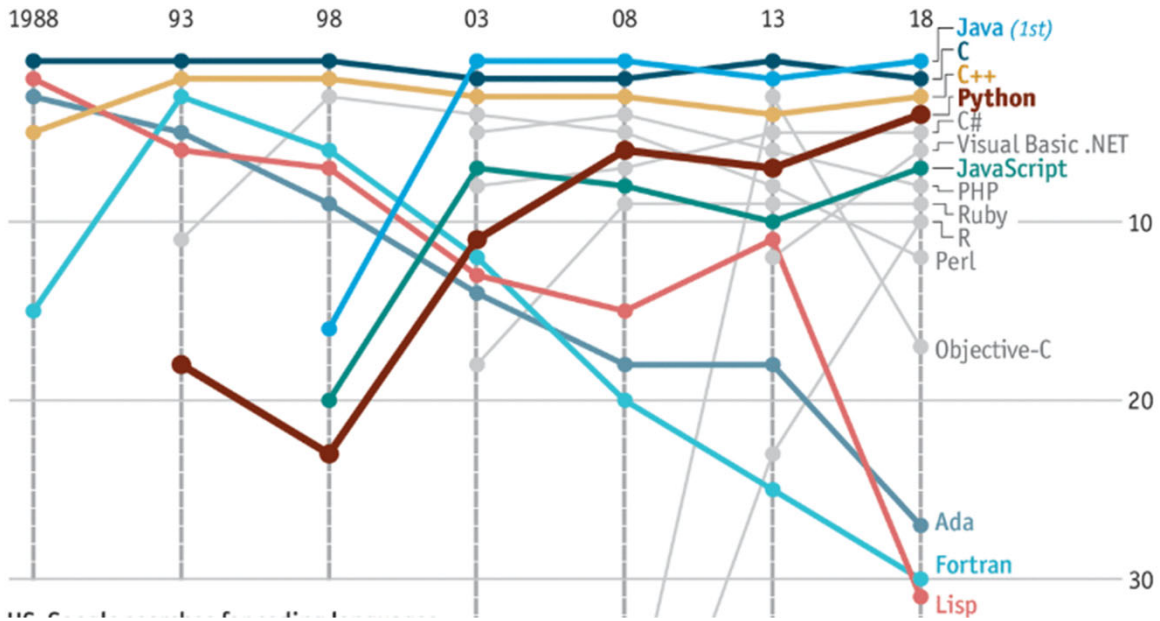
Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

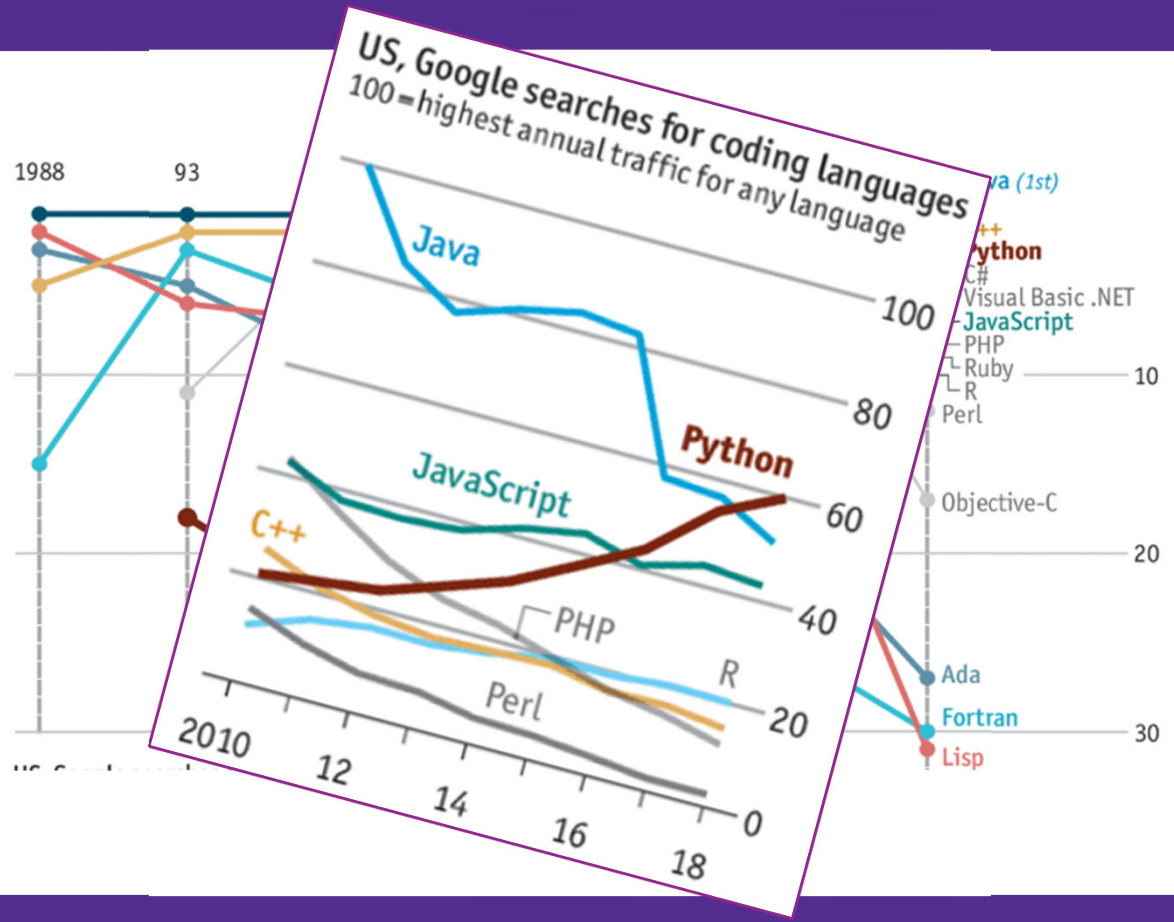
Why use Python for studying data instead of another programming language?



Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Why use Python for studying data instead of another programming language?



NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 5: Lecture 1

Tables – Joining Datasets to Link Records

NYU Center for Data Science

How can we connect the rows of different datasets using the values in the columns?

DS-UA 112
Introduction to Data Science

Week 5: Lecture 1

Tables – Joining Datasets to Link Records

Adapted from Nolan, Gonzalez, Hug, Lau

Announcements

- Please check Week 5 agenda on NYU Classes
  - Homework 2
  - Lab 4
  - Survey 2
- Remember to post to Piazza

Remember to tag your answers on Gradescope!

Review

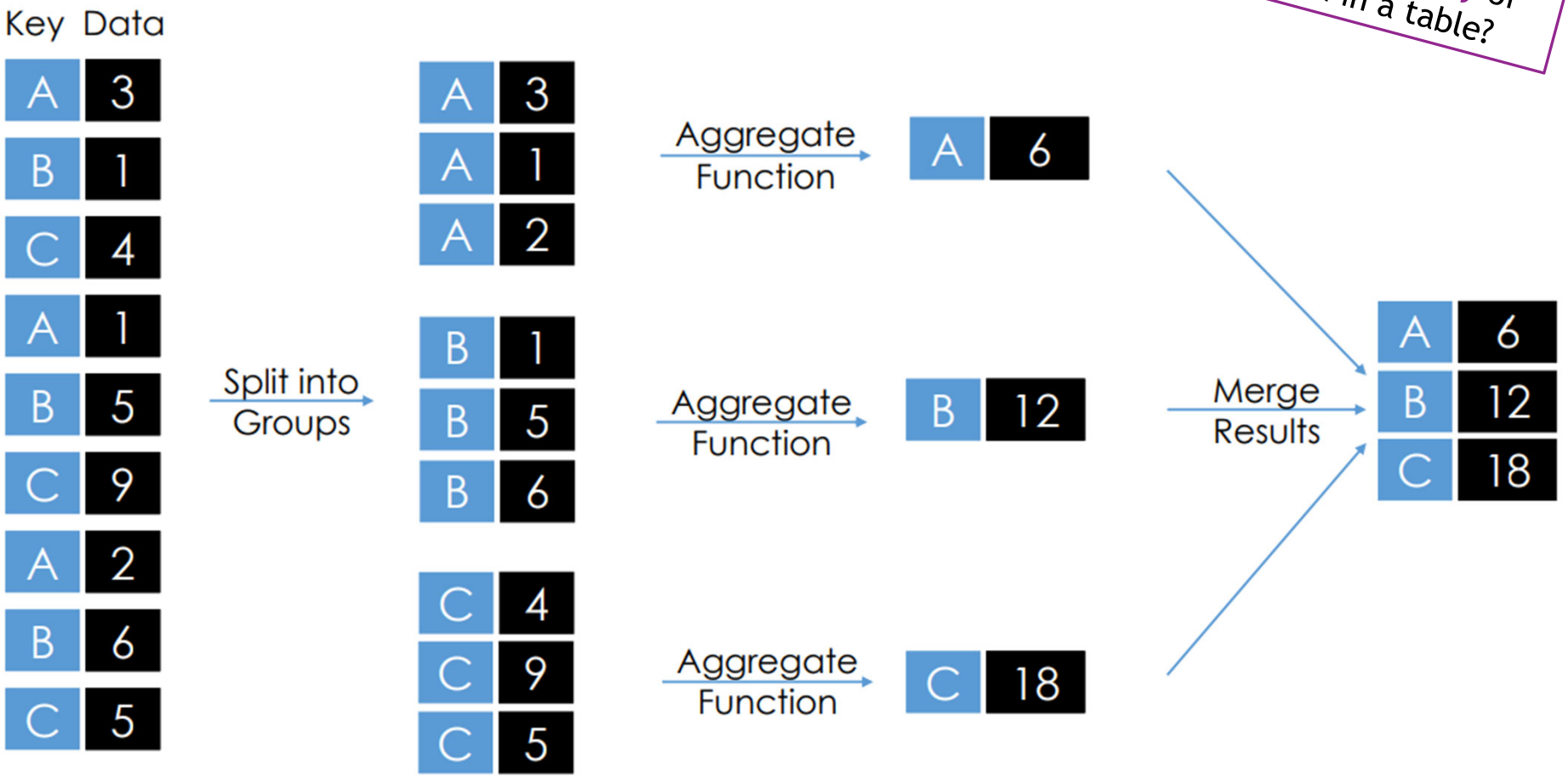
- Operations on Tables
  - Inspecting
  - Sorting
  - Summarizing
- Grouping and Pivoting
  - Find the most popular name in New York
  - Find all names that start with E.
  - Sort names by occurrence of dr and ea
  - Find the name whose popularity has changed the most.
  - Count the number of female and male babies born in each year

