VIETNAM NATIONAL UNIVERSITY, HCMC

UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY


# DATA VISUALIZATION USING D3.JS

Instructor:

Nguyễn Ngọc Minh Châu

## GROUP INFORMATION

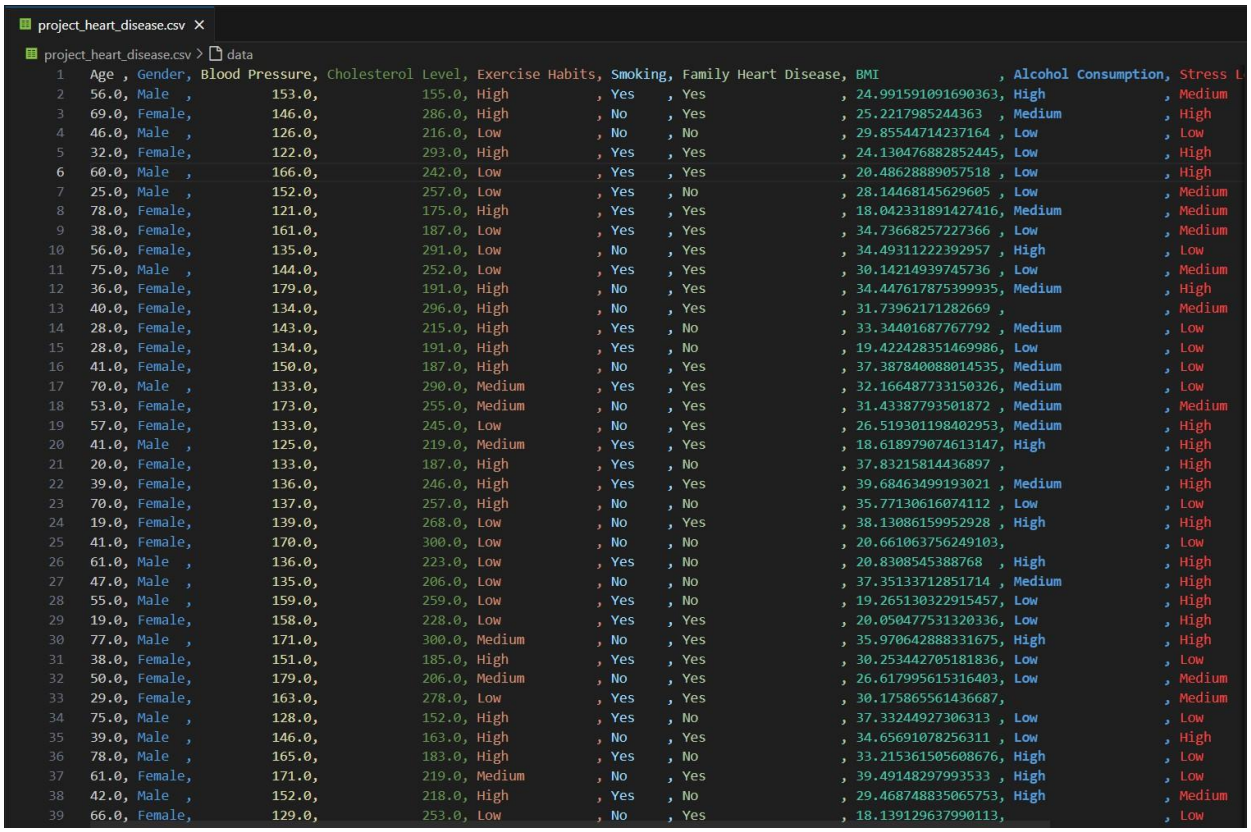
| ***Group code*** | ***MSSV*** | ***Full name*** | ***Note*** |
|---|---|---|---|
| 22HTTT Group 11 | 22127107 | Nguyễn Thế Hiển | Trưởng nhóm |
| | 22127260 | Bùi Công Mậu | |
| | 22127355 | Nguyễn Trần Đại Quốc | |
| | 22127400 | Thái Hữu Thọ | |

**NOTE: The examples provided are only for illustrative purposes in the assumption section. Do not use them as student work or assignment submissions..**

# 1. Dataprofiling

## 1.1 Data preprocessing

The dataset is provided in the file: **project_heart_disease.csv**.



project_heart_disease.csv

```
Age , Gender, Blood Pressure, Cholesterol Level, Exercise Habits, Smoking, Family Heart Disease, BMI             , Alcohol Consumption, Stress L
56.0, Male  ,          153.0,             155.0, High          , Yes    , Yes                , 24.991591091690363, High              , Medium
69.0, Female,          146.0,             286.0, High          , No     , Yes                , 25.2217985244363  , Medium            , High
46.0, Male  ,          126.0,             216.0, Low           , No     , No                 , 29.85544714237164 , Low               , Low
32.0, Female,          122.0,             293.0, High          , Yes    , Yes                , 24.130476882852445, Low               , High
60.0, Male  ,          166.0,             242.0, Low           , Yes    , Yes                , 20.48628889057518 , Low               , High
25.0, Male  ,          152.0,             257.0, Low           , Yes    , No                 , 28.14468145629605 , Low               , Medium
78.0, Female,          121.0,             175.0, High          , Yes    , Yes                , 18.042331891427416, Medium            , Medium
38.0, Female,          161.0,             187.0, Low           , Yes    , Yes                , 34.73668257227366 , Low               , Medium
56.0, Female,          135.0,             291.0, Low           , No     , Yes                , 34.49311222392957 , High              , Low
75.0, Male  ,          144.0,             252.0, Low           , Yes    , Yes                , 30.14214939745736 , Low               , Medium
36.0, Female,          179.0,             191.0, High          , No     , Yes                , 34.447617875399935, Medium            , High
40.0, Female,          134.0,             296.0, High          , No     , Yes                , 31.73962171282669 ,                   , Medium
28.0, Female,          143.0,             215.0, High          , Yes    , No                 , 33.34401687767792 , Medium            , Low
28.0, Female,          134.0,             191.0, High          , Yes    , No                 , 19.422428351469986, Low               , Low
41.0, Female,          150.0,             187.0, High          , No     , Yes                , 37.387840088014535, Medium            , Low
70.0, Male  ,          133.0,             290.0, Medium        , Yes    , Yes                , 32.166487733150326, Medium            , Low
53.0, Female,          173.0,             255.0, Medium        , No     , Yes                , 31.43387793501872 , Medium            , Medium
57.0, Female,          133.0,             245.0, Low           , No     , Yes                , 26.519301198402953, Medium            , High
41.0, Male  ,          125.0,             219.0, Medium        , Yes    , Yes                , 18.618979074613147, High              , High
20.0, Female,          133.0,             187.0, High          , Yes    , No                 , 37.83215814436897 ,                   , High
39.0, Female,          136.0,             246.0, High          , Yes    , Yes                , 39.68463499193021 , Medium            , High
70.0, Female,          137.0,             257.0, High          , No     , No                 , 35.77130616074112 , Low               , Low
19.0, Female,          139.0,             268.0, Low           , No     , Yes                , 38.13086159952928 , High              , High
41.0, Female,          170.0,             300.0, Low           , No     , No                 , 20.661063756249103,                   , Low
61.0, Male  ,          136.0,             223.0, Low           , Yes    , No                 , 20.8308545388768  , High              , High
47.0, Male  ,          135.0,             206.0, Low           , No     , No                 , 37.35133712851714 , Medium            , High
55.0, Male  ,          159.0,             259.0, Low           , Yes    , No                 , 19.265130322915457, Low               , High
19.0, Female,          158.0,             228.0, Low           , Yes    , Yes                , 20.050477531320336, Low               , High
77.0, Male  ,          171.0,             300.0, Medium        , No     , Yes                , 35.970642888331675, High              , High
38.0, Female,          151.0,             185.0, High          , Yes    , Yes                , 30.253442705181836, Low               , Low
50.0, Female,          179.0,             206.0, Medium        , No     , Yes                , 26.617995615316403, Low               , Medium
29.0, Female,          163.0,             278.0, Low           , Yes    , Yes                , 30.175865561436687,                   , Medium
75.0, Male  ,          128.0,             152.0, High          , Yes    , No                 , 37.33244927306313 , Low               , Low
39.0, Male  ,          146.0,             163.0, High          , No     , Yes                , 34.65691078256311 , Low               , High
78.0, Male  ,          165.0,             183.0, High          , Yes    , No                 , 33.215361505608676, High              , Low
61.0, Female,          171.0,             219.0, Medium        , No     , Yes                , 39.49148297993533 , High              , Low
42.0, Male  ,          152.0,             218.0, High          , Yes    , No                 , 29.468748835065753, High              , Medium
66.0, Female,          129.0,             253.0, Low           , No     , Yes                , 18.139129637990113,                   , Low
```

*File project_heart_disease.csv*

The dataset contains **10,000 records** (without headers), with a total of **11 described attributes** as shown below:

| Attribute | Description |
|---|---|
| Age | Age (in years) |
| Gender | Gender (Male or Female) |
| Blood pressure | Blood pressure (measured) |
| Cholesterol Level | Total cholesterol level |
| Exercise Habits | Physical activity level (Low, Medium, High) |
| Smoking | Whether the individual smokes (Yes or No) |
| Family Heart Disease | Whether there is a family history of heart disease (Yes/No) |

| | |
|---|---|
| BMI | Body Max Index (BMI) |
| Alcohol Consumption | Alcohol intake level |
| Stress Level | Stress level |
| Heart Disease Status | Heart Diease Status (Yes or No) |

The dataset provided is raw and contains multiple missing values, summarized as follows:

```
Số lượng giá trị null:
 Age                          29
Gender                        19
Blood Pressure                19
Cholesterol Level             30
Exercise Habits               25
Smoking                       25
Family Heart Disease          21
BMI                           22
Alcohol Consumption         2586
Stress Level                  22
Heart Disease Status           0
dtype: int64
Số lượng giá trị missing values: 0
Số lượng dòng trùng lặp: 0
```

