



## **GIỚI THIỆU DỰ ÁN 01**

---

### **GIẢNG VIÊN: NGUYỄN HUY HOÀNG**

█ **Dự án 01** là môn học làm dự án *xử lý dữ liệu hoàn chỉnh*, vận dụng kiến thức, kỹ năng từ các môn học từ kỳ 1 đến kỳ 4 đặc biệt là môn dự án mẫu.

█ Mục tiêu cung cấp kiến thức, chuẩn bị cho giai đoạn đi vào chuyên ngành.

█ Còn là môn học làm điều kiện tiên quyết cho thực tập và dự án tốt nghiệp.



 **Sản phẩm mục tiêu** là mẫu báo cáo để sinh viên hướng tới

- ❖ Không như môn học Dự án mẫu, *sinh viên* bắt buộc chủ động thực hiện tất cả các bước.
- ❖ Dựa trên các *bộ dữ liệu (dataset)* được cung cấp sẵn, *sinh viên* tự chọn với quy mô vừa sức thực hiện trong 1 block của học kỳ.
- ❖ *Sinh viên* tiến hành *xử lý dữ liệu*, sau đó tạo báo cáo report *diễn giải* và *truyền đạt* thông tin có trong bộ dữ liệu này, giúp người nghe hiểu được vấn đề nhanh và chính xác.
- ❖ *Giảng viên* sẽ hướng dẫn để sinh viên hoàn thành một báo cáo hoàn chỉnh.



- Kết thúc môn học này, bạn có khả năng
  - Vận dụng kiến thức *phân tích* và *trực quan* hóa dữ liệu (các môn học trước đó) để tự xây dựng một báo cáo hoàn chỉnh về xử lý dữ liệu.
  - *Tổ chức* được CSDL và *mã nguồn* dự án một cách có khoa học, gần với thực tế.
  - Thực hiện được dự án theo đúng *quy trình* dự án thực tế.
  - Đọc và viết được *báo cáo* về tài liệu dự án.
  - Làm quen với việc *thực hiện* một dự án sau này.





