**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC THĂNG LONG**

**---o0o---**

**Logo, company name

Description automatically generated**

**Bộ môn: Lập trình python**

**Báo cáo: Bài tập lớn**

**Giáo viên hướng dẫn: Sinh viên thực hiện:**

**Nguyễn Đức Thắng**  **A40474 Nguyễn Văn Tuấn**

**A41090 Nguyễn Chí Minh**

***HK: Kì 1 nhóm 2 năm 2022-2023***

***Hà Nội, 11/2022***

**Mục Lục**

[Phần 1:Giới thiệu về tập dữ liệu 1](#_Toc119516333)

[Phần 2. Kiểm tra và làm sạch dữ liệu 2](#_Toc119516336)

[Phần 3: Phân tích các đại lượng 5](#_Toc119516339)

[3.1 . Phân tích các dữ liệu định tính 6](#_Toc119516340)

[3.2 .Phân tích các dữ liệu định lượng 9](#_Toc119516341)

[3.3 .Phân tích đa biến 14](#_Toc119516342)

**Phần 4: Tổng kết đánh giá sự hợp tác đóng góp của các thành viên**

# Phần 1: Giới thiệu về tập dữ liệu

Tập dữ liệu về thông tin cơ bản về hơn 3 nghìn khóa học trên Udemy

Link tập dữ liệu: <https://drive.google.com/file/d/1zglnQkX756nfySz9RT8V3p1R6-69RHmS/view>

Tập dữ liệu gồm 11 cột và 3682 dòng nói về:

**course\_id: Id khóa học**

**course\_title: Tên khóa học**

**is\_paid: Khóa học miễn phí hoặc trả phí**

**price: Giá đối với khóa học trả phí**

**num\_subscribers:Số lượt người đăng ký**

**num\_reviews: Số lượt người đánh giá**

**num\_lectures: Số bài học**

**level: Level của khóa học**

**content\_duration: Thời lượng của khóa học**

**published\_timestamp: Thời gian phát hành**

**subject: Nhóm các khóa học**

# Phần 2. Làm sạch dữ liệu

Tập dữ liệu có một cột dữ liệu nếu để nguyên như vậy thì ta rất khó khai thác và phân tích nên ta cần phải làm sạch

2.1 Content\_duration

Graphical user interface, application

Description automatically generated

Đang không cùng 1 kiểu dữ liệu, cần đưa cùng về 1 kiểu dữ liệu là giờ(hour)

2.2 Published\_timestamp

Cột ngày phát hành

Graphical user interface, application

Description automatically generated

Lấy ra tháng và năm phát hành cho vào 2 cột month\_added và year\_added

Graphical user interface, application, chat or text message

Description automatically generated

2.3 Num\_lectures

Có một số khóa học không có bài học cần loại bỏ

Graphical user interface

Description automatically generated with medium confidence

2.4 Course\_id

Xóa vì không phân tích đến

# Phần 3: Phân tích dữ liệu

Chúng ta phân tích dữ liệu chủ yếu dựa trên 2 loại dữ liệu là dữ liệu định lượng và dữ liệu định tính.

Dữ liệu định tính:

is\_paid

level

subject

Dữ liệu định lượng

Price

Num\_subcribers

Num\_reviews

Num\_lecture

Content\_duaration

Do dữ liệu định lượng có khá nhiều cột nên chúng tôi chỉ chọn ra một số cột có giá trị khai thác cao để phân tích.

## 3.1. Dữ liệu định tính

Ở đây chúng tôi sẽ chủ yếu vẽ ra một số biểu đồ miêu tả dữ liệu định tính của 3 cột dữ liệu

is\_paid

level

subject

Dưới đây là một số hình ảnh tiêu biểu:

-Is\_paid

Chart, pie chart

Description automatically generated

-Level

Chart, bar chart

Description automatically generated

-Subject

Chart

Description automatically generated

## 3.2 . Phân tích dữ liệu định lượng

Chúng tôi chọn ra 2 dữ liệu để khai thác là: Price, Num\_reviews

Và dưới đây là hình ảnh về một số loại biểu đồ mà chúng tôi dùng để phân tích 2 loại dữ liệu này

Price:

Đối với Price chúng ta cần loại bỏ một số dữ liệu có price = ‘Free’ để so sánh giá cả của các khóa học với nhau.

Chart, box and whisker chart

Description automatically generated

-num\_reviews

Chart

Description automatically generated

Nhìn vào biểu đồ hộp của num\_reviews do có quá nhiều giá trị ngoại biên lớn hơn so với các dữ liệu trong cột nên chúng tôi rất khó để mà phân tích và đưa ra đánh giá về nó. Chúng tôi đã tinh chỉnh dữ liệu để có thể phân tích một cách khách quan nhất cột dữ liệu này.

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

## 3.3 . Phân tích đa biến

Ở phần này chúng tôi sẽ các biểu đồ thể hiện mối liên hệ giữa các cột dữ liệu với nhau.

Dưới đây là một số hình ảnh demo:

Chart, box and whisker chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

# Phần 4: Tổng kết

Để có được kết quả bài tập lớn như trên đã có sự đóng góp rất lớn từ các bạn cụ thể:

Nguyễn Văn Tuấn: Tìm tập dữ liệu, phân tích dữ liệu định tính, phân tích dữ liệu định lượng.

Nguyễn Chí Minh: Tìm tài liệu, phân tích đa biến, đóng góp ý kiến cùng thảo luận.