Goals

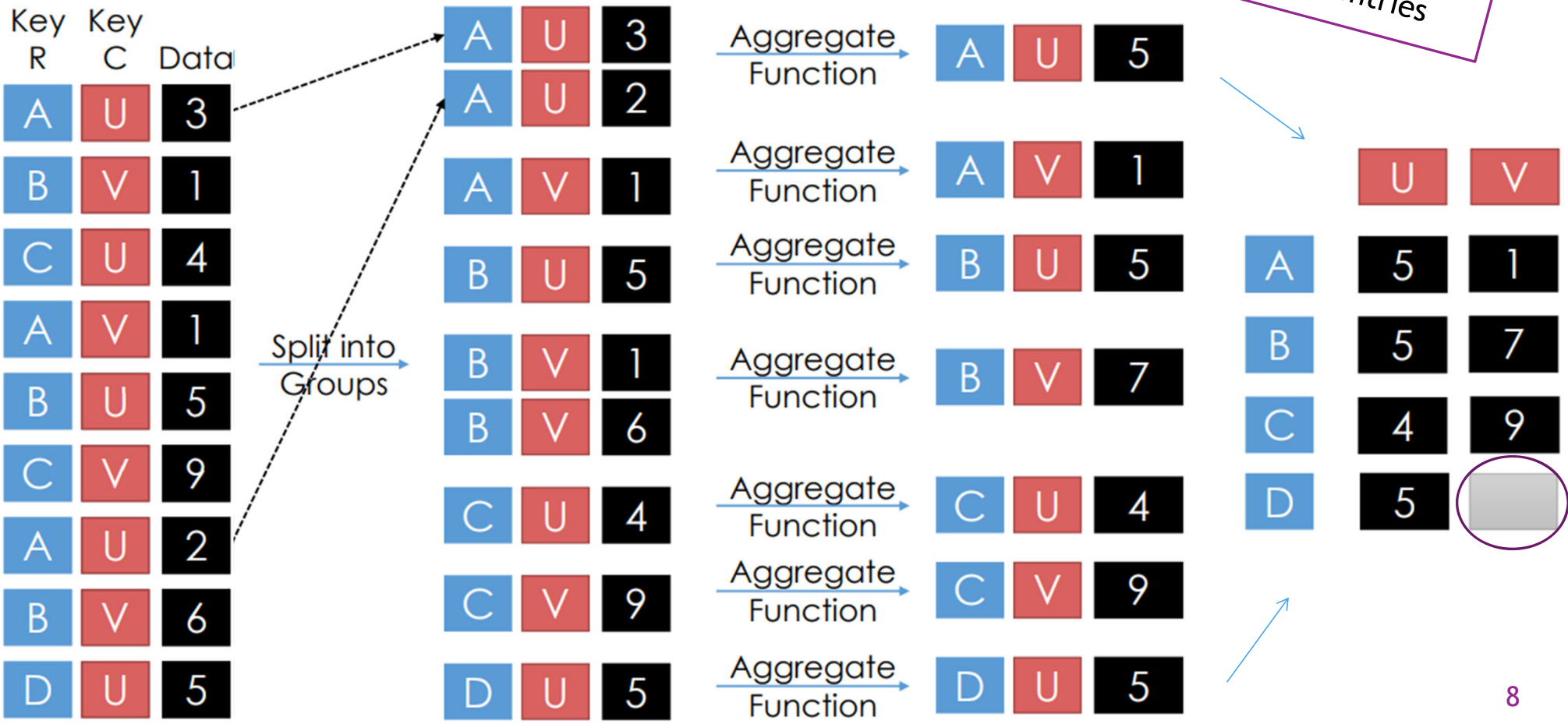
- Apply
- Group
  - agg
  - size
  - filter
- Pivot
  - stack, unstack

Review

How can we change the granularity of data in a table?



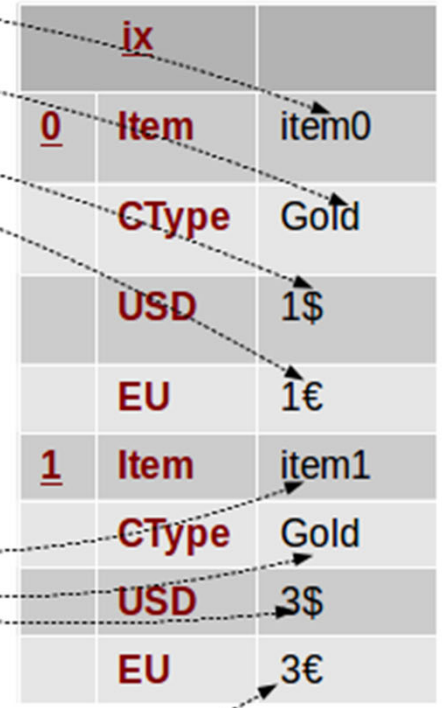
Review


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5
Aggregate Function
A U 5
Aggregate Function
A V 1
Aggregate Function
B U 5
Aggregate Function
B V 7
Aggregate Function
C U 4
Aggregate Function
C V 9
Aggregate Function
D U 5
Need to handle missing entries
U V
A 5 1
B 5 7
C 4 9
D 5