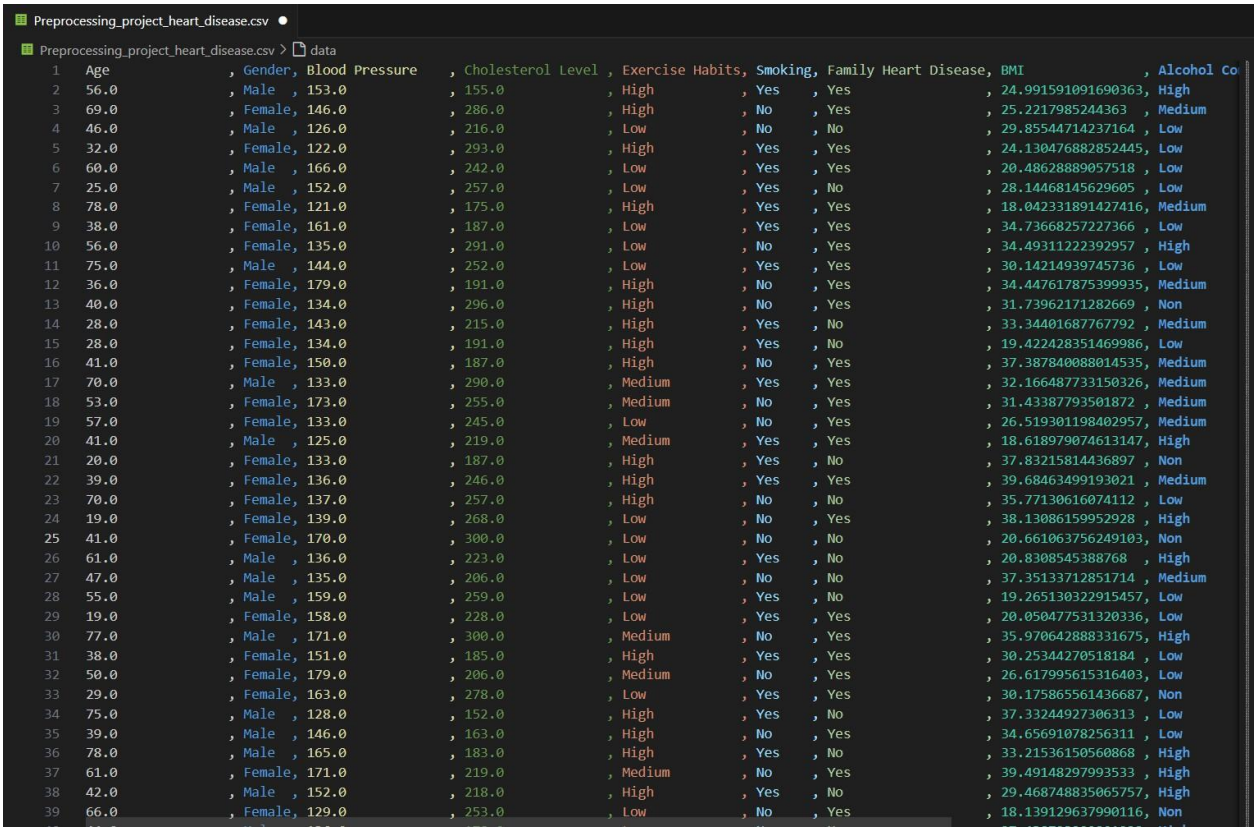
Based on the information above, the following data preprocessing steps were performed:

+ For numerical (float) data: Missing values were filled using the mean of the respective column.

+ For categorical (object/string) data: Missing values were replaced with the value “None” to indicate absence. In cases where values were not applicable or could introduce inconsistencies into the dataset, no imputation was performed to avoid data bias during processing.

After preprocessing, the cleaned dataset was saved in the file:

“Preprocessing_project_heart_disease.csv”:



Preprocessing_project_heart_disease.csv ●

Preprocessing_project_heart_disease.csv > data

```
Age , Gender, Blood Pressure , Cholesterol Level , Exercise Habits, Smoking, Family Heart Disease, BMI , Alcohol Co
56.0 , Male , 153.0 , 155.0 , High , Yes , Yes , 24.991591091690363, High
69.0 , Female, 146.0 , 286.0 , High , No , Yes , 25.2217985244363 , Medium
46.0 , Male , 126.0 , 216.0 , Low , No , No , 29.85544714237164 , Low
32.0 , Female, 122.0 , 293.0 , High , Yes , Yes , 24.130476882852445, Low
60.0 , Male , 166.0 , 242.0 , Low , Yes , Yes , 20.48628889057518 , Low
25.0 , Male , 152.0 , 257.0 , Low , Yes , No , 28.14468145629605 , Low
78.0 , Female, 121.0 , 175.0 , High , Yes , Yes , 18.042331891427416, Medium
38.0 , Female, 161.0 , 187.0 , Low , Yes , Yes , 34.73668257227366 , Low
56.0 , Female, 135.0 , 291.0 , Low , No , Yes , 34.49311222392957 , High
75.0 , Male , 144.0 , 252.0 , Low , Yes , Yes , 30.14214939745736 , Low
36.0 , Female, 179.0 , 191.0 , High , No , Yes , 34.447617875399935, Medium
40.0 , Female, 134.0 , 296.0 , High , No , Yes , 31.73962171282669 , Non
28.0 , Female, 143.0 , 215.0 , High , Yes , No , 33.34401687767792 , Medium
28.0 , Female, 134.0 , 191.0 , High , Yes , No , 19.422428351469986, Low
41.0 , Female, 150.0 , 187.0 , High , No , Yes , 37.387840088014535, Medium
70.0 , Male , 133.0 , 290.0 , Medium , Yes , Yes , 32.166487733150326, Medium
53.0 , Female, 173.0 , 255.0 , Medium , No , Yes , 31.43387793501872 , Medium
57.0 , Female, 133.0 , 245.0 , Low , No , Yes , 26.519301198402957, Medium
41.0 , Male , 125.0 , 219.0 , Medium , Yes , Yes , 18.618979074613147, High
20.0 , Female, 133.0 , 187.0 , High , Yes , No , 37.83215814436897 , Non
39.0 , Female, 136.0 , 246.0 , High , Yes , Yes , 39.68463499193021 , Medium
70.0 , Female, 137.0 , 257.0 , High , No , No , 35.77130616074112 , Low
19.0 , Female, 139.0 , 268.0 , Low , No , Yes , 38.13086159952928 , High
41.0 , Female, 170.0 , 300.0 , Low , No , No , 20.661063756249103, Non
61.0 , Male , 136.0 , 223.0 , Low , Yes , No , 20.8308545388768 , High
47.0 , Male , 135.0 , 206.0 , Low , No , No , 37.35133712851714 , Medium
55.0 , Male , 159.0 , 259.0 , Low , Yes , No , 19.265130322915457, Low
19.0 , Female, 158.0 , 228.0 , Low , Yes , Yes , 20.050477531320336, Low
77.0 , Male , 171.0 , 300.0 , Medium , No , Yes , 35.970642888331675, High
38.0 , Female, 151.0 , 185.0 , High , Yes , Yes , 30.25344270518184 , Low
50.0 , Female, 179.0 , 206.0 , Medium , No , Yes , 26.617995615316403, Low
29.0 , Female, 163.0 , 278.0 , Low , Yes , Yes , 30.175865561436687, Non
75.0 , Male , 128.0 , 152.0 , High , Yes , No , 37.33244927306313 , Low
39.0 , Male , 146.0 , 163.0 , High , No , Yes , 34.65691078256311 , Low
78.0 , Male , 165.0 , 183.0 , High , Yes , No , 33.21536150560868 , High
61.0 , Female, 171.0 , 219.0 , Medium , No , Yes , 39.49148297993533 , High
42.0 , Male , 152.0 , 218.0 , High , Yes , No , 29.468748835065757, High
66.0 , Female, 129.0 , 253.0 , Low , No , Yes , 18.139129637990116, Non
```

*File Preprocessing_project_heart_disease.csv*

## 1.2 Full Data Profiling

The entire dataset was profiled, and the results are summarized as follows:

| | Field Name | NULL | Missing | Actual | Completeness (%) | Cardinality | Uniqueness (%) | Distinctness (%) |
|---|---|---|---|---|---|---|---|---|
| 0 | Age | 0 | 0 | 10000 | 100.0 | 64 | 0.64 | 0.64 |
| 1 | Gender | 0 | 0 | 10000 | 100.0 | 3 | 0.03 | 0.03 |
| 2 | Blood Pressure | 0 | 0 | 10000 | 100.0 | 62 | 0.62 | 0.62 |
| 3 | Cholesterol Level | 0 | 0 | 10000 | 100.0 | 152 | 1.52 | 1.52 |
| 4 | Exercise Habits | 0 | 0 | 10000 | 100.0 | 4 | 0.04 | 0.04 |
| 5 | Smoking | 0 | 0 | 10000 | 100.0 | 3 | 0.03 | 0.03 |
| 6 | Family Heart Disease | 0 | 0 | 10000 | 100.0 | 3 | 0.03 | 0.03 |
| 7 | BMI | 0 | 0 | 10000 | 100.0 | 9979 | 99.79 | 99.79 |
| 8 | Alcohol Consumption | 0 | 0 | 10000 | 100.0 | 4 | 0.04 | 0.04 |
| 9 | Stress Level | 0 | 0 | 10000 | 100.0 | 4 | 0.04 | 0.04 |
| 10 | Heart Disease Status | 0 | 0 | 10000 | 100.0 | 2 | 0.02 | 0.02 |

Remarks:

- After preprocessing, all missing (NULL) values have been handled, resulting in zero missing entries across all fields.

- Only the **Heart Disease Status** attribute contains fully complete information, with a Completeness score of 100%. The remaining attributes have a Completeness rate above 99% (except for Alcohol Consumption at 74.14%), depending on how the missing values were handled.
- The attributes **Gender, Exercise Habits, Smoking, Family Heart Disease, Alcohol Consumption, Stress Level**, and **Heart Disease Status** exhibit low **Uniqueness** and **Distinctness** scores, indicating they consist of only 2 to 4 distinct values.
- The **BMI** attribute shows high **Cardinality**, **Uniqueness**, and **Distinctness**, suggesting that nearly every record has a unique value.
- A definitive **primary key** has not been identified in this dataset. Although the **BMI** attribute has a Distinctness score of 100%, duplication may still exist in real-world data.

