DATAcracyHackathon Case Study

short line

Voizfm - Journey of Content Consumption  
11-12 September 2021

# Case Study

**Voizfm** là nền tảng nghe Sách nói và Podcast chất lượng cao, 100% Bản quyền lớn nhất Việt Nam.

Trong project 4 tuần của Scala, nhóm học viên DATAcracy đã tiến hành thu thập, phân tích và thực hiện Clustering users trên bộ data của Voizfm. Kết quả được thể hiện thông qua Dashboard, với một số kiến nghị như trong [Báo Cáo](https://docs.google.com/presentation/d/1AQuo17RjGQ29k7MRzs-xWuPHWUdJaTm3Xha9K_CG-jc/edit#slide=id.ged7f4989b5_2_8) của nhóm.

Trên nền tảng kết quả các nhóm users khác nhau (kế thừa từ project của nhóm), Voizfm có nhu cầu thực hiện đào sâu và triển khai bài toán tiếp theo, như sau:

1. **Journey of Content Consumption**: Tập trung vào nhóm high-value users, phân tích tương tác giữa User và Contents (Playlist, Category) trong bộ Data của Voizfm
   1. Các nội dung thường được user nghe cùng nhau?
   2. Hành vi khác nhau giữa các nhóm users? (theo giới tính, độ tuổi, OS…)
   3. Đề xuất cách thức định lượng giá trị Content Value (attribute Revenue cho từng contents - tính tương đối) của từng contents trên từng nhóm users.
2. **Use Cases**: Chọn một trong số các use case sau, và thiết kế cách sử dụng các insights từ phần (1) cho một trong số các use cases sau:
   1. *Content Bundle*: Gộp các contents liên quan để offer cho users theo bundle
   2. *Rule-based Recommendation Engine*: Giới thiệu các nội dung mới cho users, dựa trên lịch sử nghe
   3. *Taste Clustering*: Cluster users theo khẩu vị “nghe"
3. **Streamlit Apps**: Thiết kế và triển khai streamlit apps phục vụ cho user cases trong phần (2). Đảm bảo các tiêu chí sau:
   1. Upload các data files (BTC cung cấp)
   2. Visualizations để cung cấp insights từ phần (1)
   3. Features về recommendations từ phần (2)

**Tiêu chí đánh giá**: Tính Hiệu Quả, Tính Giá Trị, Tính Bền Vững [HERE](https://docs.google.com/spreadsheets/d/1T4I3WOkpRw0VzyxbmNkZ2It9v93VP3rPelERpLUch1Q/edit#gid=0)

Tips: Các nhóm tham khảo và thực hiện tuần tự theo các [Checkpoints](#_tvquuf8hh7tc). Nhớ xem phần [Tricks & Tips](#_uh4fm6fdai3o) team nhé.

# 

# Data Dictionary

Data được lưu tại folder [Data/](https://drive.google.com/drive/u/0/folders/1A48vy1dHaOciKvkYbzMVZxA-WmUQ00wV), gồm các file:

* user\_processed.csv
* listening\_processed.csv
* order\_processed.csv
* clustering\_processed.csv

## USER

Bảng “User” ghi nhận các tin cơ bản của các user đang sử dụng voizfm

|  | Field Name | Type | Is Nullable | Description |
| --- | --- | --- | --- | --- |
| 1 | UserID (PK) | INT | No | Mã khách hàng |
| 2 | VIP hay Free | STRING | No | Phân loại khách hàng (nhóm VIP hoặc Free)  **Giá trị:** VIP hoặc Free |
| 3 | Gender | STRING | No | Giới tính của khách hàng  **Giá trị:**  female, male, no information |
| 4 | Year of Birth | STRING | Yes | Năm sinh của khách hàng |
| 5 | Age Range | STRING | No | Nhóm tuổi của khách hàng  **Giá trị:**  Below 18  18-24  25-34  35-44  45-54  55-64  Over 65 |
| 6 | Age | INT | Yes | Tuổi của khách hàng |
| 7 | Region | STRING | No | Khu vực ghi nhận của khách hàng  **Giá trị:** VN/ QT  VN: Việt Nam  QT: Quốc tế |
| 8 | Operation Systems | STRING | Yes | Platform đang sử dụng  **Giá trị:** ios/ android/ website |
| 9 | Registration Date | DATE | No | Ngày đăng ký tài khoản |
| 10 | Total Listening Time | INT | No | Tổng số lượt nghe của user |
| 11 | Total Listening Time (Min) | DECIMAL | No | Tổng thời gian nghe của user tính theo phút |

