DATA SCIENCE FOR ECONOMISTS

ECON 220 LAB

Jafet Baca-Obando

Week 6, Handling IPUMS data – 10/03/2025

Outline

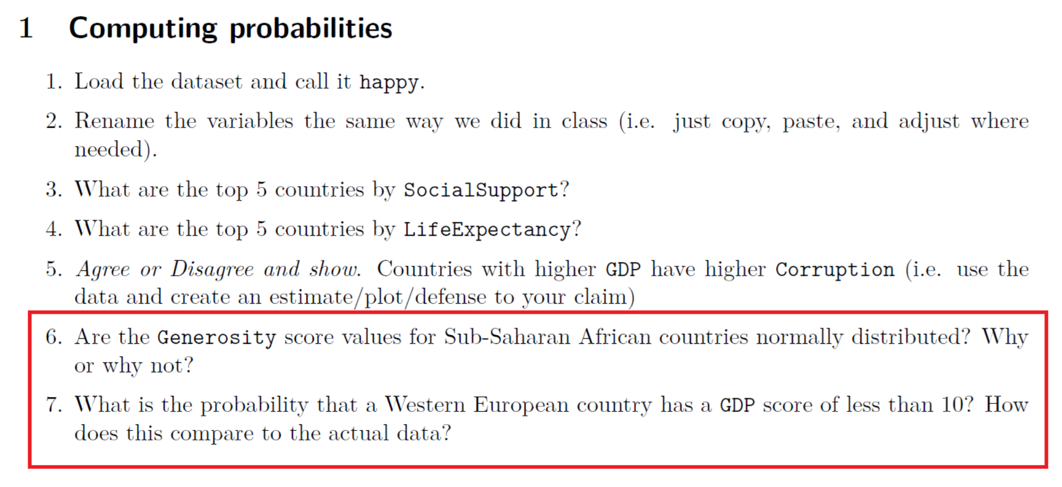
01 Introduction to IPUMS

02 Dummy variables

03 Categorical variables

Announcement

- Items 6 and 7 of Data Exercise 3 will be worth extra credit if correct (2 points)



1 Computing probabilities

1. Load the dataset and call it `happy`.
2. Rename the variables the same way we did in class (i.e. just copy, paste, and adjust where needed).
3. What are the top 5 countries by `SocialSupport`?
4. What are the top 5 countries by `LifeExpectancy`?
5. *Agree or Disagree and show.* Countries with higher `GDP` have higher `Corruption` (i.e. use the data and create an estimate/plot/defense to your claim)
6. Are the `Generosity` score values for Sub-Saharan African countries normally distributed? Why or why not?
7. What is the probability that a Western European country has a `GDP` score of less than 10? How does this compare to the actual data?

What is IPUMS?