## 1.3 Attribute-wise Profiling

## Age:

**Metadata**

| Attribute | Field Datatype | Field Length |
|---|---|---|
| Age | float64 | 1 |

**Data Profiling Summary**

| Attribute | Value |
|---|---|
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 64 |
| Uniqueness (%) | 0.64 |
| Distinctness (%) | 0.64 |

**Data Profiling Additional Statistics**

| Attribute | Value |
|---|---|
| Data Types | 1 |
| Field Length (MIN) | 2 |
| Field Length (MAX) | 2 |
| Field Value (MIN) | 18 |
| Field Value (MAX) | 80 |
| Mean | 49.3 |
| Median | 49 |
| Standard Deviation | 18.17 |

**Age Popular**

| Value | Count | Percentage (%) |
|---|---|---|
| 18 | 149 | 1.49 |
| 19 | 155 | 1.55 |
| 20 | 154 | 1.54 |
| 21 | 162 | 1.62 |
| 22 | 142 | 1.42 |
| 23 | 160 | 1.6 |
| 24 | 132 | 1.32 |
| 25 | 168 | 1.68 |
| 26 | 147 | 1.47 |
| 27 | 142 | 1.42 |
| 28 | 157 | 1.57 |
| 29 | 156 | 1.56 |
| 30 | 161 | 1.61 |
| 31 | 152 | 1.52 |
| 32 | 160 | 1.6 |
| 33 | 149 | 1.49 |
| 34 | 182 | 1.82 |
| 35 | 157 | 1.57 |
| 36 | 156 | 1.56 |
| 37 | 161 | 1.61 |
| 38 | 168 | 1.68 |
| 39 | 169 | 1.69 |
| 40 | 174 | 1.74 |
| 41 | 154 | 1.54 |
| 42 | 167 | 1.67 |

**Age Field Data Types**

| Data Type | Count | Percentage (%) |
|---|---|---|
| float64 | 10000 | 100 |
| NULL | 0 | 0 |

## Summary:

Completeness:

- The **Age** attribute is 100% complete, with no NULL values.
- The dataset is fully available and requires no imputation.

Cardinality (Number of Unique Values)

- There are **64 unique values**, indicating 64 different age levels.
- This means each age value appears multiple times in the dataset..

Uniqueness & Distinctness

- Uniqueness: 0.64% → This means each age value appears multiple times in the dataset.
- Distinctness: 0.64% → Indicates a high level of repetition across different age levels.

## Detailed Observations:

Age distribution:

- Age values range from **18 to 80 years**.
- Most common age range (highest concentration of individuals): **30–50 years**, each accounting for approximately **1.5%–1.8%** of the dataset.
- No significant outliers → Even and balanced age distribution.

Statistical Summary:

- Minimum (Min): 18 tuổi.
- Maximum (Max): 80 tuổi.
- Mean: 49.3 tuổi.
- Median: 49 tuổi → Suggests a relatively symmetrical distribution
- Standard Deviation: 18.17 → Indicates moderate dispersion and no extreme clustering in specific age groups.

Numeric Data Type:

- The **Age** attribute is of **numeric type (float64)**..
- No missing values (0% NULL rate).
- No transformation required — fully suitable for statistical analysis.

# Gender:

**Metadata**

| Attribute | Field Datatype | Field Length |
|---|---|---|
| Gender | object | 1 |

**Data Profiling Summary**

| Attribute | Value |
|---|---|
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 2 |
| Uniqueness (%) | 0.02 |
| Distinctness (%) | 0.02 |

**Data Profiling Additional Statistics**

| Attribute | Value |
|---|---|
| Data Types | 1 |
| Field Length (MIN) | 4 |
| Field Length (MAX) | 6 |

**Gender Popular**

| Value | Count | Percentage (%) |
|---|---|---|
| Male | 5022 | 50.22 |
| Female | 4978 | 49.78 |

**Gender Field Data Types**

| Data Type | Count | Percentage (%) |
|---|---|---|
| object | 10000 | 100 |
| NULL | 0 | 0 |

## Summary:

Completeness

- The **Gender** attribute is 100% complete, with no NULL values..
- Data is fully available and requires no missing value handling.

Cardinality (Number of Unique Values)

- There are **2 unique values**: "Male" and "Female".
- Indicates the data consists of two clearly defined gender categories.

Uniqueness & Distinctness

- Uniqueness: 0.02% → Each gender value appears repeatedly across the dataset.
- Distinctness: 0.02% → The data is not highly diverse and belongs to only two main groups.

## Detailed Observations:

Gender Distribution:

- Number of **Males**: 5022 (50.22% of the dataset)
- Number of **Males**: 5022 (50.22% of the dataset)
- The gender distribution is nearly balanced, with no significant skew.

Numeric Data type:

- The **Gender** field is of **object (string)** type, not numeric.
- String lengths range from **4 to 6 characters** ("Male" has 4 characters, "Female" has 6).
- No NULL values (0% missing rate).
- The data is clean and standardized, with no format inconsistencies.

# Blood Pressure:

| Metadata | | |
|---|---|---|
| Attribute | Field Datatype | Field Length |
| Blood Pressure | float64 | 1 |

| Data Profiling Summary | |
|---|---|
| Attribute | Value |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 62 |
| Uniqueness (%) | 0.62 |
| Distinctness (%) | 0.62 |

| Data Profiling Additional Statistics | |
|---|---|
| Attribute | Value |
| Data Types | 1 |
| Field Length (MIN) | 3 |
| Field Length (MAX) | 3 |
| Field Value (MIN) | 120 |
| Field Value (MAX) | 180 |
| Mean | 149.76 |
| Median | 150 |
| Standard Deviation | 17.56 |

| Blood Pressure Popular | | |
|---|---|---|
| Value | Count | Percentage (%) |
| 120 | 174 | 1.74 |
| 121 | 162 | 1.62 |
| 122 | 161 | 1.61 |
| 123 | 161 | 1.61 |
| 124 | 169 | 1.69 |
| 125 | 148 | 1.48 |
| 126 | 178 | 1.78 |
| 127 | 169 | 1.69 |
| 128 | 171 | 1.71 |
| 129 | 177 | 1.77 |
| 130 | 136 | 1.36 |
| 131 | 140 | 1.4 |
| 132 | 172 | 1.72 |
| 133 | 178 | 1.78 |
| 134 | 214 | 2.14 |
| 135 | 148 | 1.48 |
| 136 | 177 | 1.77 |
| 137 | 175 | 1.75 |
| 138 | 171 | 1.71 |
| 139 | 157 | 1.57 |
| 140 | 181 | 1.81 |
| 141 | 159 | 1.59 |
| 142 | 181 | 1.81 |
| 143 | 165 | 1.65 |
| 144 | 156 | 1.56 |

| Blood Pressure Field Data Types | | |
|---|---|---|
| Data Type | Count | Percentage (%) |
| float64 | 10000 | 100 |
| NULL | 0 | 0 |

## Summary:

Completeness

- The **Blood Pressure** attribute is 100% complete, with no NULL values.
- Data is fully available and requires no missing value handling.

Cardinality (Number of Unique Values)

- There are **62 unique values**, representing 62 distinct blood pressure levels.
- This indicates a diverse distribution of values without excessive grouping.

Uniqueness & Distinctness

- Uniqueness: 0.62% → Each blood pressure value appears multiple times in the dataset.
- Distinctness: 0.62% → There is moderate repetition across blood pressure levels.

## Deatailed Observations:

Blood Pressure Distribution:

- Blood pressure values range from **120 mmHg to 180 mmHg**.
- The most common values fall within the **125–150 mmHg** range, with a frequency of about 1.5% - 1.8% for each value.
- No abnormal outliers detected; the distribution appears even.

Descriptive Statistics:

- Minimum: 120 mmHg.
- Maximum: 180 mmHg.
- Mean: 149.76 mmHg.
- Median: 150 mmHg → The data is **symmetrically distributed**.
- Standard Deviation: 17.56 → The blood pressure data is **fairly spread out**, indicating no significant clustering around a fixed range.
- The most common blood pressure values fall within the range of **125–150 mmHg**, with each value

accounting for **1.5%–1.8%** of the data.

- The most common blood pressure values fall within the range of **125–150 mmHg**, with each value accounting for **1.5%–1.8%** of the data.

Numeric Data Type:

- All values in the **Blood Pressure** attribute are of type **float64**..
- **No NULL values** (0%).
- No data type processing required → the data is **fully suitable for numerical analysis**.

# Cholesterol Level:

**Metadata**

| Attribute | Field Datatype | Field Length |
|---|---|---|
| Cholesterol Level | float64 | 1 |

**Data Profiling Summary**

| Attribute | Value |
|---|---|
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 152 |
| Uniqueness (%) | 1.52 |
| Distinctness (%) | 1.52 |

**Data Profiling Additional Statistics**

| Attribute | Value |
|---|---|
| Data Types | 1 |
| Field Length (MIN) | 3 |
| Field Length (MAX) | 3 |
| Field Value (MIN) | 150 |
| Field Value (MAX) | 300 |
| Mean | 225.43 |
| Median | 225.43 |
| Standard Deviation | 43.51 |

**Cholesterol Level Popular**

| Value | Count | Percentage (%) |
|---|---|---|
| 150 | 61 | 0.61 |
| 151 | 57 | 0.57 |
| 152 | 72 | 0.72 |
| 153 | 62 | 0.62 |
| 154 | 65 | 0.65 |
| 155 | 66 | 0.66 |
| 156 | 62 | 0.62 |
| 157 | 49 | 0.49 |
| 158 | 61 | 0.61 |
| 159 | 66 | 0.66 |
| 160 | 64 | 0.64 |
| 161 | 67 | 0.67 |
| 162 | 78 | 0.78 |
| 163 | 65 | 0.65 |
| 164 | 56 | 0.56 |
| 165 | 78 | 0.78 |
| 166 | 77 | 0.77 |
| 167 | 75 | 0.75 |
| 168 | 69 | 0.69 |
| 169 | 47 | 0.47 |
| 170 | 65 | 0.65 |
| 171 | 63 | 0.63 |
| 172 | 67 | 0.67 |
| 173 | 63 | 0.63 |
| 174 | 61 | 0.61 |

**Cholesterol Level Field Data Types**

| Data Type | Count | Percentage (%) |
|---|---|---|
| float64 | 10000 | 100 |
| NULL | 0 | 0 |

## Summary

Completeness

- The *Cholesterol Level* attribute is **100% complete**, with **no NULL values**.
- The dataset is sufficient and does not require any missing value treatment.

Cardinality (Số lượng giá trị duy nhất)

- There are **152 unique values**, indicating **152 different cholesterol levels**.
- This shows that the cholesterol data is **highly diverse** and not clustered into a few values.

Uniqueness & Distinctness

- Uniqueness: 1.52% → Each cholesterol value appears multiple times in the dataset.
- Distinctness: 1.52% → The data has **noticeable repetition** among cholesterol levels.

## Detail Observation

Cholesterol Level Distribution

- Cholesterol values **range from 150 to 300 mg/dL**.
- The most common cholesterol levels range from 180-250 mg/dL, with a ratio ranging from 0.5% - 0.8% per cholesterol level.

Statistical Summary

- Minimum value (Min): 150 mg/dL
- **Maximum value (Max)**: 300 mg/dL
- **Mean**: 225.43 mg/dL
- **Median**: 225.43 mg/dL → The data is **fairly symmetrical**.
- **Standard Deviation**: 43.51 → Cholesterol values show a **considerable spread**

Numeric Data

- All values in the *Cholesterol Level* attribute are of **float64 type**
- **No NULL values** (0%)
- No data type transformation is needed; the data is **fully suitable for numerical analysis**

# Exercise Habits

| Input Metadata | |
|---|---|
| Field Name | Exercise Habits |
| Field Data Type | VARCHAR |
| Field Length | 10 |
| **Data Profiling Summary Statistics** | |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness | 100,00% |
| Cardinality | 3 |
| Uniqueness | 0,03% |
| Distinctness | 0,03% |
| **Data Profiling Additional Statistics** | |
| Field Data Type | 1 |
| Field Length (MIN) | 3 |
| Field Length (MAX) | 6 |
| Field Formats | 1 |

| Exercise Habits (Field Data Types) | Count | Percentage |
|---|---|---|
| VARCHAR | 10000 | 100,00% |
| **Exercise Habits (Top 5 Values)** | **Count** | **Percentage** |
| High | 3397 | 33,97% |
| Medium | 3332 | 33,32% |
| Low | 3271 | 32,71% |

## Summary:

- **Completeness** is high (100%), meaning there are **no NULL values**.
- **Cardinality** is 3, indicating the attribute only takes **3 unique values**, representing different exercise levels.
- **Uniqueness**: 0.03% → Exercise levels are **highly repetitive** in the dataset..
- **Distinctness**: 0.03% → The data shows a **high degree of repetition** among exercise categories.

## Detail Observation:

- The number of people across the **three exercise levels** is **relatively balanced**, with the **"High" level being the most frequent**.

# Smoking

| Input Metadata | |
|---|---|
| Field Name | Smoking |
| Field Data Type | VARCHAR |
| Field Length | 3 |
| **Data Profilling Summary Statistics** | |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness | 100,00% |
| Cardinality | 2 |
| Uniqueness | 0,02% |
| Distinctness | 0,02% |
| **Data Profilling Additional Statistics** | |
| Field Data Type | 1 |
| Field Length (MIN) | 2 |
| Field Length (MAX) | 3 |
| Field Formats | 1 |

| Smoking (Field Data Types) | Count | Percentage |
|---|---|---|
| VARCHAR | 10000 | 100,00% |
| **Smoking (Top 2 Values)** | **Count** | **Percentage** |
| Yes | 5148 | 51,48% |
| No | 4852 | 48,52% |

## Summary:

- **Completeness** is high (100%), with **no NULL values**.
- **Cardinality** has only **2 unique values**: "Yes" and "No", indicating the data is **well-categorized**.
- **Uniqueness**: 0.02% → Since there are only two values, the uniqueness level is **low**.
- **Distinctness**: 0.02% → The values "Yes" and "No" **occur frequently** in the dataset.

## Detail Observation:

- Among individuals diagnosed with heart disease, **the majority are smokers**, accounting for **51.48% out of the 10,000 surveyed**.

# Family Heart Disease

| Input Metadata | |
|---|---|
| Field Name | Family Heart Disease |
| Field Data Type | VARCHAR |
| Field Length | 3 |
| **Data Profilling Summary Statistics** | |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness | 100,00% |
| Cardinality | 2 |
| Uniqueness | 0,02% |
| Distinctness | 0,02% |
| **Data Profilling Additional Statistics** | |
| Field Data Type | 1 |
| Field Length (MIN) | 2 |
| Field Length (MAX) | 3 |
| Field Formats | 1 |

| Family Heart Disease (Field Data Types) | Count | Percentage |
|---|---|---|
| VARCHAR | 10000 | 100,00% |
| **Family Heart Disease (Top 2 Values)** | **Count** | **Percentage** |
| No | 5025 | 50,25% |
| Yes | 4975 | 49,75% |

## Summary:

- **Completeness** is high (100%), with **no NULL values**.

- **Cardinality** has only **2 unique values**: “Yes” and “No”, indicating the data is **well-categorized**.
- **Uniqueness**: 0.02% → Since there are only two values, the uniqueness level is **low**.
- **Distinctness**: 0.02% → The values “Yes” and “No” **occur frequently** in the dataset.

## Observation:

- It is observed that **genetic factors also play a role** among individuals with heart disease

→ If there is a family history of heart disease, the **risk of developing heart disease is significantly higher**.

## BMI

| Input Metadata | |
|---|---|
| Field Name | BMI |
| Field Data Type | FLOAT |
| Field Length | 10 |
| **Data Profilling Summary Statistics** | |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness | 100,00% |
| Cardinality | 9979 |
| Uniqueness | 99,79% |
| Distinctness | 99,79% |
| **Data Profilling Additional Statistics** | |
| Field Data Type | 1 |
| Field Value (MIN) | 18,0028369436750 |
| Field Value (MAX) | 39,9969537965812 |
| Field Formats | 1 |

| BMI (Field Data Types) | Count | Percentage |
|---|---|---|
| FLOAT | 10000 | 100,00% |
| **BMI (Top 5 Values)** | **Count** | **Percentage** |
| 29,0772689275110 | 22 | 0,22% |
| 24,9915910916903 | 1 | 0,01% |
| 25,2217985244363 | 1 | 0,01% |
| 29,8554471423716 | 1 | 0,01% |
| 24,1304768828524 | 1 | 0,01% |
| **BMI (Field Formats)** | **Count** | **Percentage** |
| XX,XXXXXXXXXXXXX | 10000 | 100,00% |

## Summary:

- **Completeness** is high (100%), with **no NULL values**.
- **Cardinality** has **9,979 unique values**, indicating **high diversity** in Body Mass Index (BMI).
- **Uniqueness**: 99.79% → There are many different BMI levels, but some duplicates still exist.
- **Distinctness**: 99.79% → The data is highly diverse, though certain values appear more than once.

## Observation:

- At a **BMI threshold > 29**, approaching the obesity level, the majority of individuals tend to fall into the group with heart disease → This suggests that **overweight and obesity are among the contributing factors to heart disease**.

# Alcohol Consumption:

| Metadata | |
|---|---|
| Attribute | Value |
| Field Name | Alcohol Consumption |
| Field Datatype | object |
| Field Length | 1 |

| Data Profiling Summary | |
|---|---|
| Attribute | Value |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 4 |
| Uniqueness (%) | 0.04 |
| Distinctness (%) | 0.04 |

| Data Profiling Additional Statistics | |
|---|---|
| Attribute | Value |
| Data Types | 1 |
| Field Length (MIN) | 3 |
| Field Length (MAX) | 6 |
| Field Value (MAX) | N/A |
| Field Formats | 3 |

| Alcohol Consumption - Popular | | |
|---|---|---|
| Value | Count | Percentage(%) |
| Non | 2586 | 25.86 |
| Medium | 2500 | 25 |
| Low | 2488 | 24.88 |
| High | 2426 | 24.26 |

| Alcohol Consumption - Field Formats | | |
|---|---|---|
| Format | Count | Percentage(%) |
| XXX | 5074 | 50.74 |
| XXXXXX | 2500 | 25 |
| XXXX | 2426 | 24.26 |

| Alcohol Consumption - Field Data Types | | |
|---|---|---|
| Data Type | Count | Percentage(%) |
| object | 10000 | 100 |

## Summary:

- **Completeness** is 100%, yet the dataset for *Alcohol Consumption* contains a significant portion of missing values, with **2,586 "Non" entries (25.86%)**, meaning only **74.14%** of the data is valid.
- **Cardinality** (Number of unique values) is **4**, representing three actual alcohol consumption levels — "Non", "Low", "Medium", and "High".
- **Uniqueness** and **Distinctness** are very low (0.04%), indicating that the data is highly repetitive and classified into only four distinct groups.

## Detailed Observations

- The four consumption levels ("Non", "Low", "Medium", "High") have **relatively equal proportions** (~25.86%, 25%, 24.88%, 24.26%), suggesting that alcohol consumption among the study population is evenly distributed across the categories.
- The largest group is **non-drinkers** ("Non"), accounting for 25.86% of the dataset.
- The smallest group is **high-level drinkers** ("High"), accounting for 24.26%.
- **Field length** ranges from **3 to 6 characters**, showing that the data is standardized with no anomalies in text length.
- The data type is **object (string)**.

# Stress Level

| Metadata | |
|---|---|
| Attribute | Value |
| Field Name | Stress Level |
| Field Datatype | object |
| Field Length | 1 |

| Data Profiling Summary | |
|---|---|
| Attribute | Value |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 4 |
| Uniqueness (%) | 0.04 |
| Distinctness (%) | 0.04 |

| Data Profiling Additional Statistics | |
|---|---|
| Attribute | Value |
| Data Types | 1 |
| Field Length (MIN) | 3 |
| Field Length (MAX) | 6 |
| Field Value (MAX) | N/A |
| Field Formats | 3 |

| Stress Level - Popular | | |
|---|---|---|
| Value | Count | Percentage(%) |
| Medium | 3387 | 33.87 |
| Low | 3320 | 33.2 |
| High | 3271 | 32.71 |
| Non | 22 | 0.22 |

| Stress Level - Field Formats | | |
|---|---|---|
| Format | Count | Percentage(%) |
| XXXXXX | 3387 | 33.87 |
| XXX | 3342 | 33.42 |
| XXXX | 3271 | 32.71 |

| Stress Level - Field Data Types | | |
|---|---|---|
| Data Type | Count | Percentage(%) |
| object | 10000 | 100 |

**Summary:**

- **Completeness** is very high (99.78%), with only **22 "Non" values (0.22%)**, indicating that the dataset is nearly complete.
- **Cardinality** (Number of unique values) is **4**, representing four categories ("Non", "Low", "Medium", "High").
- **Uniqueness** and **Distinctness** are both very low (0.04%), indicating that the data is highly repetitive and falls into only four classification groups.

**Detailed Observations:**

- The three categories "Low" (33.2%), "Medium" (33.87%), and "High" (32.71%) have **fairly balanced proportions**, showing that stress levels in the dataset are relatively evenly distributed..
- The "Non" category is minimal (**22 cases – 0.22%**) and can be easily addressed without significantly impacting the analysis.
- **Field length** ranges from **3 to 6 characters**, ensuring the data is standardized and contains no anomalies in text length.
- The data type is **object (string)**.

