Introduction to Social Data Analytics

Class 2

Arushi Kaushik

Department of Economics
UCSD

4th April, 2019

Today's Learning Objectives

From Exercise 1:

- Open Excel, save workbook, edit cells, autofill down column, apply filter, sort columns

Today's Learning Objectives

From Exercise 1:

- Open Excel, save workbook, edit cells, autofill down column, apply filter, sort columns

After today, you should be able to:

- Identify observations and variables in an Excel workbook
- Discern the unit of analysis in a data table and demonstrate how to change it
- Implement statistical and logical functions in Excel
- Understand basic Boolean logic and use logical operators

Today's Learning Objectives

From Exercise 1:

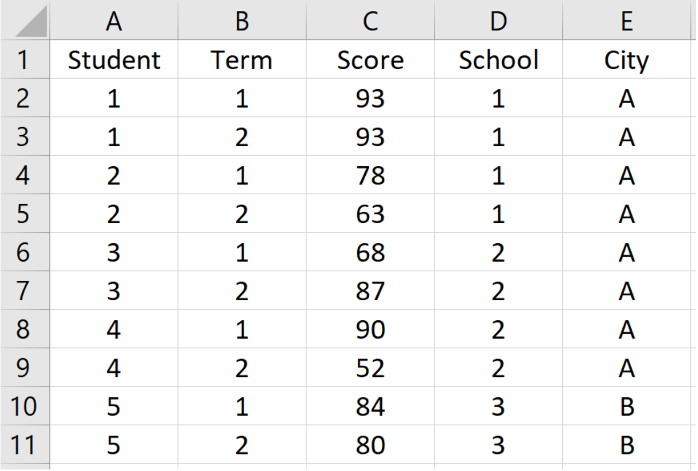
- Open Excel, save workbook, edit cells, autofill down column, apply filter, sort columns

After today, you should be able to:

- Identify observations and variables in an Excel workbook
- Discern the unit of analysis in a data table and demonstrate how to change it
- Implement statistical and logical functions in Excel
- Understand basic Boolean logic and use logical operators

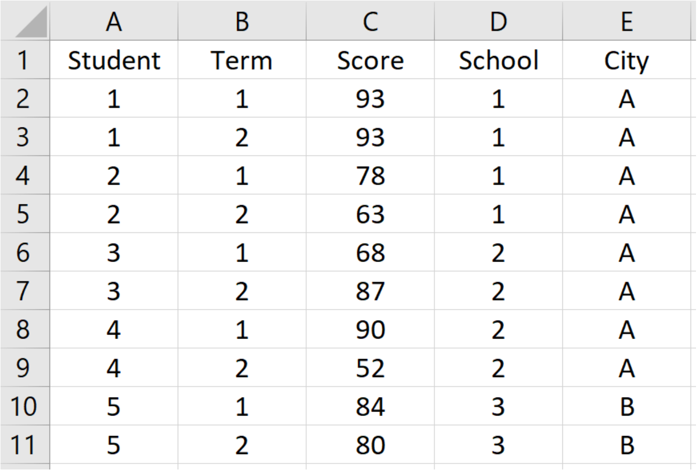
Please download and open class2.xlsx if you haven't already.

Observations $*$ Variables = Data Table



	A	B	C	D	E
1	Student	Term	Score	School	City
2	1	1	93	1	A
3	1	2	93	1	A
4	2	1	78	1	A
5	2	2	63	1	A
6	3	1	68	2	A
7	3	2	87	2	A
8	4	1	90	2	A
9	4	2	52	2	A
10	5	1	84	3	B
11	5	2	80	3	B

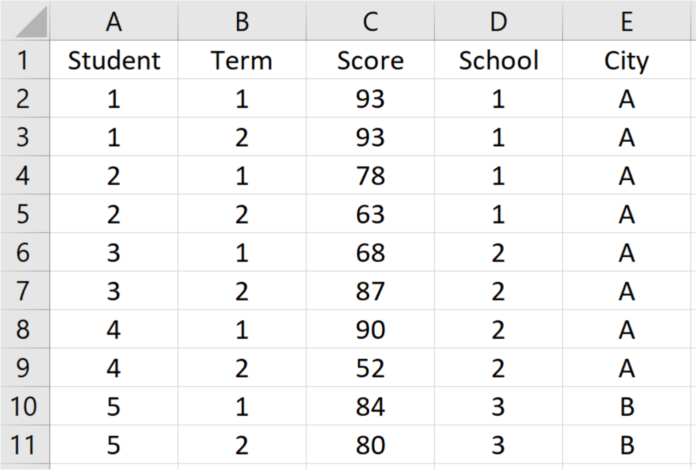
Observations $*$ Variables $=$ Data Table



	A	B	C	D	E
1	Student	Term	Score	School	City
2	1	1	93	1	A
3	1	2	93	1	A
4	2	1	78	1	A
5	2	2	63	1	A
6	3	1	68	2	A
7	3	2	87	2	A
8	4	1	90	2	A
9	4	2	52	2	A
10	5	1	84	3	B
11	5	2	80	3	B

- What does each column represent?

Observations $*$ Variables $=$ Data Table



	A	B	C	D	E
1	Student	Term	Score	School	City
2	1	1	93	1	A
3	1	2	93	1	A
4	2	1	78	1	A
5	2	2	63	1	A
6	3	1	68	2	A
7	3	2	87	2	A
8	4	1	90	2	A
9	4	2	52	2	A
10	5	1	84	3	B
11	5	2	80	3	B

- What does each column represent?
- What does each row represent?

The Unit of Analysis

- The “case” of the data set, each row
- The things to be compared, e.g. people or cities.

The Unit of Analysis

- The “case” of the data set, each row
- The things to be compared, e.g. people or cities.
- Each data table contains a consistent unit of analysis

The Unit of Analysis

- The “case” of the data set, each row
- The things to be compared, e.g. people or cities.
- Each data table contains a consistent unit of analysis
    - Usually have a unique ID for each unit
    - Similar variables can be collected on each unit

The Unit of Analysis

- The “case” of the data set, each row
- The things to be compared, e.g. people or cities.
- Each data table contains a consistent unit of analysis
  - Usually have a unique ID for each unit
  - Similar variables can be collected on each unit
- Units may have a time dimension

The Unit of Analysis

- The “case” of the data set, each row
- The things to be compared, e.g. people or cities.
- Each data table contains a consistent unit of analysis
  - Usually have a unique ID for each unit
  - Similar variables can be collected on each unit
- Units may have a time dimension
  - For example, monthly household surveys are recorded at the household-month level, so each unit is a household-month.
  - Longitudinal or panel data: observe the same sample over different points in time

The Unit of Analysis

- The “case” of the data set, each row
- The things to be compared, e.g. people or cities.
- Each data table contains a consistent unit of analysis
  - Usually have a unique ID for each unit
  - Similar variables can be collected on each unit
- Units may have a time dimension
  - For example, monthly household surveys are recorded at the household-month level, so each unit is a household-month.
  - Longitudinal or panel data: observe the same sample over different points in time
- What is the unit of analysis in class2.xlsx?

Changing the unit of analysis

A research project might examine many data tables with different units of analysis, for example:

1. A data table with each student-term (most granular)
2. A data table with each student
3. A data table with each school
4. A data table with each city (most coarse)

Changing the unit of analysis

A research project might examine many data tables with different units of analysis, for example:

1. A data table with each student-term (most granular)
2. A data table with each student
3. A data table with each school
4. A data table with each city (most coarse)

One can coarsen the unit of analysis by taking averages or counts

Changing the unit of analysis

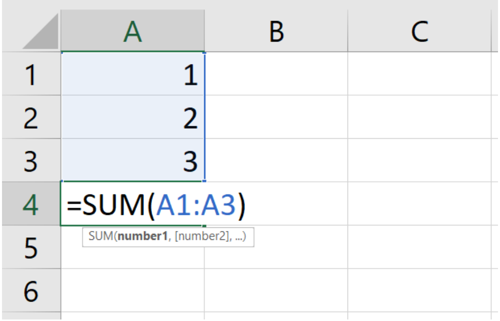
A research project might examine many data tables with different units of analysis, for example:

1. A data table with each student-term (most granular)
2. A data table with each student
3. A data table with each school
4. A data table with each city (most coarse)

One can coarsen the unit of analysis by taking averages or counts

- We'll do this in our Excel table after we learn about functions.

What is a function?

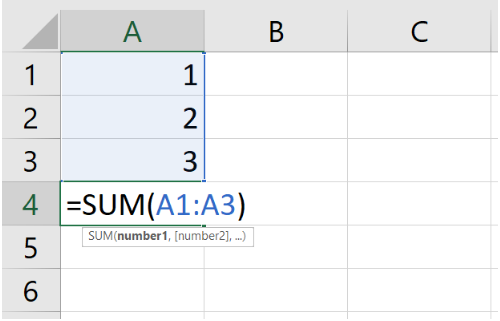


	A	B	C
1	1		
2	2		
3	3		
4	=SUM(A1:A3)		
5			
6			

SUM(**number1**, [number2], ...)

- An object that takes inputs and produces outputs

What is a function?

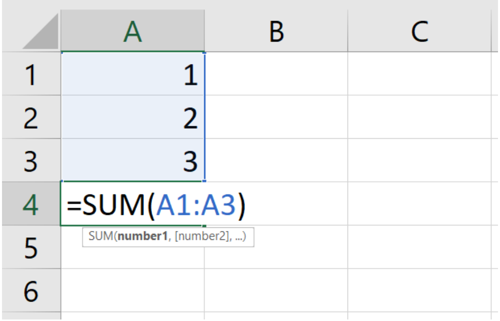


	A	B	C
1	1		
2	2		
3	3		
4	=SUM(A1:A3)		
5			
6			

SUM(**number1**, [number2], ...)

- An object that takes inputs and produces outputs
- What are the inputs in this example?

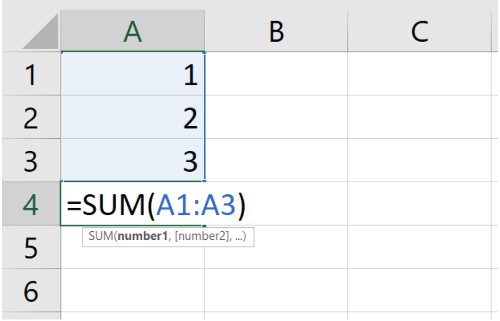
What is a function?



	A	B	C
1	1		
2	2		
3	3		
4	=SUM(A1:A3)		
5	SUM(**number1**, [number2], ...)		
6			

- An object that takes inputs and produces outputs
- What are the inputs in this example?
- What will the output be?

What is a function?



	A	B	C
1	1		
2	2		
3	3		
4	=SUM(A1:A3)		
5			
6			

SUM(**number1**, [number2], ...)

- An object that takes inputs and produces outputs
- What are the inputs in this example?
- What will the output be?
- Notice the equals sign and the text underneath the function.

Types of functions

- Statistical:
  - input: is usually a set of numbers
  - output: is usually a mathematical function of these numbers

Types of functions

- Statistical:
  - input: is usually a set of numbers
  - output: is usually a mathematical function of these numbers
- Logical:
  - input is usually a TRUE/FALSE statement
  - output is either TRUE/FALSE, or what to do if it's TRUE

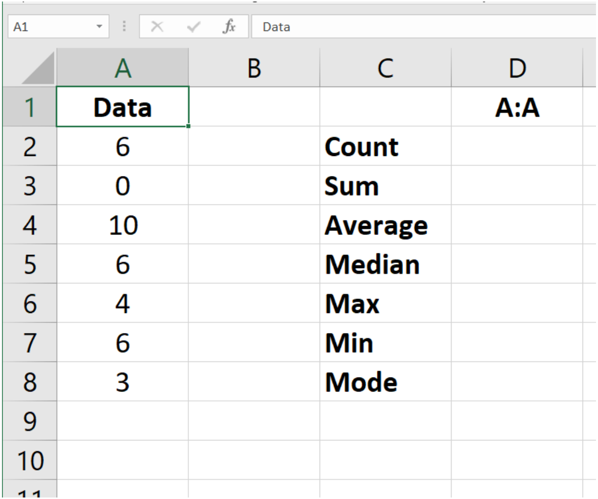
Types of functions

- Statistical:
  - input: is usually a set of numbers
  - output: is usually a mathematical function of these numbers
- Logical:
  - input is usually a TRUE/FALSE statement
  - output is either TRUE/FALSE, or what to do if it's TRUE
- Lookup:
  - We'll learn about these in our next short exercise.