NGÀY #1

---



# TỔ CHỨC NHÓM

---

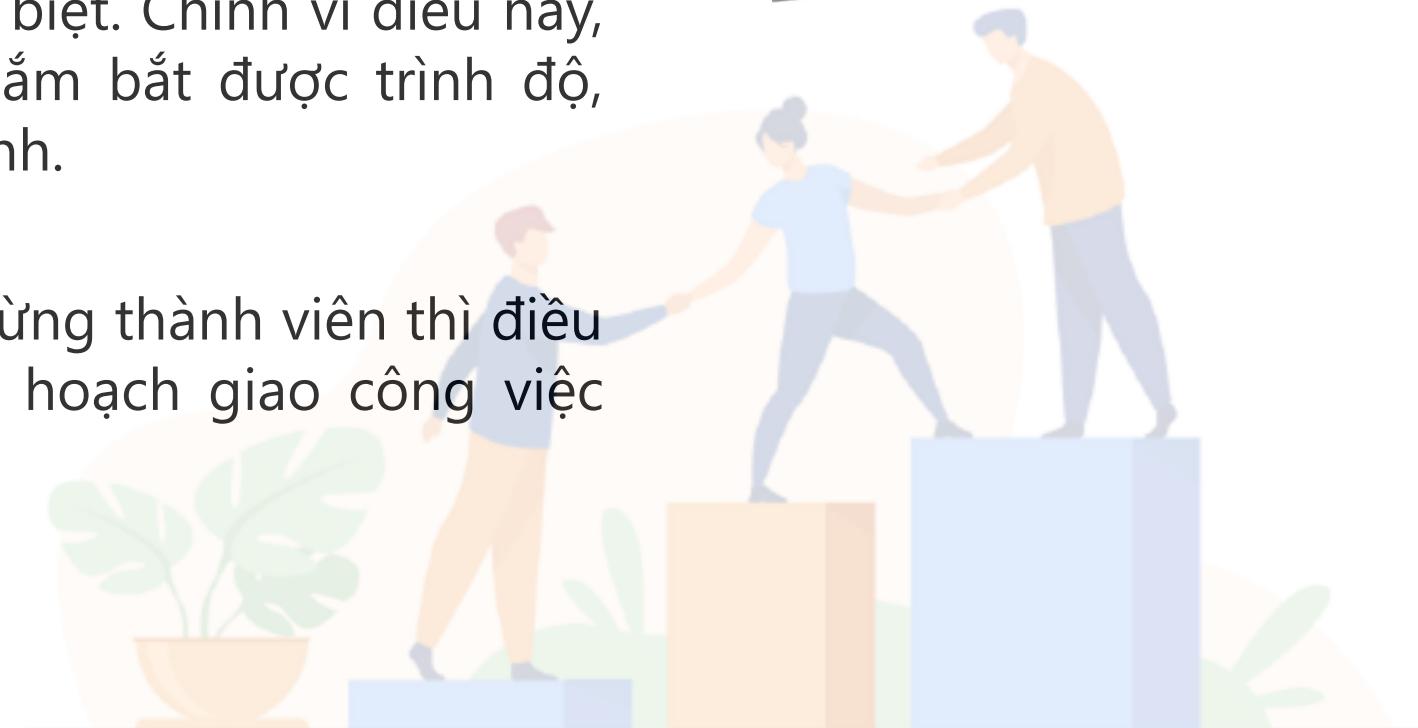
 **Phân nhóm**

- ❖ Mỗi nhóm không quá 3 sinh viên.
- ❖ Công việc được chia cho mỗi thành viên trong nhóm để thực hiện.
- ❖ Mỗi thành viên có nhiệm vụ chia sẻ kiến thức đã được phân công cho các thành viên khác trong nhóm tránh trường hợp giảng viên hỏi không biết trả lời.