# Heart Disease Status

| Metadata | |
|---|---|
| Attribute | Value |
| Field Name | Heart Disease Status |
| Field Datatype | object |
| Field Length | 1 |

| Data Profiling Summary | |
|---|---|
| Attribute | Value |
| NULL | 0 |
| Missing | 0 |
| Actual | 10000 |
| Completeness (%) | 100 |
| Cardinality | 2 |
| Uniqueness (%) | 0.02 |
| Distinctness (%) | 0.02 |

| Data Profiling Additional Statistics | |
|---|---|
| Attribute | Value |
| Data Types | 1 |
| Field Length (MIN) | 2 |
| Field Length (MAX) | 3 |
| Field Value (MAX) | N/A |
| Field Formats | 2 |

| Heart Disease Status - Popular | | |
|---|---|---|
| Value | Count | Percentage(%) |
| No | 8000 | 80 |
| Yes | 2000 | 20 |

| Heart Disease Status - Field Data Types | | |
|---|---|---|
| Data Type | Count | Percentage(%) |
| object | 10000 | 100 |

**Summary:**

- The data is **100% complete**, with no NULL or missing values
- There are only **two classification values** (Yes/No), representing heart disease status.
- **Uniqueness** and **Distinctness** are both very low (0.02%), indicating that the data is binary and highly repetitive.

**Detailed Observations:**

- **80%** of the data has the value “No” (no heart disease), while only **20%** has the value “Yes” (has heart disease).
    + This shows that the majority of individuals in the dataset **do not have heart disease**.
    + The dataset is **imbalanced**, which may affect predictive models if machine learning is applied.
- **Field length** ranges from **2 to 3 characters**, matching the expected values “Yes” or “No”, with no signs of invalid data.

# 2. Abstraction

## 2.1. Domain task 1: Relationship between Age Group and Heart Disease

### 2.1.1. Data abastraction

| | |
|---|---|
| Data: | patient_heart_disease |
| Item: | 1 dòng là 1 bệnh nhân |
| Dataset availability: | Static |

| | Semantics | Attribute Type | Hierarchical | Characteristic | Direction | Quantitative Type | Bin Number |
|---|---|---|---|---|---|---|---|
| Age | Tuổi bệnh nhân | Quantitative | None | Continuous | Sqquentia | Ratio | 8 |
| Heart Disease | Trạng thái bệnh | Categorical | None | Discrete | None | None | 0 |

### 2.1.2. Task abstraction

**Produce → Explore → Sum**

- **Produce**
  - Divide the *Age* column into age groups (20–29, 30–39, …, >80).
  - Check for missing data and clean if necessary.
- **Explore**
  - Count the number of heart disease cases by age group.
  - Plot a bar chart to observe the trend.
- **Sum**
  - **Observation:** Higher age correlates with an increased incidence of heart disease.
  - **High-risk age group:** Individuals over 50 years old have the highest risk.

## 2.2. Domain task 2: The relationship between gender and heart disease

### 2.2.1. Data abstraction

| | |
|---|---|
| Data: | gender_heart_disease |
| Item: | 1 dòng là 1 bệnh nhân |
| Dataset availability: | Static |

| | Semantics | Attribute Type | Hierarchical | Characteristic | Direction | Quantitative Type | Bin Number |
|---|---|---|---|---|---|---|---|
| Gender | Giới tính | Categorical | None | Discrete | None | None | 0 |
| Heart Disease | Trạng thái bệnh | Categorical | None | Discrete | None | None | 0 |

### 2.2.2. Task abstraction

**Produce → Explore → Sum**

- **Produce**
  - Check if the *Gender* column has formatting errors (e.g., "M", "F" vs. "Male", "Female").
  - Check for missing data and clean if necessary.
- **Explore**
  - Count the number of male and female patients with heart disease.
  - Calculate the percentage of heart disease cases by gender.
  - Create a bar chart or pie chart.
- **Sum**
  - **Observation:** The prevalence of heart disease is higher in males than in females, possibly due to physiological factors or lifestyle habits.

## 2.3. Domain task 3: The relationship between smoking status and heart disease

### 2.3.1. Data abstraction

| Attr | Abstraction |
|---|---|
| Data type | Attribute and Item |
| Dataset type | Table |
| Dataset availability | Static |

| No | Attribute Name | Attribute Type (C,O,Q) | Direction | Hierarchical | Characteristic | Direction2 | Quantitative | Bin Number | Semantic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Smoking | Categorical | None | No | Discrete | None | None | 3 | Yes. 'Có hút thuốc' hoặc 'Không hút thuốc' |
| 2 | Heart Disease Status | Categorical | None | No | Discrete | None | None | 2 | Yes. 'Có bị bệnh tim' hoặc 'Không bị bệnh tim' |

### 2.3.2. Task abstraction

**Analyze → Search → Query**

- **Analyze:**
  - **Consume:** Data on smoking status and heart disease is already available.
  - **Representation:** A statistical table or proportion chart can be used → **Present**
- **Search:**
  - No need to search for a specific individual; instead, the goal is to observe overall trends.
  - Not suitable for *Lookup* or *Locate* since the domain task does not target a specific item but to aims find relationships between attributes → **Explore.**
- **Query:**
  - Identify the relationship between attributes and summarize heart disease rates by each smoking group → **Summarize.**

**=> Present → Explore → Summary**

## 2.4. Domain task 4: The relationship between exercise status and heart disease

### 2.4.1. Data abstraction

| No | Attribute Name | Attribute Type (C,O,Q) | Direction | Hierarchical | Characteristic | Direction2 | Quantitative | Bin Number | Semantic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Exercise Habits | Ordinal | None | No | Discrete | None | None | 3 | Yes. 'Mức độ tập thể dục đo được' |
| 2 | Heart Disease Status | Categorical | None | No | Discrete | None | None | 2 | Yes. 'Có bị bệnh tim' hoặc 'Không bị bệnh tim' |

### 2.4.2. Task abstraction

**Analyze → Search → Query**

- **Analyze:**
  - **Consume**: Data on exercise frequency and heart disease status is already available and can be directly visualized → **Present**.
- **Query:**
  - Since the goal is to examine the relationship between exercise frequency and heart disease across the entire dataset, not just a specific data point → **Explore**.
- **Query:**
  - Requires analyzing the data and summarizing the trend between the two variables → **Summarize**.

**⇒ Present → Explore → Summarize**

## 2.5. Domain task 5: How do cholesterol levels vary between people with and without heart disease?

### 2.5.1. Data abstraction

| Attr | Abstraction |
|---|---|
| Data type | Attribute and Item |
| Dataset type | Table |
| Dataset availability | Static |

| No | Attribute Name | Type (C, O, Q) | Interval/Ratio | Key/Value | Direction | Hierarchical | Continuous/Discrete | Semantic |
|---|---|---|---|---|---|---|---|---|
| 1 | Cholesterol Level | Quantitative | Ratio | Value | Sequential | No | Continuous | Yes. Đây là mức Cholesterol đo được từ bệnh nhân. |
| 2 | Heart Disease Status | Categorical | None | Key | None | No | Discrete | Yes. 'Có bệnh tim' hoặc 'Không có bệnh tim' |

### 2.5.2. Task abstraction

**Analyze -> Search -> Query**

- **Analyze:**
  - **Consume:** Data on cholesterol levels and heart disease status is already available and can be directly visualized → **Present**.

- **Search:**
  - The domain task does not target a specific individual → No specific target → Cannot be **Lookup** or **Locate**.
  - The domain task requires analyzing the differences in cholesterol levels between heart disease groups → **Explore**.
- **Query:**
  - The domain task requires analyzing cholesterol levels between two groups (with heart disease and without heart disease) → Involves comparison → **Compare.**

**⇒ Present → Explore → Summarize**

## 2.6. Domain task 6: Is there a correlation between BMI and heart disease?

### 2.6.1. Data abstraction

| No | Attribute Name | Type (C, O, Q) | Interval/Ratio | Key/Value | Direction | Hierarchical | Continuous/Discrete | Semantic |
|---|---|---|---|---|---|---|---|---|
| 1 | BMI | Quantitative | Ratio | Value | Sequential | No | Continuous | Yes. Đây là chỉ số khối cơ thể của bệnh nhân. |
| 2 | Heart Disease Status | Categorical | None | Key | Sequential | Yes | Discrete | Yes. 'Có bệnh tim' hoặc 'Không có bệnh tim' |
| | | | | | | | | |

### 2.6.2. Task abstraction

**Analyze -> Search -> Query**

- **Analyze:**
  - **Consume**: Data on BMI index and heart disease status is already available and can be directly visualized → Present.
- **Search:**
  - The domain task does not target a specific individual → No specific target → Cannot be **Lookup** or **Locate**.
  - The domain task requires identifying the relationship between BMI and heart disease → **Explore**.
- **Query:**
  - The domain task requires finding the correlation between two attributes in the entire dataset → **Summarize**.

**=> Present → Explore → Summarize**

## 2.7. Domain task 7: Does a family history of heart disease increase the risk?

### 2.7.1. Data abstraction

| Attr | Abstraction |
|---|---|
| Data type | Attribute and Item |
| Dataset type | Table |
| Dataset availability | Static |

| No | Attribute Name | Type(C,O,Q) | Interval/Ratio | Key/Value | Direction | Hierarchial | Continuous/Discrete | Semantic |
|---|---|---|---|---|---|---|---|---|
| 1 | Family Heart Disease | Categorical | None | Key | None | None | Discrete | Yes. Đây là "có tiểu sử bệnh tim " hoặc " không có tiểu sử bệnh tim" |
| 2 | Heart Disease Status | Categorical | None | Value | None | None | Discrete | Yes. Đây là "có bệnh tim " hoặc " không có bệnh tim" |

### 2.7.2. Data abstraction

**Analyze -> Search -> Summarize**

- **Analyze:**
  - **Consume:** Data on family history of heart disease and current heart disease status is already available and can be directly visualized → **Present**.
- **Search:**
  - The domain task does not target a specific individual → No specific target → Cannot be **Lookup** or **Locate**.
  - The domain task requires identifying the relationship between family history of heart disease and current heart disease status → **Explore**.
- **Query:**
  - Domain task yêu cầu tìm ra mối tương quan giữa hai thuộc tính trong toàn bộ dữ liệu → **Summarize**