## LISTENING

Bảng LISTENING ghi nhận các phát sinh mở các playlist trên app của user trong thời gian từ 7/2020 - 12/2020

|  | Field Name | Type | Is Nullable | Description |
| --- | --- | --- | --- | --- |
| 1 | PlaylistID (PK) | INT | No | Mã Playlist |
| 2 | PlaylistName | STRING | No | Tên của Playlist |
| 3 | Category | STRING | No | Phân nhóm của Playlist theo hình thức  Gồm:  Sách nói, Sách tóm tắt, Podcast, Thiếu nhi |
| 4 | SubCategory | STRING | No | Phân nhóm nhỏ của Playlist theo chủ để sách |
| 5 | PlaylistType | STRING | No | Playlist thuộc vip hay free |
| 6 | Playlist Duration (min) | DECIMAL | No | Thời gian của playlist tính theo phút |
| 7 | Actual Duration (min) | DECIMAL | No | Thời gian nghe thực tế của user tính theo phút |
| 8 | UserID (FK) | INT | No | Mã khách hàng |
| 9 | Listening Time | TIME | No | Thời điểm user nghe playlist - Theo mốc giờ: hh:mm:ss AM/PM |
| 10 | Listening Date | DATE | No | Ngày mà user nghe playlist (mm/dd/yyyy) |
| 11 | Listening DateTime | DATETIME | No | Thời điểm user nghe playlist theo ngày giờ |

## 

## ORDER

Bảng ORDER ghi nhận các đơn hàng đăng ký mua các gói dịch vụ trên voizfm theo thời gian từ tháng 7/2020 đến tháng 12/2020

|  | Field Name | Type | Is Nullable | Description |
| --- | --- | --- | --- | --- |
| 1 | OrderID (PK) | INT | No | Mã Đơn hàng |
| 2 | Transaction Date | DATE | No | Ngày phát sinh đơn hàng |
| 3 | UserID (FK) | INT | No | Mã khách hàng |
| 4 | Bill Type | STRING | No | Gói dịch vụ khách hàng mua  **Giá trị**: VIP / 9K / XU |
| 5 | Total Amount | INT | No | Tổng giá trị của đơn hàng theo đơn vị VND |

## 

## CLUSTERING

Bảng CLUSTERING là bảng đánh dấu các user của voizfm theo từng nhóm khách hàng đã được phân chia dựa trên kết quả phân tích và clustering user theo RFM và RFD

|  | Field Name | Type | Is Nullable | Description |
| --- | --- | --- | --- | --- |
| 1 | User\_ID (FK) | INT | No | Mã Khách Hàng |
| 2 | Type\_user | STRING | No | Nhóm khách hàng - Free hay Paid (Đã mua) |
| 3 | MainCluster\_ID | INT | No | Mã MainCluster - Mã nhóm phân loại  Đối với Free User - 10 Nhóm  Đối với Paid User - 8 Nhóm |
| 4 | MainCluster\_Description | STRING | No | Tên mô tả của từng nhóm ([Report](https://docs.google.com/spreadsheets/d/1xeF_02ONrea9K2P4O-0BlHJwc5PKGRioVAzllveRf1g/edit#gid=1087030589)) |

# Data Modelling

## 

## 

# 

# 

# Clustering & Findings

## General *(\*)*

**Mục đích:**

Clustering nhằm tìm ra những đặc điểm chung về tính cách, hành vi của khách hàng. Từ đó, đề ra các chiến lược phù hợp cho mỗi nhóm khách hàng.

**Hướng tiếp cận:**

Vì hiện tại khách hàng của Voiz chia ra làm Free & Paid User nên nhóm sẽ tiến hành Clustering theo 2 cách, dựa trên các yếu tố:

* Đối với Free-User: Mô hình Recency - Frequency - Density (RFD)
* Đối với Paid-User: Mô hình Recency - Frequency - Monetary (RFM)
  + Recency: Thời gian kể từ lần nghe gần nhất
  + Frequency: Tần suất nghe trong một ngày
  + Density: Mật độ thời gian mỗi lần nghe
  + Monetary: Số tiền khách hàng chi trả

*(\*) Tham khảo từ* [*slide của team VoizFM*](https://docs.google.com/presentation/d/1AQuo17RjGQ29k7MRzs-xWuPHWUdJaTm3Xha9K_CG-jc/edit?usp=sharing)

## RFD - For free users

### Insights

1. General - Trong số 66146 users:
   * **96%** user free (no order) **4%** PAY user (order)
   * **58.9%** user chưa nghe lần nào (nolisten) **41.1%** user đã nghe
   * **90.4%** user đã nghe nhưng chưa order, **1%** user đã order nhưng chưa nghe, **8.6%** user đã order và có nghe.
2. Ta chỉ xét những user đã có giao dịch hoặc những user đã sử dụng app (đã nghe sách): khoảng **27458** users
3. Trong đó, có:
   * Những users đã thực hiện giao dịch đều có tổng giá trị giao dịch(order) > 0.0
   * **1%** số lượng PAY users đã giao dịch(order) nhưng chưa nghe (nolisten) (chưa bấm nút nghe của bất kì playlist nào): **275** users
   * **8.6%** PAY users đã giao dịch và đã nghe (**2363** users)
   * Còn lại **90.4%** FREE users đã nghe và CHƯA giao dịch (**24820** users), trong đó có **93** free user có tổng thời gian nghe là 0.0 phút -chưa trải nghiệm playlist nào

### Clustering

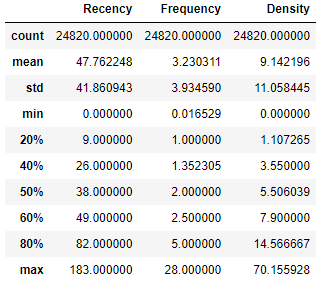
Các bước preprocessing cho clustering:

* Tạo **listening\_summary\_df** để chứa data các user đã nghe qua ít nhất 1 lần
* Tạo **order\_summary\_df** để chứa data các user đã từng order qua ít nhất 1 lần
* Tạo **user\_cluster\_df** gộp các User\_ID của 2 dataframe trên để chứa User\_ID của các user đã từng nghe qua hoặc đã từng order.
* Map các feature từ các dataframe về listening, order, user qua **user\_cluster\_df** bằng User\_ID.
* Theo nhận định của nhóm thì việc normalize features không cải thiện kết quả, vì chỉ thực hiện KMean trên 1 feature -> Thay vào đó **xử lý outliers**.
  + Vì free user nên có 1 số điểm dữ liệu không phản ánh đúng behavior của user (giả định có thể là các tester tạo account để test sản phẩm)
  + Thay thế Density và Frequency của các **sample > quantile 0.995** bằng chính nó(q995).

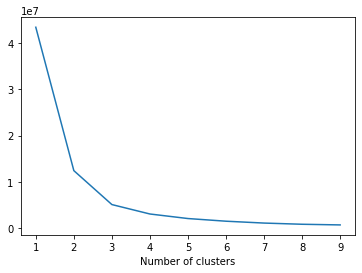
Cách tính RFD

* **Density(D):** thời lượng trung bình của 1 lượt nghe (theo phút)
* **Recency(R):** khoảng cách về ngày giữa lần nghe gần nhất của user so với lần nghe gần nhất của cả dataset
  + R càng nhỏ thì càng tốt. User với R=0 là những user active nhất.
* **Frequency(F):** Số lượt nghe trung bình trong 1 ngày nghe, có 2 loại F
  + Frequency listen week: Số tuần đã trôi qua trên hệ quy chiếu thời gian của bản thân user
  + Frequency listen day: Số ngày đã trôi qua trên hệ quy chiếu thời gian của bản thân user.
  + F càng cao thì mức độ activeness càng lớn -> Càng tốt

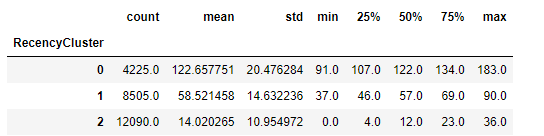
Bảng summary statistics của Free-user RFD:



**Chạy KMean clustering với Recency,**

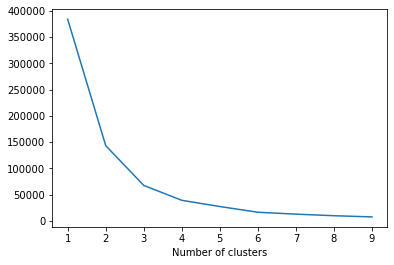


* Dựa vào phương pháp Elbow xác định K=3
* Làm 1 bước post-processing: Sắp xếp thứ tự Recency càng thấp thì số label của cluster càng cao
* Kết quả clustering:



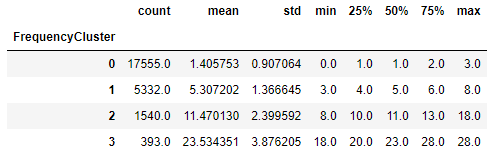
* Tổng kết, Recency chia làm 3 nhóm: Lần cuối cùng active cách khoảng thời gian:
  + Nhóm 0: Hibernate - 4 tháng (3-6 tháng)
  + Nhóm 1: Sleep - 2 tháng (1-3 tháng)
  + Nhóm 2: Active - ½ tháng (0-1 tháng)

**Chạy KMean clustering với Frequency,**

****

Dựa vào phương pháp elbow chọn số cluster K=4

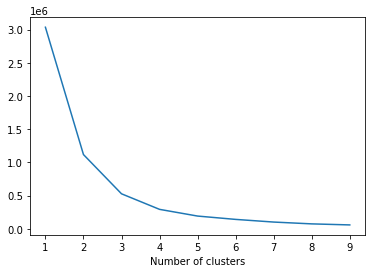
Kết quả clustering:



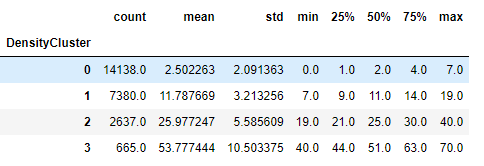
Tổng kết, Frequency chia làm 4 nhóm:

* Nhóm 0: ít dùng : 1 lần/ngày (0 - 3 ): có buổi sử dụng buổi không sử dụng
* Nhóm 1: Dùng mức độ tb:6 lần/ngày (4 - 9 ): mỗi buổi từ 1-2 lần
* Nhóm 2: Dùng mức độ thường xuyên: 13 lần/ngày (10 - 18): mỗi buổi 3-4 lần
* Nhóm 3: Dùng mức độ thường xuyên: 24 lần/ngày (19 - 29):mỗi buổi 5-7 lần

**Chạy KMean clustering với Density**

****

Dựa vào phương pháp elbow chọn K=4



Tổng kết, Frequency chia làm 4 nhóm

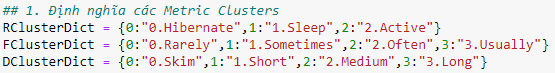
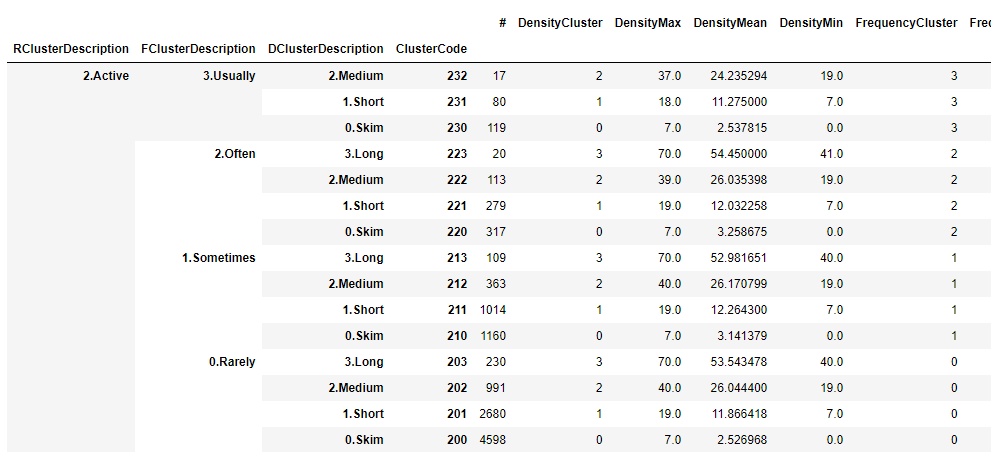
* Nhóm 0: lướt: trung bình khoảng 2' cho mỗi lượt nghe (d<=6]
* Nhóm 1: mật độ TB: trung bình 11' cho mỗi lượt nghe (7<=d<=18)
* Nhóm 2: mật độ cao: trung bình 26' () cho mỗi lượt nghe(19<=d<=40)
* Nhóm 3: mật độ rất cao: trung bình 54' (1h) cho mỗi lượt nghe (41<=d<=70)