Integrated Public Use Microdata Series

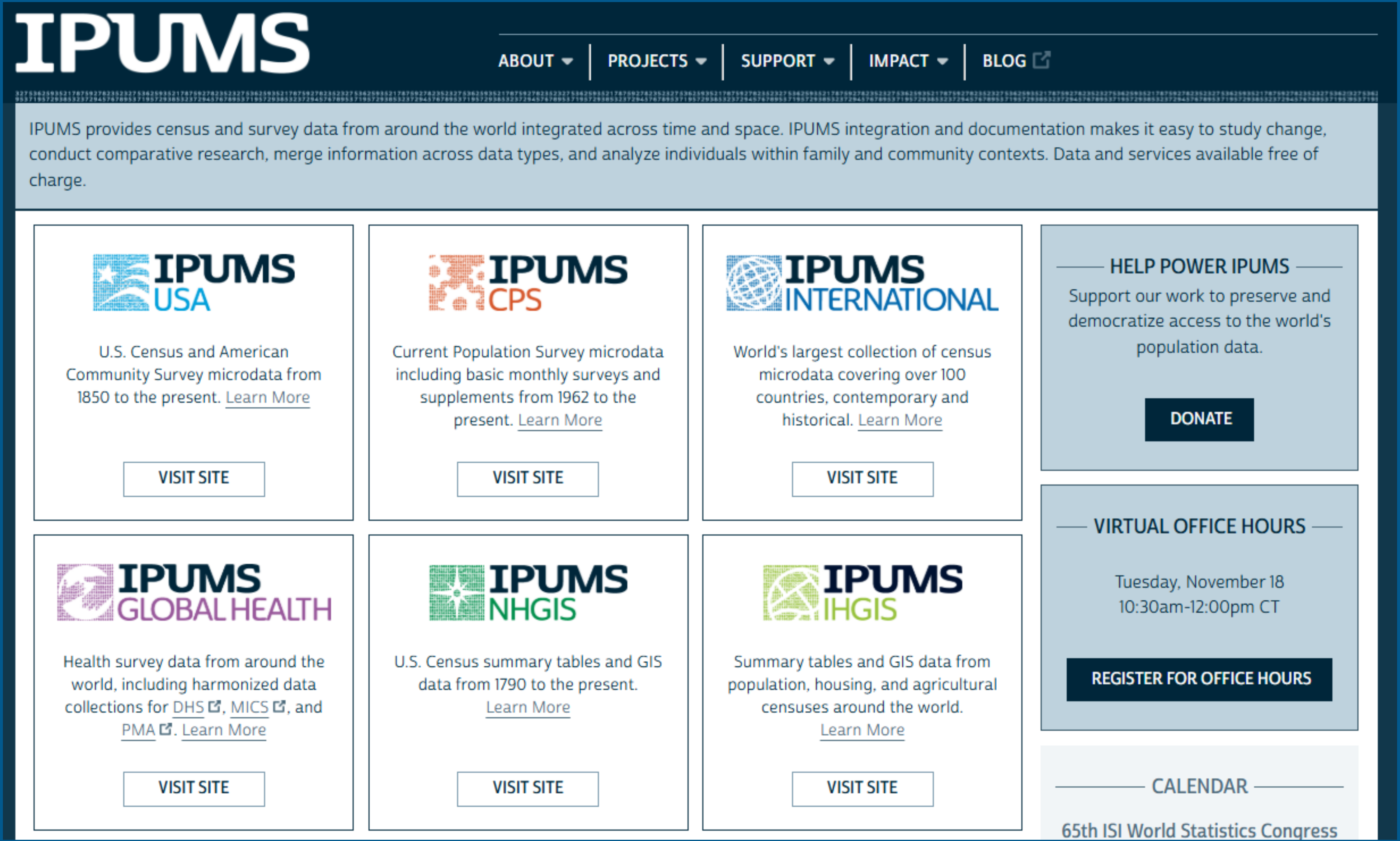
Operated by the University of Minnesota

Core mission: data harmonization

IPUMS takes datasets that were originally collected with different questions, codes, and variable names and makes them consistent. Lots of recoding!

Free access!

By providing access to detailed, anonymized individual-level data (microdata), IPUMS allows researchers to ask complex questions that can't be answered with aggregated summary tables.



IPUMS provides census and survey data from around the world integrated across time and space. IPUMS integration and documentation makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community contexts. Data and services available free of charge.

IPUMS USA

U.S. Census and American Community Survey microdata from 1850 to the present. Learn More

VISIT SITE

IPUMS CPS

Current Population Survey microdata including basic monthly surveys and supplements from 1962 to the present. Learn More

VISIT SITE

IPUMS INTERNATIONAL

World's largest collection of census microdata covering over 100 countries, contemporary and historical. Learn More

VISIT SITE

IPUMS GLOBAL HEALTH

Health survey data from around the world, including harmonized data collections for DHS, MICS, and PMA. Learn More

VISIT SITE

IPUMS NHGIS

U.S. Census summary tables and GIS data from 1790 to the present. Learn More

VISIT SITE

IPUMS IHGIS

Summary tables and GIS data from population, housing, and agricultural censuses around the world. Learn More

VISIT SITE

HELP POWER IPUMS

Support our work to preserve and democratize access to the world's population data.

