## Cài đặt phần mềm

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên phần mềm, công cụ | Link hướng dẫn | Mục đích |
| 1 | Python 3.9 | https://quantrimang.com/cach-cai-dat-python-tren-windows-macos-linux-140625 | Sử dụng ngôn ngữ lập trình python và các thư viện python để thực hiện làm sạch dữ liệu |
| 2 | Java 8 | Mở link sau, chọn 1 cái theo bản Window 32 hay 64 bit  <https://www.oracle.com/java/technologies/downloads/#java8-windows> | Môi trường để chạy các ứng dụng Java |
| 3 | Cài đặt CSDL SQL Server 2019 | <https://www.thegioididong.com/game-app/huong-dan-cai-dat-sql-server-2019-cuc-don-gian-chi-1312926>  Lưu ý: cần nhớ password của user sa (password nhập trong quá trình cài đặt) | Tạo CSDL chứa các bảng dữ liệu phục vụ việc thực nghiệm |
| 4 | Cài đặt SQL Server Management Studio (SSMS) | <https://www.thegioididong.com/game-app/huong-dan-cai-dat-sql-server-2019-cuc-don-gian-chi-1312926> ở mục III. Cài đặt SQL Server Management Studio (SSMS) | Công cụ trực quan để thuận tiện kết nối và quản lý CSDL SQL Server |
| ~~5~~ | ~~Cài đặt công cụ SPSS 20~~ |  | ~~Phân tích dữ liệu dữ liệu nguồn để nhận diện các vấn đề cần xử lý làm sạch~~ |
| 6 | Cài đặt công cụ ETL Pentaho Data Integration 9.2 | Download về máy theo link sau  <https://sourceforge.net/projects/pentaho/files/Pentaho%209.0/> | Phục vụ việc phát triển các luồng ETL tích hợp dữ liệu giữa các hệ thống |
| 7 | Cài đặt Azure Data Studio | https://docs.microsoft.com/en-us/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15#download-azure-data-studio | Tool IDE để code python |

## Sau khi cài đặt xong, thực hiện các bước sau

Bước 1: Tạo Database Khảo sát sinh viên, và tạo 1 bảng file\_khaosat\_hd\_giangday để lưu dữ liệu khảo sát hoạt động giảng day

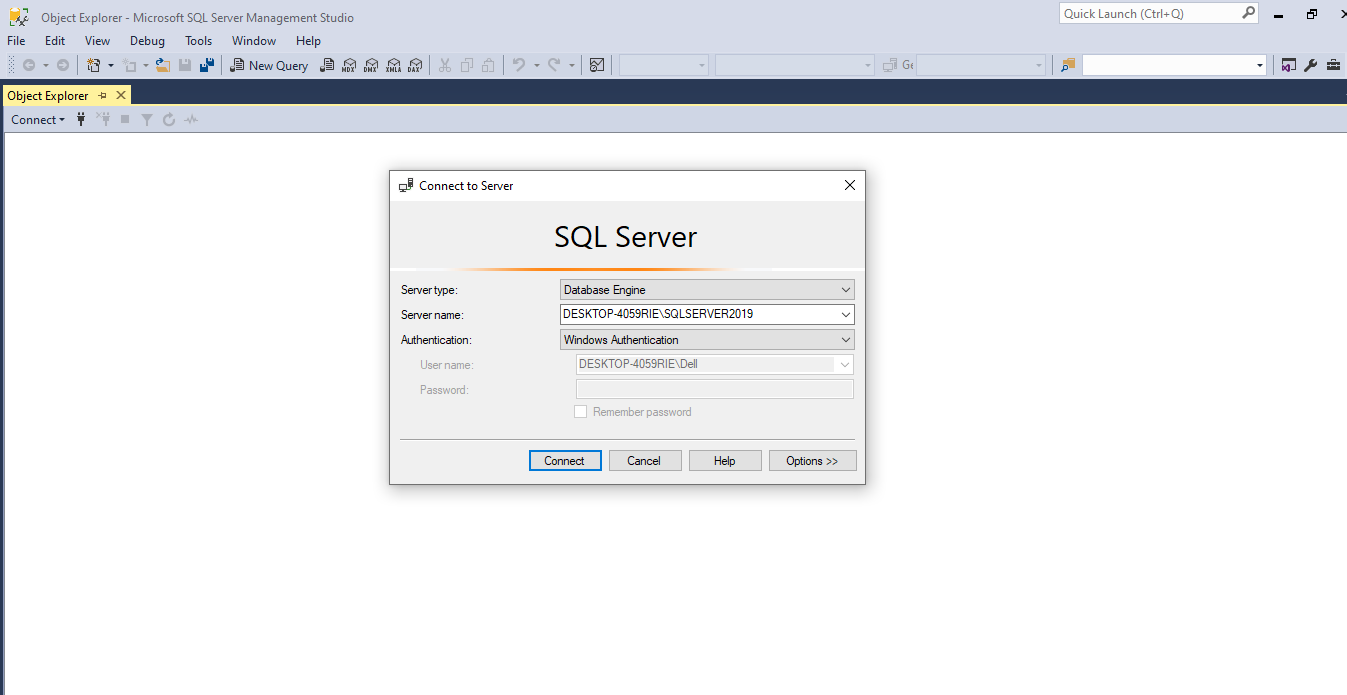
Bước 2: Chuẩn bị file khảo sát sinh viên

