LESSON 4 : DATA MODELING - DATA PROCESSING WITH PYTHON

**- Deadline : *Giảng viên / mentor set deadline và điền vào mục này trước khi gửi cho học viên***

**- Phương thức nộp bài : *Nêu rõ học viên sẽ gửi bài tập trực tiếp cho mentors hay gửi vào 1 folder đã được tạo sẵn. Nếu là folder tạo sẵn, vui lòng đính kèm link nộp bài vào đây !***

# Mục đích

*Bài tập này sẽ giúp các bạn củng cố kiến thức đã học, sau khi hoàn thiện các bài tập, các bạn có có thể đạt được :*

**Kỹ năng :**

* Xử lý dữ liệu với Python
* Thu thập dữ liệu từ nhiều nguồn / Gathering Data From Multi Source

**Kiến thức :**

Các định nghĩa, khái niệm sẽ được củng cố thông qua phần bài tập này, bao gồm :

* Dummy Variable
* Encoding Data
* Binning data
* Data Ingestion

# Bài tập

## Bài tập 1 :

Bối cảnh :

Công ty M có tổng hợp lại các dữ liệu có liên quan đến tuyển dụng nhân sự trong thời

gian gần đây. Do các nhân sự thuộc team HR chưa được phổ biến về Data nên các file

Lưu trữ dữ liệu còn khá thô và khó để có thể có thể làm báo cáo về tình trạng tuyển

Dụng, các nguồn CV/ Ứng viên chất lượng và tỉ lệ pass cao.

Link dữ liệu: [Link](https://docs.google.com/spreadsheets/d/1HHWoWQJA4Nhpm5bQWeIIaS8xXhc_37Cs/edit?usp=drive_link&ouid=110892656940326797212&rtpof=true&sd=true)

Bạn hãy thực hiện clean lại bộ dữ liệu trên với Python và tiến hành design một data

Model để đảm bảo dữ liệu được chính xác.

Biết rằng, các cột trong file được các bạn HR mô tả gồm các thông tin như sau:

**STT**: Số thứ tự

**Ngày tiếp nhận hồ sơ** : Ngày mà team HR nhận được CV của ứng viên

**Họ và tên**: Họ tên của ứng viên

**SĐT**: SĐT liên hệ của ứng viên

**Email** : Địa chỉ email mà ứng viên gửi về bộ phận HR

**Prework**: Trạng tháng làm bài kiểm tra Prework trước khi thực hiện phỏng vấn. Để

trống nếu ứng viên không làm bài Prework

**Đặc điểm ứng viên**: Các features chính mà HR Scan được trên CV của ứng viên

Công ty hiện tại: Công ty hiện tại mà ứng viên đang làm việc

**Vị trí hiện tại**: Vị trí hiện tại mà ứng viên đang làm việc

Vị trí có thể phù hợp: Vị trí mà HR gợi ý ứng viên sẽ phù hợp tại công ty M

**Unnamed: 10** : Giới tính

**Unnamed: 11** : Là dữ liệu do hệ thống xuất ra và không có thông tin gì về trường dữ

liệu này

**Tình trạng** : Trạng thái hồ sơ của ứng viên. Gồm các thông tin: PassCV, PassPV,

Onboarding.

**Nguồn**: Là nguồn/channel mà HR tuyển dụng ứng viên. Ví dụ: Facebook, Zalo, TopCV,

…

**False** : Là dữ liệu do hệ thống xuất ra và không có thông tin gì về trường dữ liệu

này

**Mức độ tiềm năng** : Đánh giá mức độ tiềm năng của ứng viên.

Bài tập 2: