WELCOME!

MPIDR DCoDe’s members, alumni and friends

Informal get together before the holidays!

14/12/2021 - Zoom

AGENDA (60 MIN)

- Welcome & information on agenda (Ali) → 3 min
- Year in review and updates (Emilio) → 5 min
- Really hard questions, *no alcohol, stay sharp!* (1 point per question, 10+1 in total → 17 min)
  - **Demography (2 questions)**
  - **Sociology (2 questions)**
  - **Data science (R, Python, SQL, 6 questions)**
  - **1 bonus question on migration to Germany!**
  - Points calculation and announcing the winner (**prize = 1 book of their choice, < 50 euro**)
  - I need a volunteer to help me (e.g., by monitoring the chat, point calculation, dropping a “next Q” in chat)
- Break out rooms with random assignment & shuffling (just chat ☺, *drinks allowed!*)
  - **1st (→ 15 min)**
  - **2nd (→ 15 min)**
  - Closing words (everyone) → 3-5 min

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

DATA SCIENCE, R (1/2)

What does “slice()” do here?

```r
library(dplyr)
library(palmerpenguins)

set.seed(2021-12-03)

sample_penguins <- penguins %>%
  group_by(species) %>%
  sample_n(3) %>%
  select(species, bill_length_mm, sex) %>%
  arrange(desc(bill_length_mm), .by_group = TRUE)

sample_penguins %>%
  slice(1)
```

DATA SCIENCE, R (1/2 → ANSWER)

Visualize %>% pipeline on last line

```
slice(1)
```

Input

	species	bill_length_mm	sex
1	Adelie	41.80	male
2	Adelie	40.30	male
3	Adelie	36.50	female
4	Chinstrap	46.50	female
5	Chinstrap	46	female
6	Chinstrap	45.40	female
7	Gentoo	50.70	male
8	Gentoo	45.10	female
9	Gentoo	43.60	female

Output

	species	bill_length_mm	sex
1	Adelie	41.80	male
2	Chinstrap	46.50	female
3	Gentoo	50.70	male

R Help

slice {dplyr}

Subset rows using their positions

Description

`slice()` lets you index rows by their (integer) locations. It allows you to select, remove, and duplicate rows. It is accompanied by a number of helpers for common use cases:

- `slice_head()` and `slice_tail()` select the first or last rows.
- `slice_sample()` randomly selects rows.
- `slice_min()` and `slice_max()` select rows with highest or lowest values of a variable.

If `.data` is a grouped df, the operation will be performed on each group, so that (e.g.) `slice_head(df, n = 5)` will select the first five rows in each group.

From: https://tidydatatutor.com/vis.html#trace=example-code/r_grouped_slice.json

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

DATA SCIENCE, SQL (1/2)

Which join is this?

migrant_id	migrant_name	region_id
m1	John	1
m2	Anne	3
m3	Robert	2
m4	Jerry	4
m5	Tim	[NULL]

region_id	region_name
1	Europe
2	Middle East
3	America
4	Africa
5	Asia

migrant_name	region_name
John	Europe
Robert	Middle East
Anne	America
Jerry	Africa
[NULL]	Asia

DATA SCIENCE, SQL (1/2 → ANSWER)

Right join
(or right outer join)

migrant_id	migrant_name	region_id
m1	John	1
m2	Anne	3
m3	Robert	2
m4	Jerry	4
m5	Tim	[NULL]

region_id	region_name
1	Europe
2	Middle East
3	America
4	Africa
5	Asia

migrant_name	region_name
John	Europe
Robert	Middle East
Anne	America
Jerry	Africa
[NULL]	Asia

```
-- # ============================
-- #### Creating 2 tables as an example ####
-- # ============================
CREATE TABLE migrants (
    migrant_id VARCHAR,
    migrant_name VARCHAR,
    region_id INTEGER
);

CREATE TABLE origin_region (
    region_id INTEGER,
    region_name VARCHAR
);

INSERT INTO migrants (
migrant_id,
migrant_name,
region_id
)
    VALUES
    ('m1', 'John', 1),
    ('m2', 'Anne', 3),
    ('m3', 'Robert', 2),
    ('m4', 'Jerry', 4),
    ('m5', 'Tim', NULL)
 ;

INSERT INTO origin_region (region_id,
region_name)
    VALUES
       (1, 'Europe'),
       (2, 'Middle East'),
       (3, 'America'),
       (4, 'Africa'),
       (5, 'Asia')
;

SELECT * from migrants ;
SELECT * from origin_region ;

SELECT migrant_name, region_name from migrants
right join origin_region
using (region_id);
```