DONATE

VIRTUAL OFFICE HOURS

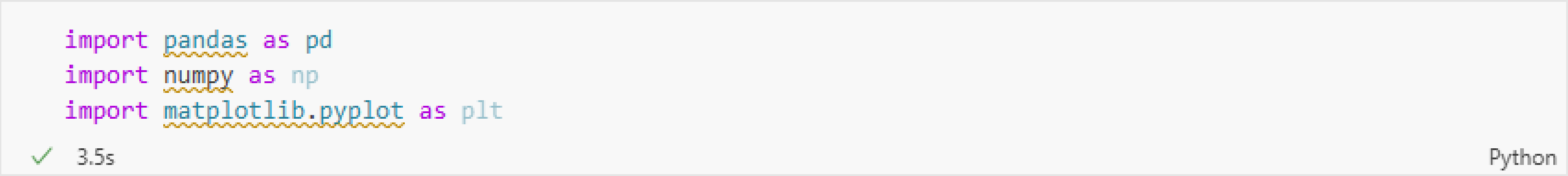
Tuesday, November 18
10:30am-12:00pm CT

REGISTER FOR OFFICE HOURS

CALENDAR

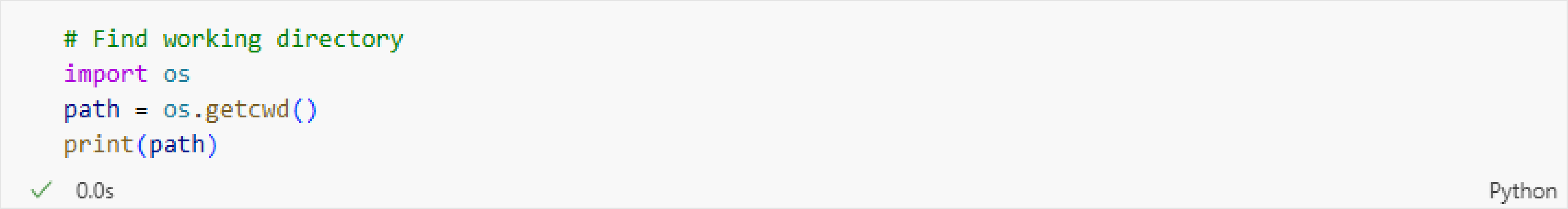
65th ISI World Statistics Congress

Importing required libraries and dataset



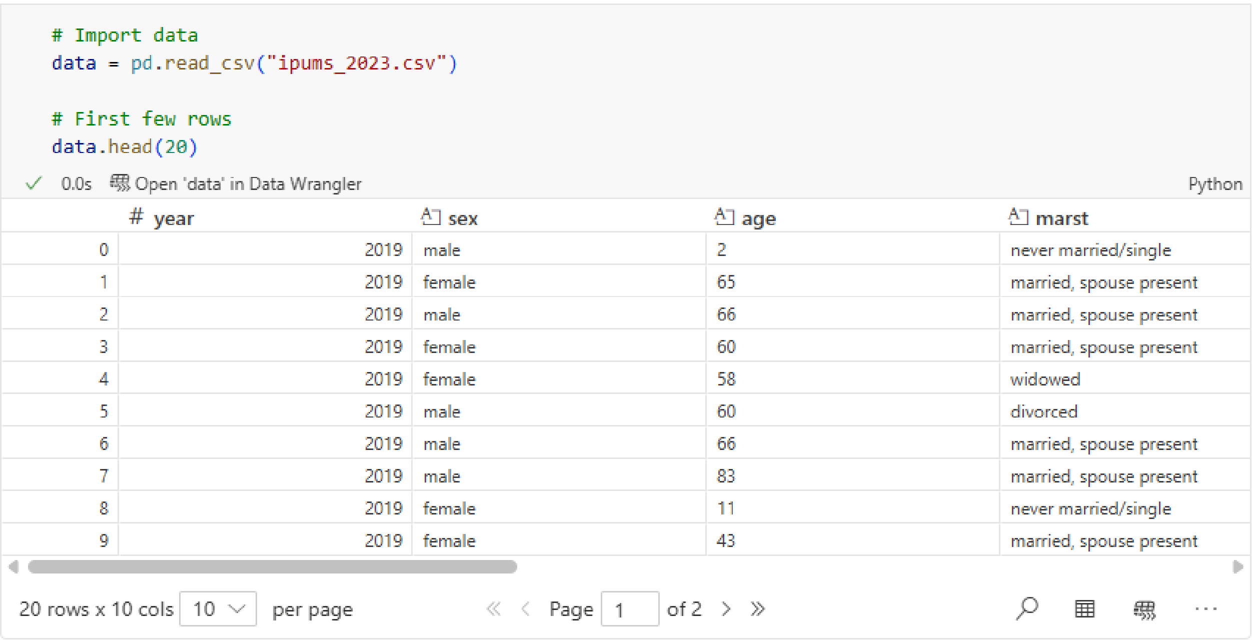