Bước 3: Lập trình Python để làm sạch dữ liệu

Bước 4: Kiểm tra kết quả làm sạch

**Bước 1: Tạo Database Khảo sát sinh viên, và tạo 1 bảng file\_khaosat\_hd\_giangday để lưu dữ liệu khảo sát hoạt động giảng day**

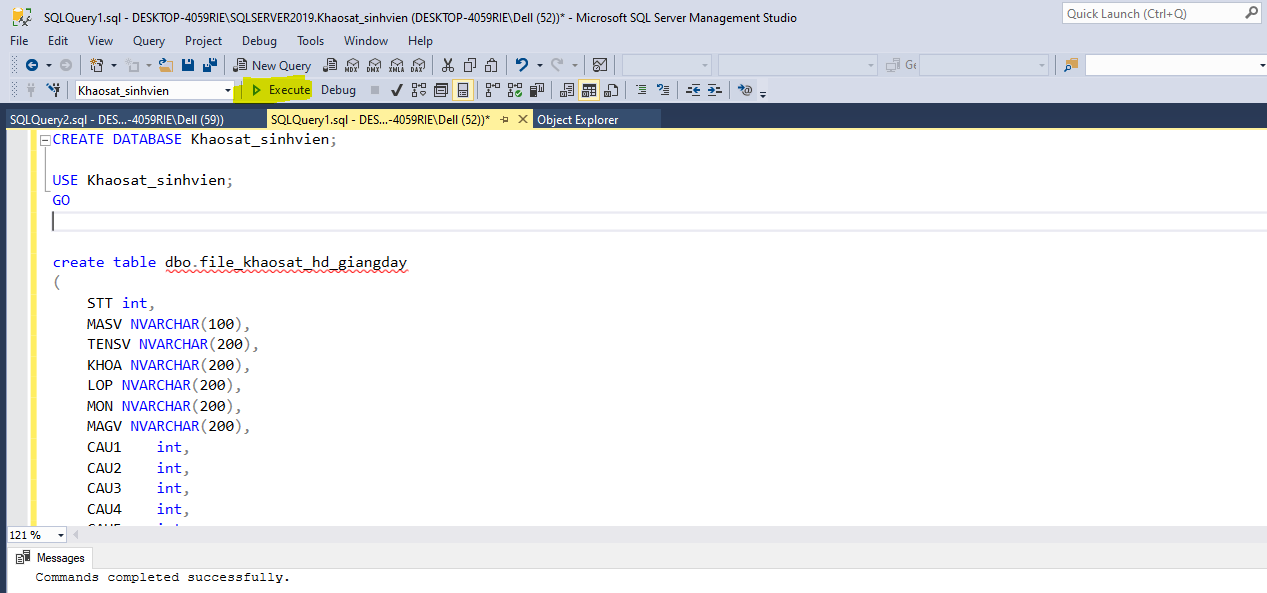
* Sử dụng công cụ SQL Server Management Studio (SSMS) để kết nối đến Database