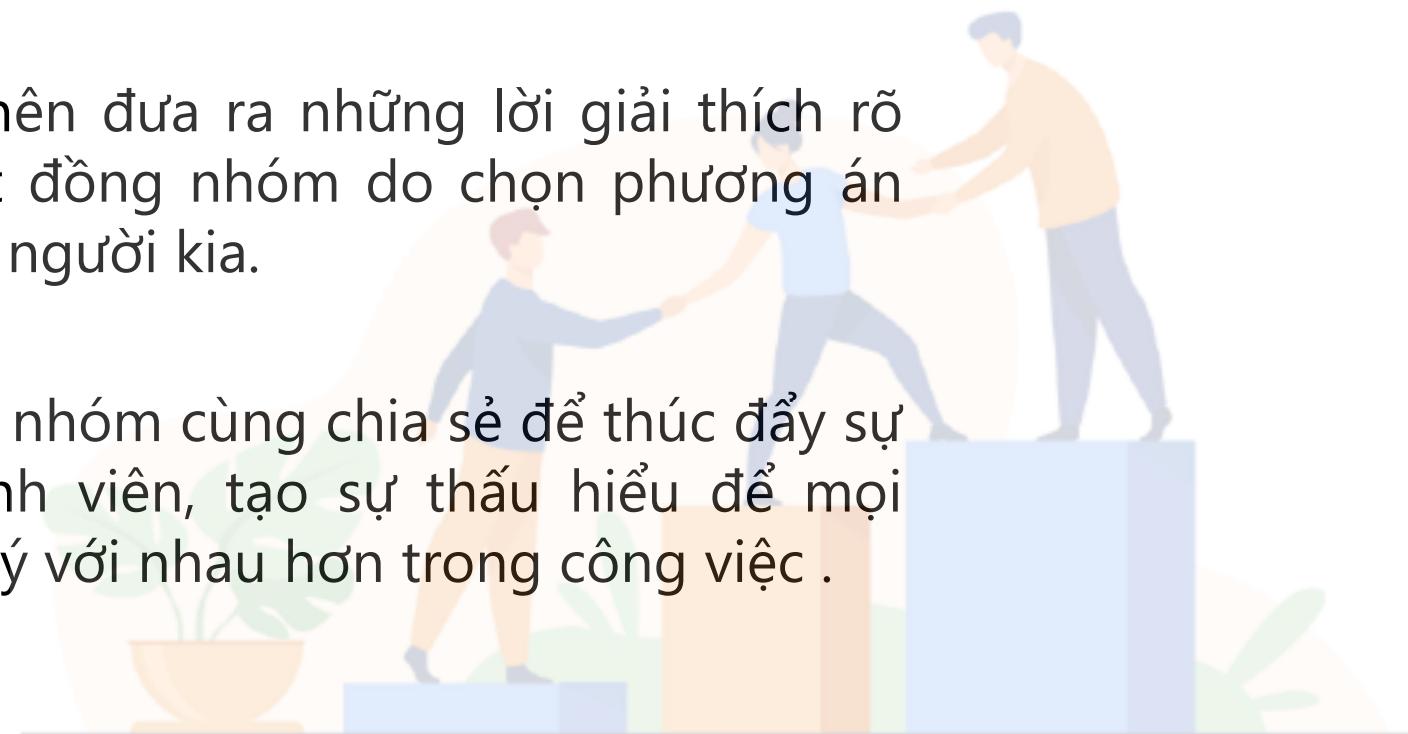


# LÀM TEAMWORK HIỆU QUẢ?

- ❑ Xây dựng mục tiêu, nguyên tắc chung
  - ❖ Cơ sở đầu tiên để có được đội nhóm gắn kết chính là có mục tiêu chung. Bạn cần tìm những người có chung chí hướng, chung suy nghĩ, từ đó tập hợp thành team.
- ❑ Hiểu được điểm mạnh, điểm yếu của từng thành viên
  - ❖ Con người không có ai hoàn hảo, mỗi người đều có những thế mạnh và hạn chế riêng biệt. Chính vì điều này, cần phải chủ động tìm hiểu và nắm bắt được trình độ, khả năng thành viên của nhóm mình.
- ❑ Phân công việc phù hợp với năng lực
  - ❖ Khi đã nắm rõ được ưu điểm của từng thành viên thì điều cần làm tiếp theo chính là có kế hoạch giao công việc phù hợp.



- ❑ Luôn có niềm tin vào đồng đội
  - ❖ Khi làm việc teamwork, bạn cần phải có sự tin tưởng với những cộng sự của mình.
- ❑ Luôn biết lắng nghe và thấu hiểu
  - ❖ Cần chủ động lắng nghe đóng góp từ các thành viên trong nhóm. tôn trọng ý kiến của nhau để có thể phối hợp và làm việc ăn ý hơn.
- ❑ Đánh giá công bằng
  - ❖ Trước khi đưa ra các quyết định nên đưa ra những lời giải thích rõ ràng, tránh những mâu thuẫn bất đồng nhóm do chọn phương án của người này mà không chọn của người kia.
- ❑ Thẳng thắn chia sẻ
  - ❖ Khuyến khích các thành viên trong nhóm cùng chia sẻ để thúc đẩy sự phát triển năng lực của mỗi thành viên, tạo sự thấu hiểu để mọi người có thể phối hợp làm việc ăn ý với nhau hơn trong công việc .





# BẢO VỆ DỰ ÁN

 **Bảo vệ**

- ❖ Sử dụng slide mẫu được cung cấp.
- ❖ Mỗi thành viên phải có hình trên slide và được phân công việc cụ thể.
- ❖ Mỗi thành viên trình bày phần việc của mình trước hội đồng giám khảo.
- ❖ Phần tài liệu dự án (*project document*) cho điểm chung.
- ❖ Phần bảo vệ cho điểm riêng.

