
7			

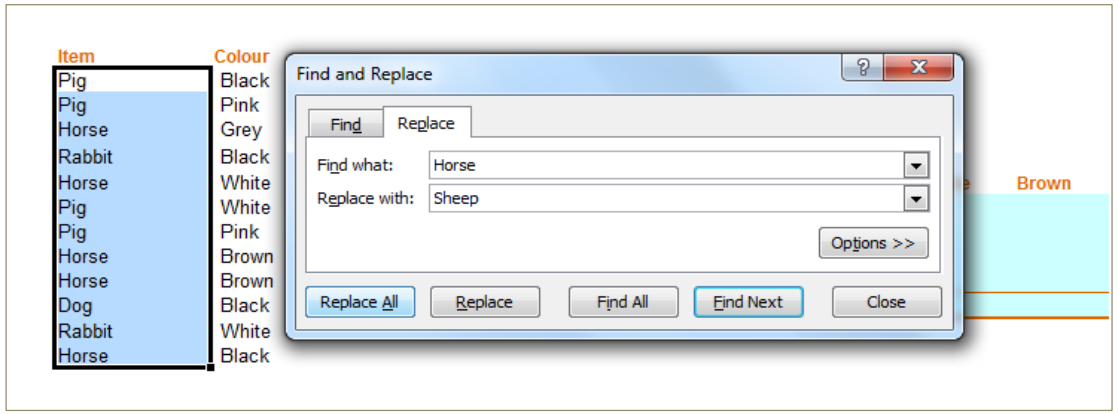
Unique values and duplicates

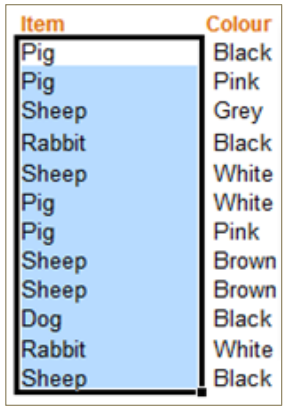
- Excel has an option to reduce a single column to its unique values
- This option can also remove duplicate rows from a table across multiple columns



Find/Replace

- You may need to find or replace values across a worksheet or workbook, which would be time-consuming if done manually
- Use Find/Replace to do this efficiently:
  - Ctrl + F: find
  - Ctrl + H: replace
- Excel will search the whole worksheet (or workbook) unless a range is selected
- The “Options >>” button enables the use of requirements such as “Match case”





Exercise #2

In ‘Payroll Data.xlsx’:

1. Address any data quality issues identified previously.
2. Create a new column with total compensation (salary plus overtime pay)… what is the average total compensation?
3. Add a column for the employee’s tenure at the company in years.
4. It turns out that all of the NJ employees work for separate division called “Tasty New Jersey”. Correct the Division column with this information.
5. Create a binary column (0/1) that identifies employees of the “Tasty” operating units.
6. Add columns for Job Position and Gender from the ‘Reference Data’ worksheet.
7. Select a random 5% sample of employees to receive a survey.