**TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC**

**🕮**

**Text

Description automatically generated**

**LAB 3**

MÔN HỌC :Phân Tích Dữ Liệu với Pandas

MÃ MÔN : DAT 205

Báo Cáo Bài Tập môn Học

**Sinh viên thực hiện :**  *Trần Gia Lạc*

**MSSV :** *Ps31064*

**Giảng viên hướng dẫn :** Văn Công Khanh

**Mục Lục**

[MỤC TIÊU: 3](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191195)

[PHẦN I 3](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191196)

[Bài 1 (2 điểm) 3](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191197)

[Bài 2 (2 điểm) 5](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191198)

[PHẦN II 6](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191199)

[Bài 3 (2 điểm) 6](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191200)

[Bài 4 (2 điểm) 7](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191201)

[Bài 5: 8](file:///H:\Kho%20dữ%20liệu\PJ\PS31064_Trần%20Gia%20Lạc_%20DP18301_lab7.docx#_Toc157191202)

# MỤC TIÊU:

# PHẦN I

## Bài 1 (2 điểm)

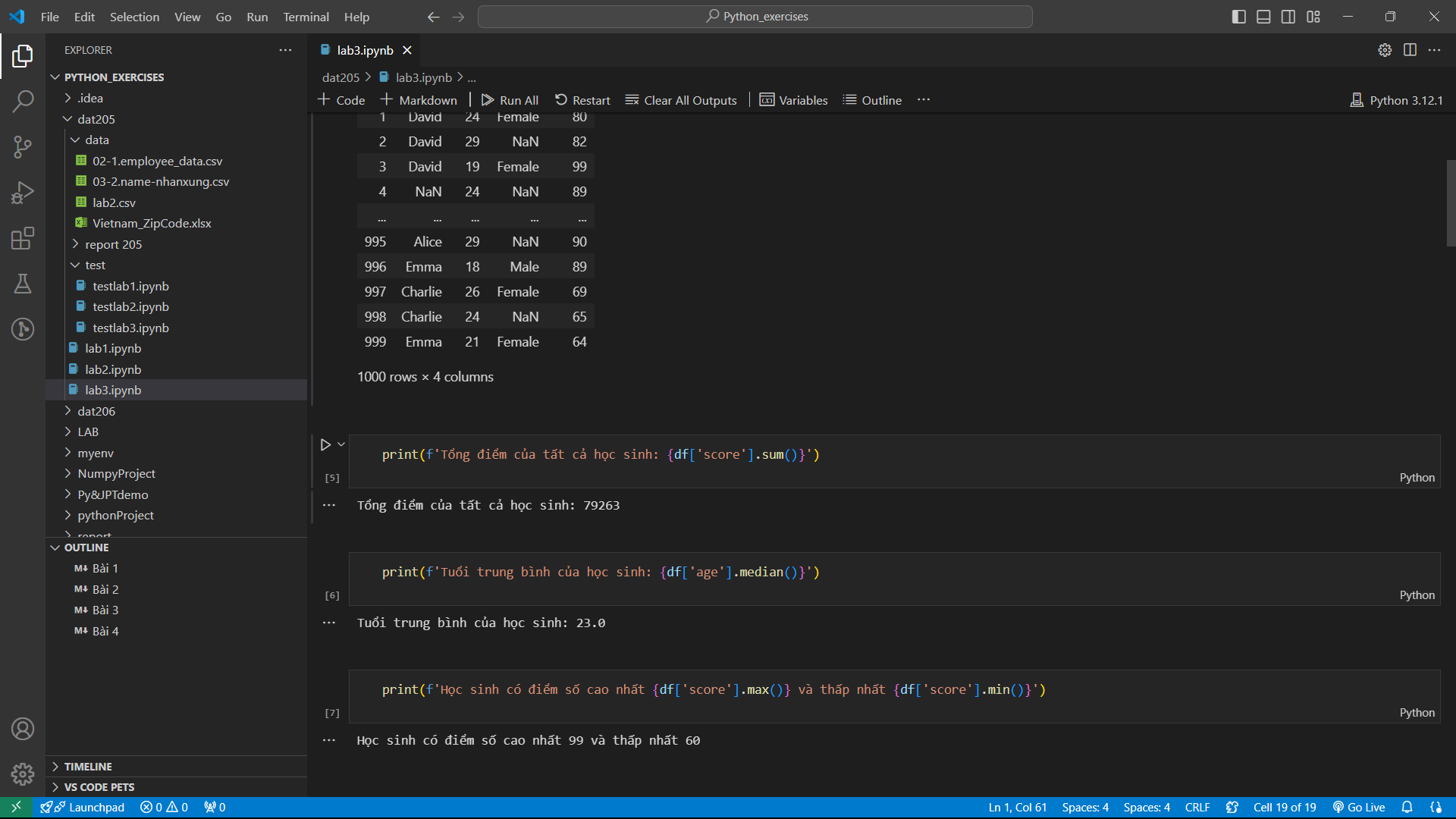
Dựa vào lab2.csv

- Tính tổng điểm của tất cả các học sinh.

- Tính tuổi trung bình của học sinh.

- Tìm học sinh có điểm cao nhất và thấp nhất.

**Bài làm:**



* Ở tuổi trung bình của học sinh vì tuổi thì không thể nào có số thập phân nên chọn tuổi ở đoạn trung vị thì sẽ hợp lí hơn