* Ấn 2 phím Ctrl + N : để mở tab query mới, copy các câu lệnh sau

|  |
| --- |
| CREATE DATABASE Khaosat\_sinhvien;  USE Khaosat\_sinhvien;  GO  create table dbo.file\_khaosat\_hd\_giangday  (  STT int,  MASV NVARCHAR(100),  TENSV NVARCHAR(200),  KHOA NVARCHAR(200),  LOP NVARCHAR(200),  MON NVARCHAR(200),  MAGV NVARCHAR(200),  CAU1 int,  CAU2 int,  CAU3 int,  CAU4 int,  CAU5 int,  CAU6 int,  CAU7 int,  CAU8 int,  CAU9 int,  CAU10 int,  CAU11 int,  CAU12 int,  CAU13 int,  CAU14 int,  CAU15 int,  CAU16 int,  CAU17 int  );  GO |

* Click nút Execute để chạy các câu lệnh SQL trên



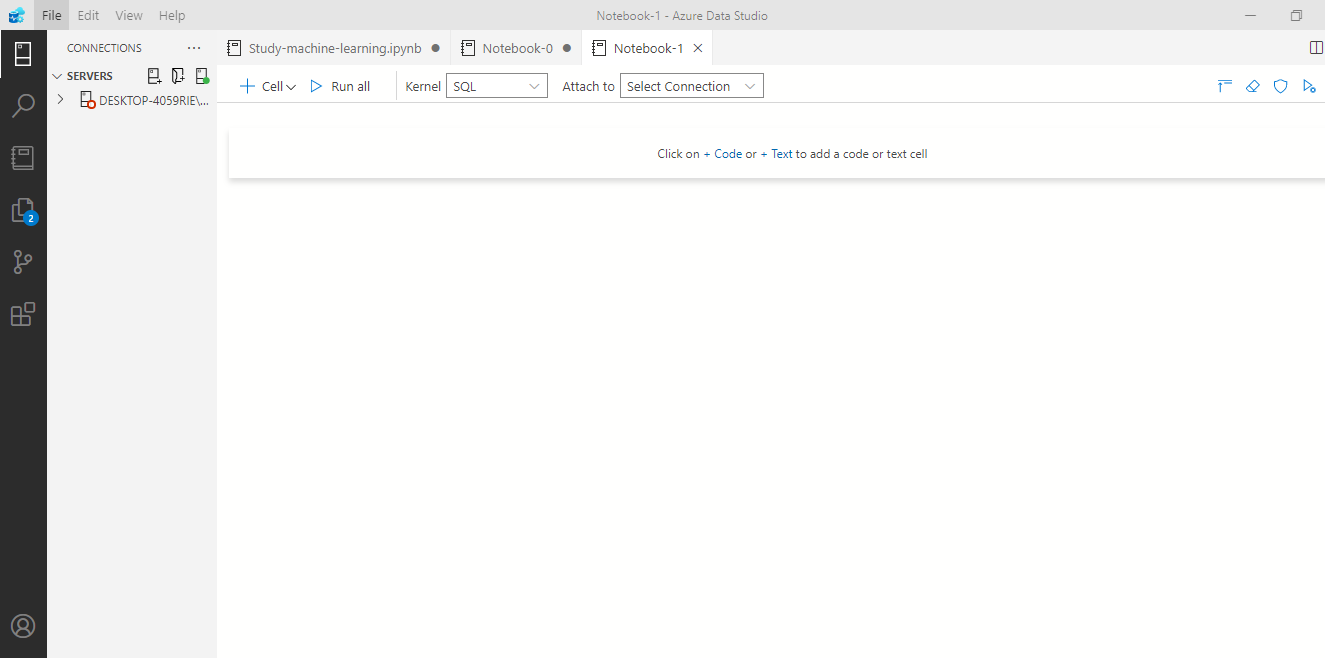
**Bước 2: Chuẩn bị file khảo sát sinh viên**