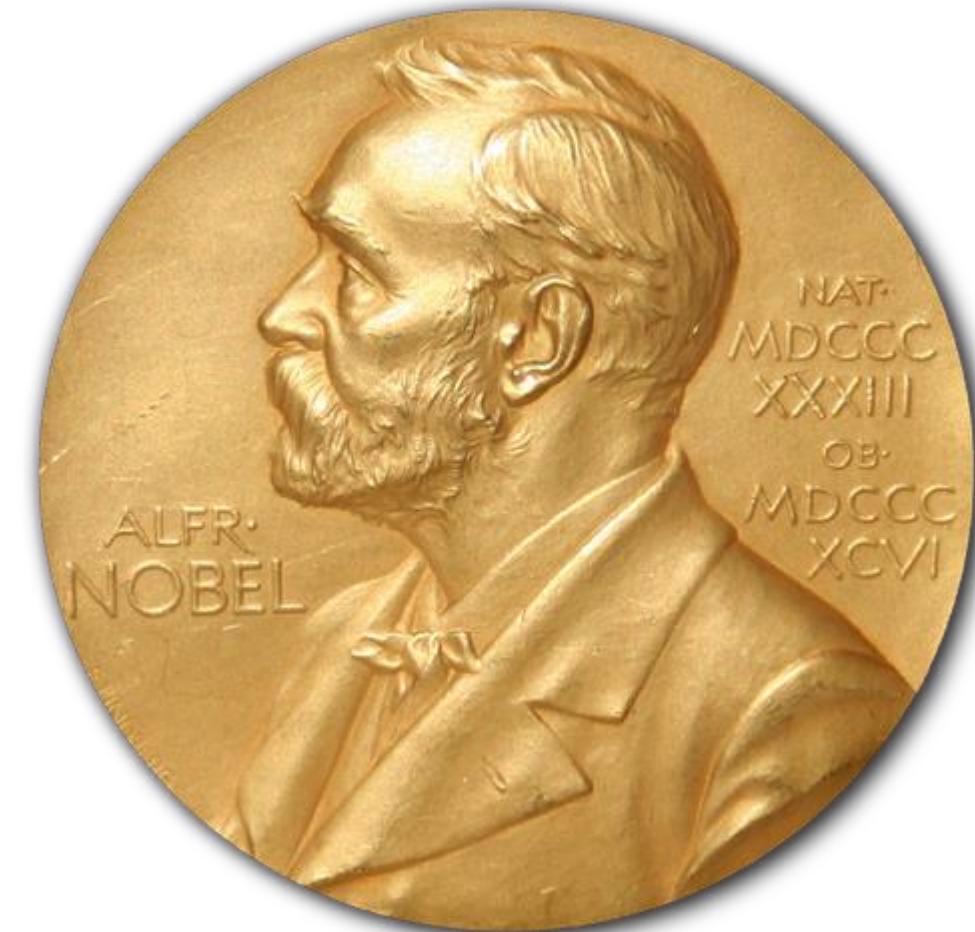




# GIỚI THIỆU BỘ DỮ LIỆU #1

## □ Giới thiệu

❖ Giải Nobel là một trong những giải thưởng quốc tế danh giá nhất được trao hàng năm để ghi nhận những thành tựu xuất sắc trong các lĩnh vực khác nhau. Được thành lập vào năm 1895 theo ý muốn của Alfred Nobel, một nhà phát minh, kỹ sư và nhà công nghiệp người Thụy Điển, giải thưởng Nobel tôn vinh các cá nhân và tổ chức có đóng góp đáng kể cho nhân loại trong sáu hạng mục: Vật lý, Hóa học, Sinh lý học hoặc Y học, Văn học, Hòa bình và Khoa học kinh tế.



## ❑ Dataset

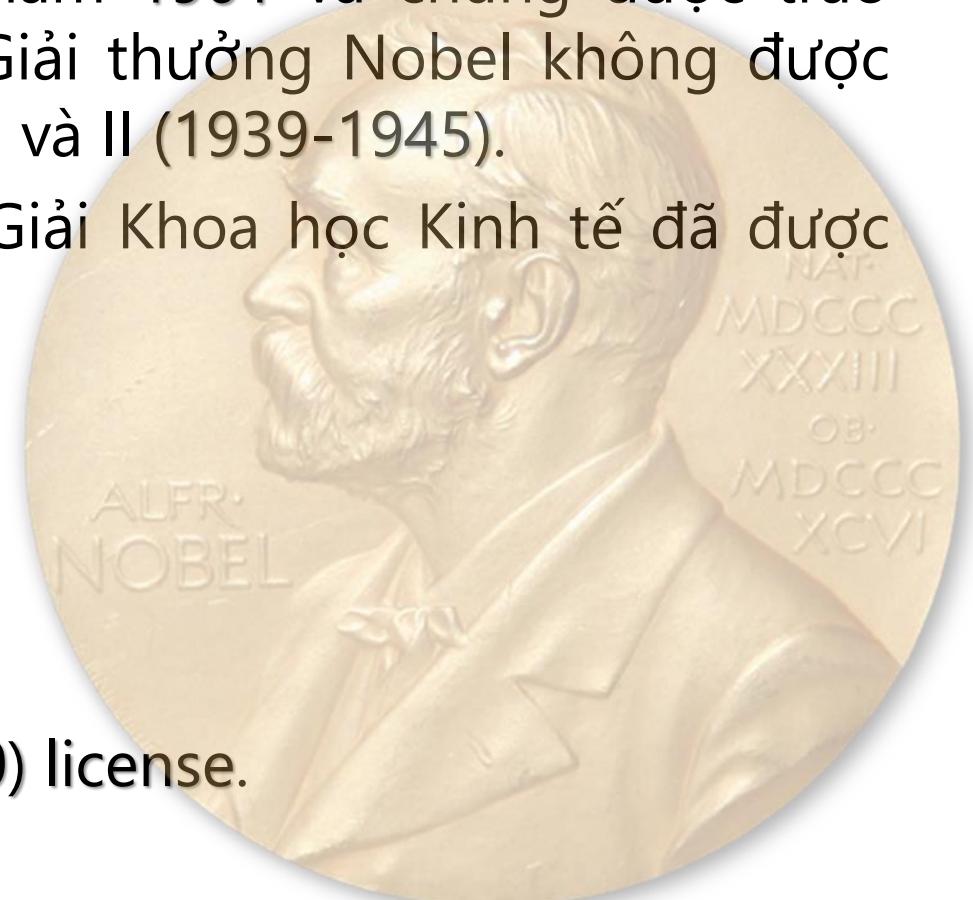
- ❖ Giải thưởng Nobel đầu tiên được trao vào năm 1901 và chúng được trao hàng năm kể từ đó. Đã có những năm mà Giải thưởng Nobel không được trao - chủ yếu là trong Thế chiến I (1914-1918) và II (1939-1945).
- ❖ Từ năm 1901 đến năm 2022, Giải Nobel và Giải Khoa học Kinh tế đã được trao cho 989 Cá nhân/Tổ chức.