## Bài 2 (2 điểm)

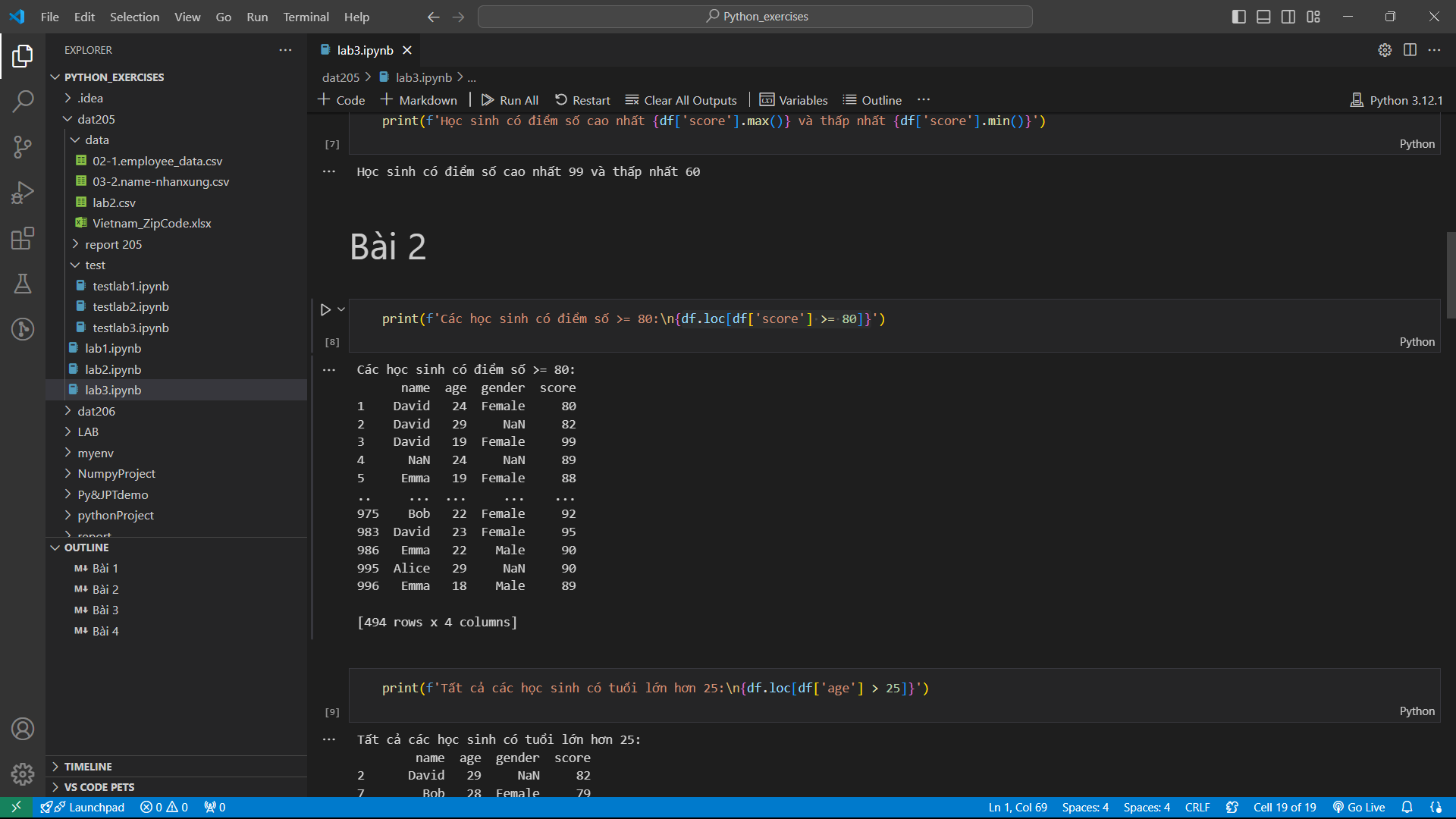
Dựa vào lab2.csv

- Lấy tất cả các hàng của các học sinh có điểm số lớn hơn hoặc bằng 80.

- Lấy các hàng của học sinh có tuổi lớn hơn 25.

- Lấy các hàng của học sinh có tên bắt đầu bằng "A".

**Bài làm:**

A screenshot of a computer

Description automatically generated

* **df.loc[df['score'] >= 80]:** hàm này sẽ dùng filter là “*df['score'] >= 80*” để chọn ra các hàng có số tuổi lớn hơn 25
* **df[df['name'].notna() & df['name'].str.startswith('A')]:** vẫn sử dụng filter để chọn hàng, ***df['name'].notna()***: chọn những hàng có dữ liệu, **df['name'].str.startswith('A')** chọn những hàng bắt đầu bằng chữ “A” và ở code trên nó áp dụng cùng lúc cả 2 điều kiện để cho ra kết quả. Nếu không **notna()** thì nó sẽ không hiểu những hàng không có dữ liệu sử lí như nào