**=> Present → Explore → Summarize**

## 2.8. Domain task 8: How does the distribution of cholesterol levels differ between men and women?

### 2.8.1. Data abstraction

| Attr | Abstraction |
|---|---|
| Data type | Attribute and Item |
| Dataset type | Table |
| Dataset availability | Static |

| No | Attribute Name | Type(C,O,Q) | Interval/Ratio | Key/Value | Direction | Hierarchial | Continuous/Discrete | Semantic |
|---|---|---|---|---|---|---|---|---|
| 1 | Cholesterol Level | Quantitative | Ratio | Value | None | None | Continuous | Yes. Đây là nồng độ Cholesterol có trong máu |
| 2 | Gender | Categorical | None | Key | None | None | Discrete | Yes. Đây là giới tính của bạn nhân nam hoặc nữ |

### 2.8.2. Task abstraction

**Analyze -> Search -> Query**

- **Analyze:**
  - **Consume:** Dữ liệu về mức độ Cholesterol và giới tính đã có sẵn, có thể biểu diễn trực tiếp.→ Present
- **Search:**
  - Domain task yêu cầu sự so sánh sự phân bố Cholesterol giữa 2 nhóm giới tính → Không có target cụ thể → Không thể là **Lookup** hay **Locate**.
  - Domain task yêu cầu tìm sự phân bố giữa Cholesterol theo từng giới tính → **Explore**
- **Query:**
  - Domain task yêu cầu tìm ra sự phân bố giữa hai thuộc tính trong toàn bộ dữ liệu → **Summarize**

**=> Present → Explore → Summarize**

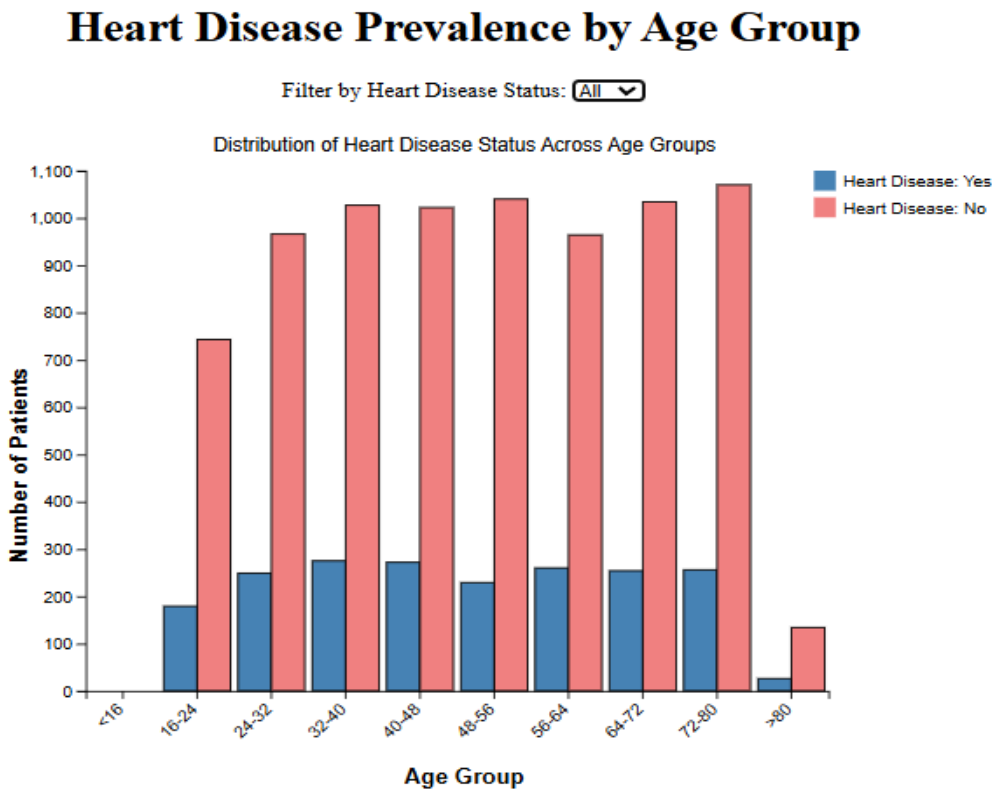
# 3. Idiom design

## 3.1. Domain task 1

### 3.1.1. Idiom

| Idiom | Bar charts | |
|---|---|---|
| **Data** | Rangting: C | |
| | Count: Q | |
| | Mark: bar chart | |
| **Encode** | Channel | Q: vertical position |
| | | C: horizontal position |
| **TASK** | Compare the proportion of heart disease cases across different age groups<br>Observe the trend of increase/decrease by age | |
| **SCALE** | Keys: 9<br>Level: Ordinal | |

### 3.1.2. Chart



### 3.1.3. Evaluate

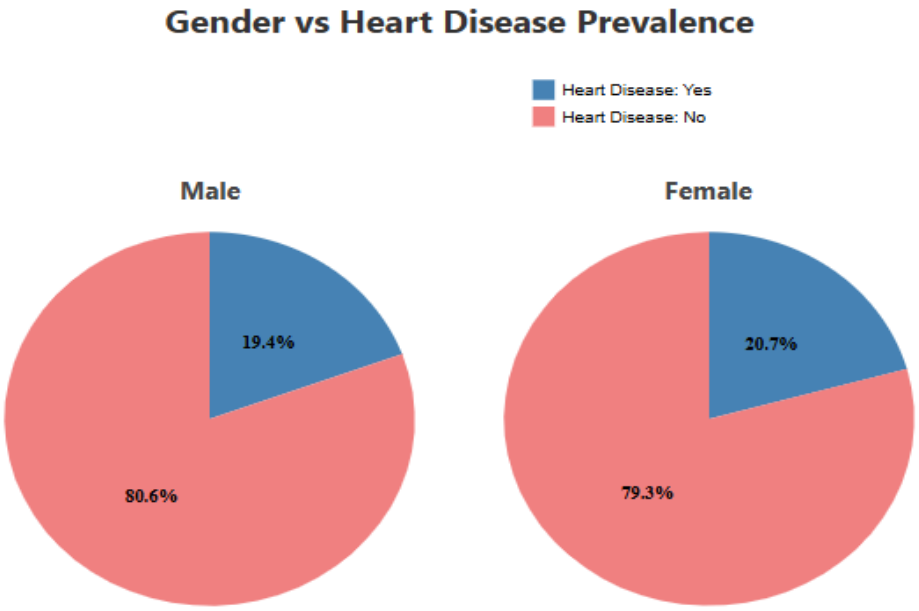
- **Expressiveness**
  - The age groups are evenly divided and easy to recognize, enabling viewers to easily compare across groups and grasp the overall trend by age.
- **Effectiveness**
  - **Accuracy:** Facilitates understanding of the number and proportion of patients with and without heart disease in each age group.
  - **Discriminablity:** The contrasting blue/red colors are easily distinguishable, helping viewers differentiate heart disease status.
  - **Separability:** không xét vì không có thuộc tính nào sử dụng 2 channel để biểu diễn.
- **Chart Analysis**
  - Identify the age groups at highest risk (ages 24–72 have the highest number of heart disease cases).

## 3.2. Domain task 2

### 3.2.1. Idiom

| Idiom | Pie charts | |
|---|---|---|
| **Data** | Gender: C | |
| | Heart Disease Status: C | |
| | Count/Percentage: Q | |
| **Encode** | Mark: pie chart | |
| | Channel | C: Separate the genders into two distinct charts |
| | | Q: Percentage of patients |
| | | C: Color |
| **TASK** | Compare the proportion of heart disease cases between males and females<br>Identify which gender has a higher risk of heart disease | |
| **SCALE** | Keys: 2<br>Level: Categorical | |

### 3.2.2. Chart




### 3.2.3. Evalute

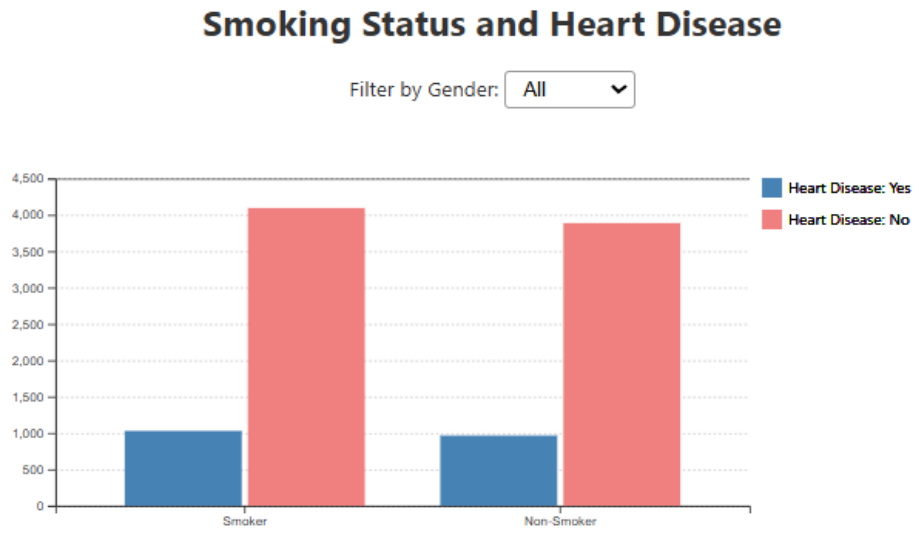
- **Expressiveness**
  - Clearly distinguishes each gender, enabling straightforward and direct comparisons.
  - Allows easy recognition of the disparity between male and female groups.
- **Effectiveness**
  - **Accuracy:** Clearly shows the number of patients and the proportion of heart disease cases by gender.
  - **Discriminablity:** Distinct colors help differentiate between groups easily.
  - **Separability:** Not considered since no attribute uses two channels for representation.
- **Chart Analysis**
  - Helps identify which gender has a higher prevalence of heart disease and the degree of difference between the two genders.

## 3.3. Domain task 3

### 3.3.1. Idiom

| Idiom | Grouped Bar Charts | |
|---|---|---|
| **Data** | Smoking Status: O | |
| | Heart Disease Status: C | |
| | Count/Percentage: Q | |
| | Mark: bar | |
| **Encode** | Channel | O: horizontal position (Smoking Status – Yes/No) |
| | | Q: vertical position (Count) |
| | | C: color (Heart Disease Status) |
| **TASK** | Compare the heart disease rates between smokers and non-smokers.<br>Observe the differences and the potential impact of smoking status on heart disease. | |
| **SCALE** | Keys: 2<br>Level: Nominal | |

### 3.3.2. Chart




### 3.3.3. Evalute

- **Expressiveness**
  - The grouped bar chart clearly displays the proportion between individuals with and without heart disease in each group.

- It is easy to distinguish between smokers and non-smokers (Yes, No).
- The chart also makes it straightforward to identify the differences in heart disease rates between groups.

