(CẤU TRÚC BÁO CÁO ĐỒ ÁN TRỰC QUAN HÓA DỮ LIỆU)

[Project\_Heart\_Disease]

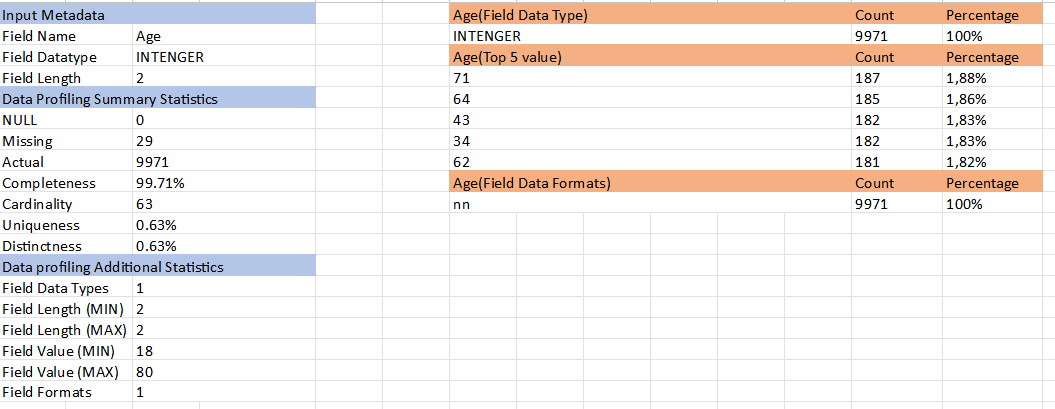
# THÔNG TIN NHÓM

|  |  |  |  |
| --- | --- | --- | --- |
| *Mã nhóm* | *MSSV* | *Họ và tên* | *Ghi chú* |
| [22HTTT]  [3] | 22127262 | Đỗ Lương Nhật Minh |  |
| 22127366 | Lê Xuân Hoàng Sơn |  |
| 22127197 | Nguyễn Đăng Khoa |  |
| 22127325 | Cao Hoàng Phong |  |

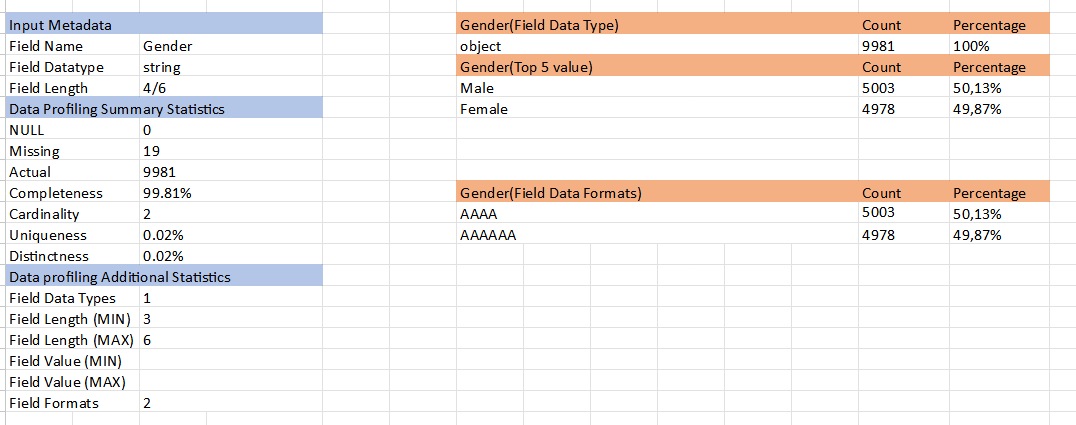
# **Dataprofiling**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **NULL** | **Missing** | **Actual** | **Completeness** | **Cardinality** | **Uniqueness** | **Distinctness** |
| **Age** | 0 | 28 | 9,971 | 99.71% | 63 | 0.63% | 0.63% |
| **Gender** | 0 | 19 | 9,981 | 99.81% | 2 | 0.02% | 0.02% |
| **Blood Pressure** | 0 | 19 | 9,981 | 99.81% | 61 | 0.61% | 0.61% |
| **Cholesterol Level** | 0 | 30 | 9970 | 99.7% | 151 | 1,51% | 1,51% |
| **Exercise Habits** | 0 | 25 | 9975 | 99.75% | 3 | 0,03% | 0,03% |
| **Smoking** | 0 | 25 | 9975 | 99.75% | 2 | 0,02% | 0,02% |
| **Family Heart Disease** | 0 | 21 | 9979 | 99.79% | 2 | 0.02% | 0.020042% |
| **BMI** | 0 | 22 | 9978 | 99.78% | 9978 | 99.78% | 100% |
| **Alcohol Consumption** | 0 | 2586 | 7414 | 74.14% | 3 | 0.0%3 | 0.040464% |
| **Stress Level** | 0 | 22 | 9978 | 99,78% | 3 | 0.03% | 0.03% |
| **Heart Disease Status** | 0 | 0 | 10000 | 100.00% | 2 | 0.02% | 0.02% |

