DATA WRANGLING/CLEANING USING EXCEL

Primitive data types:

- There are five main types of data types in Excel:
  - Booleans: (True/False, Male/Female, Default/Not Default,....)
  - Characters: (‘A’, ’c’, ‘D’,...)
  - Strings: (sets of characters: “Home”, “I like my dog”,...)
  - Integers: (-1, 2, 5, 400,...)
  - Floating point: (-13.5, 26.8, 13E6,...)

Data cleaning process

- Data cleaning is mandatory as in data analytics the main rule is:

  “Garbage in/ Garbage out”

- We need to clean our data, but this is a needed condition but not sufficient to guarantee a good model.

- **Every model may need a different preprocessing**.

Excel formulas

- Excel cells can return values from any other cell in the workbook
- Those values can be transformed using formulas
- In order to add a formula to an Excel cell, you need to add the “=” sign and then the name of the formula alongside it’s arguments
- Excel comes with many built-in formulas

Excel formulas to import data

- Data contained in another cell can be obtained using the syntax:
  - Range_fo_cells

- Data from other worksheets of the same workbook can be included in an Excel formula using the syntax:
  - **'Sheet_name'!**Range_of_cells

- In a similar way, if the data that we want to import is stored in a different workbook, we can use the syntax:
  - **'[Another_file.xlsx]Worksheet_name'!**Range_of_cells
  - **'[full_path_to_excel_file]Worksheet_name'!**Range of_cells (if the file is closed)

String functions in Excel

- **LEN(cell_name)**: provides the length of the characters found in “cell_name”
- **LEFT(cell_name, number_of_characters)**: returns “number_of_characters” of “cell_name starting from the left
- **RIGHT(cell_name, number_of_characters):** returns “number_of_characters” of “cell_name starting from the right
- **MID(cell_name, start, number_of_characters):** returns “number_of_characters” from “cell_name” starting from “start” position
- **FIND(cell_name, “string”)**: if “string” exists in “cell_name”, returns the starting position
- **IF (condition, value_if_true, value_if_false)**. It can be nested with another “IF” inside the function.

https://www.edupristine.com/blog/text-functions-excel

Data standarization

- **What** is it?
  - All the values need to follow the same format, units, code...

- **Why** is needed?
  - To allow any model to extract meaningful patterns.
  - **Different ranges** of values in each column **favor the column with the biggest range** in predicting the dependent variable. (the remaining columns are barely used in the prediction)

Functions **ImportRange** and **Indirect**

- **ImportRange:**
  This function allow us to insert data which is into another sheet of the workbook

- **Indirect(ref)**
  - This function imports the data from the cell referenced in the cell “ref”

	A	B
1	B1	1,333
2	B3	45
3	George	10
	=INDIRECT(A1)->1,333	

https://support.microsoft.com/en-us/office/indirect-function-474b3a3a-8a26-4f44-b491-92b6306fa261

Function **Vlookup**

- The VLOOKUP (Vertical lookup) is a function which given a value in one table, finds the corresponding value in another table with a common column

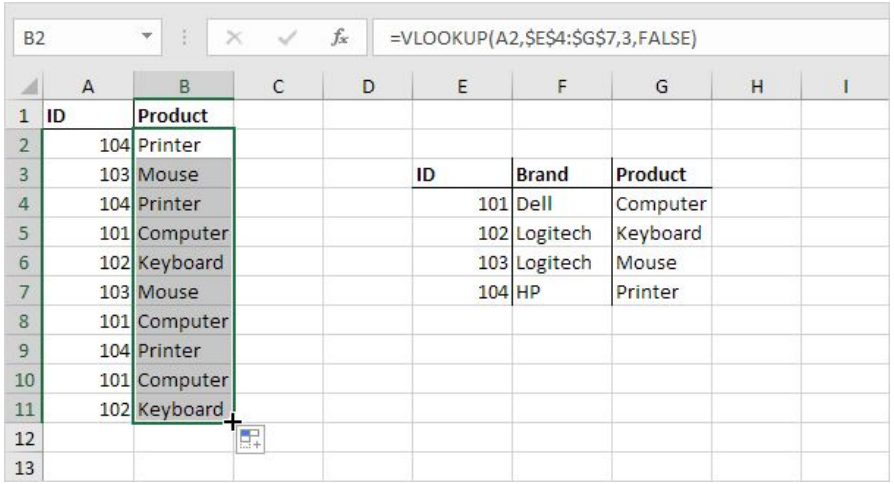
- VLOOKUP(value_to_search , range_for search, column_to_import, na_values)