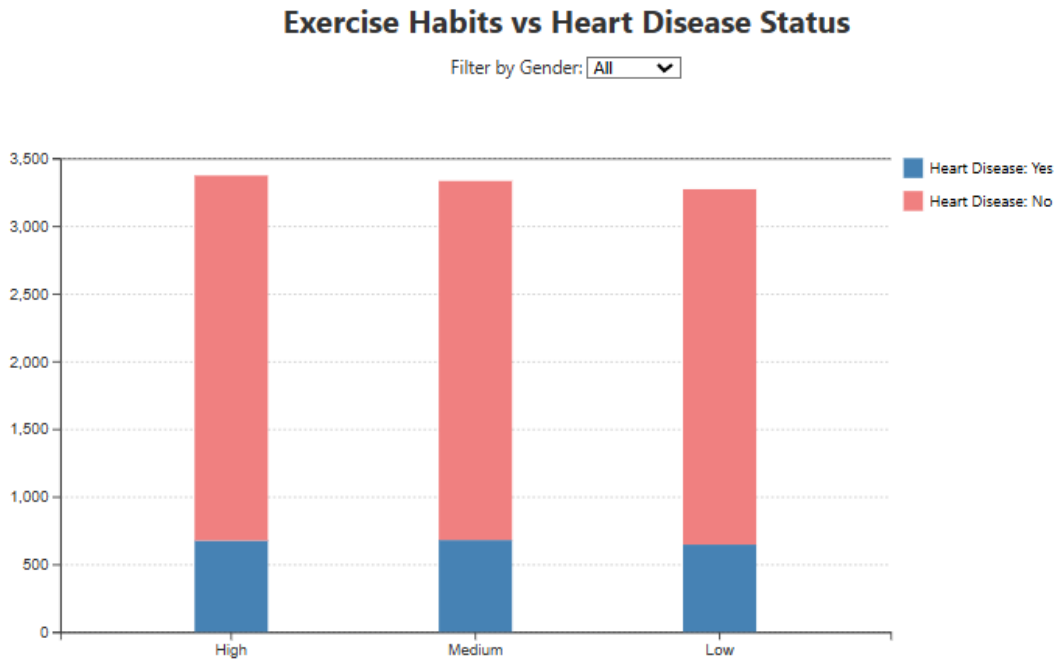
- **Effectiveness**
  - **Accuracy:** The vertical axis clearly shows both the total number and the contribution of each category, making it easy to visualize the heart disease rate between groups.
  - **Discriminability:** Color is used to differentiate between "With Heart Disease" and "Without Heart Disease," enabling readers and analysts to easily make distinctions.
- **Phân tích biểu đồ**
  - The chart facilitates comparison of heart disease rates between groups.
  - It helps identify which group has a higher risk and the magnitude of the difference.
  - It provides a clear visual basis for drawing conclusions when analyzing heart disease rates based on smoking status.

## 3.4. Domain task 4

### 3.4.1. Idiom

| Idiom | Stacked Bar Charts | |
|---|---|---|
| **Data** | Exercise Habit: O | |
| | Heart Disease Status: C | |
| | Count/Percentage: Q | |
| **Encode** | Mark: bar | |
| | Channel | O: horizontal position (Exercise Habit) |
| | | Q: vertical position (Count) |
| | | C: color (Heart Disease Status) |
| **TASK** | Compare the heart disease rate among groups with High, Medium, and Low exercise habits.<br>Compare the heart disease rate among groups with High, Medium, and Low exercise habits. | |
| **SCALE** | Keys: 3<br>Level: Nominal | |

### 3.4.2. Chart




### 3.4.3. Evaluation

- **Expressiveness**
  - The stacked bar chart clearly displays the proportion between two groups—those with and without heart disease—within each exercise habit category.
  - The horizontal axis classifies exercise habits into three levels: High, Medium, and Low.
  - The vertical axis represents the number of individuals, divided into two groups by color: With Heart Disease and Without Heart Disease.
  - The chart makes it easy to identify the corresponding heart disease rate for each exercise intensity level.
- **Effectiveness**
  - Accurary: The stacked bar chart allows viewers to assess both the total count and the relative proportion of each group, giving a clear sense of percentages.
  - Discriminability: Two contrasting colors make it easy to distinguish between individuals with and without heart disease.
  - Separability: Not applicable, as each attribute uses only one encoding channel (position or color).
- **Chart Analysis**
  - The chart shows that the heart disease rate is relatively consistent across the High, Medium, and Low exercise habit groups.

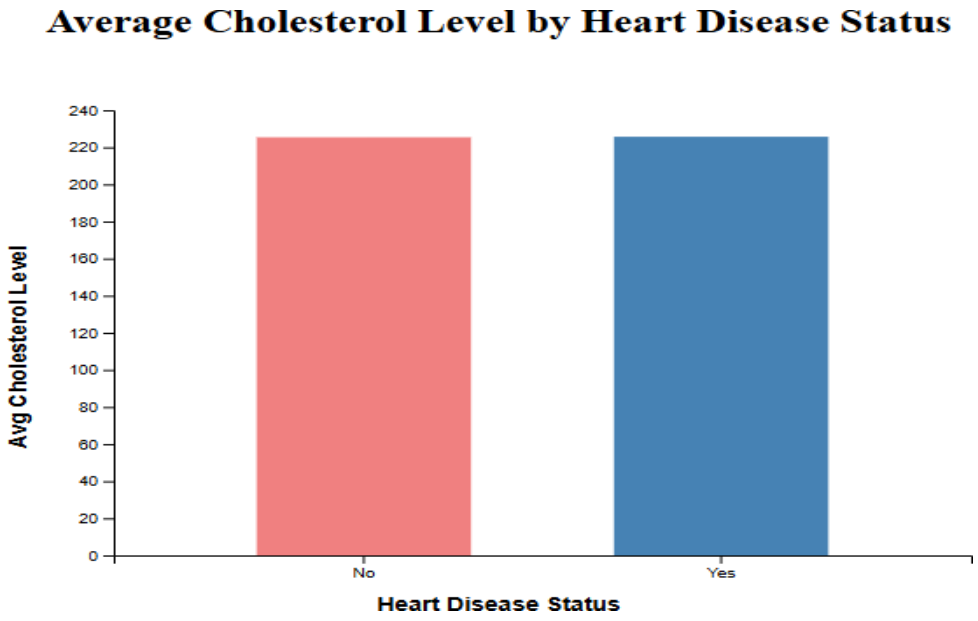
- Although the total number of individuals in each group does not differ greatly, the variation in heart disease rates is minimal, suggesting that exercise habits may not be the primary factor influencing heart disease risk in this dataset.

## 3.5. Domain task 5

### 3.5.1. Idiom

| Idiom | Bar Chart | |
|---|---|---|
| **Data** | Grouping: Heart Disease Status (Categorical – C) | |
| | Measure: Cholesterol Level (Quantitative – Q) | |
| | Count/Percentage: Q | |
| | Mark: bar | |
| **Encode** | Channel | -Cholesterol Level: pos (height) |
| | | -Heart Disease Status: color |
| **TASK** | Comparing cholesterol levels between groups with and without heart disease | |
| **SCALE** | Keys: 2 (Yes/No)<br>Level: Nominal (Status), Quantitative (Value) | |

### 3.5.2. Chart




### 3.5.3. Evaluation

- **Expressiveness**
  - The two groups are clearly defined and easily distinguishable through contrasting colors.

- The X-axis categorizes the data (Yes / No), enabling quick comparison.
- The Y-axis accurately represents the average cholesterol level.

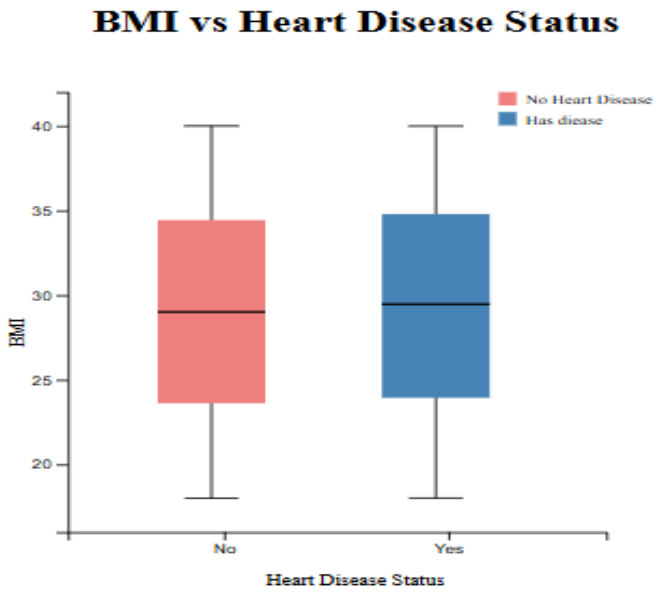
- **Effectiveness**
  - **Accurary:** Precisely displays the average cholesterol value for each group.
  - **Discriminability:** The contrast between blue and red effectively differentiates heart disease status.
  - **Separability:** Each group is represented by a separate bar, avoiding overlap and ensuring readability.
- **Chart Analysis**
  - The group with heart disease (Yes) has a slightly higher average cholesterol level compared to the group without heart disease (No).
  - This may indicate a **potential relationship** between high cholesterol levels and heart disease → further analysis using a scatter plot or correlation test is recommended if deeper insights are desired.