## ❑ Resource

- ❖ Thông tin: [nobelprize.org](https://nobelprize.org).
- ❖ Dữ liệu: [API version 2](https://api.nobelprize.org/v1/prizes.json)

## ❑ Licences

- ❖ Free được cấp từ Creative Commons Zero (CC0) license.



# NOBEL PRIZE WINNERS (1901-2022)

STT	Tên trường dữ liệu	Nội dung
1	Year	Năm nhận giải Nobel.
2	Laureate_Id	Mã nhận giải
3	Firstname	Tên
4	Lastname	Họ
5	Category	Lĩnh vực nhận giải
6	Gender	Giới tính
7	Prize_Share	Làm một mình hay nhóm
8	Motivation	Công trình nghiên cứu
9	Birth_Date	Ngày sinh
10	Birth_Country	Quốc gia nơi nhà khoa học được sinh ra (Quê hương)
11	Birth_City	Thành phố nơi nhà khoa học được sinh ra (Nguyên quán)
12	Birth_Country_Code	Mã quốc gia nơi nhà khoa học được sinh ra (mã code)
13	Death_Date	Ngày mất
14	Death_Country	Quốc gia nơi nhà khoa học mất (nơi an nghỉ)
15	Death_City	Thành phố nơi nhà khoa học mất (nơi an nghỉ)
16	Death_Country_Code	Mã quốc gia nơi nhà khoa học mất (mã code)
17	Organization_Name	Tên tổ chức (bảo trợ) nơi nhà khoa học có hoạt động nghiên cứu
18	Organization_City	Tên thành phố nơi nhà khoa học có hoạt động nghiên cứu
19	Organization_Country	Tên quốc gia nơi nhà khoa học có hoạt động nghiên cứu