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

- New package to find the path: **os**



```python
# Find working directory
import os
path = os.getcwd()
print(path)
```

Load the data



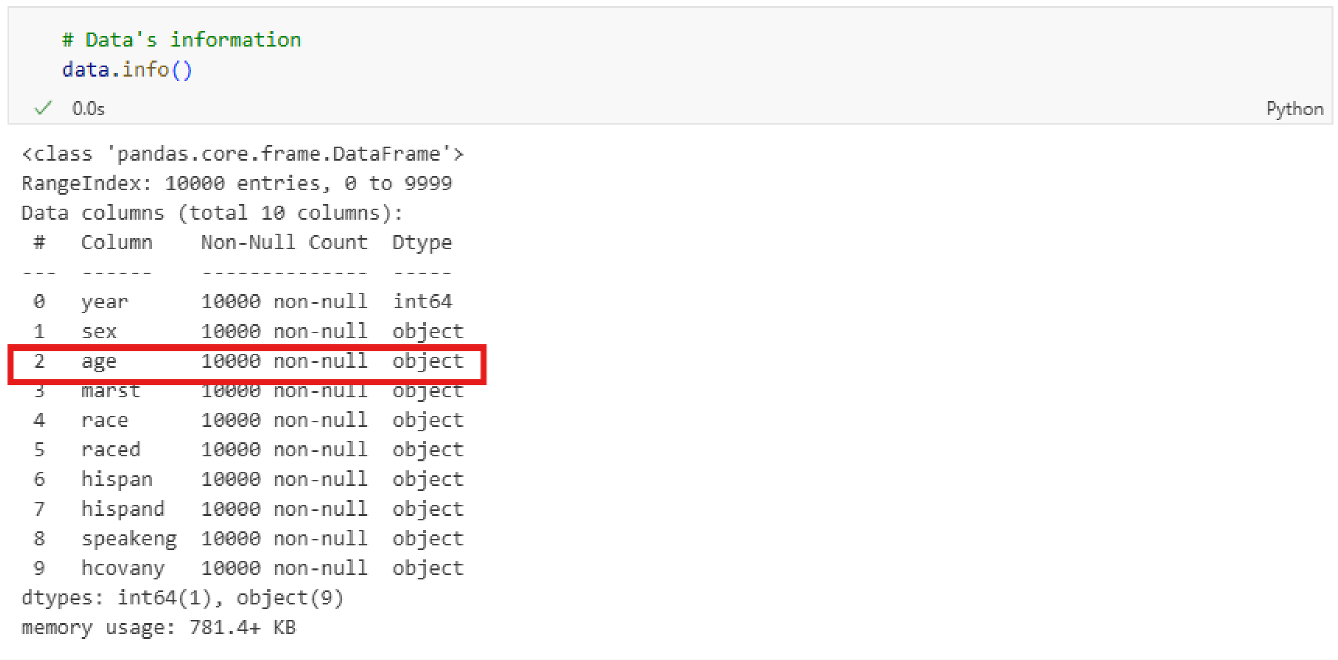
```python
# Import data
data = pd.read_csv("ipums_2023.csv")

# First few rows
data.head(20)
```