## 3.6. Domain task 6

### 3.6.1. Idiom

| Idiom | Bar Chart | |
|---|---|---|
| **Data** | BMI: Quantitative(Q) | |
| | Heart Disease Status: Categorical (C, mã hóa nhị phân) | |
| | Count/Percentage: Q | |
| **Encode** | Mark: circle | |
| | Channel | -BMI: pos X |
| | | -Status: pos Y + color |
| **TASK** | Tìm mối tương quan giữa BMI và tình trạng bệnh tim | |
| **SCALE** | BMI: Ratio<br>Status: Nominal (mã hóa Ordinal 0/1) | |

### 3.6.2. Chart




### 3.6.3. Evaluation

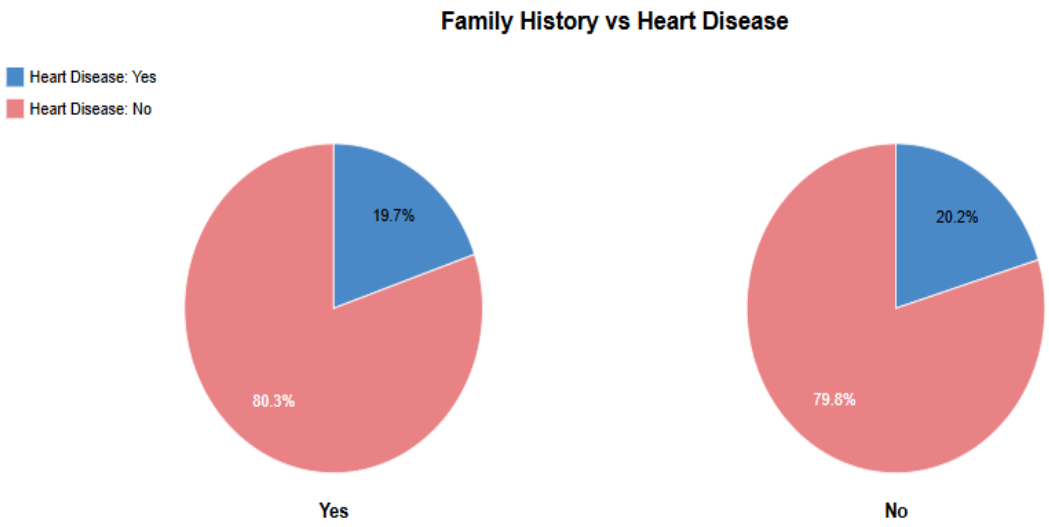
- **Expressiveness**
  - The distribution of points along the horizontal axis (BMI) is clear.
  - The Y-axis with two values (Yes/No) is easy to read thanks to color coding and position.
  - The Y-axis accurately represents the average cholesterol level.
- **Effectiveness**
  - **Accurary:** Displays each individual in detail with their specific BMI value.
  - **Discriminability:** Contrasting colors effectively distinguish heart disease status.
  - **Separability:** Groups are well separated along the Y-axis.
- **Chart Analysis**
  - No clear direct linear correlation is observed, but there is a tendency for higher BMI in the group with heart disease.
  - Further statistical analysis (e.g., correlation, logistic regression) is needed for deeper insights.

## 3.7. Domain task 7

### 3.7.1. Idiom

| Idiom | Pie charts | |
|---|---|---|
| **Data** | Family Heart Disease: C | |
| | Heart Disease Status: C | |
| | Count/Percentage: Q | |
| **Encode** | Mark: slice | |
| | Channel | C: Segmentation based on family history of heart disease |
| | | C: Color (Heart Disease Status) |
| | | Q: Slice size |
| **TASK** | Compare the rate of heart disease between people with and without a family history of heart disease | |
| **SCALE** | Keys: 2<br>Level: Nominal | |

### 3.7.2. Chart




### 3.7.3. Evaluation

- **Expressiveness**
  - The pie chart clearly displays the proportion between two groups of people with and without heart disease for each category (having a family history of heart disease, not having a family history of heart disease).
  - Based on the chart, it is easy to see the heart disease rate corresponding to the family history status (Yes/No).

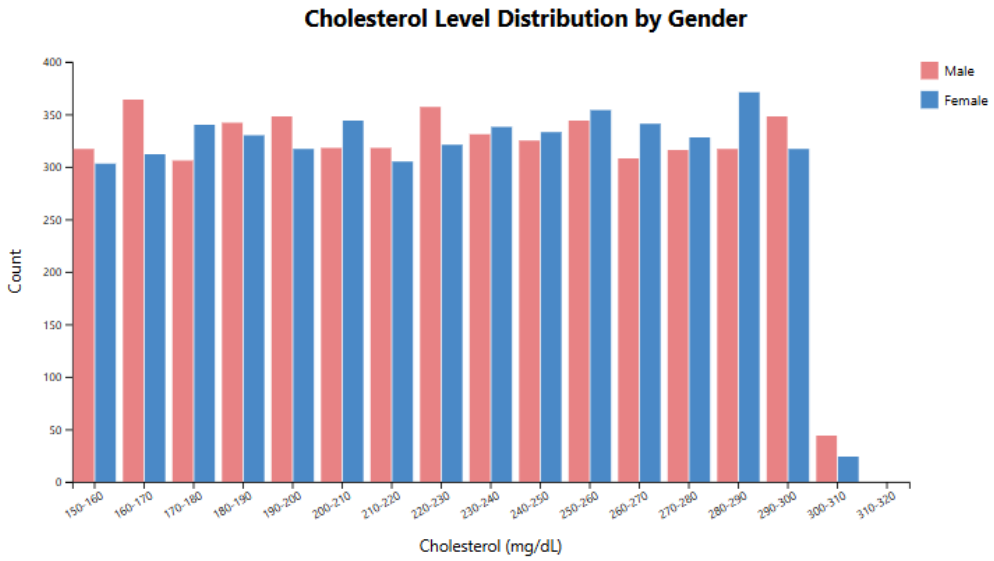
- **Effectiveness**
  - **Accurary:** The pie chart allows users to clearly observe the proportion of people with and without heart disease according to family history.
  - **Discriminability:** Two contrasting colors make it easy to distinguish between those with and without heart disease.
  - **Separability:** Not applicable, as no attribute uses two channels for encoding.
- **Chart Analysis**
  - The chart shows that the heart disease rate is relatively similar between those with and without a family history of heart disease, and the overall rate is quite low.
  - The chart shows that the heart disease rate is relatively similar between those with and without a family history of heart disease, and the overall rate is quite low.

## 3.8. Domain task 8

### 3.8.1. Idiom

<table>
<tr><th>Idiom</th><th colspan="2">Histogram</th></tr>
<tr><td rowspan="3">Data</td><td colspan="2">Cholesterol Level: Q</td></tr>
<tr><td colspan="2">Gender: C</td></tr>
<tr><td colspan="2">Count/Percentage: Q</td></tr>
<tr><td rowspan="4">Encode</td><td colspan="2">Mark:</td></tr>
<tr><td rowspan="3">Channel</td><td>Q: pos ngang</td></tr>
<tr><td>C: màu sắc (Gender)</td></tr>
<tr><td>Q:  pos dọc (count)</td></tr>
<tr><td>TASK</td><td colspan="2">Mức cholesterol theo giới tính (nam/nữ)</td></tr>
<tr><td>SCALE</td><td colspan="2">Keys: 16<br>Level: Nominal</td></tr>
</table>

### 3.8.2. Chart



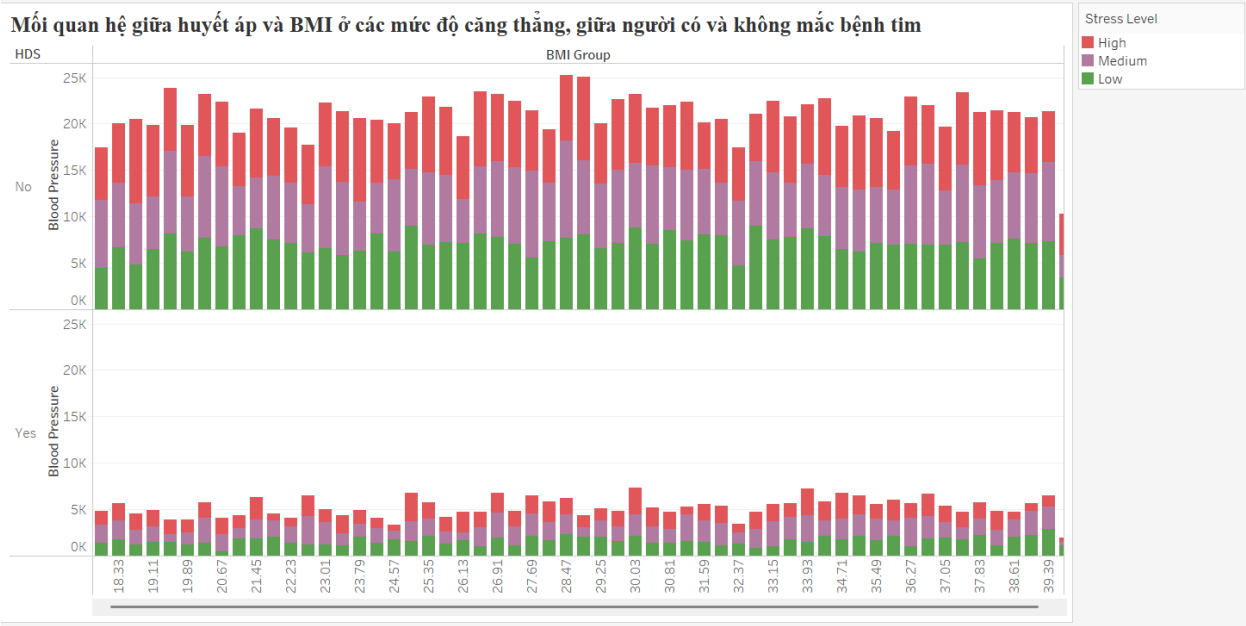

### 3.8.3. Evaluation

- **Expressiveness**
  - The histogram clearly illustrates the distribution of cholesterol levels by gender (male/female).
  - Based on the chart, it is easy to identify the cholesterol levels corresponding to each gender (male/female).
- **Effectiveness**
  - Accurary: The histogram allows users to clearly observe cholesterol levels for both males and females
  - Discriminability: The histogram allows users to clearly observe cholesterol levels for both males and females
  - Separability: Not applicable, as no attribute uses two channels for encoding.
- **Chart Analysis**
  - The chart shows that the number of males and females at each cholesterol level is relatively similar. However, at the 300–310 mg/dL range, the total number of males and females is the lowest and differs significantly from other cholesterol levels.
  - The chart shows that the number of males and females at each cholesterol level is relatively similar. However, at the 300–310 mg/dL range, the total number of males and females is the lowest and differs significantly from other cholesterol levels.

# 4. Bonus

## 4.1. Bonus 1

## 4.2. Bonus 2




**Impact of Stress Level on Blood Pressure and BMI**

- Individuals with high stress levels (High) tend to have higher blood pressure, regardless of BMI..
- Stress level has a stronger effect on the group without heart disease: In the non-heart disease group, high stress levels (red) account for a larger proportion of total blood pressure compared to the heart disease group.
- The heart disease group shows a relatively even distribution of stress across all levels (Low, Medium, High)

**Comparison Between People With and Without Heart Disease**

- The non-heart disease group has higher total blood pressure, which may be due to the larger number of people in this group.
- The heart disease group has lower total blood pressure but still follows a similar trend: higher BMI is associated with higher blood pressure and higher stress levels.
- The difference in stress levels between the two groups is not very large, but the non-heart disease group has more cases of high stress.