Statistical Functions in Excel

- COUNT
- SUM
- AVERAGE
- MEDIAN
- MAX, MIN
- MODE

Statistical Functions in Excel

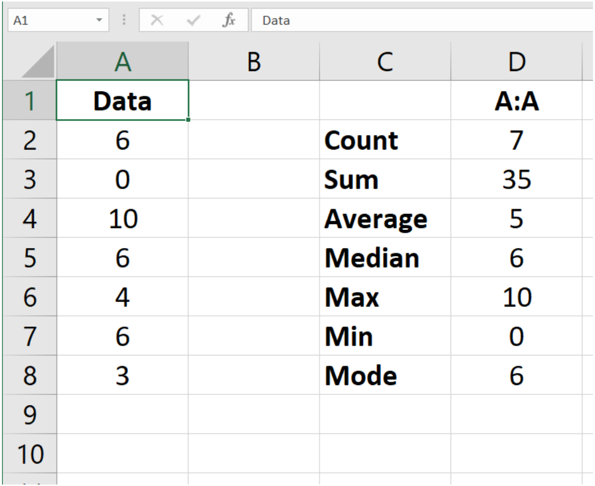


A1

fx Data

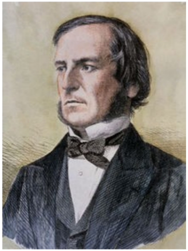
	A	B	C	D
1	**Data**			**A:A**
2	6		**Count**	
3	0		**Sum**	
4	10		**Average**	
5	6		**Median**	
6	4		**Max**	
7	6		**Min**	
8	3		**Mode**	
9				
10				
11				

Statistical Functions in Excel



	A	B	C	D
1	Data			A:A
2	6		Count	7
3	0		Sum	35
4	10		Average	5
5	6		Median	6
6	4		Max	10
7	6		Min	0
8	3		Mode	6
9				
10				

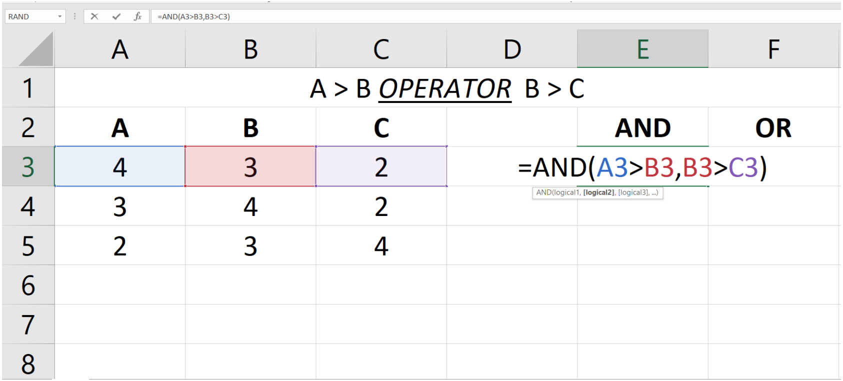
Boolean Logic



George Boole
November 2, 1815 - December 8, 1864

- A statement can be TRUE or FALSE
- Use these to form other statements using:
  - AND
  - OR
  - NOT

Boolean Logic in Excel

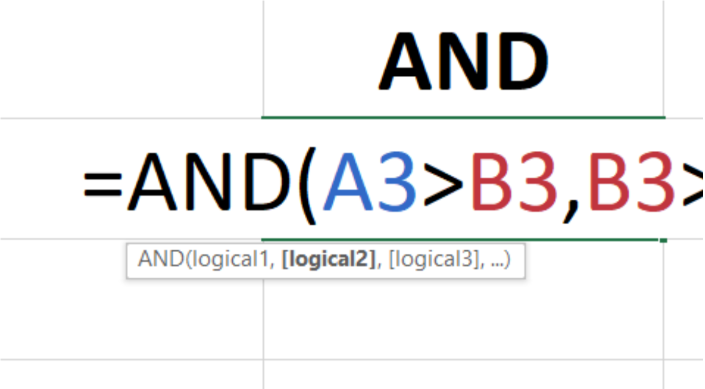


RAND =AND(A3>B3,B3>C3)

	A	B	C	D	E	F
1	A > B *OPERATOR* B > C					
2	**A**	**B**	**C**		**AND**	**OR**
3	4	3	2		=AND(A3>B3,B3>C3)	
4	3	4	2			
5	2	3	4			
6						
7						
8						

AND(logical1, **[logical2]**, [logical3], ...)