Review

stack will combine the row labels and columns labels

ix	Item	CType	USD	EU
0	Item0	Gold	1$	1€
1	Item1	Gold	3$	3€



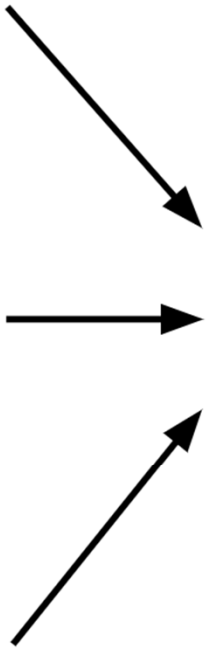
The opposite operation called unstack will break up the indices into rows and columns

Review

1992	3	ak
1996	1	tx
2000	4	fl
1996	1	hi
1992	5	mi
2000	9	ak
2000	2	ca
2000	6	sd

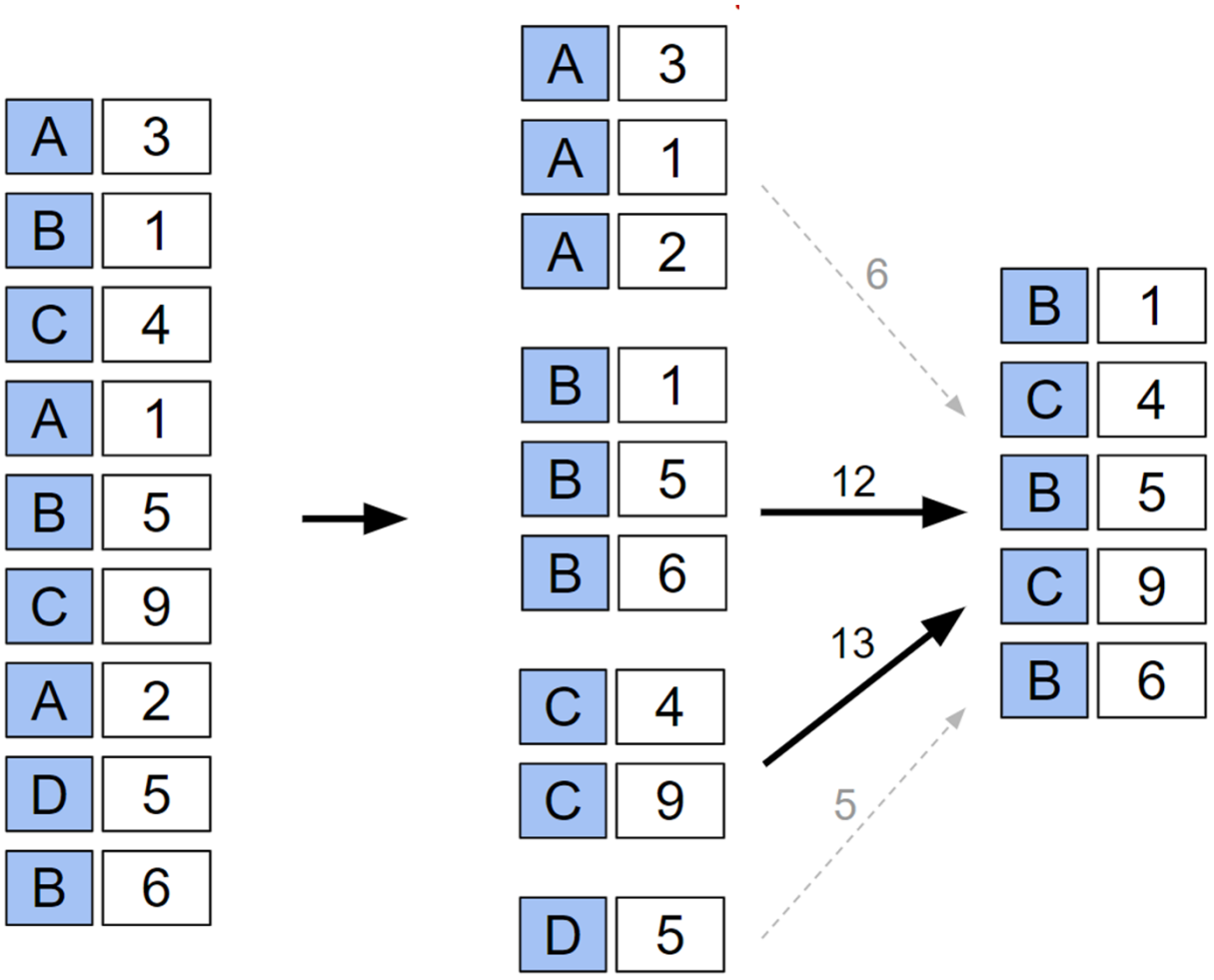
1992	3	ak
1992	5	hi
1996	1	tx
1996	1	mi
2000	4	fl
2000	9	ak
2000	2	ca
2000	6	sd

size will calculate the number of rows in each group



1992	2
1996	2
2000	4

Review



filter removes any groups that fail to satisfy conditions

Note that filter does not affect the entries of the group

Agenda

- Compressing Files
- Joining
  - Inner, Outer
  - Left, Right
  - Cross
- Properties of Data
  - Qualitative or Quantitative
  - Scope
  - Granularity
  - Temporality
  - Faithfulness