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

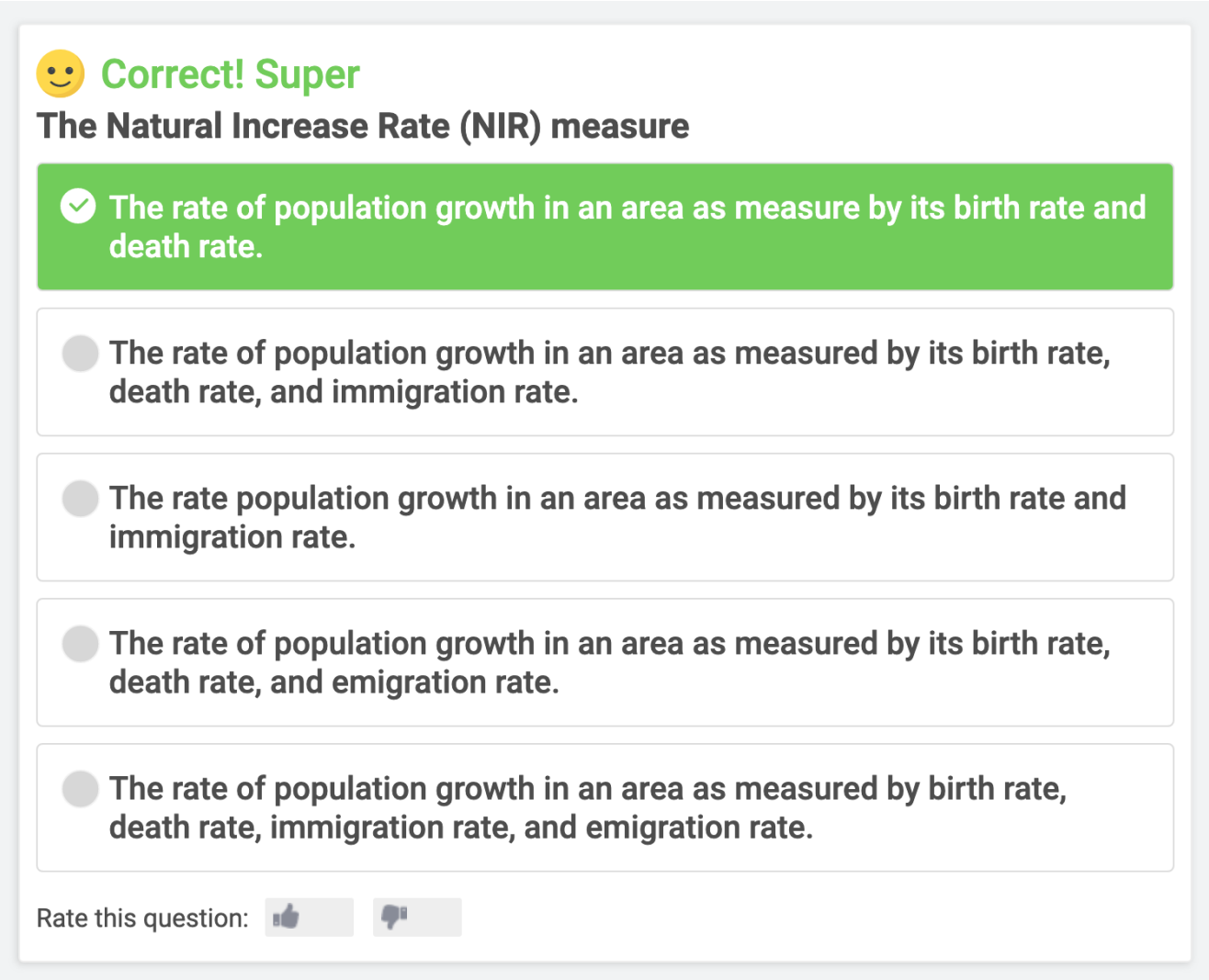
The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

DEMOGRAPHY (1/2)

The Natural Increase Rate (NIR) measure

- ○ The rate of population growth in an area as measure by its birth rate and death rate.
- ○ The rate of population growth in an area as measured by its birth rate, death rate, and immigration rate.
- ○ The rate population growth in an area as measured by its birth rate and immigration rate.
- ○ The rate of population growth in an area as measured by its birth rate, death rate, and emigration rate.
- ○ The rate of population growth in an area as measured by birth rate, death rate, immigration rate, and emigration rate.

DEMOGRAPHY (1/2 → ANSWER)