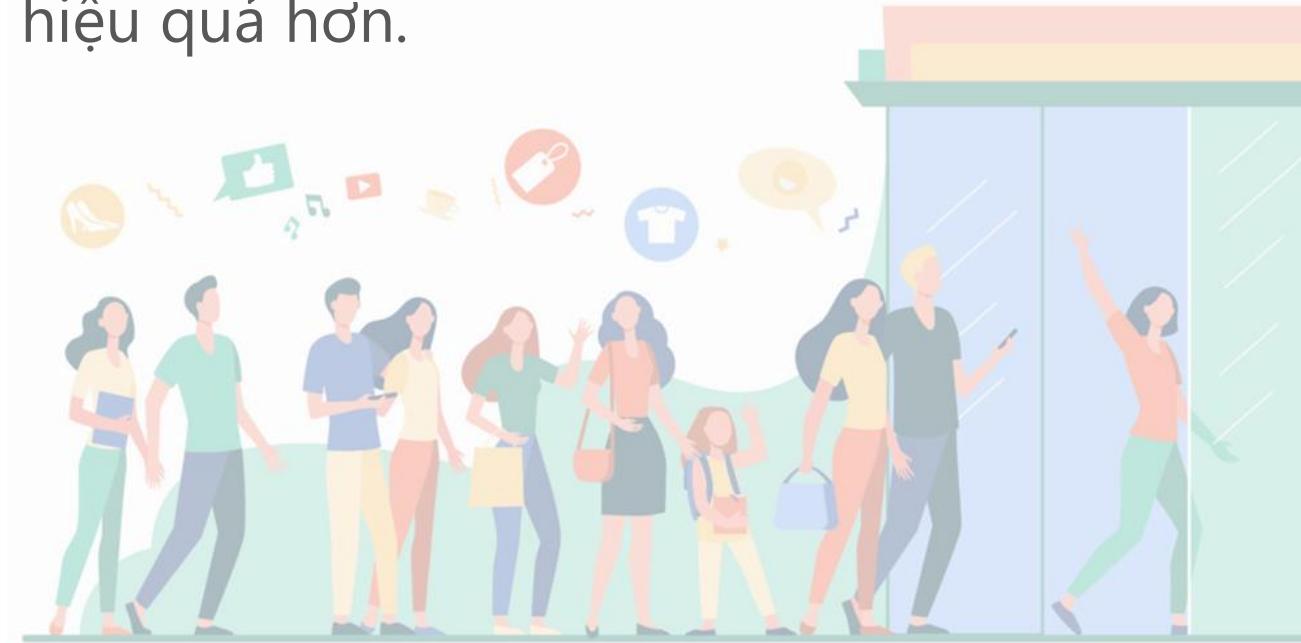




## GIỚI THIỆU BỘ DỮ LIỆU #2

## □ Giới thiệu

- ❖ Phân tích tính cách khách hàng là phân tích chi tiết về khách hàng lý tưởng của công ty. Nó giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình và giúp họ dễ dàng sửa đổi sản phẩm theo nhu cầu, hành vi và mối quan tâm cụ thể của các loại khách hàng khác nhau.
- ❖ Phân tích tính cách khách hàng giúp doanh nghiệp sửa đổi sản phẩm của mình dựa trên khách hàng mục tiêu từ các loại phân khúc khách hàng khác nhau. Giúp bán hàng và tiếp thị hiệu quả hơn.



## ❑ Dataset

- ❖ Thông tin cá nhân cơ bản (không phải sensitive data) và lịch sử mua sắm của một số khách hàng, được khảo sát tại một siêu thị ở Mỹ.
- ❖ Dữ liệu được khảo sát từ năm 2020 đến năm 2021.

## ❑ Resource

- ❖ Thông tin: [Dr. Omar Romero-Hernandez](#).
- ❖ Dữ liệu: Dr. Omar Romero-Hernandez [cung cấp](#)

## ❑ Licences

- ❖ Free được cấp từ Creative Commons Zero (CC0) license.



# CUSTOMER PERSONALITY ANALYSIS

STT	Tên trường dữ liệu	Nội dung
1	ID	Mã định danh KH
2	Year_Birth	Năm sinh KH
3	Education	Trình độ học vấn KH
4	Marital_Status	Tình trạng hôn nhân
5	Income	Thu nhập hộ gia đình hằng năm của KH
6	Kidhome	Số con trong hộ gia đình của KH
7	Teenhome	Số lượng thanh thiếu niên trong hộ gia đình KH
8	Dt_Customer	Ngày đăng ký thành viên của KH với siêu thị
9	Recency	Lần gần nhất KH mua hàng là bao nhiêu ngày
10	Complain	Khiếu nại: 1 nếu khách hàng phàn nàn trong 2 năm qua, 0 là không than phiền.
11	MntWines	Số tiền chi cho rượu trong 2 năm qua
12	MntFruits	Số tiền chi cho trái cây trong 2 năm qua
13	MntMeatProducts	Số tiền chi cho thịt trong 2 năm qua
14	MntFishProducts	Số tiền chi cho cá trong 2 năm qua
15	MntSweetProducts	Số tiền chi cho đồ ngọt bánh kẹo trong 2 năm qua