✓ 0.0s Open 'data' in Data Wrangler Python

	# year	sex	age	marst
0	2019	male	2	never married/single
1	2019	female	65	married, spouse present
2	2019	male	66	married, spouse present
3	2019	female	60	married, spouse present
4	2019	female	58	widowed
5	2019	male	60	divorced
6	2019	male	66	married, spouse present
7	2019	male	83	married, spouse present
8	2019	female	11	never married/single
9	2019	female	43	married, spouse present

20 rows x 10 cols 10 per page Page 1 of 2

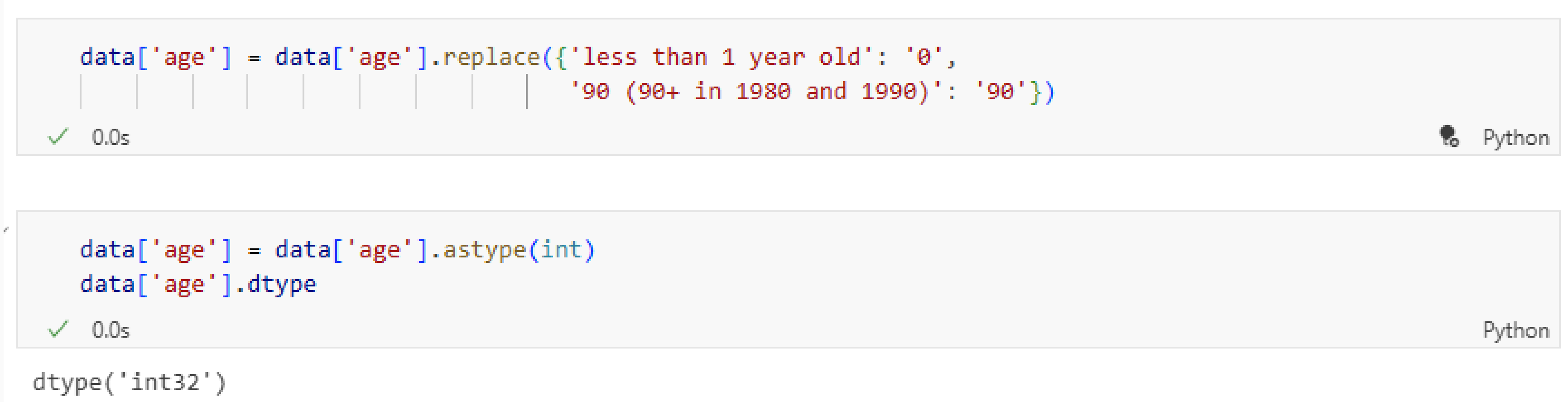
Debugging 101: Why is **age** an *object*?



```python
# Data's information
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   year      10000 non-null  int64 
 1   sex       10000 non-null  object
 2   age       10000 non-null  object
 3   marst     10000 non-null  object
 4   race      10000 non-null  object
 5   raced     10000 non-null  object
 6   hispan    10000 non-null  object
 7   hispand   10000 non-null  object
 8   speakeng  10000 non-null  object
 9   hcovany   10000 non-null  object
dtypes: int64(1), object(9)
memory usage: 781.4+ KB
```