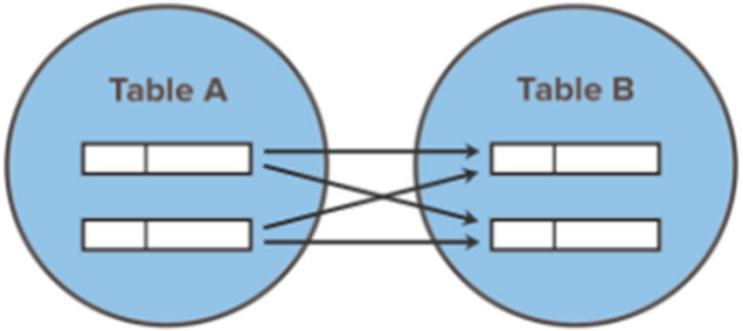
Discussed in Lab 4

References

- Nolan, Lau, Gonzalez (Chapter 5,6)

Cross Join

- Cross Join pairs each of the rows in the tables regardless of the entries in the columns.
- All pairs of rows appear in the Join.



s

M	U
1	W
2	X
3	X
4	Y

t

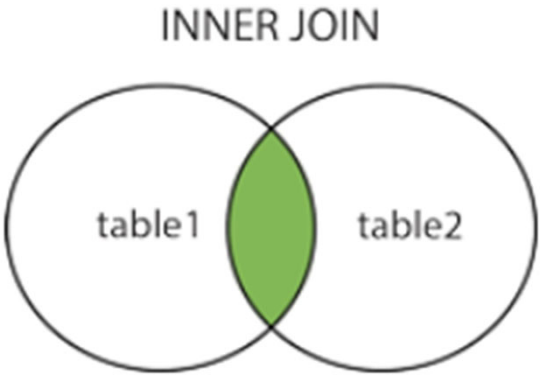
N	V
A	X
B	X
C	Y
D	Z

M	U	N	V
1	W	A	X
2	X	A	X
3	X	A	X
4	Y	A	X
1	W	B	X
2	X	B	X
3	X	B	X
4	Y	B	X

M	U	N	V
1	W	C	Y
2	X	C	Y
3	X	C	Y
4	Y	C	Y
1	W	D	Z
2	X	D	Z
3	X	D	Z
4	Y	D	Z

Inner Join

- Inner Join pairs each of the rows in the tables depending on the entries in specific columns.
- The entries of columns must match for the pair to appear in the Join.



s

M	U
1	W
2	X
3	X
4	Y

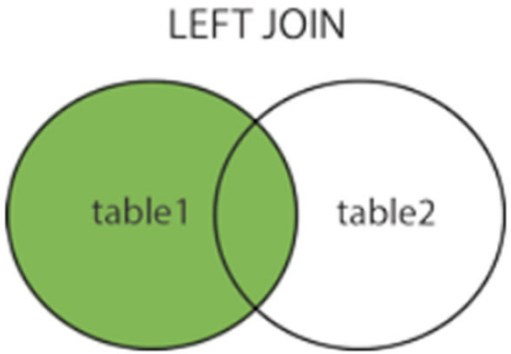
t

N	V
A	X
B	X
C	Y
D	Z

M	U	N	V
2	X	A	X
3	X	A	X
2	X	B	X
3	X	B	X
4	Y	C	Y

Left Join

- Left Join pairs each of the rows in the left table to rows in the right table depending on the entries in specific columns.
- The entries of columns in the right table must match for the pair to appear in the Join.



s

M	U
1	W
2	X
3	X
4	Y

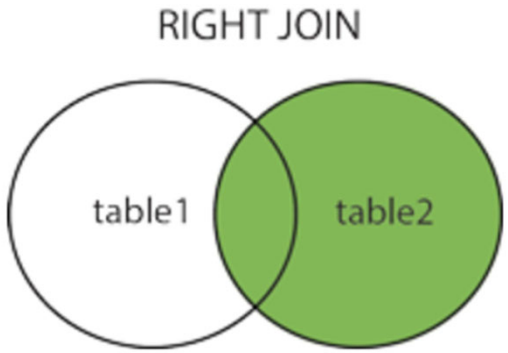
t

N	V
A	X
B	X
C	Y
D	Z

1	W	null	null
2	X	A	X
3	X	A	X
2	X	B	X
3	X	B	X
4	Y	C	Y

Right Join

- Right Join pairs each of the rows in the right table to rows in the left table depending on the entries in specific columns.
- The entries of columns in the left table must match for the pair to appear in the Join.



s

M	U
1	W
2	X
3	X
4	Y

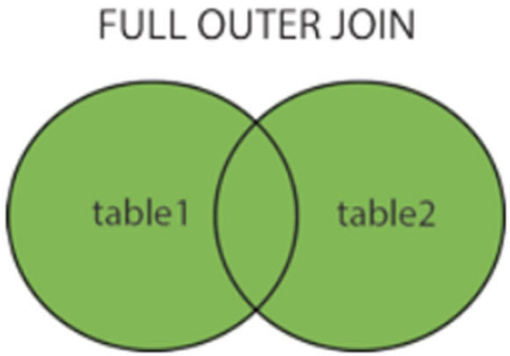
t

N	V
A	X
B	X
C	Y
D	Z

2	X	A	X
3	X	A	X
2	X	B	X
3	X	B	X
4	Y	C	Y
null	null	D	Z

Outer Join

- Outer Join combines Left Join and Right Join
- Note that Outer Join does not contain duplicate entries for the matching rows, that is, the rows contained in the Inner Join.