Notice the arguments listed underneath the function

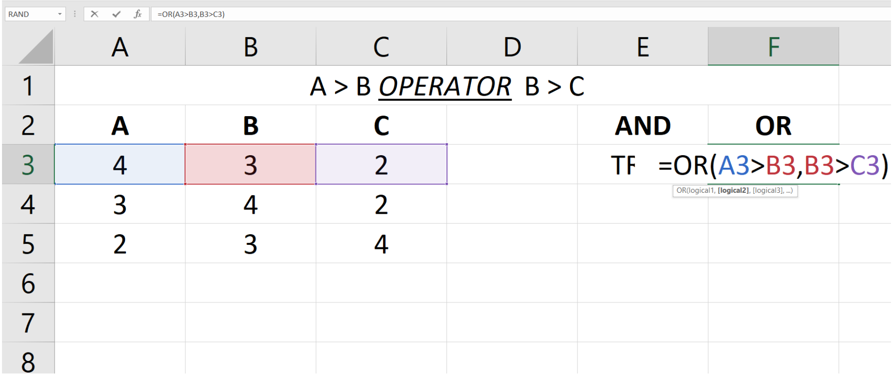


AND

=AND(A3>B3,B3>

AND(logical1, **[logical2]**, [logical3], ...)

Boolean Logic in Excel



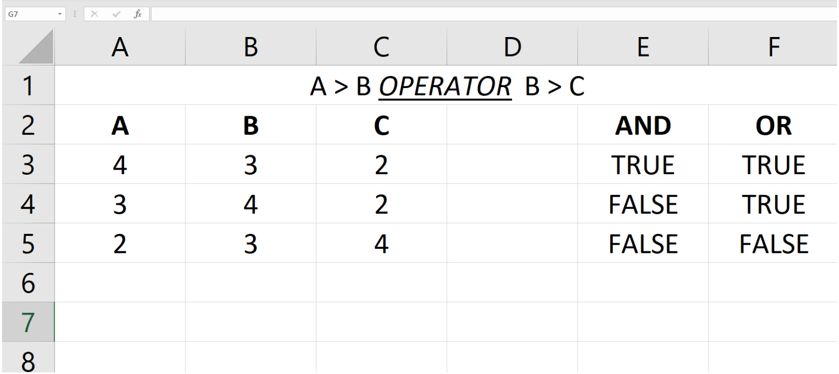
RAND

=OR(A3>B3,B3>C3)

	A	B	C	D	E	F
1			A > B *OPERATOR* B > C			
2	**A**	**B**	**C**		**AND**	**OR**
3	4	3	2		TF	=OR(A3>B3,B3>C3)
4	3	4	2			
5	2	3	4			
6						
7						
8						

OR(logical1, **[logical2]**, [logical3], ...)

Boolean Logic in Excel



	A	B	C	D	E	F
1	A > B *OPERATOR* B > C					
2	**A**	**B**	**C**		**AND**	**OR**
3	4	3	2		TRUE	TRUE
4	3	4	2		FALSE	TRUE
5	2	3	4		FALSE	FALSE
6						
7						
8						

IF statements in Excel

=IF(logical_statement, [value_if_true], [value_if_false])

If the logical statement is TRUE, do one thing
If the logical statement is FALSE, do another thing

e.g. =IF(A1=B1, 1, 0)

If Statements and Statistics

We can combine if statements and statistical functions!

- AVERAGEIF:
  - input: a set of numbers
  - output: the average only for numbers that satisfy the condition

If Statements and Statistics

We can combine if statements and statistical functions!

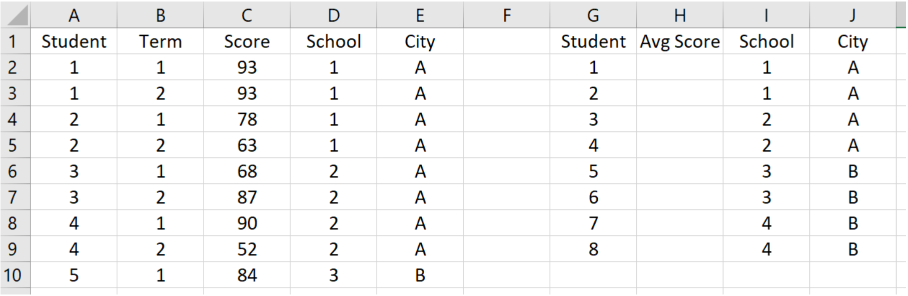
- AVERAGEIF:
  - input: a set of numbers
  - output: the average only for numbers that satisfy the condition
- AVERAGEIFS:
  - input: a set of numbers
  - output: the average only for numbers that satisfy multiple conditions

If Statements and Statistics

We can combine if statements and statistical functions!