From: https://www.proprofs.com/quiz-school/playquiz/?title=njgxntm1znki

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

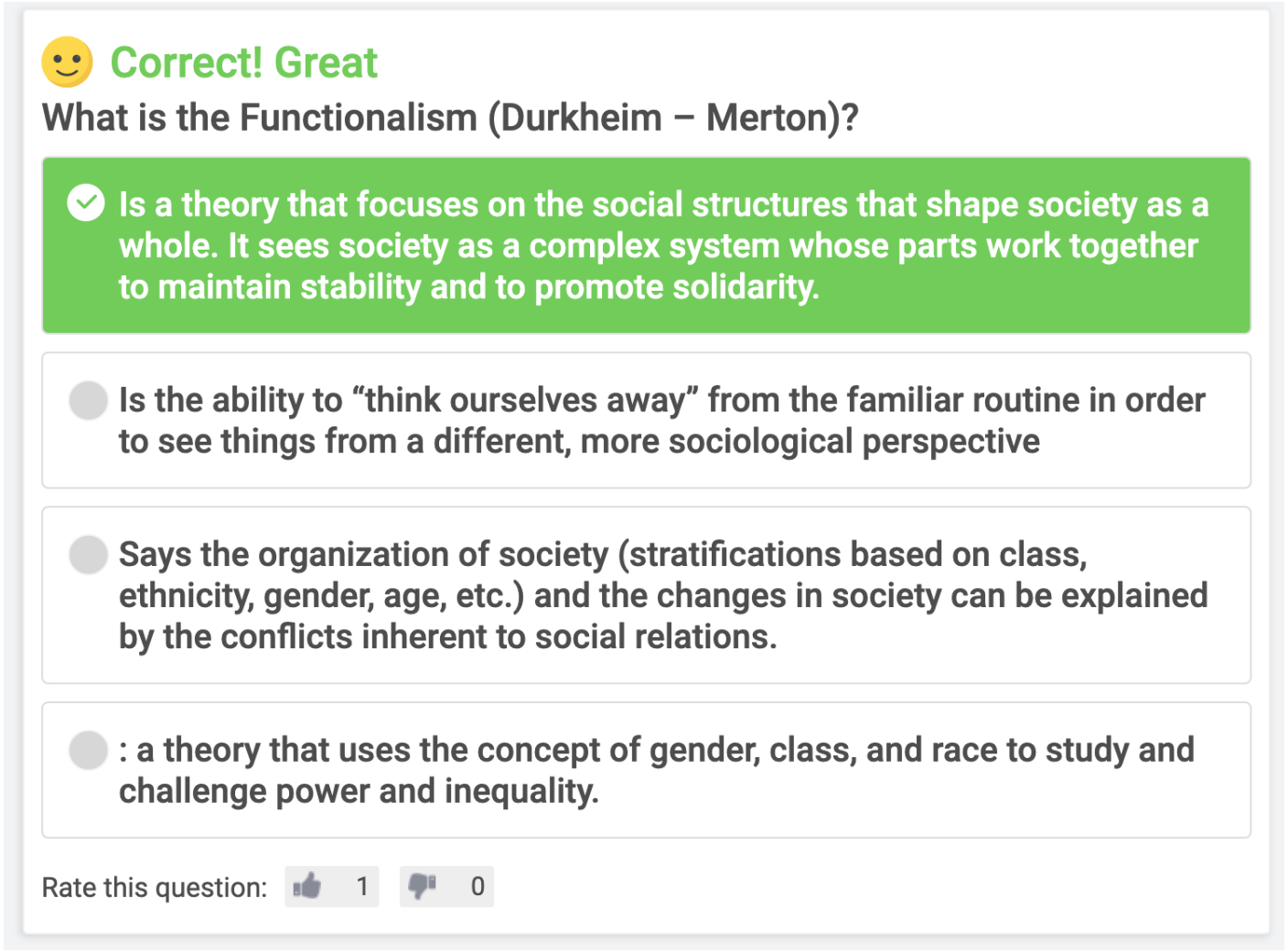
The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

SOCIOLOGY (1/2)

What is the Functionalism (Durkheim – Merton)?

- ○ Is a theory that focuses on the social structures that shape society as a whole. It sees society as a complex system whose parts work together to maintain stability and to promote solidarity.
- ○ Is the ability to “think ourselves away” from the familiar routine in order to see things from a different, more sociological perspective
- ○ Says the organization of society (stratifications based on class, ethnicity, gender, age, etc.) and the changes in society can be explained by the conflicts inherent to social relations.
- ○ : a theory that uses the concept of gender, class, and race to study and challenge power and inequality.

SOCIOLOGY (1/2 → ANSWER)



Correct! Great

What is the Functionalism (Durkheim – Merton)?

- ✓ Is a theory that focuses on the social structures that shape society as a whole. It sees society as a complex system whose parts work together to maintain stability and to promote solidarity.
- Is the ability to "think ourselves away" from the familiar routine in order to see things from a different, more sociological perspective
- Says the organization of society (stratifications based on class, ethnicity, gender, age, etc.) and the changes in society can be explained by the conflicts inherent to social relations.
- : a theory that uses the concept of gender, class, and race to study and challenge power and inequality.