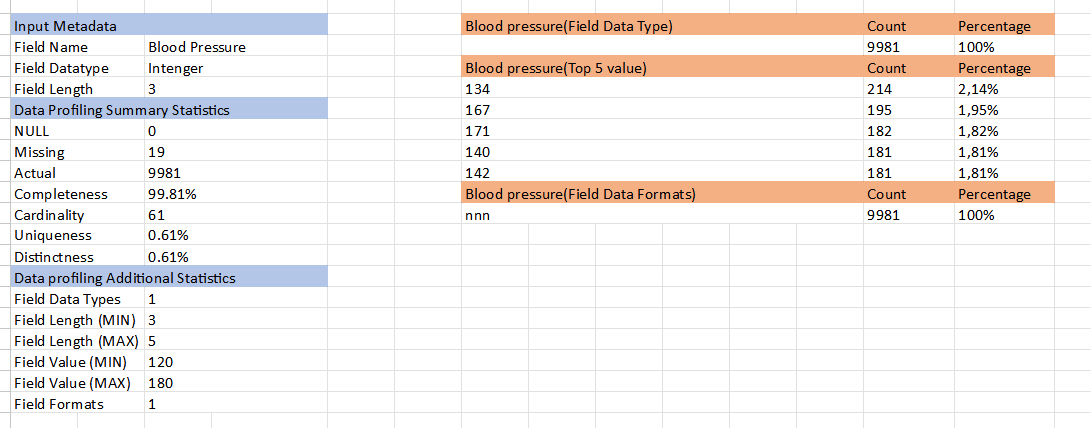
* **Age**

****

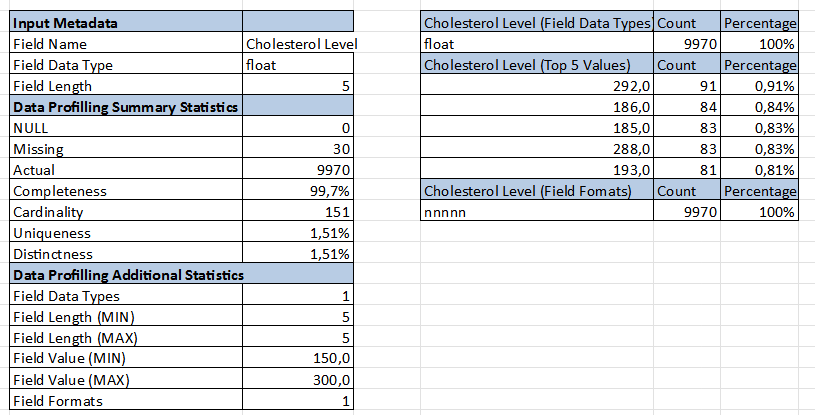
* **Gender**

****

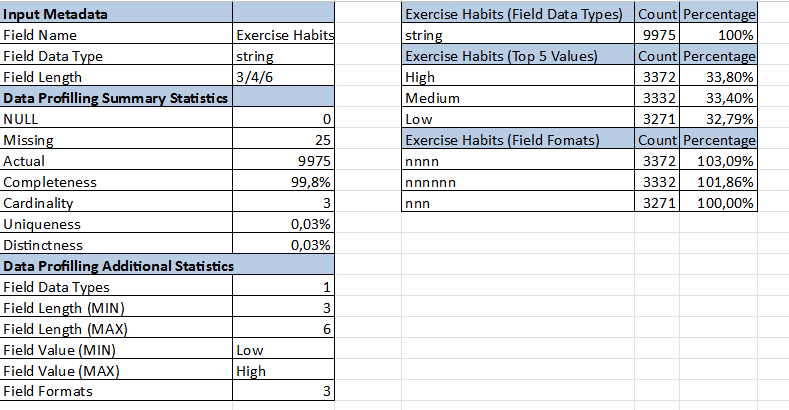
* **Blood Pressure**

****

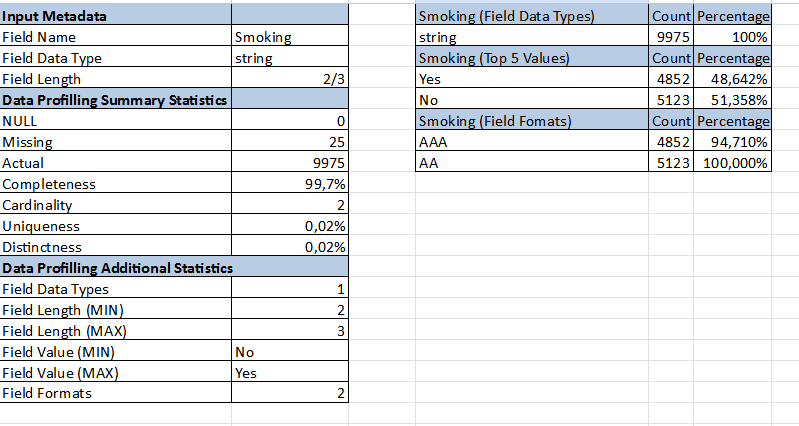
* **Cholesterol Level**

****

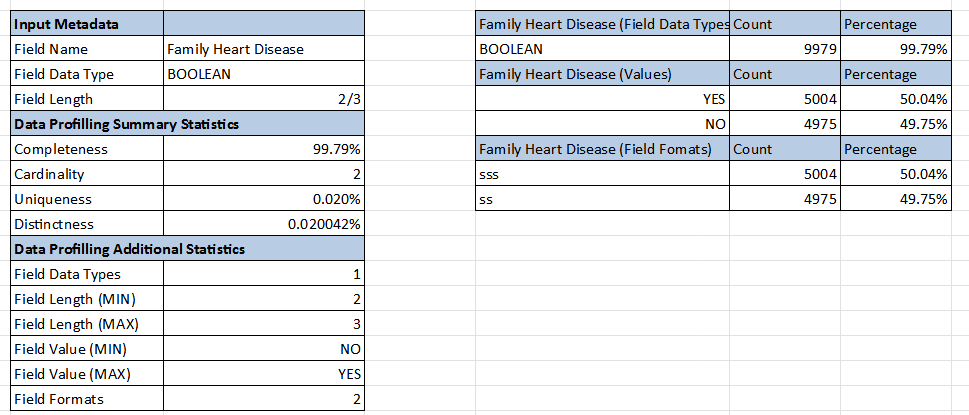
* **Exercise Habits**

****

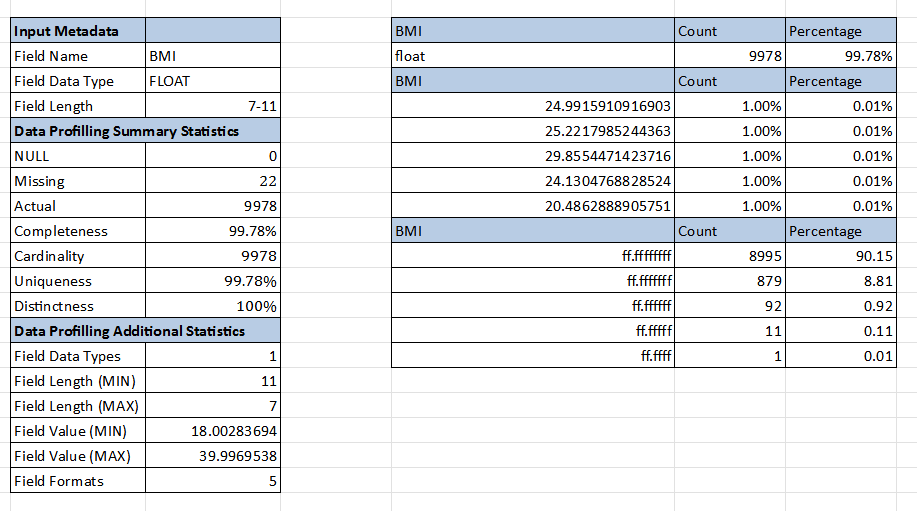
* **Smoking**

****

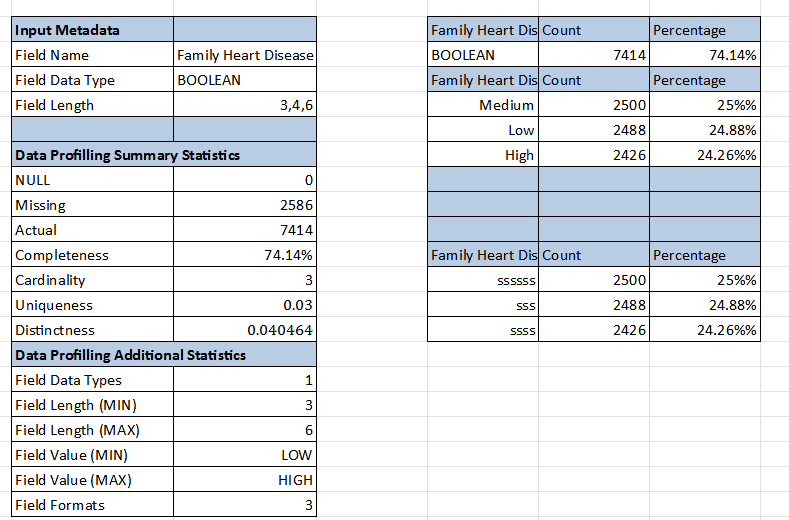