- AVERAGEIF:
  - input: a set of numbers
  - output: the average only for numbers that satisfy the condition
- AVERAGEIFS:
  - input: a set of numbers
  - output: the average only for numbers that satisfy multiple conditions
- Same with SUMIF, SUMIFS, COUNTIF, COUNTIFS.

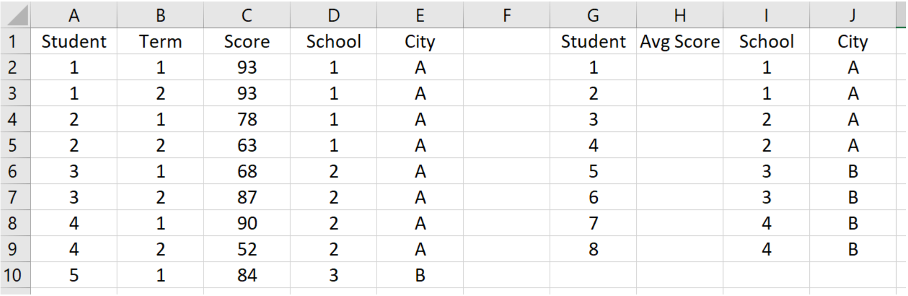
Change unit of analysis using functions



	A	B	C	D	E	F	G	H	I	J
1	Student	Term	Score	School	City		Student	Avg Score	School	City
2	1	1	93	1	A		1		1	A
3	1	2	93	1	A		2		1	A
4	2	1	78	1	A		3		2	A
5	2	2	63	1	A		4		2	A
6	3	1	68	2	A		5		3	B
7	3	2	87	2	A		6		3	B
8	4	1	90	2	A		7		4	B
9	4	2	52	2	A		8		4	B
10	5	1	84	3	B					

- How can we coarsen the unit of analysis from 'student-term' to 'student'?

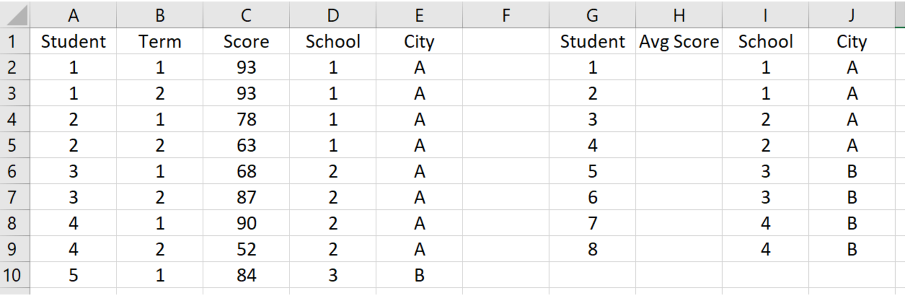
Change unit of analysis using functions



	A	B	C	D	E	F	G	H	I	J
1	Student	Term	Score	School	City		Student	Avg Score	School	City
2	1	1	93	1	A		1		1	A
3	1	2	93	1	A		2		1	A
4	2	1	78	1	A		3		2	A
5	2	2	63	1	A		4		2	A
6	3	1	68	2	A		5		3	B
7	3	2	87	2	A		6		3	B
8	4	1	90	2	A		7		4	B
9	4	2	52	2	A		8		4	B
10	5	1	84	3	B					

- How can we coarsen the unit of analysis from ‘student-term’ to ‘student’?
- Use AVERAGEIF(range, criteria, average_range).

Change unit of analysis using functions



	A	B	C	D	E	F	G	H	I	J
1	Student	Term	Score	School	City		Student	Avg Score	School	City
2	1	1	93	1	A		1		1	A
3	1	2	93	1	A		2		1	A
4	2	1	78	1	A		3		2	A
5	2	2	63	1	A		4		2	A
6	3	1	68	2	A		5		3	B
7	3	2	87	2	A		6		3	B
8	4	1	90	2	A		7		4	B
9	4	2	52	2	A		8		4	B
10	5	1	84	3	B					

- How can we coarsen the unit of analysis from ‘student-term’ to ‘student’?
- Use AVERAGEIF(range, criteria, average_range).
- Try coarsening to the ‘school’ and ‘city’ units of analysis

Next class

Friday we will practice using these functions.

See you then!