

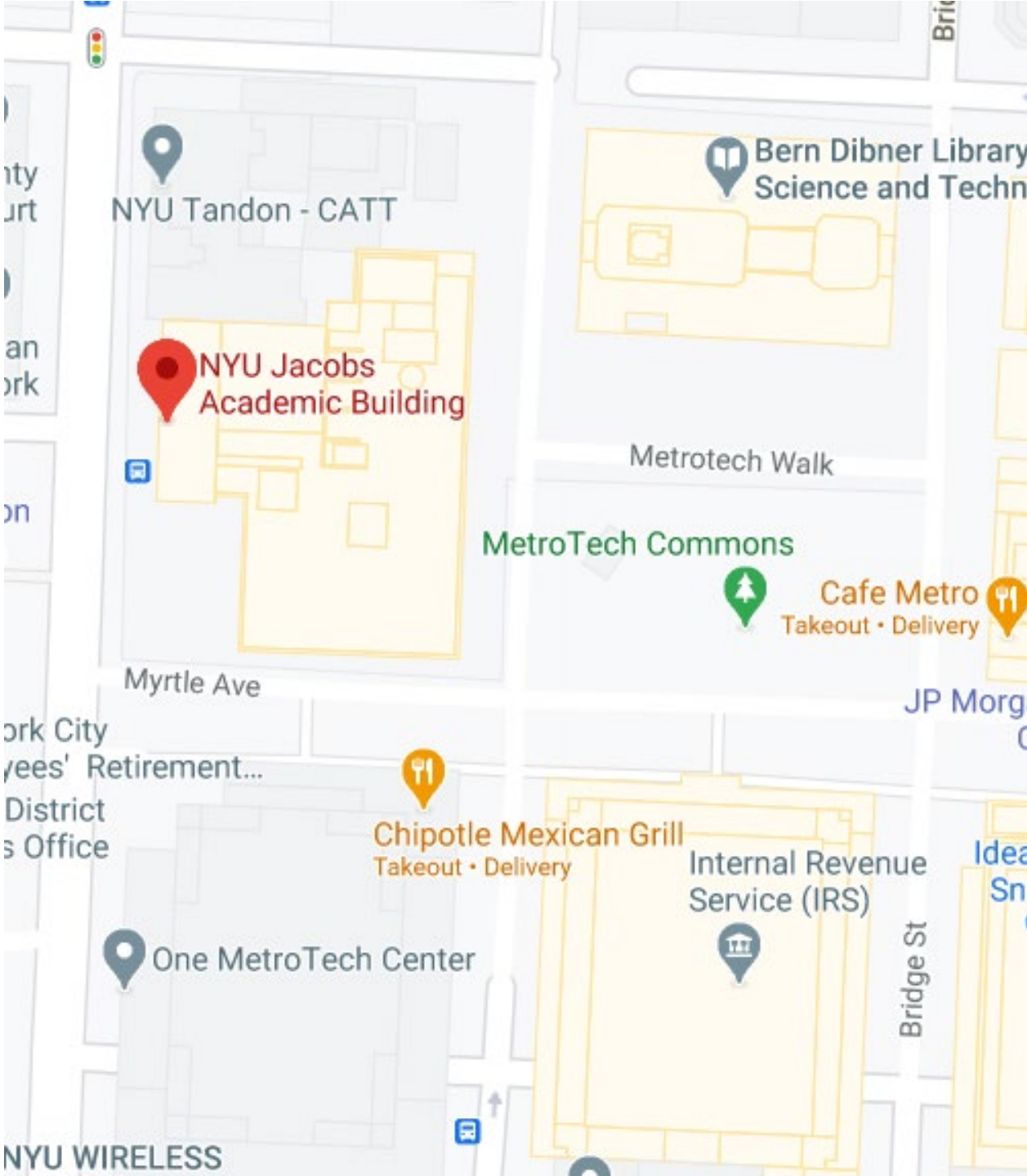


NYU | TANDON SCHOOL
OF ENGINEERING

Lecture 6

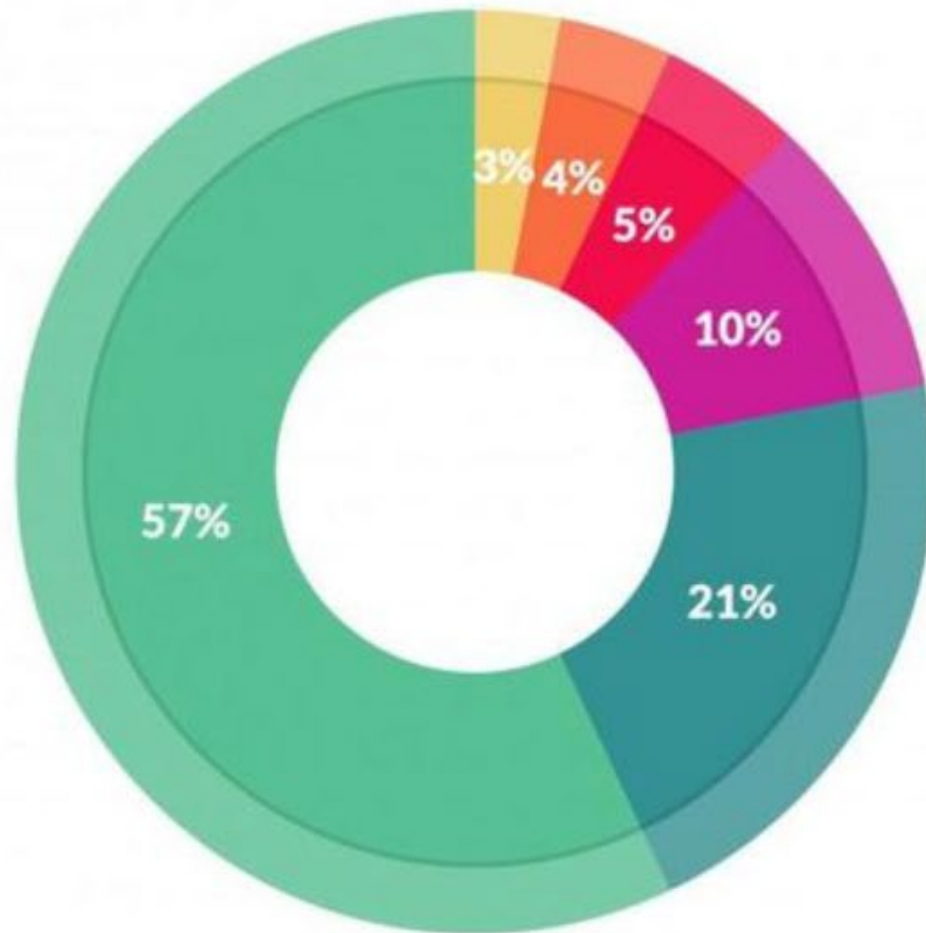
Tables

Programming for Business Analytics
MG-GY 8401



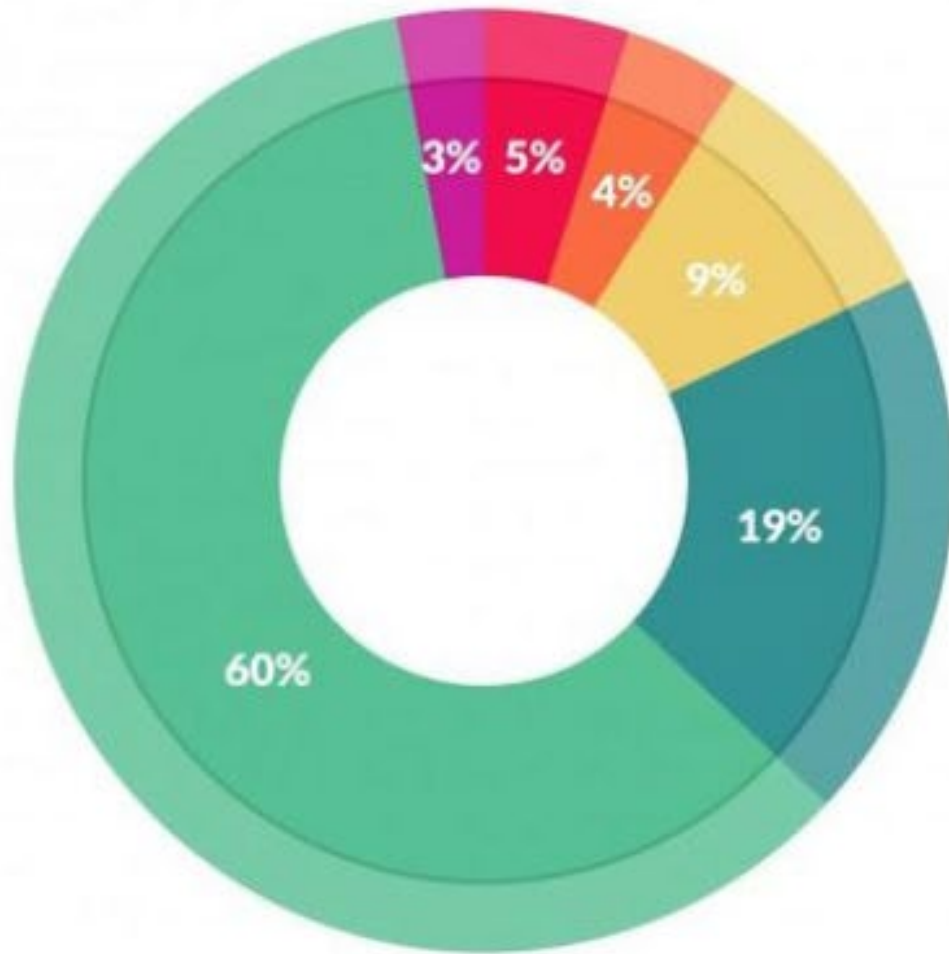
Logistics

- Office Hours
 - Monday's 5-6PM ET
- Homework
 - Project
 - Homework 6
 - Homework 5



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



THE WORLD FACTBOOK 1990

Country: Afghanistan

- Geography

Total area: 647,500 km²; land area: 647,500 km²

Comparative area: slightly smaller than Texas

Land boundaries: 5,826 km total; China 76 km, Iran 936 km,
Pakistan 2,430 km, USSR 2,384 km

Coastline: none--landlocked

Maritime claims: none--landlocked

Disputes: Pashtun question with Pakistan; Baloch question with Iran
and Pakistan; periodic disputes with Iran over Helmand water rights;
insurgency with Iranian and Pakistani involvement; traditional tribal
rivalries

Climate: arid to semiarid; cold winters and hot summers

Terrain: mostly rugged mountains; plains in north and southwest

Natural resources: natural gas, crude oil, coal, copper, talc, barites,
sulphur, lead, zinc, iron ore, salt, precious and semiprecious stones

Land use: 12% arable land; NEGL% permanent crops; 46% meadows and
pastures; 3% forest and woodland; 39% other; includes NEGL% irrigated

Environment: damaging earthquakes occur in Hindu Kush mountains;
soil degradation, desertification, overgrazing, deforestation, pollution

Note: landlocked

1990





Title :Afghanistan
Text :

Afghanistan

Geography

1995

Location:

Southern Asia, north of Pakistan

Map references:

Asia

Area:

total area:

647,500 sq km

land area:

647,500 sq km

comparative area:

slightly smaller than Texas

Land boundaries:

total 5,529 km, China 76 km, Iran 936 km, Pakistan 2,430 km, Tajikistan 1,206 km, Turkmenistan 744 km, Uzbekistan 137 km

Coastline:

0 km (landlocked)

Maritime claims:

none; landlocked

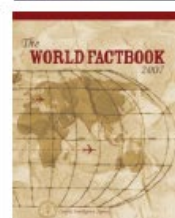
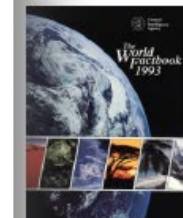
International disputes:

periodic disputes with Iran over Helmand water rights; Iran supports clients

in country, private Pakistani and Saudi sources also are active; power struggles among various groups for control of Kabul, regional rivalries among emerging warlords, traditional tribal disputes continue; support to Islamic fighters in Tajikistan's civil war; border dispute with Pakistan (Durand Line); support to Islamic militants worldwide by some factions

Climate:



arid to semiarid; cold winters and hot summers







Geography





Afghanistan

[Top of Page](#)**Location:**  





Southern Asia, north and west of Pakistan, east of Iran

Geographic coordinates:  

33 00 N, 65 00 E

Map references:  [Asia](#)**Area:**   *total:* 647,500 sq km*water:* 0 sq km*land:* 647,500 sq km**Area - comparative:**  



slightly smaller than Texas

Land boundaries:  *total:* 5,529 km*border countries:* China 76 km, Iran 936 km, Pakistan 2,430 km, Tajikistan 1,206 km, Turkmenistan 744 km, Uzbekistan 137 km**Coastline:**  

0 km (landlocked)

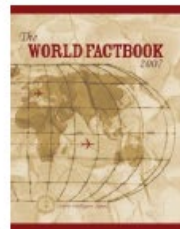
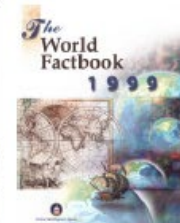
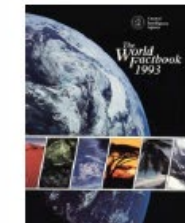
Maritime claims:  

none (landlocked)

Climate:  

arid to semiarid; cold winters and hot summers

2005

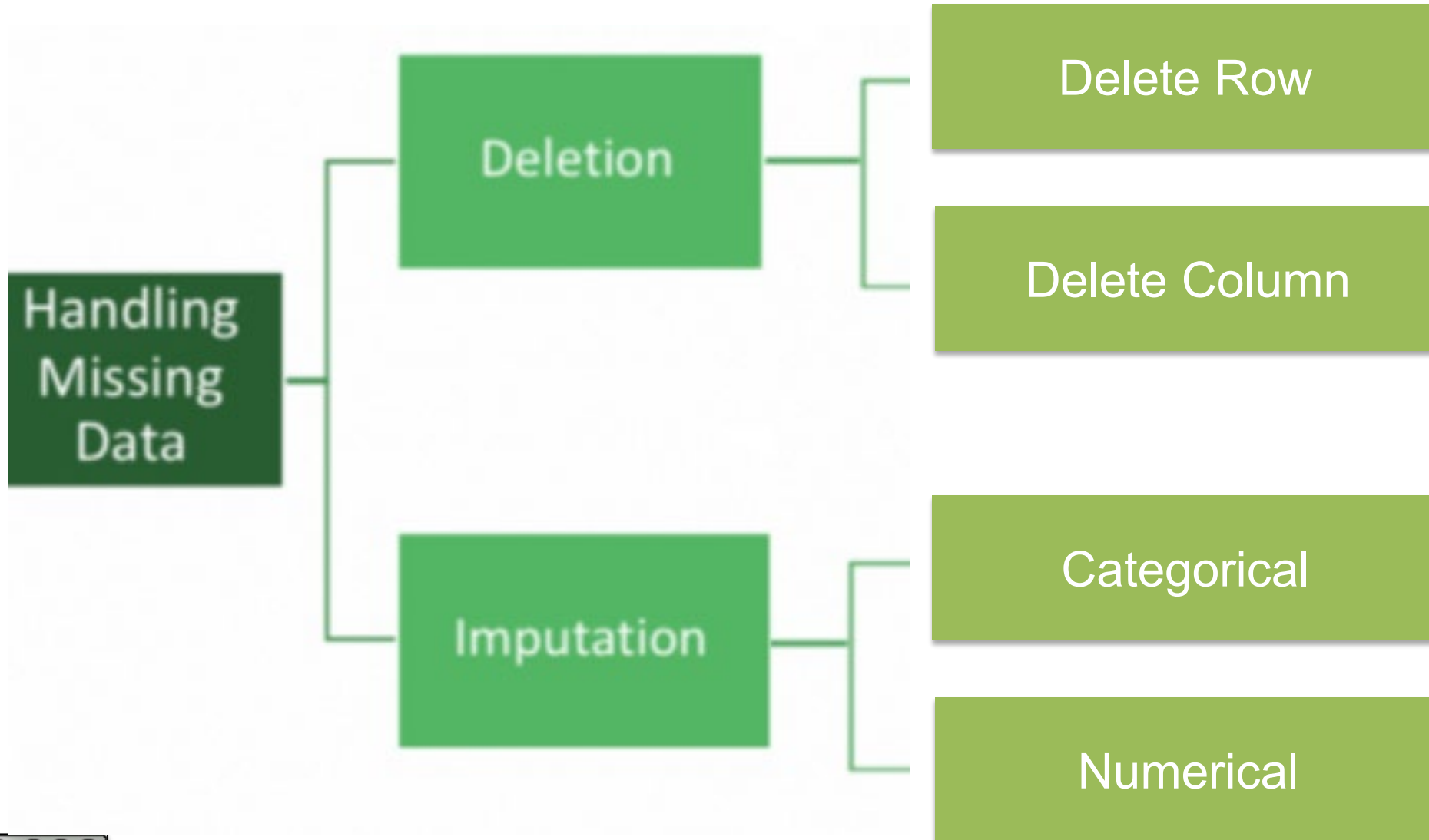




		Data
Index	0	0.0
	1	2.0
	2	3.5
	3	5.0

Series

dtype: float64



Which of the following expressions can represent a missing value in a dataset?

1. NULL
2. -1
3. #NA
4. 999

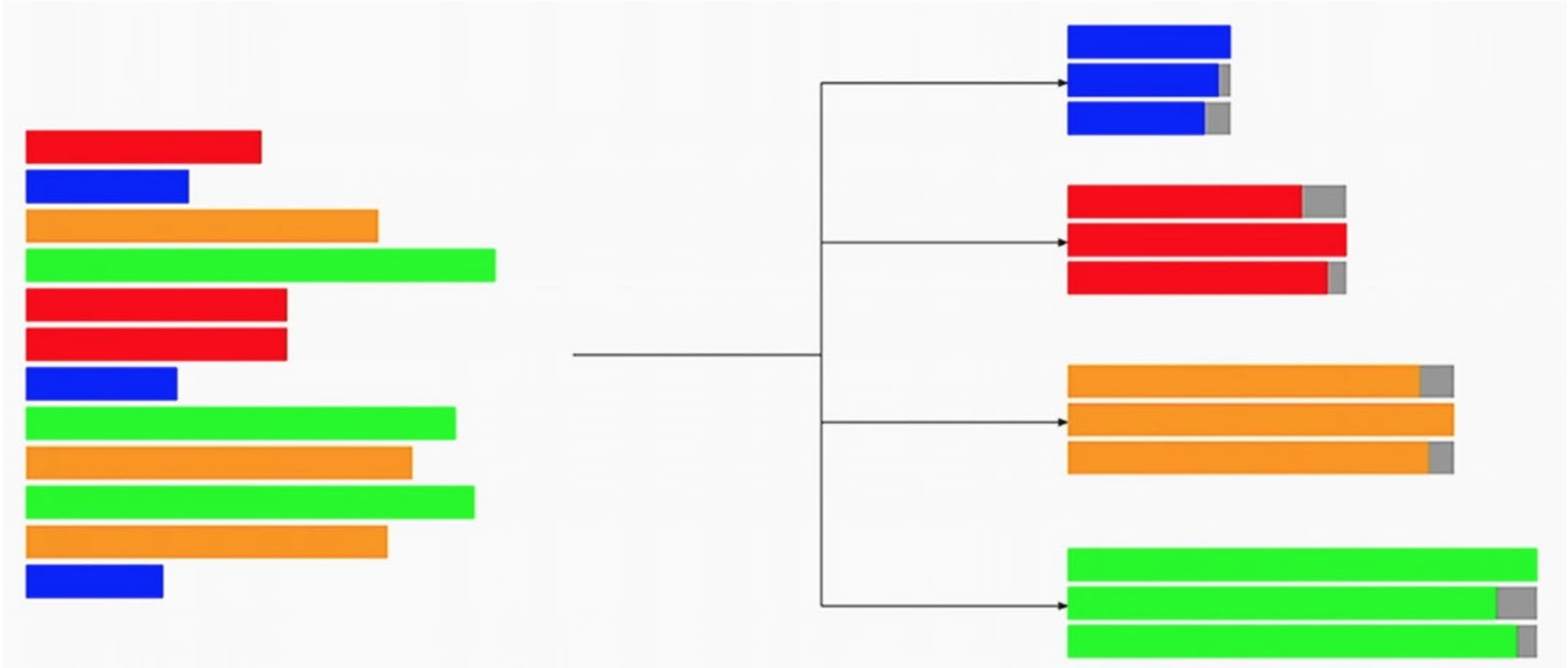
Which of the following expressions can represent a missing value in a dataset?

1. NULL

2. -1

3. #NA

4. 999



Least enjoyable part of Data Science?

Collecting data (**21%**)

Cleaning and organizing data (**57%**)

Spend most time doing

Collecting data (**19%**)

Cleaning and organizing data (**60%**)

Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

A	3
A	1
A	2



Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Split into
Groups →

A	3
A	1
A	2
B	1
B	5
B	6
C	4
C	9
C	5



Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Split into
Groups

A	3
A	1
A	2
B	1
B	5
B	6
C	4
C	9
C	5

Aggregate
Function

A	6
---	---



Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Split into
Groups

A	3
A	1
A	2
B	1
B	5
B	6
C	4
C	9
C	5

Aggregate
Function

A	6
---	---

Aggregate
Function

B	12
---	----

Aggregate
Function

C	18
---	----



Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Split into
Groups

A	3
A	1
A	2
B	1
B	5
B	6
C	4
C	9
C	5

Aggregate
Function

A	6
---	---

Aggregate
Function

B	12
---	----

Aggregate
Function

C	18
---	----

Merge
Results

A	6
B	12
C	18

The grouping operation consists of three steps:
split, apply, combine.

True or False: The pandas package will perform
one of these steps for us

1. True
2. False

The grouping operation consists of three steps:
split, apply, combine.

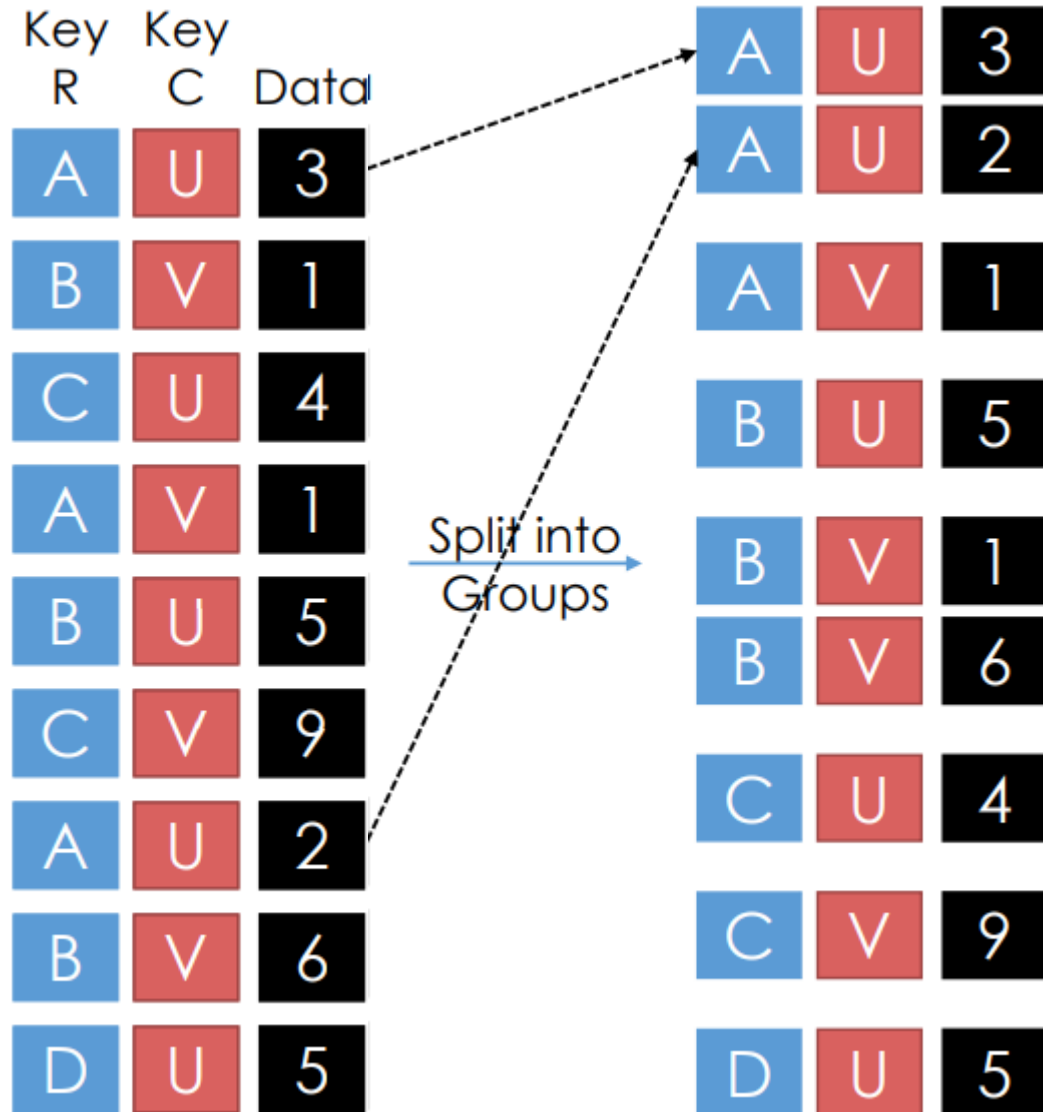
True or False: The pandas package will perform
one of these steps for us

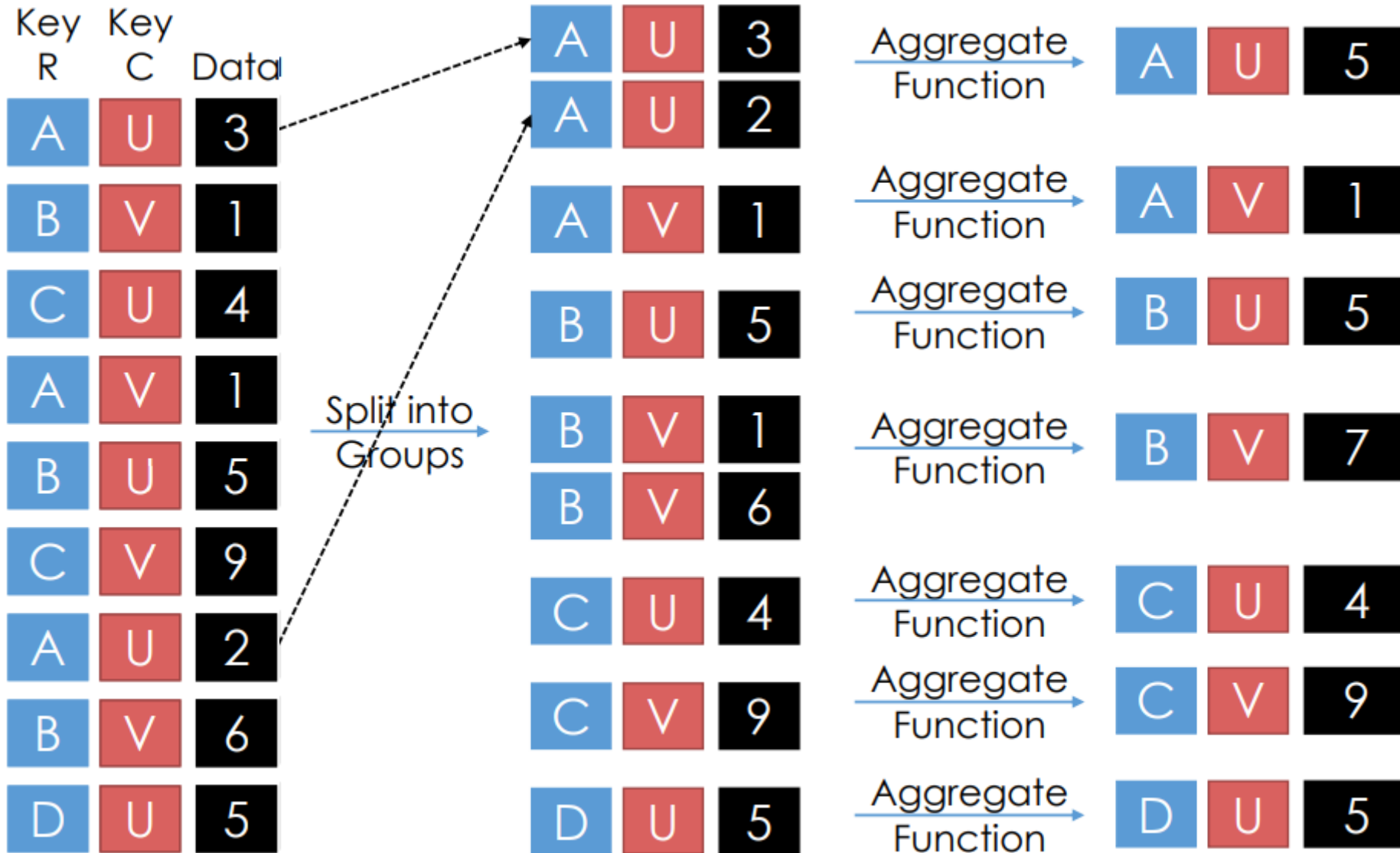
1. True

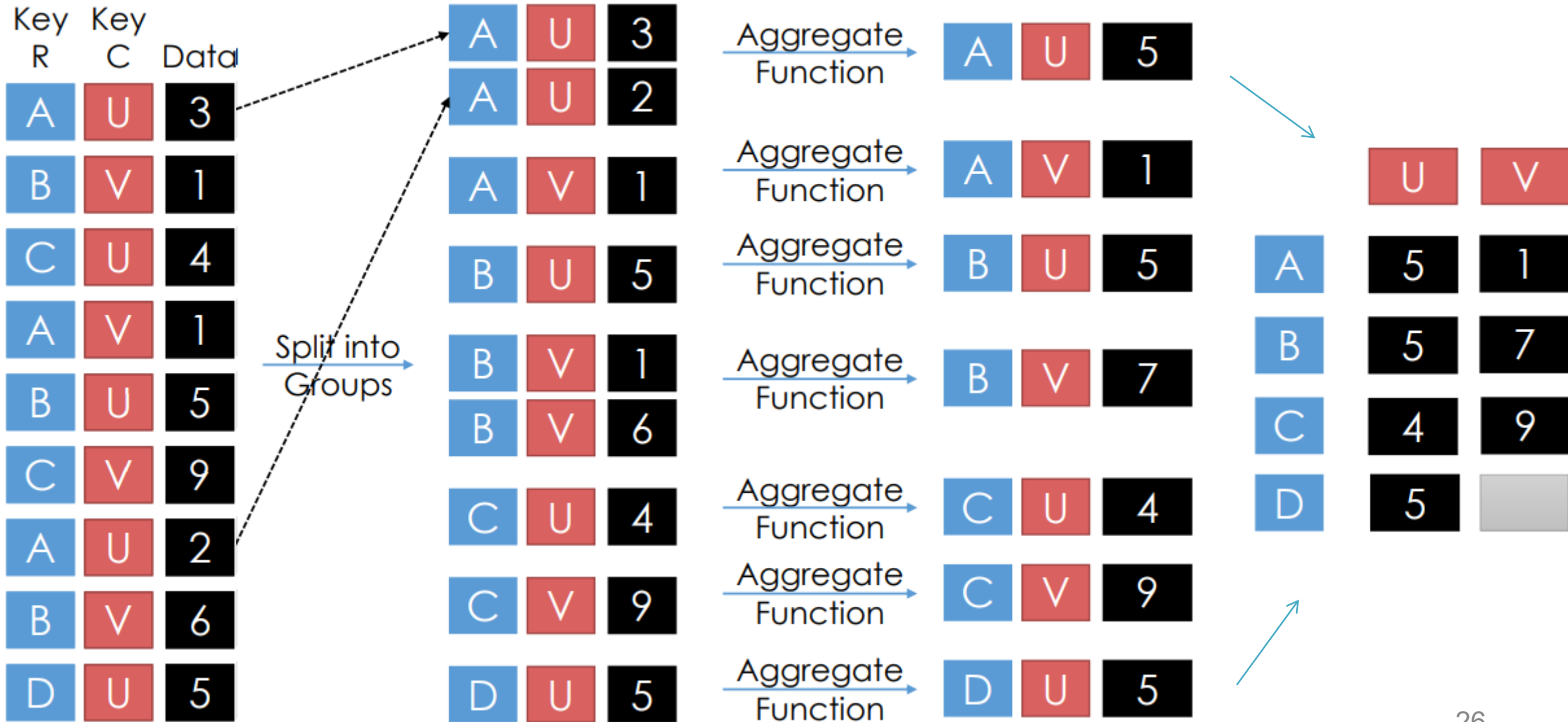
2. False



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



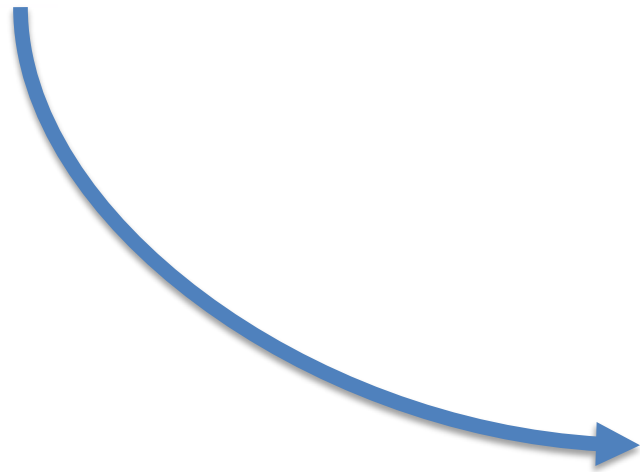








	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150



	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150



NYU

TANDON SCHOOL
OF ENGINEERING



Review

- Missing Entries
- Grouping Records
- Reshaping Tables



NYU

TANDON SCHOOL
OF ENGINEERING



References

- McKinney, Python for Data Analysis
(10.1-10.4 + 7.1-7.3)

Questions

- Describe the learning objectives.
- Summarize the relevant take-aways.
- Ask about unclear information.