### **Overall # of Clusters:**

* **F: 4**
* **D: 4**
* **R: 3**

**Overall Final RFD Clusters**

Cluster cuối cùng của user: Aggregate các sub cluster (R, F, D) thành 1 dataframe

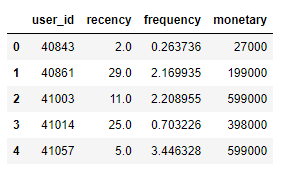
* Gắn các label đã định nghĩa của các cluster vào
* 
* Sắp xếp lại RFD cluster dataframe và export ra file.
* 
* Map từng cluster description vào từng User\_ID và export ra file

## RFM - For Paid users

**Cách tính RFM:**

* **Recency(R)** - Khoảng cách từ ngày cuối cùng của 2020 (31/12/2020) đến ngày gần nhất mà user đó nghe trên app (nếu user chưa nghe lần nào thì sử dụng ngày order gần nhất)
* **Frequency(F)** - Tổng số lần nghe / Time window
  + Time window (đơn vị ngày) được tính bằng ngày nghe cuối cùng (last\_listening\_date) - ngày nghe đầu tiên (first\_listening\_date). Nếu user chưa nghe lần nào, giá trị sẽ là NaN
* **Monetary(M)** - Tổng số lần order

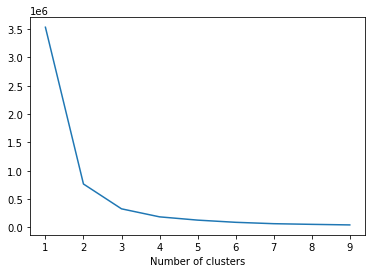
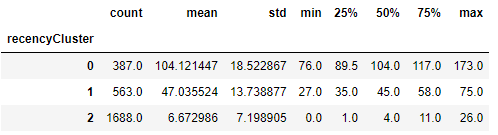
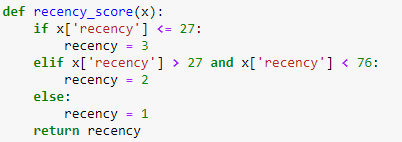
Gộp các dataframe R,F,M lại thành 1:

* 

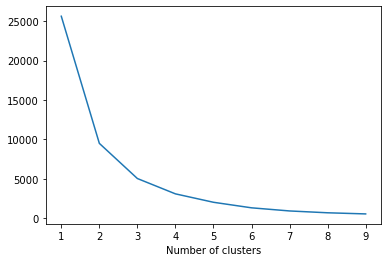
**Các bước preprocessing trước khi clustering:**

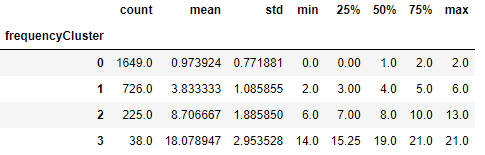
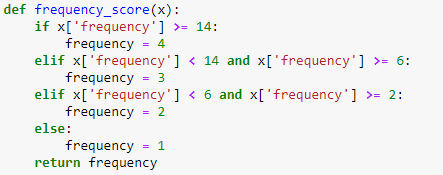
* Thay thế Monetary và Frequency của các **sample > quantile 0.995** bằng chính nó (q995)