Correcting issue with .replace()



```python
data['age'] = data['age'].replace({'less than 1 year old': '0',
                                   '90 (90+ in 1980 and 1990)': '90'})
```

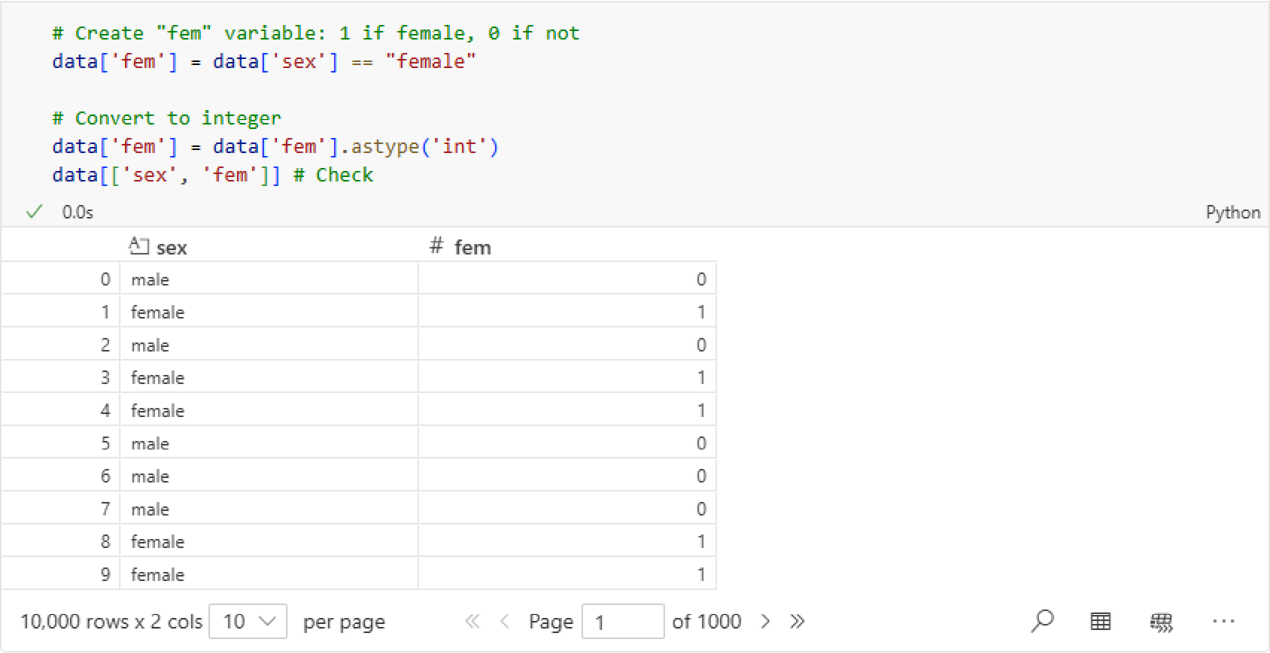
```python
data['age'] = data['age'].astype(int)
data['age'].dtype
```

```
dtype('int32')
```

Dummy variables

- **Binary** variable used to represent **categorical** data.
- It takes a value of 1 if a **certain characteristic is present** and 0 if it is not.
- For example, in a dataset of workers, we could create a female dummy where female = 1 for women and female = 0 otherwise.