s

M	U
1	W
2	X
3	X
4	Y

t

N	V
A	X
B	X
C	Y
D	Z

1	W	null	null
2	X	A	X
3	X	A	X
2	X	B	X
3	X	B	X
4	Y	C	Y
null	null	D	Z

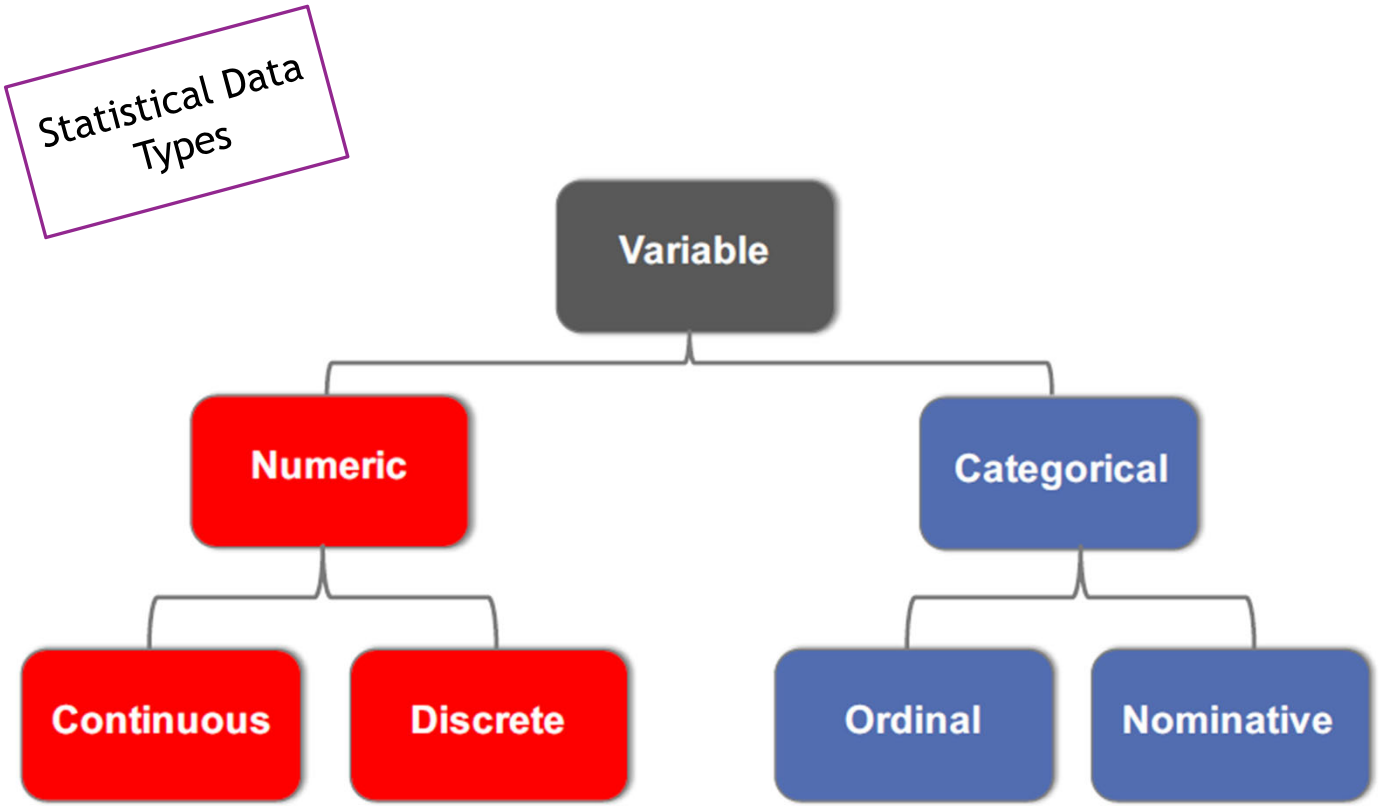
Data Types

Computational Data Types

- We store data as variables with different types.
- While Python does not require us to specify the types, each variable has a type.
- We should know about types because they determine available operations on the variable