https://www.excel-easy.com/functions/lookup-reference-functions.html

- In the example on the right image we would like fill out the “Product” column “G” according to the product “ID” with the values on column “B”

  VLOOKUP(E4,$A$1:$B$11,2,0)

- In case of duplicates, vlookup returns the first match.



	A	B	C	D	E	F	G	H	I
1	ID	Product							
2	104	Printer							
3	103	Mouse			ID	Brand	Product		
4	104	Printer			101	Dell	Computer		
5	101	Computer			102	Logitech	Keyboard		
6	102	Keyboard			103	Logitech	Mouse		
7	103	Mouse			104	HP	Printer		
8	101	Computer							
9	104	Printer							
10	101	Computer							
11	102	Keyboard							
12									
13									

Function **Hlookup**

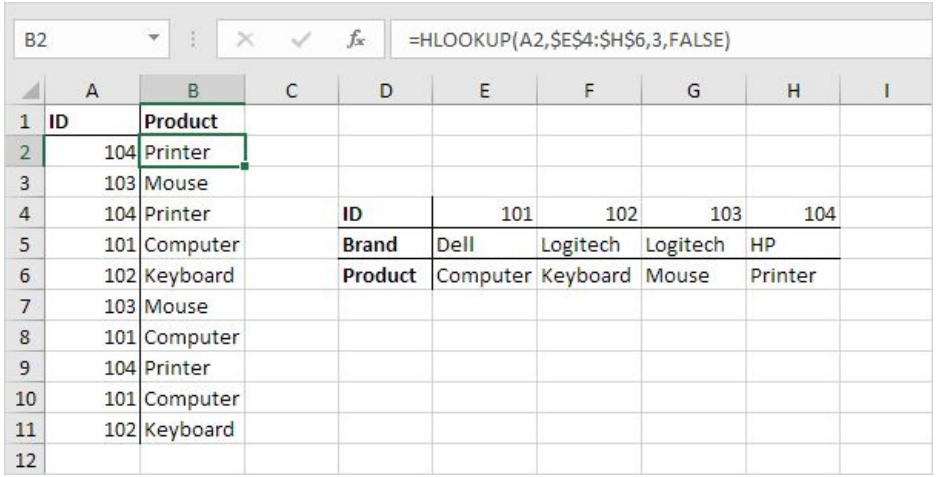
- The HLOOKUP (Horizontal lookup) is a function which given a value in one table, finds the corresponding value in another table horizontally.

- HLOOKUP(value_to_search , range_to_search, column_to_import, na_values)

https://www.excel-easy.com/functions/lookup-reference-functions.html

- In the example on the right, we want to fill the second table with the product name.

HLOOKUP(E4;$A1:$B11;2;0)



	A	B	C	D	E	F	G	H	I
1	ID	Product							
2	104	Printer							
3	103	Mouse							
4	104	Printer		ID	101	102	103	104	
5	101	Computer		Brand	Dell	Logitech	Logitech	HP	
6	102	Keyboard		Product	Computer	Keyboard	Mouse	Printer	
7	103	Mouse							
8	101	Computer							
9	104	Printer							
10	101	Computer							
11	102	Keyboard							
12									

Excel limitations

- Small files
- Total number of rows and columns on a worksheet
  - **1,048,576** rows by **16,384** columns
- Bigger files can’t be read from Excel and they are quite common nowadays


IRON
HACK