# PHẦN II

## Bài 3 (2 điểm)

Dựa vào lab2.csv

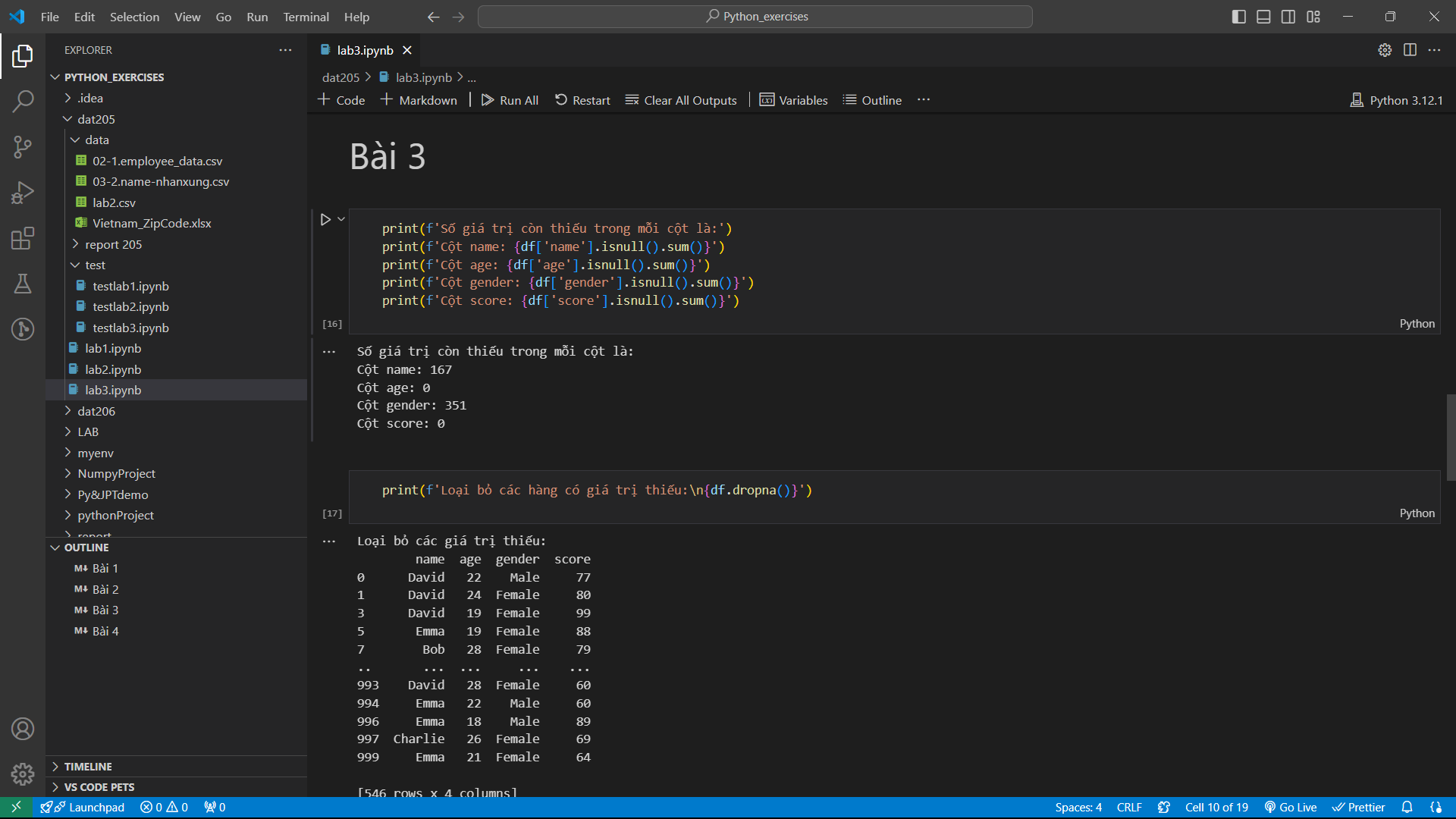
- Đếm số lượng giá trị thiếu trong mỗi cột.

- Loại bỏ các hàng có giá trị thiếu trong DataFrame.

- Loại bỏ các hàng trùng lặp trong DataFrame.

- Tìm và loại bỏ trong cột 'score' (dưới 60 và trên 100).

**Bài làm:**

 A screenshot of a computer

Description automatically generated

* **isnull().sum():** chọn các ô chứa giá trị trống và đếm tổng lại
* **dropna():** loại bỏ các hàng còn thiếu
* **drop.duplicates():** xóa bỏ hàng trùng lặp
* **df[(df['score'] >= 60) & (df['score'] <= 100)]:** áp dụng filter với 2 điều kiện cùng lúc. (do cột **score** trong data thỏa điều kiện nên dữ liệu không thay đổi)

## Bài 4 (2 điểm)

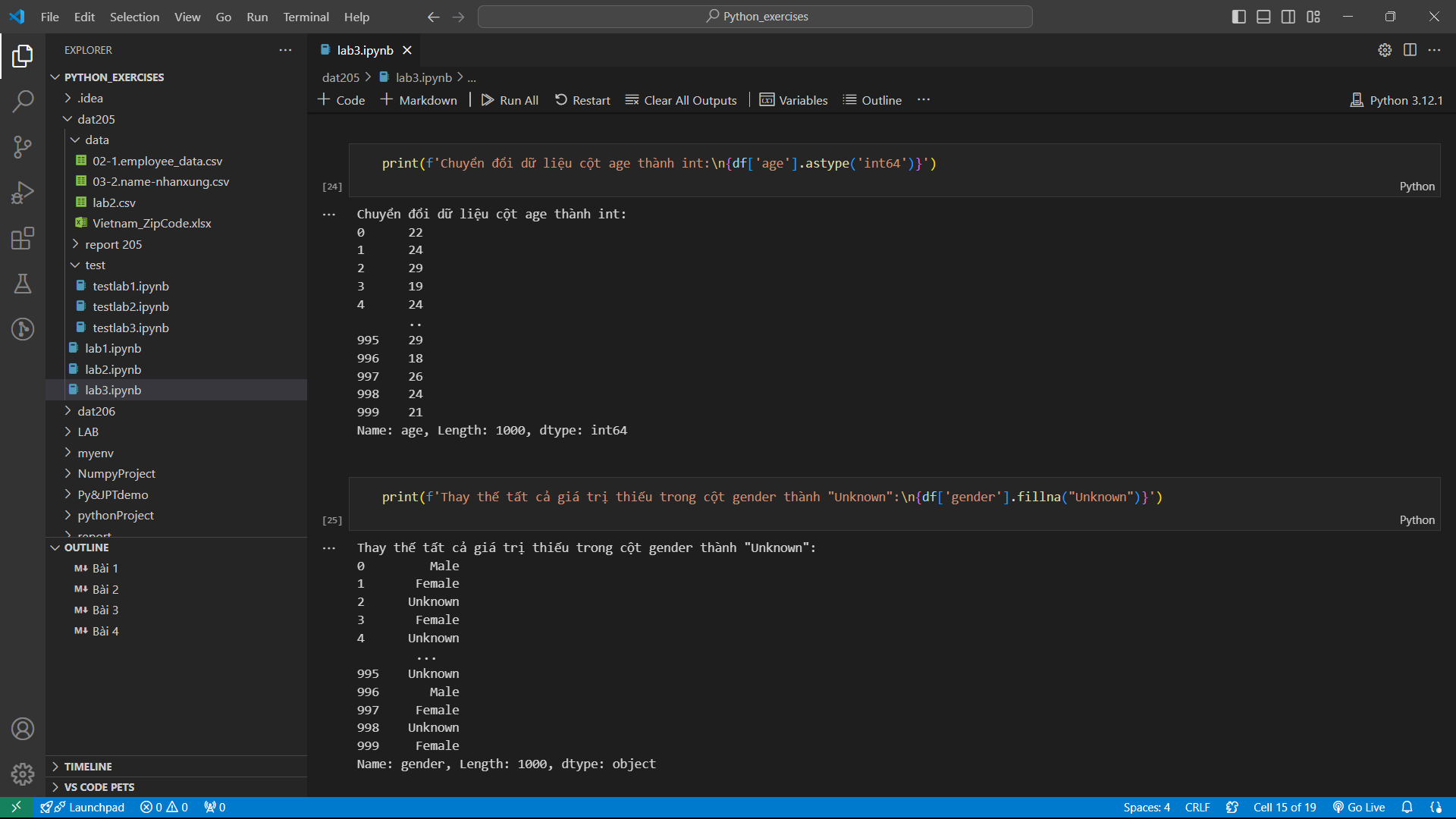
Dựa vào lab2.csv

- Chuyển đổi kiểu dữ liệu của cột 'age' thành kiểu dữ liệu số nguyên.

- Thay thế tất cả các giá trị thiếu trong cột 'gender' bằng giá trị 'Unknown'.

- Thay thế tất cả các giá trị “na” trong cột 'score' bằng giá trị trung bình của cột đó.

**Bài làm:**

 A screenshot of a computer

Description automatically generated

* **.astype():** chuyển đổi kiểu dữ liệu của cột theo kiểu đưa ra
* **.fillna():** thay thế các dữ liệu trống bằng điều kiện cho trước