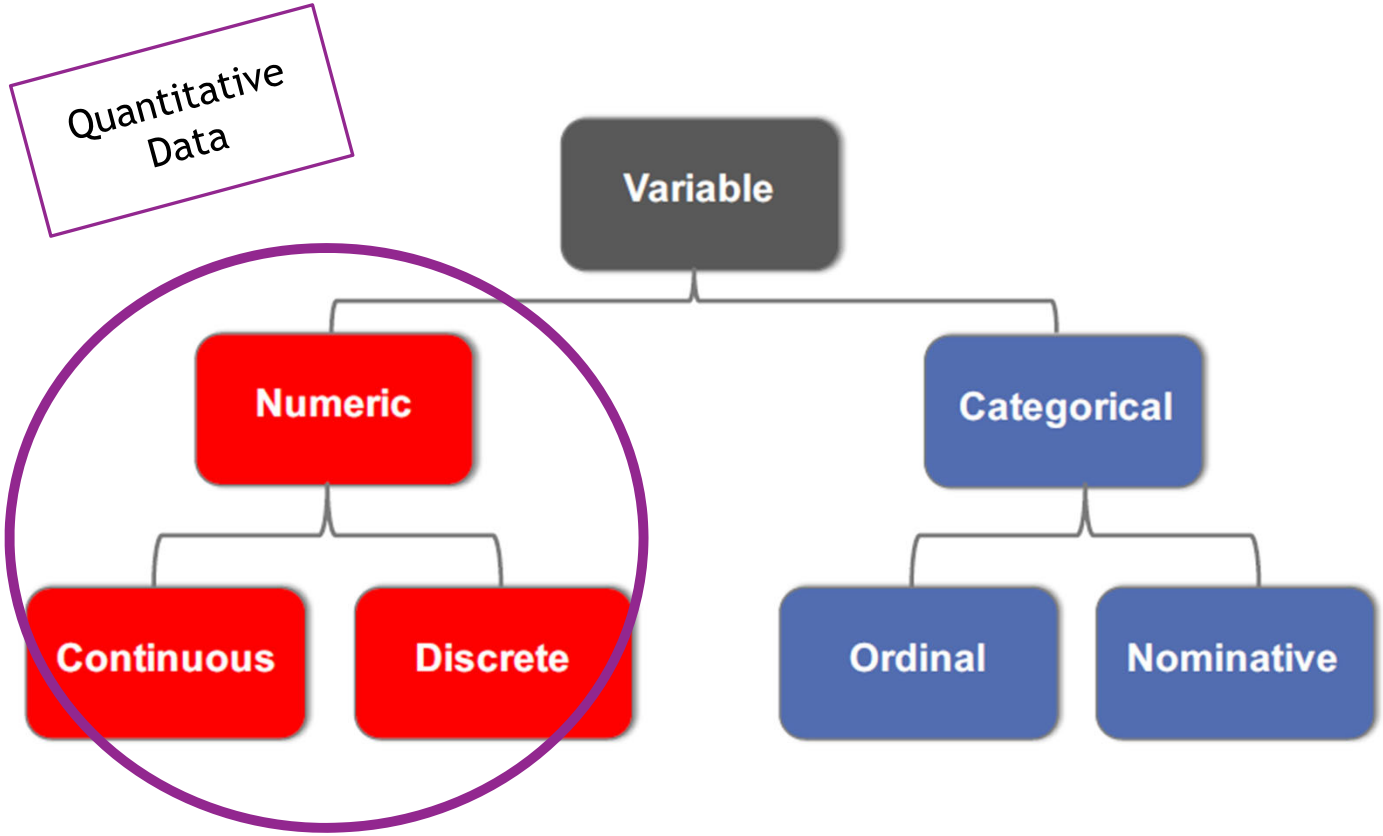
Example	Data Type
`x = "Hello World"`	str
`x = 20`	int
`x = 20.5`	float
`x = {"name" : "John", "age" : 36}`	dict
`x = {"apple", "banana", "cherry"}`	set
`x = ["apple", "banana", "cherry"]`	list
`x = ("apple", "banana", "cherry")`	tuple
`x = True`	bool