* Đổi tên các cột trong file gốc thành các cột như sau (chỉ cần copy và ghi đè lại tên cột tiêu đề)

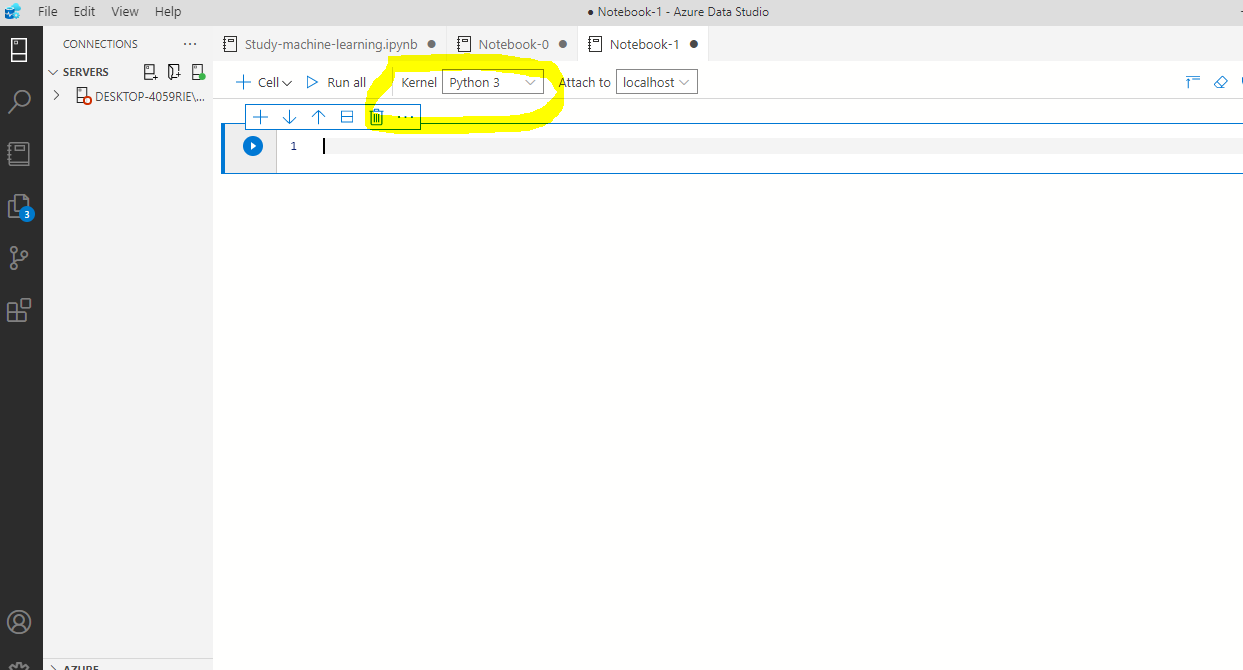
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **MASV** | **TENSV** | **KHOA** | **LOP** | **MON** | **MAGV** | **CAU1** | **CAU2** | **CAU3** | **CAU4** | **CAU5** | **CAU6** | **CAU7** | **CAU8** | **CAU9** | **CAU10** | **CAU11** | **CAU12** | **CAU13** | **CAU14** | **CAU15** | **CAU16** | **CAU17** |

**Bước 3: Lập trình Python để làm sạch dữ liệu**

* Mở tool Azure Data Studio



* Chọn File -> New Notebook. Ở trong tab Notebook mới mở, chọn Kernel là Python 3



* Copy code python
* Lưu ý 1 số mục cần chỉnh sửa:
* df = pd.read\_excel(**r'D:/Freelancer/file\_khaosat\_giangday.xlsx'**,
* sheet\_name='Sheet1')
* Sửa đường dẫn lưu file khảo sát sinh viên vào đây (lưu ý lưu ở thư mục ko có tên Tiếng việt hay có dấu space)

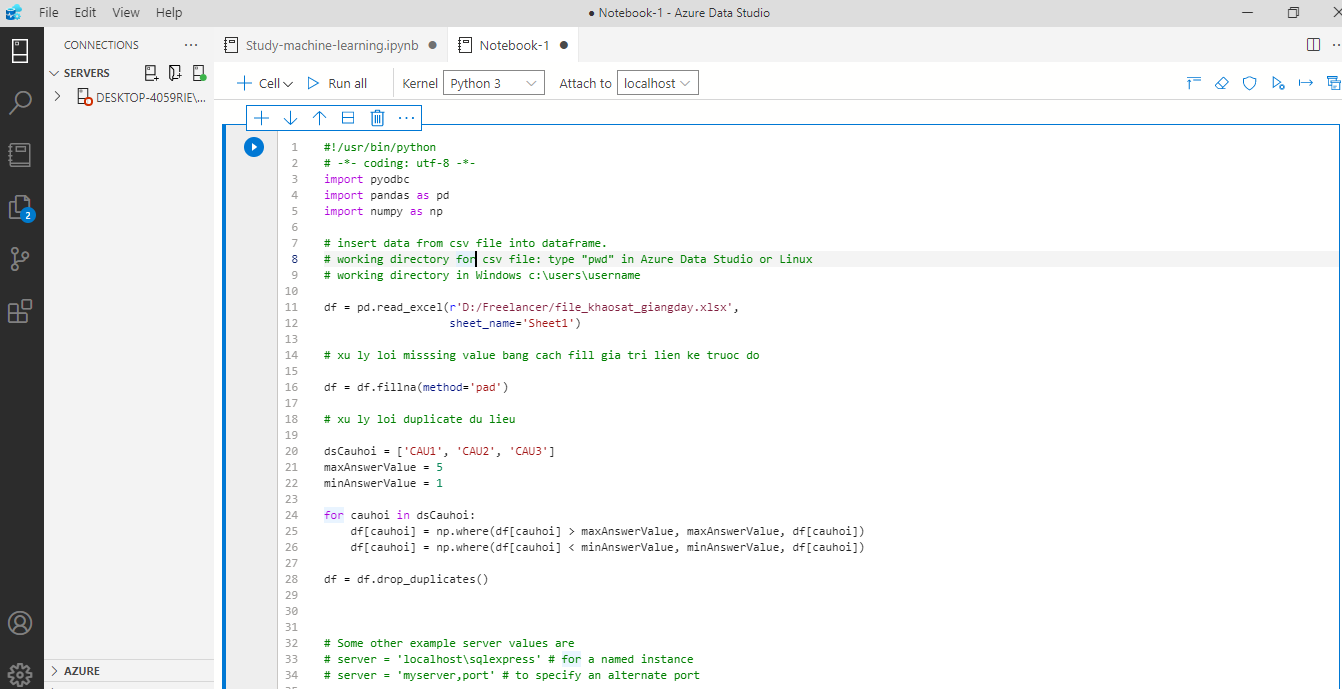
server = 'DESKTOP-4059RIE\SQLSERVER2019'

database = 'Khaosat\_sinhvien'

username = 'sa'

password = 'Dert1234'

* Sửa tên server và username/password theo máy tính cá nhân



* sau đó ấn nút **Run all** để chạy các lệnh làm sạch dữ liệu và đẩy vào bảng file\_khaosat\_hd\_giangday trên Database Khảo sát sinh viên

**Bước 4: Kiểm tra kết quả làm sạch**

* Sau khi chạy xong Python, quay lại tool SSMS và chạy lệnh SQL select để kiểm tra dữ liệu đã được làm sạch thành công