Dummy variable **female**



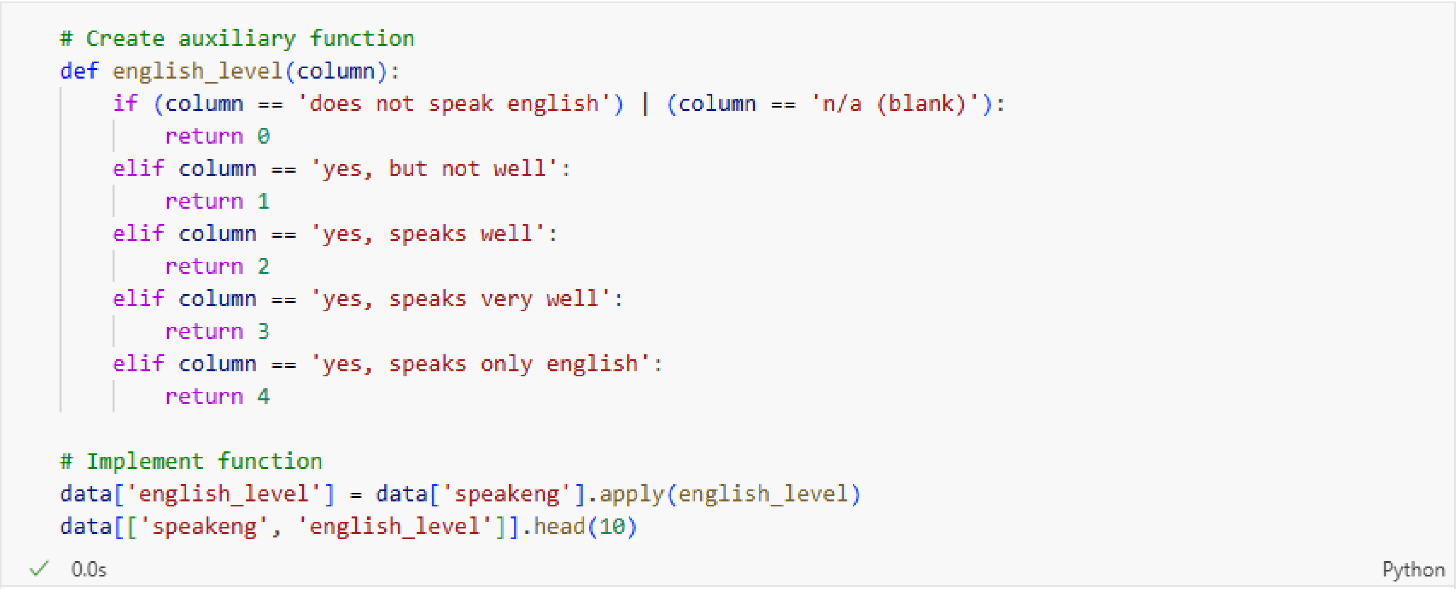
```python
# Create "fem" variable: 1 if female, 0 if not
data['fem'] = data['sex'] == "female"

# Convert to integer
data['fem'] = data['fem'].astype('int')
data[['sex', 'fem']] # Check
```

	sex	fem
0	male	0
1	female	1
2	male	0
3	female	1
4	female	1
5	male	0
6	male	0
7	male	0
8	female	1
9	female	1

Categorical variables

- They represent distinct groups or categories.
- These variables can be **nominal** (no natural order), or **ordinal** (with a meaningful order)
- Examples:
  - Education Level: "High School," "Bachelor's," "Master's"
  - Region: "North," "South," "East," "West“
  - Credit Rating: "Poor," "Fair," "Good," "Excellent"

Example – Categorizing English proficiency



```python
# Create auxiliary function
def english_level(column):
    if (column == 'does not speak english') | (column == 'n/a (blank)'):
        return 0
    elif column == 'yes, but not well':
        return 1
    elif column == 'yes, speaks well':
        return 2
    elif column == 'yes, speaks very well':
        return 3
    elif column == 'yes, speaks only english':
        return 4

# Implement function
data['english_level'] = data['speakeng'].apply(english_level)
data[['speakeng', 'english_level']].head(10)
```

Recap

- We handled IPUMS data.
- Introduced the concepts of dummy and categorical variables.
- Implemented a few coding examples.

To-do list

- **Complete Data Exercise 3**
  - Upload Jupyter notebook (.ipynb file) and HTML file on **October 5**
- **Complete Data Exercise 4**
  - Upload Jupyter notebook (.ipynb file) and HTML file on **October 12**