Data Types



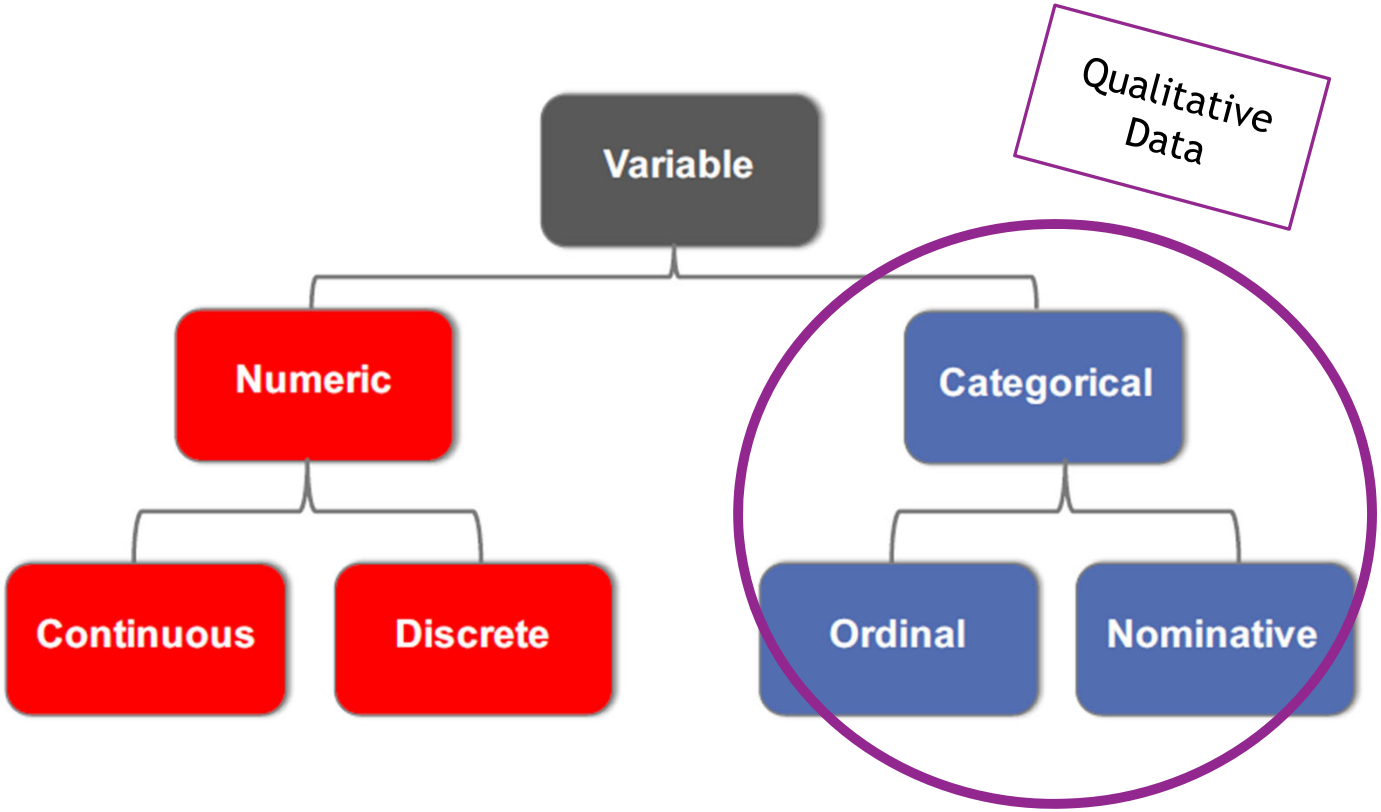
- We study data with different properties.
- Dividing these properties into types helps us to communicate the information behind the data
- We split between properties involving number for calculations and non-numbers for labels

Data Types



- Discrete
  - We can count discrete variables because they can take finitely many values
  - For example 1,2,3
- Continuous
  - We cannot count continuous variables because they can take infinitely many values
  - For example 1.54, 2.43, 3.14

Data Types



- Nominal
  - We use nominal data for labels to distinguish between different categories
  - For example blue, green
- Ordinal
  - If we can rank nominal data, then we have an order to the variables
  - For example high, medium, low

Data Types

NYC Police Reports

Column	Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

- Granularity
  - Amount of detail in the dataset
- Scope
  - Coverage of the dataset
- Temporality
  - Date and time of the information and the collection of information
- Faithfulness
  - Accuracy of the information

Summary

- Compressing Files
- Joining
  - Inner, Outer
  - Left, Right
  - Cross
- Properties of Data
  - Qualitative or Quantitative
  - Scope
  - Granularity
  - Temporality
  - Faithfulness

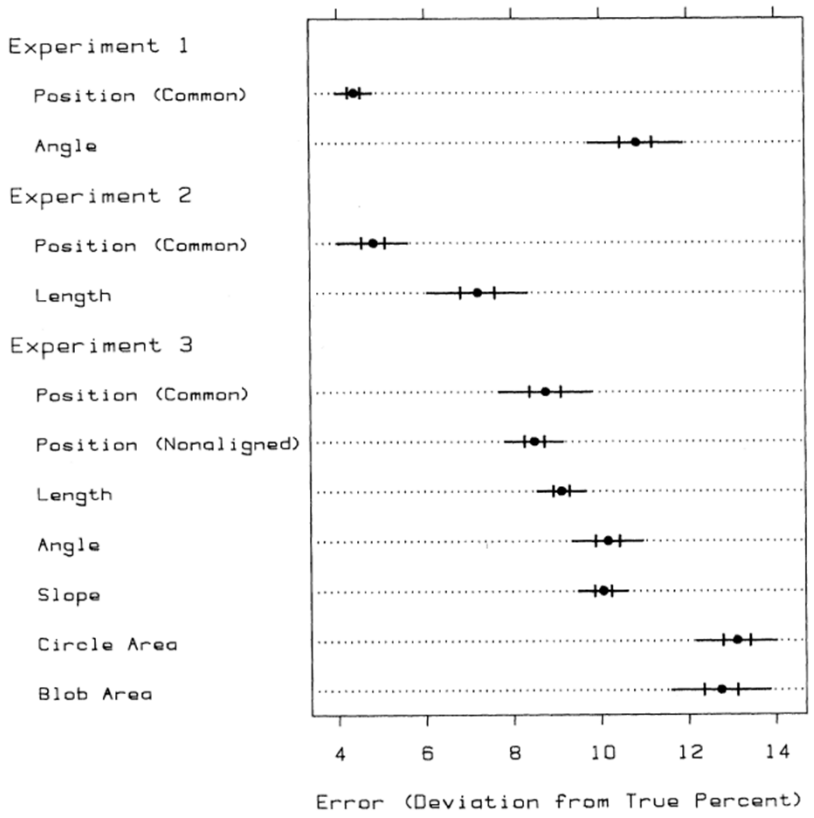
Goals

- Zip
- Merge
- Data Types
- Describing Tables

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

What aspects of charts are most understandable to you?



Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

What aspects of charts are most understandable to you?