Rate this question: 1 0

From: https://www.proprofs.com/quiz-school/playquiz/?title=intro-to-sociology-test-1_1

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

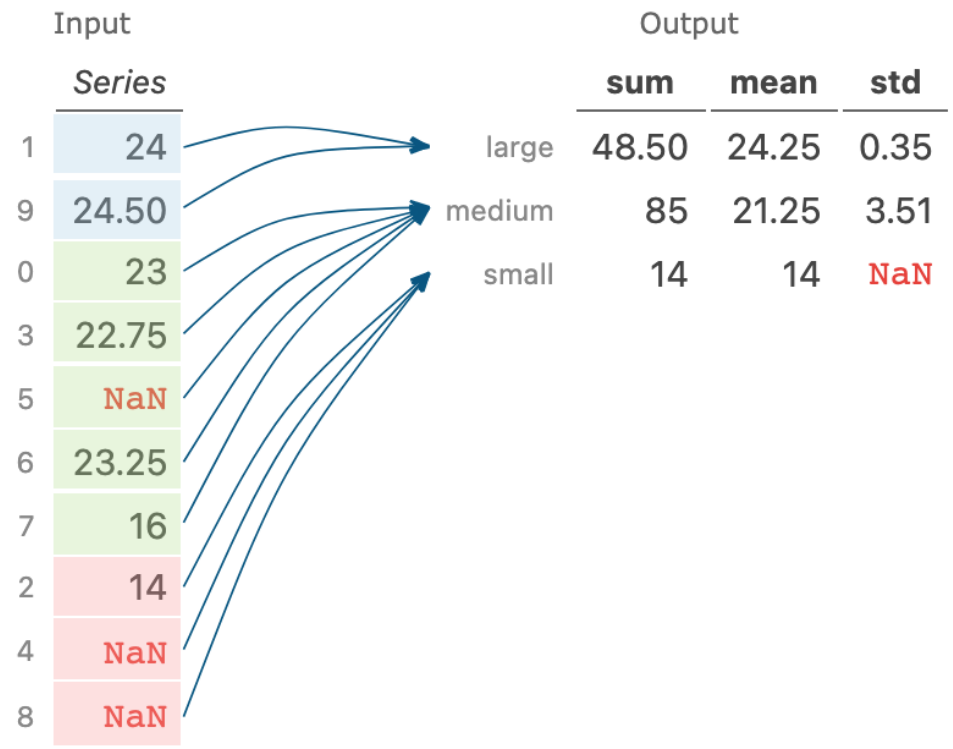
DATA SCIENCE, PYTHON (1/2)

Whad does ".agg" do here?

```
1  import pandas as pd, io
2  csv = '''
3  breed,size,weight,height
4  Labrador Retriever,medium,67.5,23.0
5  German Shepherd,large,,24.0
6  Beagle,small,,14.0
7  Golden Retriever,medium,60.0,22.75
8  Yorkshire Terrier,small,5.5,
9  Bulldog,medium,45.0,
10 Boxer,medium,,23.25
11 Poodle,medium,,16.0
12 Dachshund,small,24.0,
13 Rottweiler,large,,24.5
14 '''
15 dogs = pd.read_csv(io.StringIO(csv))
16 (dogs
17   .sort_values('size')
18   .groupby('size')['height']
19   .agg(['sum', 'mean', 'std'])
20 )
```

DATA SCIENCE, PYTHON (1/2 → ANSWER)

```
(dogs
  .sort_values('size')
  .groupby('size')['height']
  .agg(['sum', 'mean', 'std'])
)
```



Signature:
pd.DataFrame.agg(self, func=None, axis: 'Axis' = 0, *args, **kwargs)

Docstring:
Aggregate using one or more operations over the specified axis.

From: https://pandastutor.com/vis.html#trace=example-code/py_sort_groupby_agg.json

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

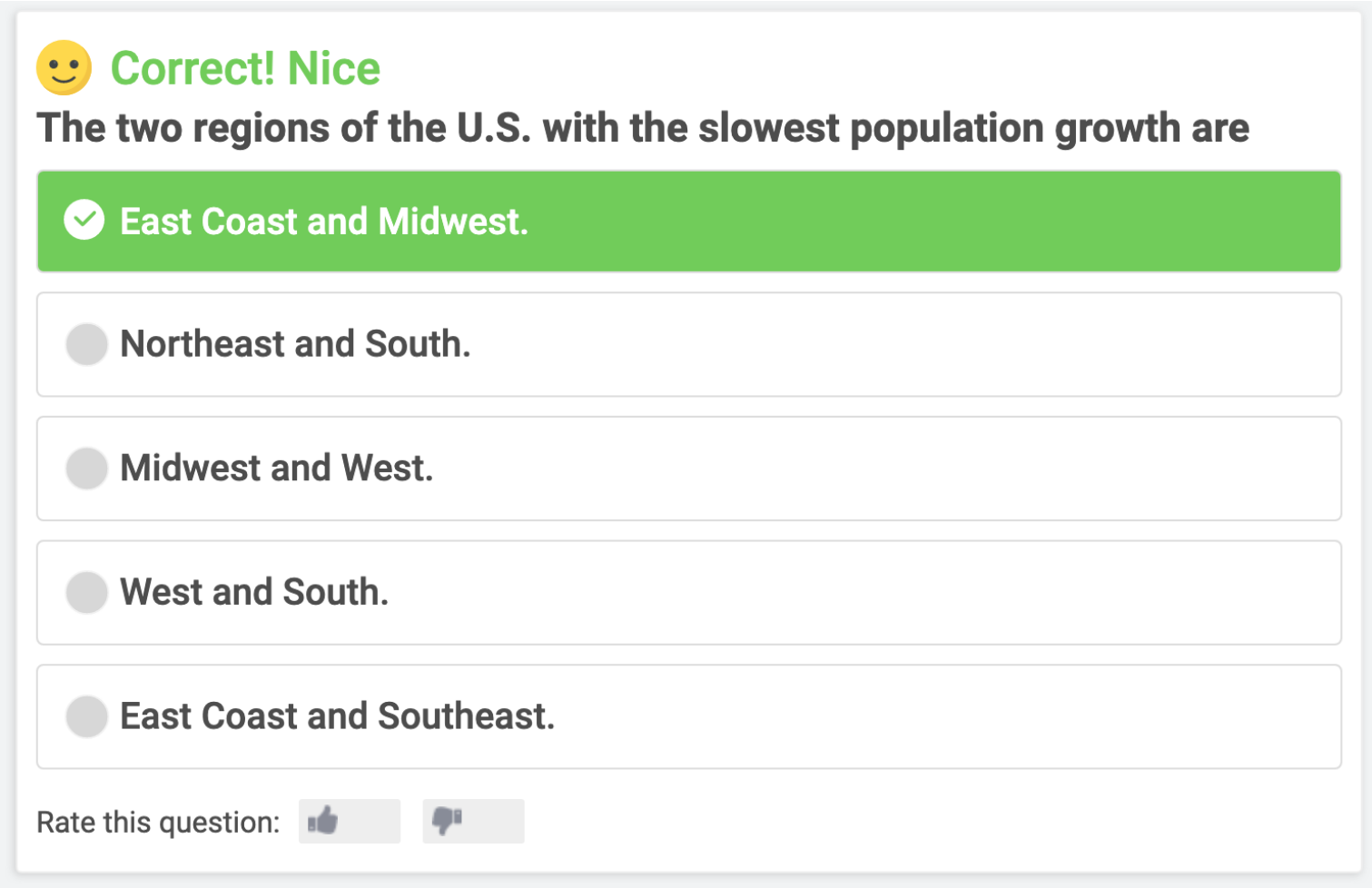
The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

DEMOGRAPHY (2/2)

The two regions of the U.S. with the slowest population growth are

- ◯ East Coast and Midwest.
- ◯ Northeast and South.
- ◯ Midwest and West.
- ◯ West and South.
- ◯ East Coast and Southeast.

DEMOGRAPHY (2/2 → ANSWER)



From: https://www.proprofs.com/quiz-school/playquiz/?title=njgxntm1znki

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

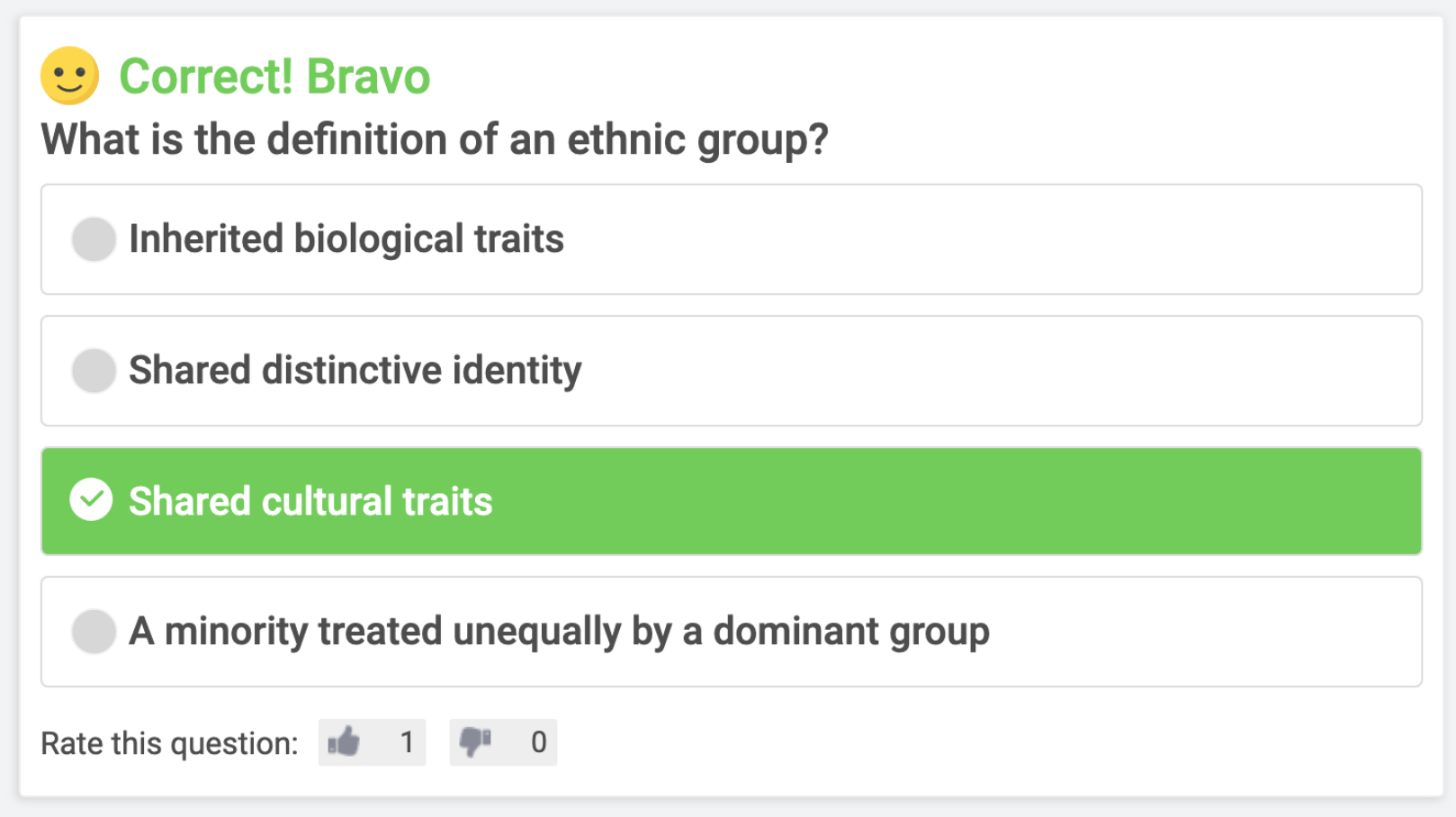
The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

SOCIOLOGY (2/2)

What is the definition of an ethnic group?

- ◯ Inherited biological traits
- ◯ Shared distinctive identity
- ◯ Shared cultural traits
- ◯ A minority treated unequally by a dominant group

SOCIOLOGY (2/2 → ANSWER)



From: https://www.proprofs.com/quiz-school/playquiz/?title=sociology-practice-test

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

DATA SCIENCE, R (2/2)

What does “nest” do here?

```
1  library(tidyverse)
2
3  df <- tibble(x = c(1, 1, 1, 2, 2, 3), y = 1:6, z = 6:1)
4  df %>% nest(y_z = c(y, z))
5
```

DATA SCIENCE, R (2/2 → ANSWER)

```r
library(tidyverse)

df <- tibble(x = c(1, 1, 1, 2, 2, 3), y = 1:6, z = 6:1)
df %>% nest(y_z = c(y, z))

```

Visualize %>% pipeline on last line

```
nest(y_z = c(y, z))
```

Input

	x	y	z
1	1	1	6
2	1	2	5
3	1	3	4
4	2	4	3
5	2	5	2
6	3	6	1

Output

	x	y_z
1	1	[{"y":1,"z":6},{"y":2,"z":5},{"y":3,"z":4}] [-]
2	2	[{"y":4,"z":3},{"y":5,"z":2}] [-]
3	3	[{"y":6,"z":1}]

R Help

nest {tidyr}

Nest and unnest

Description

Nesting creates a list-column of data frames; unnesting flattens it back out into regular columns. Nesting is implicitly a summarising operation: you get one row for each group defined by the non-nested columns. This is useful in conjunction with other summaries that work with whole datasets, most notably models.

Learn more in `vignette("nest")`.

Usage

```
nest(.data, ..., .names_sep = NULL, .key = deprecated())
```

From: https://tidydatatutor.com/vis.html#trace=example-code/r_grouped_slice.json

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

DATA SCIENCE, PYTHON (2/2)

What does ".xs" do here?

```
In [79]: coords = [("AA", "one"), ("AA", "six"), ("BB", "one"), ("BB", "two"), ("BB", "six")]

In [80]: index = pd.MultiIndex.from_tuples(coords)

In [81]: df = pd.DataFrame([11, 22, 33, 44, 55], index, ["MyData"])

In [82]: df
Out[82]:
        MyData
AA one      11
   six      22
BB one      33
   two      44
   six      55
```

```
In [84]: df.xs("six", level=1, axis=0)
```

DATA SCIENCE, PYTHON (2/2 → ANSWER)

```
In [79]: coords = [("AA", "one"), ("AA", "six"), ("BB", "one"), ("BB", "two"), ("BB", '

In [80]: index = pd.MultiIndex.from_tuples(coords)

In [81]: df = pd.DataFrame([11, 22, 33, 44, 55], index, ["MyData"])

In [82]: df
Out[82]:
        MyData
AA one      11
   six      22
BB one      33
   two      44
   six      55
```

To take the cross section of the 1st level and 1st axis the index:

```
# Note : level and axis are optional, and default to zero
In [83]: df.xs("BB", level=0, axis=0)
Out[83]:
     MyData
one      33
two      44
six      55
```

...and now the 2nd level of the 1st axis.

```
In [84]: df.xs("six", level=1, axis=0)
Out[84]:
    MyData
AA      22
BB      55
```

Signature:
pd.DataFrame.xs(self, key, axis=0, level=None, drop_level: 'bool_t' = True)

Docstring: Return cross-section from the Series/DataFrame.