# CUSTOMER PERSONALITY ANALYSIS

STT	Tên trường dữ liệu	Nội dung
16	MntGoldProds	Số tiền chi cho mua vàng trang sức trong 2 năm qua
17	NumDealsPurchases	Số lần mua hàng được giảm giá
18	NumWebPurchases	Số lần mua hàng được thực hiện thông qua trang web của công ty
19	NumCatalogPurchases	Số lần mua hàng được thực hiện bằng danh mục
20	NumStorePurchases	Số lần mua hàng được thực hiện trực tiếp tại các cửa hàng
21	NumWebVisitsMonth	Số lượt truy cập vào trang web của công ty trong tháng trước
22	AcceptedCmp1	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch đầu tiên, 0 nếu không
23	AcceptedCmp2	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ 2, 0 nếu không
24	AcceptedCmp3	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ 3, 0 nếu không
25	AcceptedCmp4	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ 4, 0 nếu không
26	AcceptedCmp5	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ 5, 0 nếu không
27	Response	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch trước, 0 nếu không
28	Z_CostContact	Không dùng
29	Z_Revenue	Không dùng



## GIỚI THIỆU BỘ DỮ LIỆU #3

## □ Giới thiệu

❖ Lốc xoáy thường xuyên xảy ra ở Hoa Kỳ, dẫn đến sự tàn phá lớn và thường gây thương tích và tử vong. Chúng xảy ra thường xuyên hơn ở Hoa Kỳ và Canada so với các quốc gia khác có khu vực dễ bị lốc xoáy nhất ở Hoa Kỳ là các bang miền trung và đông nam dọc theo một hành lang đôi khi được gọi là "Hẻm lốc xoáy".



## The Enhance Fujita Scale

- ❖ Sức tàn phá của một cơn lốc xoáy phần lớn bắt nguồn từ tốc độ gió bên trong nó. Vì lý do này, các nhà khí tượng đánh giá lốc xoáy bằng cách sử dụng thang đo dựa trên tốc độ gió. Ở Hoa Kỳ, lốc xoáy ban đầu được đánh giá theo Thang đo Fujita và kể từ tháng 2 năm 2007 theo Thang đo Fujita Nâng cao. Hai thang đo bao gồm các phạm vi tốc độ hơi khác nhau, nhưng đối với các mục đích thực tế thì giống nhau. Thang đo Fujita nâng cao được hiển thị bên dưới.
- ❖ mph (Mile per hour): dặm/giờ
- ❖ kph (Kilometers per hour): km/giờ
- ❖ 1 MPH = 1.61 KPH

Rating	Wind Speed	Damage
EF0	65–85 mph	Light damage
EF1	86–110 mph	Moderate damage
EF2	111–135 mph	Considerable damage
EF3	136–165 mph	Severe damage
EF4	166–200 mph	Devastating damage
EF5	>200 mph	Incredible damage

## ❑ Dataset

- ❖ Thông tin các cơn lốc xoáy xảy ra ở các tiểu bang ở Mỹ.
- ❖ Dữ liệu được ghi nhận từ năm 1950 đến năm 2022.
- ❖ Last Update: 25 April 2023

## ❑ Resource

- ❖ Thông tin: [NOAA's National Weather Service.](#)
- ❖ Dữ liệu: NOAA's Storm Prediction Center cung cấp

## ❑ Licences

- ❖ Free được cấp từ Creative Commons Zero (CC0) license.



STT	Tên trường dữ liệu	Nội dung
1	yr	Năm có 4 chữ số
2	mn	Tháng (1-12)
3	dy	Ngày trong tháng
4	date	Ngày
5	time	Giờ
6	st	Mã tiểu bang nơi bắt nguồn cơn lốc xoáy; viết tắt 2 chữ số
7	stf	Mã tiểu bang theo số thứ tự
8	mag	xếp hạng mag là F trước 01/2007; là EF sau 01/2007 (-9 nếu xếp hạng không xác định) dựa vào The Enhanced Fujita Scale
9	inj	Số người bị thương
10	fat	Số người chết
11	loss	Số người mất tích
12	slat	Vĩ độ bắt đầu, tính bằng độ thập phân
13	slon	Kinh độ bắt đầu theo độ thập phân
14	elat	Vĩ độ kết thúc, theo độ thập phân (giá trị 0 nếu thiếu)
15	elon	Kinh độ kết thúc theo độ thập phân (giá trị 0 nếu thiếu)

STT	Tên trường dữ liệu	Nội dung
16	len	Chiều dài đường đi của lốc xoáy tính bằng dặm
17	wid	Chiều rộng của lốc xoáy tính bằng yard
18	ns	Không dùng
19	sn	Không dùng
20	sg	Không dùng
21	f1	Không dùng
22	f2	Không dùng
23	f3	Không dùng
24	f4	Không dùng
25	fc	Không dùng



NGÀY #2

## Nội dung

- ❖ Dựa vào bộ dữ liệu và đề tài nhóm đã đăng kí.
- ❖ Xây dựng câu chuyện dữ liệu bằng cách trả lời các câu hỏi. (Data Storytelling)
- ❖ Cập nhật project document.