**Clustering cho Recency**

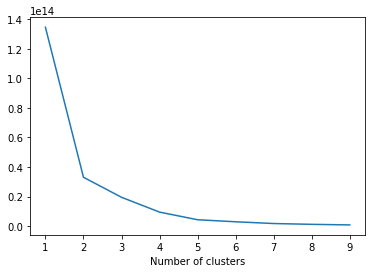
* 
* Dựa vào elbow method chọn K=3
* Kết quả clustering:
* 
* Dựa vào kết quả của clustering chia Recency làm 3 nhóm:
* 

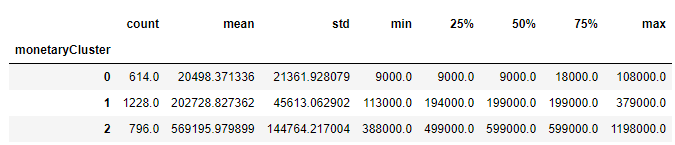
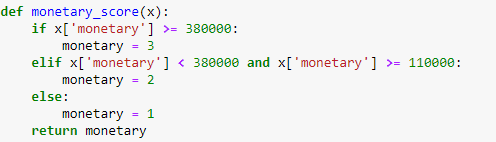
**Clustering cho Frequency**



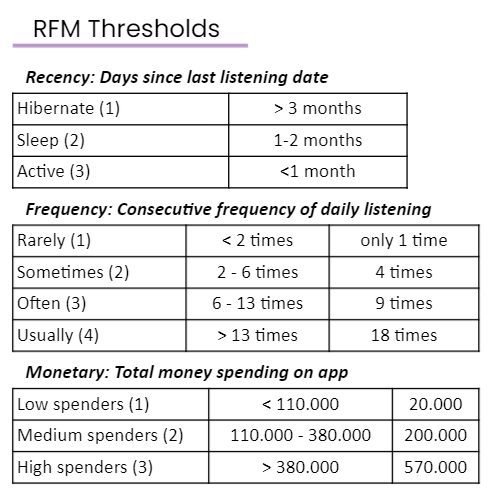
* Dựa vào elbow method chọn K=4
* Kết quả clustering
* 
* Dựa vào kết quả cluster frequency chia làm 4 nhóm:
* 

**Clustering cho Monetary**

****

* Dựa vào elbow method chọn k=3
* Kết quả clustering:
* 
* Dựa vào kết quả Monetary chia làm 3 nhóm
* 

Ý nghĩa cách chia này match với cách bạn trình bày trong [slide](https://docs.google.com/presentation/d/1AQuo17RjGQ29k7MRzs-xWuPHWUdJaTm3Xha9K_CG-jc/edit?usp=sharing):



# Checkpoints

# Tips: Đây là các checkpoints gợi ý, nhóm có thể đi theo lộ trình khác tuỳ ý.

|  | Checkpoints | Outputs | Guidance |
| --- | --- | --- | --- |
| 1 | Develop a Plan | 1. Backlog (List of Tasks) 2. Potential Approaches/ Techniques 3. Potential Hypotheses | - Đọc kỹ và hiểu case study, báo cáo của nhóm SCALA, data dictionary. Hỏi câu hỏi nếu cần thiết  - Google các techniques có thể áp dụng (nếu cần thiết)  - Brainstorms các cách tiếp cận, giả thuyết => Shortlist dựa trên mức độ khả thi và giá trị đem lại  - Backlog các task, deliverables ở mỗi checkpoints, cách đánh giá chất lượng |
| 2 | Data Explore | 1. Kết quả EDA 2. Bộ Data đã được xử lý, thêm các features để thực hiện các bước tiếp theo | - Map các bảng data  - Thực hiện EDA  - Data có cần xử lý: outlier, missing values, hay thêm các trường dữ liệu không (features) |
| 3 | Journey of Contents Consumptions | <Theo plan của nhóm> | - Nhóm định nghĩa thế nào là high-value users - Bắt đầu hypotheses, chứng minh bằng data |
| 4 | Use Cases | 1. Goals của Use Cases 2. Blue-print  3. Evaluation Metrics | - Goals của use cases là gì: Phục vụ cho ai? Cung cấp thông tin gì? Đem lại giá trị gì (về doanh thu)? - Thiết kế (diagram): input data gì, có các features gì, components nào, output là gì?  - Evaluation Metrics của use cases (Đọc: [Tricks & Tips](#_uh4fm6fdai3o)) |
| 5 | Finishing | 1. Presentation 2. Streamlit Apps 3. User Guide | - Dựa vào Blue-print thiết kế Streamlit Apps - Chuẩn bị User Guide để hướng dẫn sử dụng Streamlit Apps + Presentations |