This method takes a `key` argument to select data at a particular level of a MultiIndex.

From:
https://pandas.pydata.org/docs/user_guide/cookbook.html

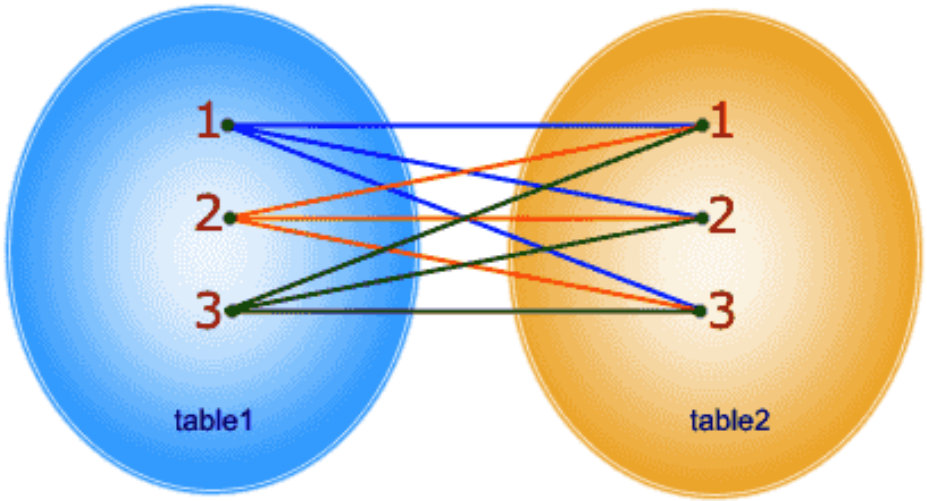
READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

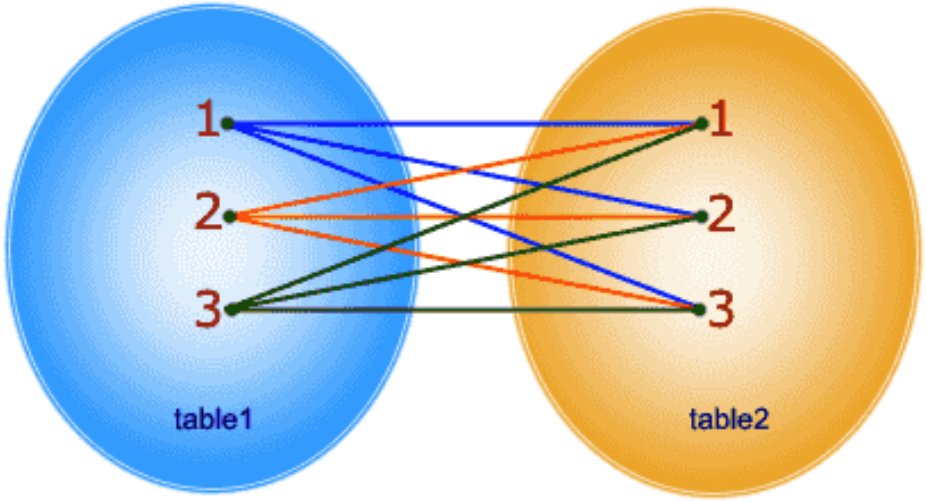
DATA SCIENCE, SQL (2/2)

Which join is this?



DATA SCIENCE, SQL (2/2 → ANSWER)

SELECT * FROM table1 CROSS JOIN table2;



In CROSS JOIN, each row from 1st table joins with all the rows of another table.
If 1st table contain x rows and y rows in 2nd one the result set will be x * y rows.

migrant_id	migrant_name	region_id
m1	John	1
m2	Anne	3
m3	Robert	2
m4	Jerry	4
m5	Tim	[NULL]

region_id	region_name
1	Europe
2	Middle East
3	America
4	Africa
5	Asia

migrant_id	migrant_name	region_id	region_id	region_name
m1	John	1	1	Europe
m2	Anne	3	1	Europe
m3	Robert	2	1	Europe
m4	Jerry	4	1	Europe
m5	Tim	[NULL]	1	Europe
m1	John	1	2	Middle East
m2	Anne	3	2	Middle East
m3	Robert	2	2	Middle East
m4	Jerry	4	2	Middle East
m5	Tim	[NULL]	2	Middle East
m1	John	1	3	America
m2	Anne	3	3	America
m3	Robert	2	3	America
m4	Jerry	4	3	America
m5	Tim	[NULL]	3	America
m1	John	1	4	Africa
m2	Anne	3	4	Africa
m3	Robert	2	4	Africa
m4	Jerry	4	4	Africa
m5	Tim	[NULL]	4	Africa
m1	John	1	5	Asia
m2	Anne	3	5	Asia
m3	Robert	2	5	Asia
m4	Jerry	4	5	Asia
m5	Tim	[NULL]	5	Asia

From: https://www.w3resource.com/sql/joins/cross-join.php

READY?!