NGÀY #3

## Nội dung

- ❖ Nhóm trình bày ngắn gọn Câu Chuyện Dữ Liệu của đề tài. (bằng PowerPoint)
- ❖ Giảng viên góp ý.
- ❖ Sinh viên hoàn chỉnh document.



NGÀY #4

## ❑ Nội dung

- ❖ Hướng dẫn cài đặt công cụ thực hiện dự án.
- ❖ Chuẩn bị dữ liệu
- ❖ Làm sạch dữ liệu (Data Cleaning)



NGÀY #5

---

## ❑ Nội dung

- ❖ Chuyển đổi dữ liệu (Data Transformation)
- ❖ Thêm phần chuyển đổi dữ liệu vào project document.



NGÀY #6

## ❑ Nội dung

- ❖ Nhóm trình bày ngắn gọn phần chuyển đổi dữ liệu. (bằng PowerPoint)
- ❖ Nộp giai đoạn 1
- ❖ File backup database Staging qua các giai đoạn xử lý.
- ❖ File Powerpoint trình bày tới thời điểm này.
- ❖ File Project Document cập nhật tới thời điểm này.



NGÀY #7

---

## ❑ Nội dung

- ❖ Sử dụng Power Query để chuẩn hóa dữ liệu.
- ❖ Dùng Power BI Mô hình hóa dữ liệu theo hình SAO (Data Modeling)



NGÀY #8

---

## Nội dung

- ❖ Nhóm trình bày ngắn gọn giải thích mô hình hóa dữ liệu (bằng PowerPoint)
- ❖ Thêm phần mô hình hóa dữ liệu vào project document.



NGÀY #9

## ❑ Nội dung

- ❖ Xử lý dữ liệu với DAX 1
- ❖ Measures



NGÀY #10

## ❑ Nội dung

- ❖ Xử lý dữ liệu với DAX 2
- ❖ Calculated columns



NGÀY #11

## Nội dung

- ❖ Trực quan hóa dữ liệu thông qua các biểu đồ (Data Visualization)



NGÀY #12

## ❑ Nội dung

- ❖ Trực quan hóa dữ liệu thông qua các biểu đồ
- ❖ Cập nhật PowerPoint báo cáo và project document



NGÀY #13

---

## Nội dung

- ❖ Nhóm trình bày ngắn gọn giải thích các xử lý trên dữ liệu và biểu đồ (bằng PowerPoint)
- ❖ Nộp giai đoạn 2
- ❖ File backup database Spend qua các giai đoạn xử lý.
- ❖ File Powerpoint trình bày tới thời điểm này.
- ❖ File Project Document cập nhật tới thời điểm này.



NGÀY #14

## ❑ Nội dung

- ❖ Xuất bản báo cáo (Report Visualization)
- ❖ Dashboard và chia sẻ



NGÀY #15

---

Nội dung

- ❖ Viết tài liệu dự án



NGÀY #16

## ☐ Nội dung

- ❖ Hoàn thiện tài liệu dự án
- ❖ Chuẩn bị hoàn chỉnh nộp lên LMS trước buổi thi 3 ngày
- ❖ Nộp giai đoạn 3 (file project document)
- ❖ File backup database Spend qua các giai đoạn xử lý.
- ❖ File Powerpoint trình bày hoàn chỉnh.
- ❖ File Project Document hoàn chỉnh.
- ❖ File Power BI project hoàn chỉnh (Lưu project với format tên: DAT109-CLASS-MSSV.pbix (CLASS: tên lớp, MSSV: mã số sinh viên hoặc mã nhóm).)
- ❖ Tạo thư mục tên <NHOM> chứa tất cả file trên.
- ❖ Nén thư mục StockReport và đặt tên của file nén được quy định là <Mã sinh viên>\_DAT111\_Assignment.zip (ví dụ: TyNVPS10008\_DAT111\_Assignment.zip).
- Nộp bài đúng nơi yêu cầu của giảng viên hướng dẫn.



# BẢO VỆ DỰ ÁN



Cảm ơn