# 

# Tricks & Tips

Mentors Hotline: Các nhóm tận dụng tối đa sự trợ giúp của mentor thông qua Slack channel #hackathon-mentors-hotline. Tất cả các nhóm đều cùng một channel nên việc “thám thính" hỏi-đáp của các nhóm khác đôi khi sẽ có ích cho nhóm mình 😛 Kể cả khi không có câu hỏi gì, hãy “kêu gọi" mentors ở các [checkpoints](#_tvquuf8hh7tc) để đảm bảo mình đi đúng hướng nhé.

Goals-and-Evaluation: Thành phẩm nào cũng cần theo dõi (để bảo hành). Dựa vào goals nhóm đặt ra cho giải pháp của mình, thiết kế CÁC evaluation metrics để đánh giá performance của giải pháp theo: (1) Ultimate Metrics (để đánh giá cuối cùng); (2) Sub-metrics: Đánh giá các tiêu chí của goals; (3) Đánh giá theo các components của solutions (để xem phần nào đang hoạt động hiệu quả)

Ít-mà-chất: Đừng “theo đuổi" quá nhiều hướng/hypothesis cùng một lúc. Team chọn ra 1 hướng chính và 1 hướng back-up. Tập trung làm hướng tiềm năng nhất. Đào sâu và cụ thể hoá. Nếu “bế tắc", mới chuyển qua các hướng back-up. Nên có =< 3 hypotheses để phân tích.

Backlog (Task Management): Dựa vào các [checkpoints](#_tvquuf8hh7tc), team chia nhỏ thành components và tasks. Lúc bắt đầu không cần quá chi tiết, về sau team có thể cụ thể hoá từng tasks (chia nhỏ xuống subtasks). Lưu các tasks vào Backlog (có thể dùng google spreadsheet, hoặc Trello). Khi “triển" xong hết Backlog, là team hoàn thành Hackathon.

Sprint: Chia thành các phiên làm việc (Sprint) kéo dài 2-3 tiếng (tuỳ nhóm quyết định). Trước mỗi sprint, cùng nhau set goals mỗi người sẽ hoàn thành gì trong sprint (chọn các tasks từ backlog). Kết thúc mỗi sprint, team review và thảo luận với nhau. Hints: Mọi người sẽ thấy tinh thần này khá giống Pomodoro (hay Agile)

Giản lược và Chọn lọc: Nhóm có thể làm rất nhiều thứ, nhưng không nhất thiết phải trình bày hết lên Streamlit Apps hay trong bài thuyết trình. Chọn những thông tin đặc sắc nhất, và trình bày sao cho hiệu quả, dễ hiểu nhất.

Collaborative Coding: Đảm bảo code được share kịp thời và hiệu quả, team thoả thuận trước sẽ dùng công cụ gì, cách thức kết hợp ra sao: [Github Desktop](https://desktop.github.com/), [Teletype Atom](https://teletype.atom.io/), Google Colab. Team nên dành tối thứ 6 để họp và thảo luận trước, cũng như đảm bảo các thành viên trong nhóm đã cài đặt các tool và biết cách sử dụng.

Các tools không cần Code: Không phải members nào cũng có khả năng code bằng python. Team có thể tham khảo một số tools kéo thả mà vẫn vô cùng lợi hại: [Exploratory.io](https://exploratory.io/)

# Cấu trúc bài Trình bày Mỗi nhóm 20’ (trình bày: 10’, QA: 10’)

1. **Goals**: Giá trị gì cho Doanh Nghiệp
2. **Evaluation**: Cách thức đánh giá (solutions có đạt các goals đã đề ra) và Kết quả
3. **Solution Design:** Diagram của Streamlit Apps
4. **Demo:** End-user (Doanh Nghiệp) sẽ sử dụng Streamlit Apps như thế nào?

**Appendix (KHÔNG thuyết trình, nhưng trình bày trong slide):**

* Insights & Findings
* Techniques & methodologies