Fingers on the keyboard, type an “a” in the chat if you have the answer!

The first person typing gets to answer and collects the point (if correct) and receives a negative point (if incorrect)

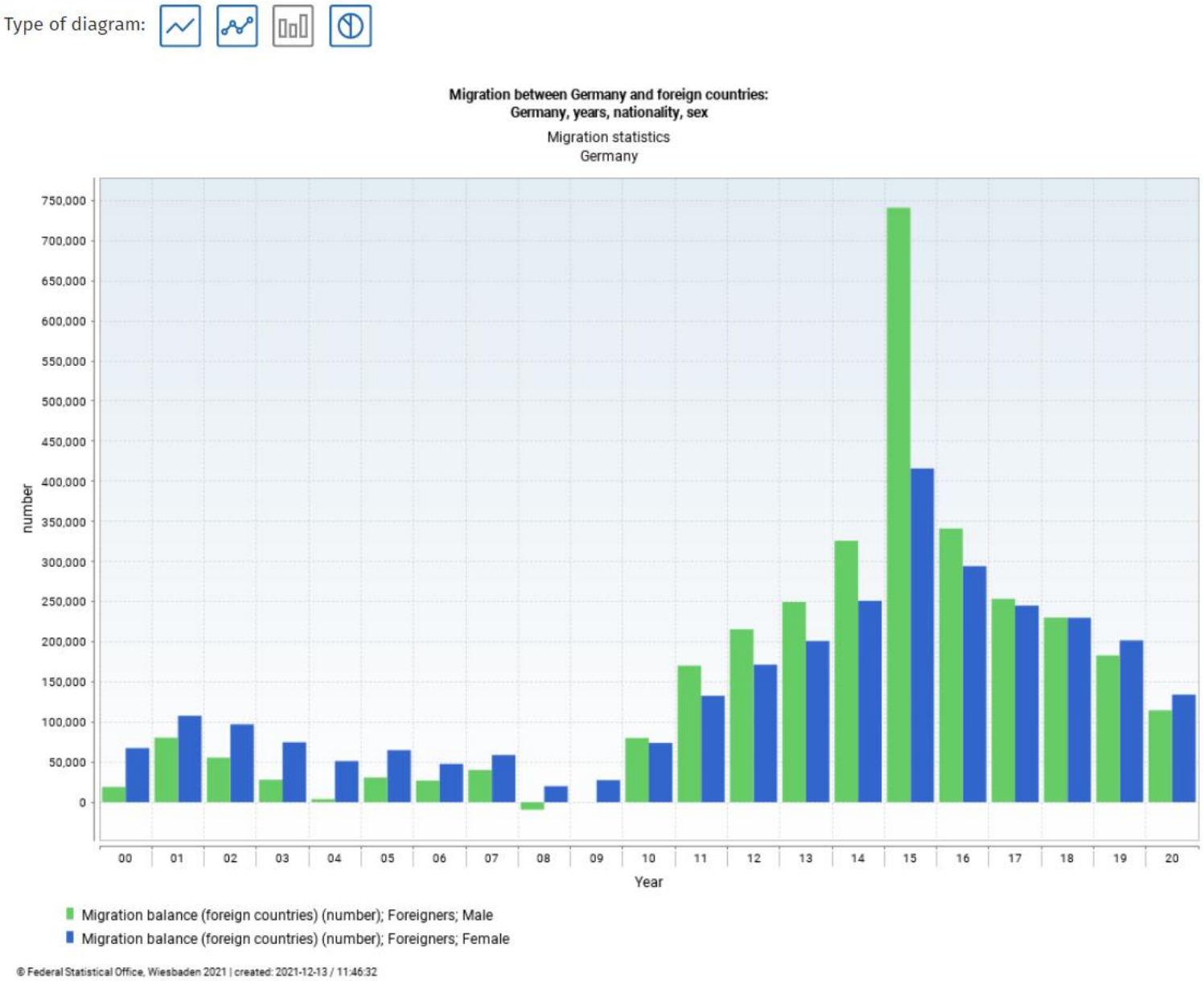
1 BONUS QUESTION ON MIGRATION TO GERMANY!

Do you think more foreign **males** are migrating to Germany or more **females** (from 2000 – 2020, general population)?

1. In **most** years more foreign **males** have arrived to Germany
2. In **most** years more foreign **females** have arrived to Germany
3. **Equal** rates of **males** and **females** have arrived to Germany

1 BONUS QUESTION ON MIGRATION TO GERMANY! (→ ANSWER)

- In **12** years, more females have arrived (blue bars)
- In **8** years, more males have arrived (green bars)
- *Is this data correct?!*



From: https://www-genesis.destatis.de

12711-0005: Migration between Germany and foreign countries: Germany, years, nationality, sex

THE WINNER IS?!