* **Family Heart Disease**

****

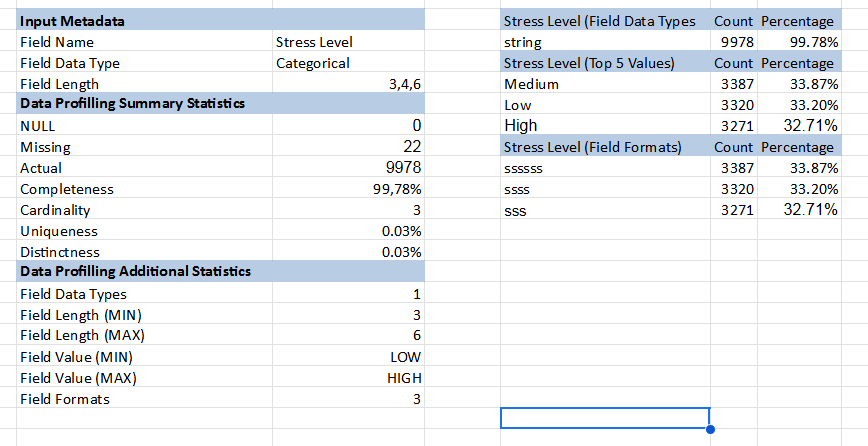
* **BMI**

****

* **Alcohol Consumption**

****

* **Stress Level**

****

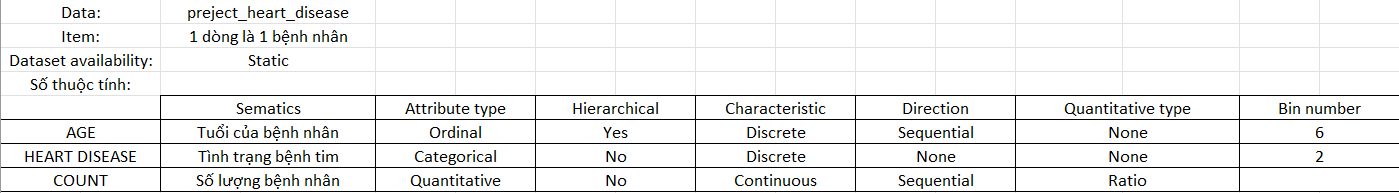
* **Heart Disease Status**

# 

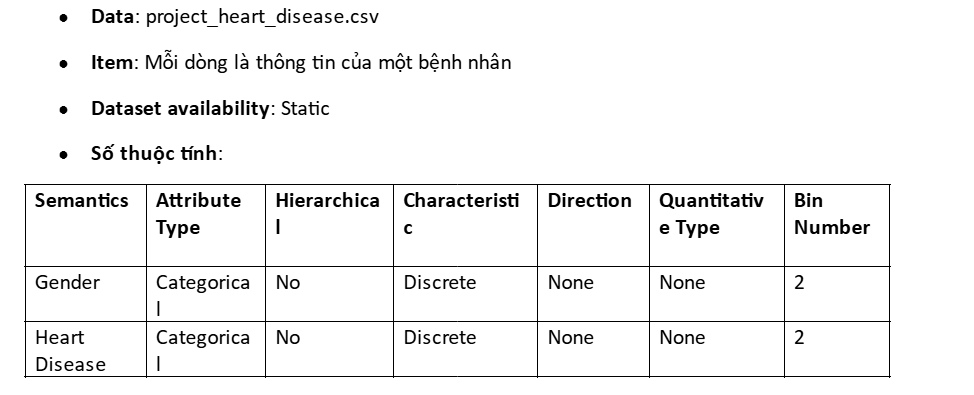
* **Đánh giá kết quả profiling**
* **Tổng quát trên bảng dữ liệu:**
* Dựa trên kết quả profiling, ta nhận thấy không có giá trị NULL trong tất cả các thuộc tính. Tuy nhiên, một số thuộc tính có giá trị bị thiếu (Missing). Đáng chú ý nhất là Alcohol Consumption, với 2,586 giá trị bị thiếu, làm giảm mức độ hoàn chỉnh của thuộc tính này xuống còn 74.14%.
* Ngoài ra, thuộc tính Cholesterol Level và Age cũng có một lượng nhỏ giá trị bị thiếu (30 và 28 giá trị), nhưng tỷ lệ hoàn chỉnh vẫn trên 99%, do đó không ảnh hưởng quá lớn đến chất lượng dữ liệu.
* Nhận xét trên từng thuộc tính:
* Độ tuổi được khảo sát trong thí nghiệm trải dài từ 18 đến 80 tuổi, trong đó chiếm nhiều nhất là các người thuộc nhóm tuổi từ trung niên trở lên (trên 30 tuổi)
* Tỉ lệ nam nữ tham gia khảo sát khá đồng đều ( ~50% mỗi bên)
* Tỉ lệ người bị mắc bệnh tim < so với người không mắc (¼)
* Các thuộc tính mang category (Yes/No hoặc Low/Med/High) luôn được phân phối khá đồng đều giữ các mức (trừ thuộc tính Heart Disease Status)

# **Abstraction**

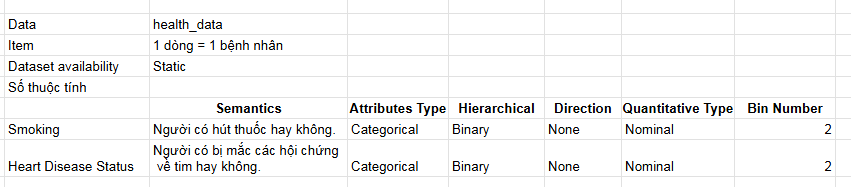
* 1. **Domain task 1: What is the distribution of heart disease status across different age groups?** 
     1. **Data abstraction**



* + 1. **Task abstraction**
* **group → compare → count**
  + **Group (Nhóm tuổi)**: Phân chia dữ liệu theo các nhóm tuổi khác nhau (ví dụ: 20-30, 31-40,...).
  + **Compare (So sánh tỷ lệ mắc bệnh)**: Xác định sự khác biệt về tỷ lệ bệnh tim giữa các nhóm tuổi.
  + **Count (Đếm số lượng trường hợp)**: Tính số lượng bệnh nhân mắc và không mắc bệnh tim trong từng nhóm tuổi.
  1. **Domain task 2: How does gender correlate with heart disease prevalence?** 
     1. **Data abstraction**

****

* + 1. **Task abstraction**
* **Consume → Explore → Compare**
  + **Produce**: Xây dựng bảng thống kê mối tương quan giữa giới tính và tỷ lệ mắc bệnh tim.
  + **Explore**: Kiểm tra tỷ lệ mắc bệnh tim theo từng nhóm.
  + **Compare**: So sánh tỉ lệ mắc bệnh tim giữa nam và nữ.
  1. **Domain task 3: Is there a relationship between smoking status and heart disease occurrence?**
     1. **Data abstraction**



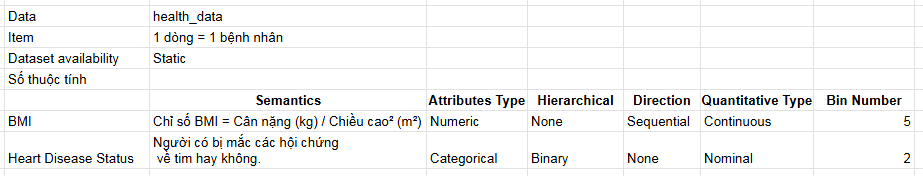
* + 1. **Task abstraction**
* **Procedure → Explore → Compare**
  + **Procedure:** Tạo bảng mới với 2 nhóm: (1) Nhóm hút thuốc và (2) Nhóm không hút thuốc, kèm theo số ca mắc bệnh tim trong mỗi nhóm
  + **Explore:** Kiểm tra tỷ lệ mắc bệnh tim giữa hai nhóm hút thuốc và không hút thuốc + Sử dụng grouped/Stacked Bar Chart.
  + **Compare:**So sánh tỉ lệ mắc bệnh tim giữa hai nhóm để xem hút thuốc có làm tăng nguy cơ hay không
  1. **Domain task 4: How do exercise habits influence heart disease status?** 
     1. **Data abstraction**

### 

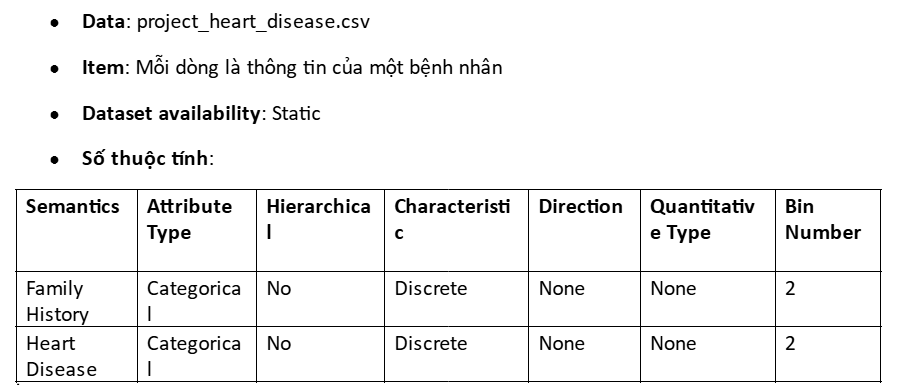
* + 1. **Task abstraction**
* **Task Type:** Comparative (so sánh)  
  **Produce → Filter → Compare**
  + **Produce:** Chọn ra cột "Exercise Habits" và "Heart Disease Status".
  + **Filter:** Phân nhóm theo từng mức độ "Exercise Habits".
  + **Compare:** So sánh tỷ lệ mắc bệnh tim trong từng nhóm thói quen tập thể dục.
  1. **Domain task 5: How does cholesterol level vary between individuals with and without heart disease?** 
     1. **Data abstraction**

### 

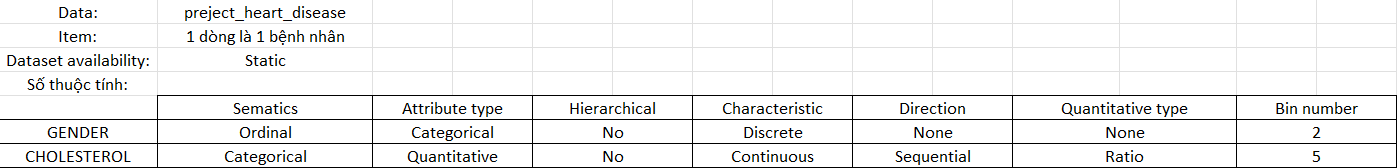
* + 1. **Task abstraction**
* **Task Type:** Comparative (so sánh phân bố)
* **Produce → Filter → Compare Distribution**
  + **Produce:** Lấy cột "Cholesterol" và "Heart Disease Status".
  + **Filter:** Phân nhóm theo "Heart Disease Status" (Yes/No).
  + **Compare Distribution:** So sánh sự khác biệt trong phân bố cholesterol giữa 2 nhóm (bằng boxplot hoặc violin plot).
  1. **Domain task 6: Is there a correlation between BMI and heart disease status?** 
     1. **Data abstraction**

****

* + 1. **Task abstraction**
* **Procedure → Explore → Analyze**
  + **Procedure:** Tạo bảng chứa cột BMI và Heart Disease Status
  + **Explore:** Tính hệ số tương quan (Pearson hoặc Spearman) giữa BMI và Heart Disease Status + Vẽ biểu đồ scatter plot hoặc boxplot để trực quan hóa mối quan hệ
  + **Analyze:** Kiểm tra xem BMI có mối liên hệ chặt chẽ với bệnh tim không
  1. **Domain task 7: Does a family history of heart disease increase the risk?**
     1. **Data abstraction**

****

* + 1. **Task abstraction**
* **Produce → Explore → Analyze**
  + **Produce**: Xây dựng bảng thống kê mối tương quan giữa tiền sử gia đình và tỷ lệ mắc bệnh tim.
  + **Explore**: Khám phá mối quan hệ giữa tiền sử gia đình và bệnh tim.
  + **Analyze**: Đánh giá ảnh hưởng của tiền sử gia đình đến nguy cơ mắc bệnh tim.
  1. **Domain task 8: How does the distribution of cholesterol levels differ between males and females**
     1. **Data abstraction**



* + 1. **Task abstraction**
* **group → compare → summarize**
  + **Group (Nhóm theo giới tính)**: Chia dữ liệu thành hai nhóm **Male (Nam)** và **Female (Nữ)**.
  + **Compare (So sánh sự khác biệt)**: Kiểm tra sự khác biệt về phân bố mức cholesterol giữa hai nhóm.
  + **Summarize (Tóm tắt dữ liệu)**: Sử dụng các thống kê như trung bình, trung vị, khoảng tứ phân vị (IQR) để mô tả sự khác biệt.